Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# Characterisation of dairy strains of *Geobacillus stearothermophilus* and a genomics insight into its growth and survival during dairy manufacture

A thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Microbiology at Massey University, Palmerston North, New Zealand

**Sara Burgess**

**2016**

*This thesis is dedicated to my sons, Samuel and James.*

# Abstract

The thermophilic bacilli, such as *G. stearothermophilus*, are an important group of contaminants in the dairy industry. Although these bacilli are generally not pathogenic, their presence in dairy products is an indicator of poor hygiene and high numbers are unacceptable to customers. In addition, their growth may result in milk product defects caused by the production of acids or enzymes, potentially leding to off-flavours. These bacteria are able to grow in sections of dairy manufacturing plants where temperatures reach 40 – 65 °C. Furthermore, because they are spore formers, they are difficult to eliminate. In addition, they exhibit a fast growth rate and tend to readily form biofilms. Many strategies have been tested to prevent the formation of thermophilic bacilli biofilms in dairy manufacture, but with limited success. This is, in part, because little is known about the diversity of strains found in dairy manufacture, the structure of thermophilic bacilli biofilms and how these bacteria have adapted to grow in a dairy environment.

In Chapters 2 and 3, phenotypic approaches were taken to understand the diversity of strains within a manufacturing plant. Specifically in Chapter 2, strains of the most dominant thermphilic bacilli, *G. stearothermophilus*, were isolated from the surface of various locations within the evaporator section and ten strains were evaluated for different phenotypic characteristics. Biochemical profiling, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and fatty profiling demonstrated that the population was diverse. In Chapter 3, it was shown that the same ten strains varied in their ability to form biofilms and produce spores. Three strains of *G. stearothermophilus*, A1, P3 and D1, were selected for further analysis. SEM demonstrated that there were differences in biofilm morphologies between the three strains, particularly D1 versus the other two strains, A1 and P3.

In Chapters 4, 5 and 6 a comparative genomics approach was taken to determine how these bacteria are able to grow and survive within a dairy manufacturing environment, as well as how they differ from other strains of *Geobacillus*. In Chapter 4 draft genome sequences were generated for three strains of *G.stearothermophilus.* Identification of a putative lactose operon in the three dairy strains provided evidence of dairy adaptation. In Chapter 5 a phylogenomics approach was taken to resolve relationships within the *Geobacillus* genus and to identify differences within the *G. stearothermophilus* group itself. Finally in Chapter 6 comparison with the model organism *B. subtilis,* gave a genomics insight into the potential mechanisms of sporulation for *Geobacillus* spp.

# List of Publications

**Burgess S A**, Flint S H, Lindsay D, Cox M P and Biggs P J (2016). An updated analysis of *Geobacillus* taxonomy based on phylogenomic principles. Submitted to *BMC Microbiology*.

**Burgess S A**, Cox M P, Flint S H, Lindsay D and Biggs P J (2015). Draft genome sequences of three strains of *Geobacillus stearothermophilus* from a milk powder manufacturing plant. *Genome Announcements*. doi:10.1128/genomeA.00939-15.

**Burgesss S A**, Lindsay D and Flint S H (2014). Biofilms of thermophilic bacilli isolated from dairy processing plants and efficacy of sanitisers. In: *Microbial Biofilms: Methods and Protocols*, Methods in Molecular Biology, vol. 1147, pp 367-377, Springer, New York.

**Burgess S A**, Flint S H and Lindsay D (2014). Characterization of thermophilic bacilli from a milk powder processing plant. *Journal of Applied Microbiology*. 11: 350-359.

**Burgess S A,** Lindsay D, Flint S H (2010). Thermophilic bacilli and their importance in dairy processing. *International Journal of Food Microbiology*. 144: 215 – 225.

# List of Presentations

**Burgess S A**, Lindsay D, Flint S H and Biggs P J (2015). Genomic insights into the adaptation of *Geobacillus stearothermophilus* to a milk powder manufacturing plant. Presented at: *New Zealand Microbiological Society Conference, Rotorua, November 2016*.

**Burgess S A**, Lindsay D, and Flint S H (2012). Biofilm formation and spore production of *Geobacillus stearothermophilus*. Presented at: *New Zealand Microbiological Society Conference, Dunedin, November 2012*.

# Acknowledgements

To my supervisors Steve Flint, Denise Lindsay, and Roberto Kolter thank-you for sticking with me over the 6 years it took to complete this journey. Steve, you always have an amazing amount of enthusiasm. Denise, thank-you for your thorough reviews and providing such a supportive environment when things got tough. Roberto, I will be forever grateful for taking me on even though you had never heard of me. Your probing questions at meetings kept me on my toes and the time I spent in your lab opened my eyes to a different world of microbiology.

To my additional supervisor Patrick Biggs, who came on board part-way through my PhD, thank-you for taking me on. Your expertise was a lifesaver and your enthusiasm gave me the boost I needed. I have learnt an enormous amount from you.

Steve Holroyd, without your support I never would have started this PhD. Thank-you for being there as a sounding-board when things got tough.

Hera Vlamakis, you are one of the most wonderful people I've met. Thank-you for making me so welcome in the Kolter lab and for not rolling your eyes when going through some basic microbial genetics.

To everyone in the Kolter lab (Erin, Pascale, Beth, Matt, Diana, Robe, Majse, Christian, Liraz, Maerite, Marta, Diego, and Ilana) thank-you for making me so welcome and providing advice when needed.

Thank-you to Fulbright New Zealand for giving me the opportunity to go to Boston through the Fulbright Science and Innovation Graduate Award.

Ashlee Earle, thank-you for getting me started with the comparative genomics.

Murray Cox, thank-you for generating the draft assemblies. I also really appreciated the thorough reviews of some of my draft manuscripts with some probing questions that got me thinking more outside the square.

Teng Yi Sin, the Massey Singapore student, thank-you for your dedication and hard-work during your project.

Doug and Jordan from the Massey University microscopy unit thank-you for your help and advice with the microscopy imaging.

Jasna Rakonjac and Wesley Wen thank-you for your advice and help with DNA extractions. Jasna I also really appreciated being able to come to your laboratory meetings and for giving advice when needed.

Anne-Marie Jackson and Julia Good I really appreciated you making me feel at ease when moving to the SEAT microbiology lab and helping me out with equipment.

Matt and Steve thank-you for the IT support.

Thank-you to the various people at the Fonterra Cooperative Group who helped out with laboratory work and sampling, including Donna Crowe, Matt Lin, Jacinda Aplin, Saloni Gokarn, Mark Reynolds, Helaina Colle, Micro-servicing staff and plant operators from Kauri and Te Rapa.

To my family I could not have done this without your love and support. To my Dad for initiating some ideas, being there when I needed a sounding board or advice. To my Mum for looking after the kids sometimes, for the dinners every Tuesday and hugs when I needed them.

To my husband Mitch, thank-you so so so much. You have been so supportive, you gave up your job so that I could spend time in the Kolter lab, you have been there for the kids at night so that I could get my sleep and you even built me a computer that made the genomic analyses so much easier. I could not have done this without you. Finally to my two sons Samuel and James, who tolerated many mornings without me.

# Significant contributions from others

- Fatty acid profiling by Saloni Gokarn and Mark Reynolds, Fonterra Research and Development Centre
- 18-h biofilm assays by Massey University Singapore student Teng Yi Sin
- Biofilm SEM imaging by Massey University microscopy unit (MMIC)
- Mi-Seq™ sequencing by New Zealand Genomics Ltd., Massey Genome Service, Massey University
- Draft genome assemblies by Murray Cox
- In-house Perl scripts and reputer analysis showing repetative regions in the genome of *G. kaustophilus* HTA426 prepared by Patrick Biggs

# Table of Contents

Abstract ..... III

List of Publications ..... IV

List of Presentations ..... V

Acknowledgements ..... VI

Significant contributions from others ..... VIII

Table of Contents ..... IX

List of Figures ..... XIV

List of Tables ..... XVII

Non-standard abbreviations ..... XIX

Definitions ..... XXI

Description of computer programs and on-line genomic tools ..... XXIII

1 General introduction ..... 1
- 1.1 Classification of the thermophilic bacilli ..... 1
  - 1.1.1 Taxonomy of the obligate thermophilic bacilli ..... 1
  - 1.1.2 Specific characteristics of the *Geobacillus* genus ..... 2
  - 1.1.3 Specific characteristics of *Anoxybacillus flavithermus* ..... 2
- 1.2 Importance of thermophilic bacilli in dairy processing ..... 3
  - 1.2.1 Spoilage potential of thermophillic bacilli ..... 3
  - 1.2.2 Thermophillic bacilli as hygiene indicators ..... 3
  - 1.2.3 Controlling thermophilic bacteria in dairy processing ..... 4
- 1.3 Endospores of thermophillic bacilli ..... 5
  - 1.3.1 Endospore structure ..... 6
  - 1.3.2 Spore formation ..... 7
  - 1.3.3 Spore resistance ..... 8
  - 1.3.4 Germination of spores ..... 10
- 1.4 Biofilm formation ..... 11
  - 1.4.1 Attachment of spores and vegetative cells ..... 12
  - 1.4.2 Growth of attached populations ..... 13

1.4.3 Spore development within biofilms ... 14

1.5 Enumeration and identification of the thermophillic bacilli in milk powder product ... 14

1.5.1 Plate count methods used to enumerate thermophilic bacilli ... 14

1.5.2 Alternative rapid methods ... 15

1.5.3 Characterization and identification of milk powder isolates ... 16

1.6 Conclusions ... 17

1.7 Objectives of this study ... 17

2 Isolation, identification and characterisation of strains of thermophilic bacilli from a milk powder manufacturing plant ... 19

2.1 Introduction ... 19

2.2 Materials and methods ... 20

2.2.1 Operation of the evaporators ... 20

2.2.2 Sampling regime ... 21

2.2.3 Identification of isolates ... 23

2.2.4 Biochemical characterisation ... 24

2.2.5 Analyses of fatty acid methyl esters (FAMEs) ... 24

2.2.6 Analysis by MALDI-TOF MS ... 24

2.3 Results ... 25

2.3.1 Isolation of strains ... 25

2.3.2 Biochemical characterisation ... 28

2.3.3 Analyses of FAMEs ... 30

2.3.4 MALDI-TOF MS analysis ... 32

2.4 Discussion ... 33

2.5 Conclusions ... 35

3 *In vitro* biofilm formation by *Geobacillus stearothermophilus* on stainless steel ... 38

3.1 Introduction ... 38

3.2 Materials and methods ... 40

3.2.1 Source of strains ... 40

3.2.2 Media and culture preparation ... 40

3.2.3 16-h biofilm and spore production assay ... 41

3.2.4 Plate counts of attached cells ... 41

3.2.5 Crystal violet microtitre plate assay ... 41

3.2.6 80-h biofilm and spore production assay ... 42

3.2.7 Scanning electron microscopy (SEM) ... 43

3.3 Results ... 43

3.3.1 16-h biofilm and spore production assay ... 43

3.3.2 Crystal violet microtitre plate assay ... 44

3.3.3 80-h biofilm and spore production assay ... 47

3.4 Discussion ........ 52
3.5 Conclusions ........ 54

4 Genome sequencing of dairy strains of *Geobacillus stearothermophilus* ........ 55
4.1 Introduction ........ 55
4.2 Materials and methods ........ 57
4.2.1 Genome sequencing of three *G. stearothermophilus* strains isolated from a milk powder plant ........ 57
4.2.2 Genomic DNA extraction ........ 57
4.2.3 Next Generation sequencing and assembly ........ 58
4.2.4 Genome sequences ........ 58
4.2.5 Gene prediction and annotation ........ 60
4.2.6 Prediction of rRNA genes and generation of consensus gene sequences ........ 62
4.2.7 Assigning predicted proteins to Clusters of Orthologous Groups (COGs) ........ 63
4.2.8 Identification of differences in gene presence within the *Geobacillus* genus ........ 63
4.2.9 Identification of genes important for lactose metabolism ........ 66
4.2.10 CRISPR arrays ........ 66
4.3 Results ........ 67
4.3.1 Next generation sequencing and assembly outputs ........ 67
4.3.2 General genomic features ........ 68
4.3.3 Assembly and analysis of the rRNA genes ........ 69
4.3.4 Gene prediction and annotation ........ 74
4.3.5 COG for assigning protein function ........ 77
4.3.6 Differentiation of the *G. stearothermophilus* dairy strains ........ 79
4.3.7 Lactose metabolism ........ 89
4.3.8 Differences between the three dairy strains ........ 93
4.3.9 CRISPR arrays ........ 94
4.4 Discussion ........ 100
4.4.1 Fragmentation of genomes ........ 100
4.4.2 Diversity of the rRNA operons ........ 101
4.4.3 Differentiation of the *G. stearothermophilus* taxon ........ 102
4.4.4 Niche adaptation of the dairy strains of *G. stearothermophilus* ........ 103
4.4.5 Differences between the dairy strains ........ 104
4.4.6 CRISPR arrays ........ 105
4.5 Conclusions ........ 107

5 Comparative analysis of *Geobacillus* taxonomy using phylogenomic tools ........ 108
5.1 Introduction ........ 108

5.2 Materials and methods ..... 109
5.2.1 Genome sequences ..... 109
5.2.2 Phylogenetic tree of 16S rRNA sequences ..... 110
5.2.3 Ribosomal multi-locus sequence typing (rMLST) ..... 110
5.2.4 Average nucleotide identity (ANI) and Tetra nucleotide frequency calculations ..... 110
5.2.5 Core genome comparisons ..... 111
5.2.6 Identification of strains based on the 16S rRNA gene ..... 111
5.3 Results ..... 111
5.3.1 16S rRNA gene sequencing and rMLST ..... 111
5.3.2 ANI and TETRA calculations ..... 115
5.3.3 Core genome comparisons ..... 117
5.3.4 Classification of *Geobacillus* species ..... 121
5.4 Discussion ..... 122
5.5 Conclusions ..... 125

6 Genomic insights into biofilm formation and sporulation of *Geobacillus* spp. ..... 127
6.1 Introduction ..... 127
6.2 Materials and methods ..... 132
6.2.1 Genome sequences ..... 132
6.2.2 Gene lists ..... 135
6.2.3 Identification of homologues ..... 136
6.2.4 Catergorization of histidine kinases ..... 138
6.3 Results and discussion ..... 138
6.3.1 Activation of the phosphorelay ..... 139
6.3.2 Biofilm formation ..... 146
6.3.3 Sporulation ..... 156
6.3.4 Summary of results ..... 185
6.3.5 Reliability of results ..... 187
6.4 Conclusions ..... 188

7 Conclusions and summary ..... 190
7.1 Strain diversity within a dairy manufacturing plant ..... 190
7.2 Taxonomic classification ..... 191
7.3 Genomics insight into growth and survival in a dairy manufacturing environment ..... 191
7.4 Summary ..... 192
7.5 Practical contributions to the dairy industry ..... 193
7.6 Future directions ..... 193

8 Bibliography ..... 195

Appendix 1 Preparing a BLAST database ........ 235

Appendix 2 Mapping using Bowtie 2 ........ 237

Appendix 3 Identifying orthologous clusters ........ 238

Appendix 4 Cumulative frequency of trimmed read lengths ........ 239

Appendix 5 General genomic features of published *Geobacillus* strains ........ 243

Appendix 6 Core genome sequence comparisons ........ 244

Appendix 7 Heat-maps to visualize the presence/absence of homologues in *Geobacillus* spp. ........ 245

Appendix 8 Gene organisation of the *skin* elements ........ 247

Appendix 9 BLASTX analysis of the *rsfA* and *gerR* genes ........ 248

# List of Figures

Figure 1.1. Structure of a *B. subtilis* spore, taken from McKenney *et al.* (2013). 7
Figure 1.2. Schematic diagram of the sporulation process, taken from McKenney *et al.* (2013). 7
Figure 2.1. Schematic diagram of the milk powder processing plant used for this study. 21
Figure 2.2. Schematic diagram of one MVR effect from an evaporator. 22
Figure 2.3. Thermophile plate counts (cfu/ml) of milk samples taken at the start (black bars) and end (grey bars) of a nutrional milk powder processing cycle. 26
Figure 2.4. Variation in the percentage of six fatty acids across five *Geobacillus stearothermophilus* strains. 32
Figure 2.5. Dendrogram based on MALDI-TOF spectra of the *G. stearothermophilus* isolates. 33
Figure 3.1. Total thermophile counts (solid filled bars) and thermophilic spore counts (pattern filled bars) attached to the surface of a stainless steel coupon after 16 h of incubation at 55°C in RSM. 43
Figure 3.2. SEM images of three selected *G. stearothermophilus* strains: P3 (A), D1 (B) and A1 (C); grown on the surface of stainless steel in RSM. 44
Figure 3.3. Biofilm forming index based on the $OD_{590nm}$ obtained with the CV assay for the three selected strains of *G. stearothermophilus*. 45
Figure 3.4. Planktonic cell count of three selected strains of *G. stearothermophilus* after 16-h growth in modified milk medium with the CV assay. 46
Figure 3.5. Biofilm formation visualised by CV staining of three selected strains of*G. stearothermophilus* on the surface of a 6-well microtitre plate, grown in modified milk medium for 18 h. 46
Figure 3.6. Total thermophile counts (A) and thermophilic spore counts (B) of three selected *G. stearothermophilus* strains attached to the surface of a stainless steel coupon over an incubation period of 80 h at 55°C in 1/5 diluted RSM based medium. 48
Figure 3.7. pH readings of the medium, before it was replaced throughout an 80 h incubation period, used for growing biofilms of three selected strains of *G. stearothermophilus.* 49
Figure 3.8. SEM images with a magnification of 1300x of three selected *G. stearothermophilus* strains: A1, P3 and D1; grown on the surface of stainless steel in 1/5 diluted RSM based medium over an 80 h incubation period at 55°C. 51
Figure 3.9. SEM images with a magnification of 5000x of biofilms after 56 h of incubation of the three selected *G. stearothermophilus* strains strains: A1, P3 and D1; grown on the surface of stainless steel. 52

Figure 3.10. SEM images with a magnification of 5000x showing spore formation within biofilms of the three selected *G. stearothermophilus* strains: A1 P3 and D1; after 8 h of incubation. ........ 52
Figure 4.1 Lactose utilisation in LAB........ 56
Figure 4.2. Overall process of mode 1 from the in-house Prokka script for annotating genomes ........ 62
Figure 4.3. Overall process of the in-house OrthoMCL script........ 64
Figure 4.4. Reputer analysis showing repetative regions greater than or equal to 150bp in the genome of *G. kaustophilus* HTA426........ 68
Figure 4.5. Alignment of the 5S rRNA copies. (A) A1 and (B) D1. ........ 70
Figure 4.6. Genetic context of a 5S rRNA gene from strain D1, which sits by itself, compared with the split rRNA operon found in *G. kaustophilus* HTA426........ 70
Figure 4.7. Gene organisation of a region of the genome for *G.stearothermophilus* strain P3 containing putative biofilm genes. ........ 75
Figure 4.8. Gene organisation of a region of the genome for *G.stearothermophilus* strain P3 containing a CRISPR array. ........ 76
Figure 4.9. Proportion of coding sequences assigned to a protein function for the three *G.stearothermophilus* dairy strains (coloured in red) compared with other strains of *Geobacillus* spp. and *B. subtilis*. ........ 77
Figure 4.10. The conservation of selected genes involved in the utilisation of lactose and galactose. ........ 90
Figure 4.11. Comparison of the organisation of the *lac* genes. ........ 92
Figure 4.12. Organisation of the CRISPR arrays and associated genes in the dairy strains A1, P3 and D1........ 97
Figure 4.13. Organisation of the spacers in each CRISPR array. ........ 98
Figure 5.1. Un-rooted maximum likelihood phylogenetic trees for the *Geobacillus* genus........ 113
Figure 5.2. Heat-map comparison of 16S rRNA gene percentage identities. ........ 114
Figure 5.3. Heat map comparison of the ANIm values. ........ 116
Figure 5.4 Core genome sequence comparisons. ........ 120
Figure 6.1. Initiation of biofilm formation and sporulation through activation of Spo0A. ........ 128
Figure 6.2. Simplified schematic of the sporulation sigma factor regulatory network. ........ 130
Figure 6.3. Conservation of those genes involved in the phosphorelay and their equivalent COGs in *Geobacillus* spp. ........ 140
Figure 6.4. Domain organization of the orphan HKs from (A) *B. subtilis* and (B) *Geobacillus* strains A1 (Types 2, 3, 4 and 5) and NUB3621 (Type 1, 6 and 7). ........ 144
Figure 6.5. Alignment of the sequence surrounding the phosphorylated histidine of orphan HKs from *B. subtilis* 168, *G. stearothermophilus* A1 and *Geobacillus* sp. NUB3621........ 144
Figure 6.6. Conservation of selected *B. subtilis* biofilm genes in *Geobacillus* spp. and their equivalent COGs. ........ 148

Figure 6.7. Gene organisation of biofilm genes in *B. subtilis* and putative exopolysaccharide genes in selected *Geobacillus* strains and *B. cereus* ATCC 14579. ............................ 152
Figure 6.8. Gene organisation of the *ica* operon in *S. aureus* and the homologue of *icaA* in the dairy strains *G. stearothermophilus* A1 and D1...................................................... 155
Figure 6.9. Conservation of selected *Bacillus subtilis* sporulation genes in the *Geobacillus* genus...................................................................................................................... 158
Figure 6.10. Conservation of selected *B. subtilis* regulatory network genes in *Geobacillus* spp. and their equivalent COGs. ..................................................................... 161
Figure 6.11. Gene organisation of *sigK* in *B. subtilis* 168 and selected *Geobacillus* strains. ................. 164
Figure 6.12. Organisation of selected *B. subtilis* genes encoding for proteins involved with engulfment and their homologues in *Geobacillus*. ....................................................... 167
Figure 6.13. Conservation of selected *B. subtilis* cortex genes in *Geobacillus* spp. and their equivalent COGs. ........................................................................................... 168
Figure 6.14. Gene organisation of *coxA* (coloured in green) and *safA* (coloured in blue) in *B. subtilis* and their homologues in two strains of *Geobacillus*: *G. stearothermophilus* A1 and *Geobacillus sp.* Y4.1MC1 (Taxon group 10, *G. thermoglucosidasius*)........................................................................................... 170
Figure 6.15. Conservation of selected *B. subtilis* spore coat genes in *Geobacillus* spp. and their equivalent COGs. .................................................................................... 172
Figure 6.16. Gene organisation of the *sps* operon in *B. subtilis* and putative *sps* operons in selected strains of *Geobacillus* spp. ............................................................... 176
Figure 6.17. Collagen-like glycoproteins........................................................................................... 178
Figure 6.18. Gene organisation of the *spoIIAA-sigF* and *spoVA* operons of *B. subtilis* and their homologues (as depicted by the different colours) in *Geobacillus*. ...................... 180
Figure 6.19. Presence/absence of genes encoding for SASPs.............................................................. 183
Figure 6.20. Conservation of *B. subtilis* mother-cell lysis genes in *Geobacillus* spp. and their equivalent COGs. ........................................................................................... 184

# List of Tables

Table 2.1. Identification of strains isolated from the evaporator ... 27
Table 2.2. Biochemical reactions of thermophile *G. stearothermophilus* isolates based on the Microgen Bacillus-ID, BBL Crystal and API 50CHB identification kits ... 29
Table 2.3 FAME percentage composition[*] of the thermophile *G. stearothermophilus* isolates compared with the type strain *G. stearothermophilus* ATCC 12980 ... 31
Table 3.1. Bacterial strains used in this study. ... 40
Table 3.2. Percentage of the total thermophile counts that were spores attached to the surface of a stainless steel coupon for the three selected *G. stearothermophilus* strains. ... 49
Table 4.1. Bacterial genomes[a] used in this study ... 59
Table 4.2. Options for the in-house Prokka script ... 61
Table 4.3. Description of the MySQL tables generated from the in-house Prokka script ... 62
Table 4.4. Options for the in-house OrthoMCL script ... 64
Table 4.5. Description of the MySQL tables generated by the in-house OrthoMCL script ... 65
Table 4.6. Coverage for each G. *stearothermophilus* genome sequenced ... 67
Table 4.7. Assembly statistics ... 67
Table 4.8. General features of the draft genomes for the sequenced *G. stearothermophilus* strains A1, D1 and P3. ... 69
Table 4.9. Variant frequency of polymorphisms in the 5S rRNA gene sequences of the *G. stearothermohilus* dairy strains. ... 70
Table 4.10. Diversity of the 5S rRNA genes. ... 71
Table 4.11. Variant frequency of polymorphisms in the 16S rRNA gene sequences of the *G. stearothermophilus* dairy strains. ... 72
Table 4.12. Variant frequency of polymorphisms in the 23S rRNA gene sequences of the *G. stearothermophilus* dairy strains. ... 73
Table 4.13. Diversity of the 16S rRNA genes. ... 74
Table 4.14. Diversity of the 23S rRNA genes. ... 74
Table 4.15. Comparison of annotation descriptions between RAST and Prokka for a region of the genome of strain P3 containing putative biofilm genes. ... 75
Table 4.16. Comparison of annotation descriptions between RAST and Prokka for a region of the genome of strain P3 containing a CRISPR array. ... 76
Table 4.17. Selected genes only found in *G. stearothermophilus* A1, P3 and D1 ... 80
Table 4.18. Selected genes absent in *G. stearothermophilus* A1, P3 and D1 ... 81
Table 4.19. COGs absent in *G. stearothermophilus* A1, P3 and D1 and present in the other *Geobacillus* spp. ... 82
Table 4.20. Selected genes absent in *G. stearothermophilus* A1, P3, D1 or ATCC 7953 but present in the other *Geobacillus* spp ... 83

Table 4.21. COGs absent in *G. stearothermophilus* and present in the other *Geobacillus* spp. .... 88
Table 4.22. BLASTX of *G. stearothermophilus* A1 *lac* cluster components .... 93
Table 4.23. CRISPR arrays found in the *Geobacillus* genus. .... 94
Table 4.24. Characteristics of the CRISPR arrays .... 99
Table 5.1. Number of genes and amino acids used in the OrthoMCL clustering. .... 118
Table 5.2. Identification of *Geobacillus* strains using the RDP. .... 121
Table 5.3. Identity of *Geobacillus* strains. .... 122
Table 6.1. Genes involved in production of the extracellular matrix in *B. subtilis*. .... 129
Table 6.2. *Geobacillus* genomes used in this study. .... 133
Table 6.3. Reference genomes used in this study. .... 134
Table 6.4. Genes encoding for exosporium proteins of *B. anthracis*. .... 136
Table 6.5. Description of each COG, which contained a protein involved in the *B. subtilis* phosphorelay. .... 141
Table 6.6. Description of each COG, which contained a protein involved in biofilm formation of *B. subtilis*. .... 150
Table 6.7. Putative *eps* cluster in *B. cereus* and their homologues in *B. subtilis* and presence in *Geobacillus* spp. .... 153
Table 6.8. Description of each COG, which contained a protein involved in *B. subtilis* spore regulation. .... 162
Table 6.9. Genomes containing a split *sigK* gene. .... 163
Table 6.10. Description of each COG, which contained proteins involved with cortex formation. .... 169
Table 6.11. Description of each COG which contained proteins involved with *B. subtilis* coat formation. .... 173
Table 6.12. Homologues of *B. cereus* group genes encoding for exosporium proteins. .... 177
Table 6.13. Description of COGs which contain proteins involved with mother-cell lysis. .... 184
Table 6.14. Summary of the key components of the phosphorelay, biofilm and sporulation pathways conserved in *Geobacillus* spp. .... 186

# Non-standard abbreviations

| | |
|---|---|
| $A_w$ | Water activity |
| ANI | Average nucleotide identity |
| BDBH | Bidirectional best hit |
| CIP | Clean-in-place |
| COG | Clusters of orthologous groups |
| DPA | Dipicolinic acid |
| DSI | direct steam injection |
| EOR | End-of-run |
| GFF | General file format |
| HK | Histidine kinase |
| MALDI-TOF | Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry |
| MCL | Markov clustering |
| MLVA-HRM | Multi-locus variable-number analysis - high-resolution melt analysis |
| MVR | Mechanical vapour recompression (evaporator) |
| n/d | Not determined |
| PHE | Plate heat exchanger |
| rMLST | Ribosomal multi-locus sequence typing |
| SFB | Static fluid bed |
| SOR | Start-of-run |
| THE | Tubular heat exchanger |
| TM | Transmembrane |

| | |
|---|---|
| T/S | Total solids |
| TVR | Thermal vapour recompression (evaporator) |

# Definitions

| | |
|---|---|
| Accessory genome | Additional genes that are present in some members and absent from others within a group of isolates under investigation. |
| Clean-in-place (CIP) | Cleaning regime after a manufacturing run. |
| Conditioning layer | The thin layer of proteins and exopolysaccharides that forms immediately on a surface when it is submersed in a liquid. |
| Core genome | A set of genes shared by all members in a group of isolates under investigation. |
| Direct steam injection (DSI) | A direct method of heat treatment where steam is injected into the milk. |
| Effect | A section of an evaporator that has the same boiling temperature. |
| Engulfment | Part of the sporulation process where degradation of the septal membrane (between the mother-cell and forespore), and relocation of the mother-cell membrane around the forespore occurs. |
| Forespore | The immature form of the spore when it is being formed within the mother cell. |
| Foulant | The build-up of milk proteins and calcium phosphate salts on equipment surfaces in dairy manufacturing plants. |
| Homologue | Genes that are descendents of the same ancestral gene but were separated by either speciation or gene duplication. |
| Orifice pans | Located at the top of the evaporator to distribute milk into the pass tubes. |
| Mother cell | The cell which houses the forespore as it matures into an endospore. |
| Orthologue | Genes in different species which were derived from the same ancestral gene and were separated by speciation. |
| Paralogue | Genes that are descendents of the same ancestral gene, but were separated by gene duplication. |
| Pass | A section of the effect, in the evaporator, that is made up of a set of |

| | |
|---|---|
| | tubes that the milk passes through. |
| Plate heat exchanger (PHE) | An indirect method of heat treatment that consists of a series of plates where the heating or cooling medium passes on one side, and the milk on the other. |
| Pseudogene | A “gene” which has lost its ability to code for a protein, generally through the accrual of mutations such as stop codons or frameshifts within the gene. |
| Sliding | The passive movement of bacteria across a surface. This process does not make use of bacterial appendages such as flagella or pilli. |
| Spore coat | The outer layers of the endospore. |
| Spore cortex | The layer between the inner and outer membranes of the spore and is composed of peptidoglycan. |
| Spore crust | The outer layer of the coat in spores of *Bacillus subtilis.* |
| Spore exosporium | The outermost layer of spores in some species of *Bacillus*. It is composed of glycoprotein and separated from the coat by a large irregular space. |
| Swarming | The coordinated movement, through the use of flagella, of a bacterial population across a surface. |
| Water activity | In the dairy context this refers to the amount of water not bound to food molecules. This water can enable the growth of bacteria. When milk powder is made the water activity decreases through the evaporators as the milk is concentrated and once dried reaches levels that no longer supports bacterial growth. |

# Description of computer programs and on-line genomic tools

| | |
|---|---|
| Bowtie 2 | An alignment program used for aligning short sequences (e.g. sequence reads from a genome sequencer) to long sequences (e.g. genome sequences) (Langmead & Salzberg, 2012). The output generated by Bowtie 2 is a SAM file. |
| COGnitor | A software tool designed to assign predicted proteins to the already established COGs (Tatusov *et al.*, 2000, Galperin *et al.*, 2015). |
| CRISPRDetect | An on-line tool (http://brownlabtools.otago.ac.nz/CRISPRDetect/predict_crispr_array.html), designed to detect the presence of CRISPR arrays (Biswas *et al.*, 2014). |
| CRISPRTarget | An on-line tool (http://brownlabtools.otago.ac.nz/CRISPR_WEB/crispr_analysis.html), designed to determine the target of CRISPR spacers (Biswas *et al.*, 2013). |
| GET_HOMOLOGUES | A software package that incorporates three different algorithms (BDBH, COGtriangles and OrthoMCL) for clustering homologous genes (Contreras-Moreira & Vinuesa, 2013). |
| Jspecies | A software package designed for comparing the similarity of two or more bacterial species (Richter & Rossello-Mora, 2009). Synthetic DNA-DNA hybridisations can be carried out using three methods: Average nucleotide identity (ANI) calculated using BLAST, ANI calculated using MUMmer and calculation of tetra nucleotide frequencies (TETRA). |
| OrthoMCL | A software program which uses an algorithm incorporating both BLASTP and the Markov clustering algorithm to determine orthologous groups of proteins within a group of genomes (Li *et al.*, 2003). |
| Pfam | Pfam (http://pfam.xfam.org/) is a database of protein families. In this present study it was used for identifying domains in predicted protein sequences. |
| Prokka | Prokka is a software package used for rapidly annotating prokaryotic |

| | |
|---|---|
| | genomes (Seemann, 2014). |
| Rapid Annotation using Subsystem Technology (RAST) | A web based server (http://rast.nmpdr.org/), which can carry out automated annotations on bacterial genomes (Aziz *et al.*, 2008). |
| RNAmmer | A web based server (http://www.cbs.dtu.dk/services/RNAmmer/), used to predict prokaryotic and eukaryotic rRNA gene sequences in genome sequences (Lagesen *et al.*, 2007). |
| Velvet | An algorithm package used for *de novo* genome assembly (Zerbino & Birney, 2008). In assembling, the sequence reads are broken into shorter sequences called k-mers and used to generate de Bruijn graphs. A range of k-mer lengths are tested to generate the best assembly. |

# Chapter 1

This chapter is partially based on the publication by Burgess *et al.* (2010).

# 1 General introduction

The thermophilic bacilli are potential contaminants in a variety of industries where elevated temperatures (40–65 °C) prevail during the manufacturing process or when product is stored. These industries include paper mills, canning, juice pasteurisation, sugar refining, gelatin production, dehydrated vegetable manufacture and dairy product manufacture (Denny, 1981, Hayes, 1985, Splittstoesser *et al.*, 1998, De Clerck *et al.*, 2004a, Suihko *et al.*, 2004, Tai *et al.*, 2004, Chen *et al.*, 2006, Scott *et al.*, 2007). The focus of this review is the growth and spoilage potential of thermophilic bacilli in dairy manufacturing plants, with particular emphasis on milk powder processing.

## 1.1 Classification of the thermophilic bacilli

In the dairy industry, the thermophilic bacilli are usually enumerated using an aerobic plate count (APC) incubated at 55 °C. Those that have been isolated from dairy products at this incubation temperature can be divided into two groups: the obligate thermophiles and the facultative thermophiles (also known as thermotolerant microorganisms). The obligate thermophiles grow only at elevated temperatures (approximately 40-68 °C) and include *Anoxybacillus flavithermus* and *Geobacillus* spp. (Flint *et al.*, 2001a, Ronimus *et al.*, 2003, Scott *et al.*, 2007). The facultative thermophiles belong to the *Bacillus* genus and tend to grow at both mesophilic (approximately 20-45 °C) and thermophilic temperatures, depending on the strain. Some examples of species include *Bacillus licheniformis*, *Bacillus coagulans*, *Bacillus pumilus*, *Bacillus sporothermodurans* and *Bacillus subtilis* (Crielly *et al.*, 1994, Flint *et al.*, 2001a, Ronimus *et al.*, 2003, Scheldeman *et al.*, 2006). The main focus of this review will be on the obligate thermophiles.

### 1.1.1 Taxonomy of the obligate thermophilic bacilli

Taxonomic studies have provided evidence that the genus *Bacillus* is very diverse (Ash *et al.*, 1991, Rossler *et al.*, 1991, Rainey *et al.*, 1994). Throughout the 1980s, there was debate about whether the genus *Bacillus* was truly one genus or many, and 16S rRNA sequence analysis as well as DNA−DNA hybridisation studies became key techniques for defining species (Fox *et al.*, 1977, Sneath, 1986, Tindall *et al.*, 2010). Work throughout the 1990s resulted in the reclassification of species within the *Bacillus* genus into new genera (Wisotzkey *et al.*, 1992, Heyndrickx *et al.*, 1996, Shida *et al.*, 1996, Heyndrickx *et al.*, 1997, Heyndrickx *et al.*, 1998, Waino *et al.*, 1999, Yoon *et al.*, 2002, Heyrman *et al.*, 2003, Goto *et al.*, 2004, Yoon *et al.*, 2004, Ren & Zhou, 2005). A taxonomic study led to *Bacillus stearothermophilus* being reclassified into the new genus *Geobacillus* (Nazina *et al.*, 2001). The isolation of *Anoxybacillus pushchinoensis* led to *Bacillus flavothermus* being reclassified as *A. flavithermus* (Pikuta *et al.*, 2000). Prior to its

reclassification, *A. flavithermus* was not recognised as a valid species. In the dairy industry *A. flavithermus* may have been mistakenly identified as *B. stearothermophilus* prior to 2001.

### 1.1.2 Specific characteristics of the *Geobacillus* genus

The *Geobacillus* genus includes the thermophilic species that were previously classified in rRNA Group 5 of the *Bacillus* genus (Ash *et al.*, 1991). Individual species within the genus are very closely related, with a 96.5–99.5% 16S rDNA sequence similarity (White *et al.*, 1994, Nazina *et al.*, 2001, Weng *et al.*, 2009). The G+C content ranges between 45 and 69%, they produce sub-terminal or terminal spores, with the optimum growth temperature of this genus being above 50 °C. *Geobacillus* strains have been isolated from temperate areas, as well as hot environments, such as hot springs, oilfields, deep sea sediments, sugar refineries, canned foods, dehydrated vegetables and dairy factories (Zarilla & Perry, 1987, Nazina *et al.*, 2001, Marchant *et al.*, 2002, Ronimus *et al.*, 2003, Tai *et al.*, 2004, Takami *et al.*, 2004a, Scott *et al.*, 2007, Postollec *et al.*, 2012, Luecking *et al.*, 2013). *G. stearothermophilus* is the predominant species to have been isolated from milk powder (Stadhouders *et al.*, 1982, Rueckert *et al.*, 2004, Hill & Smythe, 2012, Seale *et al.*, 2012, Yuan *et al.*, 2012, Luecking *et al.*, 2013, Sadiq *et al.*, 2016) [For simplicity, all studies previously referring to as *B. stearothermophilus* are referred to in this review as *G. stearothermophilus*]. *Geobacillus kaustophilus* and *Geobacillus thermoglucosidasius* have also been isolated from milk and milk manufacturing plants (Priest *et al.*, 1988, Zhao *et al.*, 2013).

### 1.1.3 Specific characteristics of *Anoxybacillus flavithermus*

This bacterium was first isolated from hot springs in New Zealand (Heinen *et al.*, 1982), and has since been isolated from hot springs around the world (Nold *et al.*, 1996, Belduz *et al.*, 2000, Matsutani *et al.*, 2013, Rozanov *et al.*, 2014, Jiang *et al.*, 2015, Khalil *et al.*, 2015), as well as from gelatin extracts (De Clerck *et al.*, 2004b) and milk powder (Flint *et al.*, 2001a, Ronimus *et al.*, 2003, Scott *et al.*, 2007, Reginensi *et al.*, 2011, Caspers *et al.*, 2013, Sadiq *et al.*, 2016). The type species, *A. pushchinoensis*, was first described as a strict anaerobe, hence the name *Anoxybacillus* (Pikuta *et al.*, 2000). This description has since been revised as members of this genus appear to be aerotolerant or facultative anaerobes (Pikuta *et al.*, 2003).

*A. flavithermus* is described as a facultatively anaerobic thermophile, is motile, has terminal endospores, has a G+C content of 61% and grows in the pH range 6−9 (Heinen *et al.*, 1982). Characterisation of the New Zealand hot spring strain showed that the temperature growth range is between 30 and 70 °C with an optimum growth temperature of 60 °C. However, isolates obtained from milk powder tend to have an optimum growth temperature ranging between 50 and 65 °C (Cucksey, 2002, Ronimus *et al.*, 2003, Scott, 2005). It has been demonstrated that incubation at 68 °C is a useful approach for differentiating between *A. flavithermus* and *Geobacillus* strains isolated from milk powder, with *A. flavithermus* strains generally being unable to grow at 68 °C (Hill and Smythe, 2003, unpublished results, Fonterra Research Centre, Palmerston North, New Zealand).

## 1.2 Importance of thermophilic bacilli in dairy processing

Within the dairy processing context, thermophilic bacilli are used as hygiene indicators in processed product. This is because the presence of these organisms is usually an indicator of inadequate cleaning. In addition, thermophiles are potential spoilage organisms, as they are capable of producing enzymes and acids that may lead to off-flavours in final product. Thermophilic bacilli can grow in a variety of dairy manufacturing plants, but their growth in milk powder manufacturing plants has been the main focus of previous studies.

### 1.2.1 Spoilage potential of thermophillic bacilli

Strains of obligate and facultative themophiles are capable of producing acids, as well as a variety of heat-stable enzymes, including proteinases and lipases, which could result in the spoilage of dairy products (Basappa, 1974, Chopra & Mathur, 1984, Cosentino *et al.*, 1997, Chen *et al.*, 2004, Gundogan & Arik, 2004, Murugan & Villi, 2009). The real potential for the obligate thermophiles to spoil dairy products is thought to be low, as dairy products are generally stored at temperatures below 37 °C, temperatures at which obligate thermophiles will not grow. In the case of milk powder the water activity is too low for growth. However, heat-stable enzymes produced during the manufacture of milk powder have the potential to cause off flavours in the final product. *G. stearothermophilus* has been associated with ‘flat-sour’ spoilage in a variety of canned food products, including evaporated milk (Kalogridouvassiliadou, 1992).

In the case of facultative thermophiles, some strains of *B. licheniformis* are also capable of producing a slimy extracellular substance that can affect the quality of pasteurised milk and cream (Gilmour & Rowe, 1990). *B. subtilis* has been associated with ropiness in raw and pasteurised milk as well as the spoilage of UHT and canned milk products (Heyndrickx & Scheldeman, 2002). *B. coagulans* has been connected to the spoilage of UHT and canned milk products due to the production of lactic acid (Gilmour & Rowe, 1990).

The obligate thermophiles are not known to be pathogenic. However, some of the facultative thermophiles including *B. licheniformis*, *B. pumilis* and *B. subtilis* can produce toxins, although this has been studied only at mesophilic temperatures (Oh & Cox, 2009, De Jonghe *et al.*, 2010). On rare occasions, these three microorganisms have been implicated in food poisoning incidents (Griffiths, 1995, Pavic *et al.*, 2005).

### 1.2.2 Thermophillic bacilli as hygiene indicators

The presence of high numbers (> $10^4$ cfu/g) of thermophiles in finished dairy products, like milk powders, is an indicator of poor hygiene during processing. Initial thermophile contaminants are believed to enter a dairy manufacturing process via the raw milk in the form of endospores. However, the quality of the raw milk in terms of thermophile numbers is unrelated to the number of thermophiles present in the final milk powder product (Muir *et al.*, 1986). In fact, the levels of thermophiles in raw milk are usually very low (e.g. < 10 cfu/ml) (Hill & Smythe, 1994, McGuiggan *et al.*, 2002, Miller *et al.*, 2015a). On rare occasions when higher levels do occur (e.g. > 100 cfu/ml), facultative thermophiles such as *B. licheniformis* , *B. pumilus* and *B.*

*coagulans* tend to be the predominant thermophilic species (Hill B, 1998, unpublished results, Fonterra Research Centre, Palmerston North, New Zealand; Miller et al., 2015b). However, when final product from raw milk of this quality is processed, the predominant bacterial population changes to *A. flavithermus* and *Geobacillus* spp (Miller *et al.*, 2015b).

Low numbers of spores in the incoming raw milk and a short residence time (< 30 min) of milk in a dairy processing plant cannot account for resulting high numbers ($10^4$–$10^6$ cfu/g) of thermophilic bacilli in final dairy products. However, growth of these bacteria within the manufacturing plant itself does account for this phenomenon. In particular, thermophillic bacilli have been found growing as biofilms on the surfaces of processing equipment or within milk foulant layers. Extensive growth may occur when production cycles are too long, the manufacturing equipment is not cleaned properly between production cycles, recycle loops are used and ingredients or by-products that contain thermophiles are used. Some examples of dairy processes in which thermophile growth has been an issue include milk powder, pasteurised milk, buttermilk and whey (Muir *et al.*, 1986, Langeveld *et al.*, 1990, Flint *et al.*, 1997, Scott *et al.*, 2007).

### 1.2.3 Controlling thermophilic bacteria in dairy processing

Current methods for controlling thermophilic bacilli and their biofilm growth in dairy manufacturing plants include shorter production lengths (therefore increasing the cleaning frequency), the use of sanitisers (disinfectants), altering temperatures, reducing the surface area in the optimal temperature growth zone and the use of dual equipment.

Limiting the production length of milk treatment to 6–8 h has reduced thermophile growth in centrifugal separators and plate heat exchangers (PHEs) used for pasteurisation. In milk powder manufacturing plants, production lengths are usually limited to 18–24 h to limit thermophile growth. Between each production cycle, the wet sections of dairy manufacturing plants are cleaned, using a cleaning-in-place (CIP) regime which consists of a warm water rinse, a 1.5 % caustic wash at 75 °C for 30 min, followed by a water rinse, then a 0.5 % nitric acid wash at 70 °C for 20 min, and finishes with a second water rinse (Bylund, 1995). The caustic wash is used to remove organic matter and the nitric acid wash is used to remove inorganic matter. In some dairy manufacturing plants, a sanitiser is also used to inactivate any remaining microbial populations on the equipment surfaces.

The benefit of using sanitisers and the ability of CIP procedures to remove biofilms in milk processing lines are still largely unknown and much debated. After a CIP procedure microorganisms may remain on the surface even though it may appear visibly clean (Watkinson, 2008). The use of a sanitiser is important to remove and kill remaining microbial cells attached to the surface. Opinion on the ability of CIP procedures without sanitisers to remove biofilms and spores is mixed. For example, Lindsay *et al.* (2000) isolated various *Bacillus* spp., in particular *Bacillus cereus*, from the caustic soda wash solutions re-used in dairy CIP procedures. Parkar *et al.* (2003) demonstrated that a combined caustic and acid wash consisting of 2% sodium hydroxide solution (75 ºC for 30 min) and 1.8% nitric acid (75 ºC for 30

min) optimally removed biofilms of thermophilic bacilli. However, changing the temperatures of application, or the concentrations of the caustic/acid combination drastically reduced the ability of the cleaning procedure to remove biofilm cells. The above described combination of sodium hydroxide and acid washes are reportedly also sporicidal (Knight & Weeks, 2008, Seale *et al.*, 2011).

A non-chemical approach to controlling dairy biofilms has been shown by altering the temperature of a process. For example, cold (approximately 15–30 °C) processing in ultra-filtration plants and in centrifugal separators, used to separate cream from whole milk, has resulted in extended production cycles and reduced thermophile contamination. Temperature disruption has been shown to control the formation of *Streptococcus thermophilus* biofilms in PHE pasteurisers of milk used in cheese production (Knight *et al.*, 2004). A similar approach may also be feasible for controlling thermophlic biofilms, as Burgess *et al.* (2009) demonstrated that lowering the temperature of the growth environment from 55 to 48 °C prevented the formation of spores in biofilms of *A. flavithermus*. A temperature of 55°C is often encountered *in situ* in dairy manufacturing.

Reducing the surface area and using dual preheating systems are two further methods that are commonly used in milk powder manufacturing plants to reduce thermophile growth (Refstrup, 1998, Refstrup, 2000). Reducing the surface area in the optimal temperature growth zone of thermophiles is achieved by using direct contact heating systems, such as direct steam injection (DSI). A DSI system involves injecting steam into the milk, which rapidly increases the temperature of the milk, avoiding the optimum growth temperature for thermophiles. This system is very expensive compared with indirect heating systems such as PHEs. The use of DSI also results in a dilution of the milk due to added water; therefore flash evaporation is required after the DSI to remove the additional water. An alternative is to use a dual preheating system, in which the milk is directed from one preheater to a second preheater after 8–12 h of processing. This allows the first preheater to undergo a CIP procedure without disrupting the manufacturing cycle.

Future developments for preventing biofilms of thermophilic bacilli in dairy manufacturing plants are likely to focus on altering the manufacturing conditions (such as temperature), manipulating the surface of stainless steel to reduce bacterial attachment and developing novel sanitisers (Bremer *et al.*, 2009, Seale *et al.*, 2015).

## 1.3 Endospores of thermophillic bacilli

Sporulation is a natural process in the growth cycle of certain groups of bacteria such as *Bacillus* species. It is a survival mechanism, generally regarded as a process that occurs when the organism is under stress. In the dairy environment, thermophilic bacilli endospore (henceforth called spore) formation occurs readily, but the factors contributing to this process are not clearly understood. Most of the work carried out on spores has focused on the mesophilic bacilli, *B. cereus* and *B. subtilis*. The structure, sporulation and germination process,

as well as the resistance mechanisms of spores, is presumed to be similar for the thermophilic bacilli. Genomic comparisons have shown that those genes regarded as essential for sporulation in *B. subtilis* are generally conserved in *Geobacillus* spp. (de Hoon *et al.*, 2010, Galperin *et al.*, 2012, Abecasis *et al.*, 2013).

Spore formation of the thermophilic bacilli enables them to survive the harsh conditions of a dairy manufacturing environment as well as long periods in milk powder. For example, viable spores of *B. subtilis* and *B. licheniformis* were isolated from 90-year-old milk powder (Ronimus *et al.*, 2006). Understanding how spore formation occurs is important as this knowledge may help to control spores in dairy processing plants.

### 1.3.1 Endospore structure

Spores consist of a core, which contains the nuclear material, calcium dipicolinate (Ca-DPA) and small acid-soluable proteins (SASPs), surrounded by the cortical membrane and cortex, which is in turn enclosed in the spore coat (Figure 1.1). In *B. subtilis* the outermost layer is called the crust (McKenney *et al.*, 2013). Other *Bacillus* species have a layer over the spore coat called the exosporium. The exosporium has been most well characterised in *B. cereus* and *B. anthracis*. It consists of a base layer covered with “hair-like” protrusions called the nap (Stewart, 2015). Transmission electron microscopy has also provided evidence of an exosporium in a dairy strain of *G. stearothermophilus* (Seale *et al.*, 2010). Some species such as *B. cereus* have appendages (Ankolekar & Labbe, 2010). The major difference in spore structure between species is the structure and number of layers in the spore coat, whereas the cortex and core are very similar (Aronson & Fitzjames, 1976, Atrih & Foster, 2001).

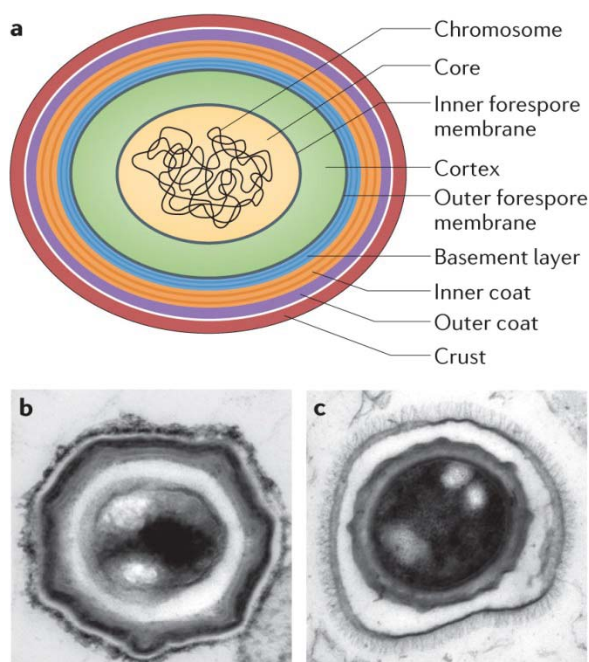


**Figure 1.1. Structure of a *B. subtilis* spore, taken from McKenney *et al.* (2013).** (**A**) Schematic diagram of the spore structure. Electron micrographs of a (**B**) *B. subtilis* spore stained with ruthenium red and (**C**) a *B. cereus* spore stained with osmium tetroxide.

### 1.3.2 Spore formation

Sporulation is a complicated process, divided into a series of seven steps (Figure 1.2), which is regarded as being very similar amongst aerobic and facultative anaerobic spore-forming bacteria in the *Bacilli* and *Clostridia* orders (Galperin *et al.*, 2012, Al-Hinai *et al.*, 2015). Initiation of sporulation is seen as a last resort for cell survival in response to stress as a result of circumstances, such as starvation. The early stages of sporulation can be reversed if the culture is transferred into a nutrient-rich environment (Greene & Slepecky, 1972, Cooney & Freese, 1976, Babu *et al.*, 2006).

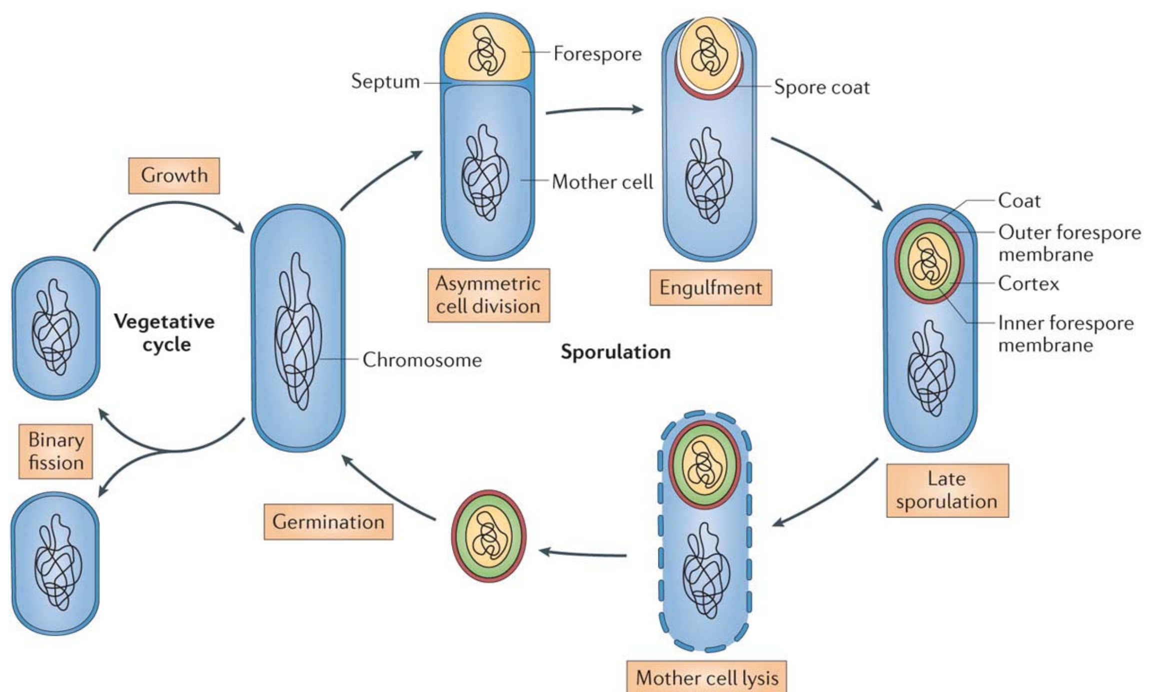



**Figure 1.2. Schematic diagram of the sporulation process, taken from McKenney *et al.* (2013).**

The conditions that trigger spore formation in thermophilic bacteria within a dairy environment are poorly defined. It seems unlikely that starvation would be the trigger for sporulation in a milk powder manufacturing plant given there is a continuous supply of nutrients in the incoming milk. It has been suggested that perhaps these nutrients are unable to reach a subpopulation within a biofilm (Seale *et al.*, 2015). Alternatively, a lack of oxygen within a biofilm or foulant layer could trigger sporulation. Recently, it was shown that decreased oxygen levels triggered both matrix

production in a biofilm (as will be discussed further in Section 1.4) and sporulation of *B. subtilis* (Kolodkin-Gal *et al.*, 2013). The formation of spores within a biofilm will be discussed in more detail in Section 1.4.3. The presence of minerals such as calcium, manganese, and magnesium, which are also present in milk (Holt, 1985), could also play a role in triggering sporulation. Evidence for this possibility is found in the formation of *Geobacillus* spores in the laboratory, which is influenced by the development medium and the presence of salts (Seale *et al.*, 2008). Mineral salts are commonly used in sporulation media (Cooney & Freese, 1976, Kaul & Singh, 1982, Warriner & Waites, 1999). These minerals, in particular calcium, are important for the development of a mature spore (Sanchez-Salas *et al.*, 2011) and they also accumulate in the forespore, resulting in the core containing most of these minerals (Johnstone *et al.*, 1980, Mallidis & Scholefield, 1987, Huang *et al.*, 2007). Calcium is the main salt found in the core and it has been shown that higher concentrations of calcium in the sporulation medium increase the expression of genes involved in the sporulation pathway (Oomes *et al.*, 2009).

### 1.3.3 Spore resistance

Spores are resistant to heat, mechanical disruption and a wide variety of chemicals, making it very hard to destroy them in dairy manufacturing processes (Chandler *et al.*, 2001, Cortezzo *et al.*, 2004, Jones *et al.*, 2005, Scheldeman *et al.*, 2006). In the case of mesophillic bacilli, combinations of several properties reportedly contribute to the overall resistance of *Bacillus* spores, including their low water content, the impermeability of the inner membrane, the spore coat, the peptidoglycan cortex, α/β type small acid-soluble proteins (SASPs) and dipicolinic acid (DPA) (Nicholson *et al.*, 2000, Driks, 2002, Cortezzo *et al.*, 2004, Cortezzo & Setlow, 2005, Setlow *et al.*, 2006, Setlow, 2006, Setlow, 2007, Setlow, 2014b). These properties have been well reviewed by Setlow (2014b). In a dairy manufacturing environment, it is their resistance against CIP chemicals and particularly wet heat that would result in their survival in final product.

There is no one mechanism that contributes to chemical resistance, instead a number of characteristics contribute, such as protection of DNA through binding of the α/β type SASPs, the considerable impermeability of the inner membrane and the detoxification nature of the coat. The resistance mechanisms against acids and bases, such as sodium hydroxide and nitric acid, used during CIP regimes in dairy manufacturing plants, are not well understood. As already discussed spores have the potential to suvive NaOH solutions used for CIP. The mechanisms of spore resistance against hydrogen peroxide, which is sometimes used as a processing aid, are better understood, where α/β type SASPs and the coat have both been shown to play a role (Setlow, 2006, Granger *et al.*, 2011). The coat can contain the enzyme catalase, which breaks down hydrogen peroxide (Checinska *et al.*, 2012).

The main characteristic that is believed to be involved in wet heat resistance is the core water content (Beaman *et al.*, 1989, Setlow, 2014b). The core water content can be influenced by the surrounding environment. Other characteristics that contribute to heat resistance include DPA, divalent metal ions, and α/β type SASPs (Setlow *et al.*, 2006). Uptake of DPA during sporulation

results in dehydration of the core; therefore, DPA-less spores are more susceptible to heat (Paidhungat *et al.*, 2000). Minerals in the form of divalent cations are located in the spore core, mainly chelated with DPA. The predominant divalent cation bound with DPA is $Ca^{2+}$ (Setlow, 2014b), although other divalent cations and some monovalent cations can also bind to DPA. Demineralization results in decrease in heat-resistance (Palop *et al.*, 1999b). $Ca^{2+}$ appears to contribute most to heat-resistance (Scheldeman *et al.*, 2006), the contribution of other cations is contradictory and appears to be strain dependent (Bender & Marquis, 1985, Marquis & Bender, 1985, Setlow, 2006, Ghosh *et al.*, 2011, Granger *et al.*, 2011, Setlow, 2014b). Mutant spores that were unable to produce DPA were also unable to accumulate calcium (Hintze & Nicholson, 2010).

Environmental factors, such as sporulation temperature, pH, solid versus liquid medium, biofilm versus planktonic growth, and suspension medium, can also affect the heat resistance of spores (Voort & Abee, 2013). Studies have shown that spores formed at temperatures greater than the optimum growth temperature of a particular strain tend to be more heat resistant (Beaman & Gerhardt, 1986, Beaman *et al.*, 1988, Etoa & Michiels, 1988, Palop *et al.*, 1999a, Sang & Sim, 2006, den Besten *et al.*, 2010, Guizelini *et al.*, 2012). The heat shocks received during the manufacture of milk powder could influence the resistant nature of the spores in the final product. For example, it was demonstrated that the structure of the spores of *B. sporothermodurans* changed when the spores were isolated from different locations i.e. spores isolated from UHT milk were denser, smaller and more heat-resistant than those isolated from raw milk (Scheldeman *et al.*, 2006). It has also been found that spores produced by thermophilic bacilli in a milk powder factory are more heat-resistant than those produced on laboratory media (Hill, 2004).

The pH of the sporulation medium influences heat resistance, with an alkaline medium resulting in more heat-resistant spores of *G. stearothermophilus* (Guizelini *et al.*, 2012). This may have implications for the dairy industry where reusing NaOH may result in the spread of spores with a higher heat resistance. As will be discussed in Section 1.4.3 spores can also form in biofilms. It has been shown that spores obtained from biofilms have a higher heat resistance compared with those derived from a planktonic culture (Hayrapetyan *et al.*, 2016). Spores suspended in milk or a high salt concentration are more heat resistant compared with those in water.

The heat resistance profiles of spores of thermophilic bacilli vary greatly between strains and between studies. In the New Zealand dairy industry thermophilic spores are categorized into heat-resistant spores (survival after 30 min at 100 °C) and highly heat-resistant spores (> 106 °C for 30 min). Initiatlly it was found that, of the thermophilic bacilli (obligate and facultative), only spores of *Geobacillus* spp. were categorized as highly heat-resistant spores and therefore have the potential to survive UHT treatments (140 °C for approximately 3 s) and retort-treated (110−120 °C for 15–20 min) products (Schwarzenbach & Hill, 1999). Of the mesophilic bacilli *B. sporothermodurans* also has the potential to survive UHT treatments (Hammer *et al.*, 1996b, Pettersson *et al.*, 1996, Hammer *et al.*, 2000, Scheldeman *et al.*, 2006). This microorganism is

also an indicator of poor hygiene and has been isolated from recycle loops immediately after the heating section within UHT processing plants (Hammer *et al.*, 1996a, Pettersson *et al.*, 1996) Scheldeman *et al.* (2006) found that out of a number of *Bacillus* strains obtained from milk products, only *B. sporothermodurans* survived high heat treatments of ≥ 120 °C for 5 s (Scheldeman *et al.*, 2006). However, other studies have found that a number of *Bacillus* strains such as *B. cereus*, *B. subtilis* and *B. licheniformis* can produce highly heat-resistant spores (Mostert *et al.*, 1979, Luecking *et al.*, 2013, Berendsen *et al.*, 2015). It is difficult to make comparisons between studies because of differences in time/temperature combinations used, and as already discussed differences in the sporulation conditions such as temperature, pH and medium composition and suspension medium, which can alter the spore resistance properties. In addition, there may be inaccuracies in ensuring the spore suspension has reached the correct temperature within a set time and in typing methods, leading to the misidentification of a strain that is able to survive a high heat-treatment. The production of heat-resistant (80-100 °C for 10-30 min), and particularly highly heat-resistant (> 106 °C for 30 min) spores by thermophillic bacilli is an issue of particular importance to milk powder producers, as these powders are used as ingredients in other types of dairy foods, such as ultra high temperature (UHT)-treated products.

### 1.3.4 Germination of spores

When the conditions are right for growth, germination of a spore can occur. Three stages are involved in the process of changing from a spore to a vegetative cell: activation, germination and outgrowth (Setlow, 2014a). Activation involves preparation of the spore for germination, of which the mechanisms are not well understood. However, it is generally accepted that spores must be activated before germination can occur. Heat, chemicals and a decrease in pH to 2−3 can activate spores (Kim & Foegeding, 1990, Iciek *et al.*, 2006, Vries, 2006, Ghosh & Setlow, 2009). In the dairy industry, heat is the most likely mechanism of thermophilic spore activation, because of the extensive use of heat as a preservation technology. This process appears to be strain and species specific. For example, spores of *G. stearothermophilus* were shown to be activated at temperatures as high as 110 °C, whereas spores of *B. subtilis* reportedly have a lower activation temperature of 65−70 °C (Finley & Fields, 1962, Leuschner & Lillford, 1999).

Once activated, germination is initatiated on recognition of a germinant. The main type of germinant is likely to be various kinds of nutrients, such as amino acids and sugars (Paredes-Sabja *et al.*, 2011), which are recognised by, and then bind to, germination receptors (GRs). In *B. subtilis* germination is a three-step process involving monovalent cation release, Ca-DPA release and break-down of the cortex (Setlow, 2014a). Ca-DPA release probably occurs via the SpoVA channel. This channel will be discussed more in Chapter 6. Once the cortex is degraded by the cortex lytic enzymes CwlJ and SleB, the cortex can take up water and then outgrowth can begin.

Two other mechanisms that do not involve GRs include: (1) release of Ca-DPA from the core through the use of cationic surfactants, of which dodecylamine has been the best studied

(Vepachedu & Setlow, 2007b), and (2) triggering degradation of the cortex by the cortex lytic enzymes through the addition of high concentrations of Ca-DPA (Paidhungat *et al.*, 2001). Spore populations are heterogeneous in their ability to germinate with a small proportion being very slow to germinate or are unable to germinate (Yi & Setlow, 2010, Setlow *et al.*, 2012, Perez-Valdespino *et al.*, 2013, Zhang *et al.*, 2013). These spores are referred to as 'super-dormant' spores (Ghosh & Setlow, 2009).

Germination has been poorly studied in the obligate thermophilic bacilli. Although, recently, the kinetics of *G. stearothermophilus* spore germination was examined using Raman Spectroscopy and Differential Interference Contrast Microscopy (Zhou *et al.*, 2013). In this study, the overall process of spore germination was found to be similar to that of *B. subtilis*. However, there were some differences such as requirement of a higher activation temperature and a more variable lag phase between addition of a germinant and Ca-DPA release.

## 1.4 Biofilm formation

Bacterial adhesion to surfaces and subsequent biofilm development is very common in many environments such as medical, marine and fresh water, water distribution units and food processing. Biofilms are defined as micro colonies of microorganisms enclosed in an extracellular polymeric substances (EPS) matrix (Costerton *et al.*, 1995, Jones *et al.*, 2013). In this present study a more general definition was used to include bacteria, which are attached and growing on a surface.

Biofilm development occurs through a number of stages, which is generally conserved between species (Stoodley *et al.*, 2002, Johnston, 2004). First, cells attach to a surface. Then, the initial reversible attachment is followed by irreversible attachment. Once the cells have attached, micro colony formation occurs followed by production of EPS. These micro colonies expand, forming multilayers of bacterial cells embedded in the EPS matrix. Water channels may develop within the EPS matrix (Stoodley *et al.*, 1999). This matrix provides a source of nutrients for bacterial cells. The process of attachment and the production of EPS render bacterial cells within a biofilm far more resistant than the planktonic cells to harsh conditions.

The molecular mechanisms of biofilm formation do differ between species and sometimes even between strains of the same species (Kirisits *et al.*, 2005, Stanley & Lazazzera, 2005, O'Neill *et al.*, 2007, Ryder *et al.*, 2007). Environmental conditions may also influence the structure and composition of a biofilm. For example, biofilms formed under turbulent flow conditions can generate "streamers" in the direction of the flow (Stoodley *et al.*, 1999). In the dairy context, biofilms can be grouped into environmental and process biofilms (Bremer *et al.*, 2015). Environmental biofilms form in the surrounding environment, such as around drains and do not have direct contact with the dairy product being manufactured. Biofilms of the thermophilic bacilli belong to the second group, of process biofilms. Three key features differentiate process from environmental biofilms. (1) process biofilms form a monolayer and therefore do not form the final stage of a traditional biofilm, consisting of a multilayer of bacterial cells embedded in an

EPS matrix with water channels; (2) they are generally dominated by one to two species probably because of selective pressures from the surrounding environment such as heat, product composition, pH and water activity; (3) they have a fast growth rate, reaching numbers of $10^6$ cfu/cm$^2$ within 6 – 12 h (Flint *et al.*, 2001b, Scott *et al.*, 2007, Burgess *et al.*, 2009, Seale *et al.*, 2012).

In dairy manufacturing plants biofilm formation of the thermophilic bacilli may occur on the surface of equipment in sections of dairy manufacturing plants at elevated temperatures of 40–65 °C (Stadhouders *et al.*, 1982, Flint *et al.*, 1997, Murphy *et al.*, 1999, Warnecke, 2001, Lewis, 2003, Scott *et al.*, 2007). Examples include preheating and evaporation sections of milk powder plants, plate heat exchangers (PHEs) used during the pasteurisation process, centrifugal separators (used to separate cream from whole milk) operated at warm temperatures (45–55 °C), recycle loops in butter manufacturing plants, cream heaters in anhydrous milkfat (AMF) plants and ultrafiltration (UF) plants operated at warm temperatures (approximately 50 °C).

The process of thermophile contamination and subsequent biofilm formation in dairy manufacturing plants is still not clearly understood. However, information has been gained from a number of laboratory studies on the adhesion and growth of biofilms of both *A. flavithermus* and *Geobacillus* spp.

### 1.4.1 Attachment of spores and vegetative cells

The attachment of bacterial cells to a surface is a two-step process, of reversible and irreversible attachment. Reversible attachment is mediated by physio-chemical interactions such as van der Waal forces and electrostatic interactions as well as specific interactions between the bacterial cell and the surface (Bos *et al.*, 1999). Interactions also occur between the cell and the conditioning layer, which is a layer of proteins and polysaccharides that forms on a surface as soon as it comes into contact with a liquid (Lennox, 2011). In the dairy context this conditioning layer is likely to contain proteins, fat and salts, the main components of milk. Milk proteins have been shown to inhibit attachment (Parkar *et al.*, 2001, Han *et al.*, 2011). Pre-conditioning the cells through growth in a cation rich environment can also increase the ability of vegetative cells to attach to a stainless steel surface for some strains of *A. flavithermus* and *G. stearothermophilus* (Somerton *et al.*, 2013). In milk powder manufacturing plants a foulant layer can also form over time, which has been shown to increase thermophile attachment and growth compared with un-fouled surfaces (Flint *et al.*, 2001b, Hinton *et al.*, 2002). Milk foulant is the build-up of milk proteins and calcium phosphate salts on equipment surfaces in dairy manufacturing plants (Jeurnink *et al.*, 1996). Fouling is particularly a problem at high temperatures (> 70 °C) e.g. in DSI units and heat exchangers as well as in recirculating or dead end zones such as under the orifice plates in evaporators (Visser & Jeurnink, 1997a, Visser & Jeurnink, 1997b, Scott *et al.*, 2007). Irreversible attachment occurs upon production of the EPS matrix.

The formation of a biofilm of thermohilic bacilli is likely to be initiated by the attachment of both vegetative cells and spores. Parkar *et al.* (2001) demonstrated that a variety of thermophilic

bacilli strains, including species of *Geobacillus*, *Anoxybacillus*, *B. licheniformis*, *B. coagulans* and *B. pumilus* were able to attach to the surface of stainless steel and milk foulant in similar numbers. In the same study spores were shown to have a greater propensity than vegetative cells for attachment, where the spores attached in numbers of approximately $10^6$ – $10^7$ cells/$cm^2$, whereas the vegetative cells attached in numbers of approximately $10^4$ – $10^5$ cells/$cm^2$.

The attachment process appears to be a multifactorial process; therefore the ability for an organism to attach cannot be directly attributed to one factor. For example cell surface proteins may play a role in attachment for some species but not others. Faille *et al.* (2007) demonstrated a slight reduction in the ability of *B. cereus* spores, without an exosporium, to attach to stainless steel. In contrast, when the outer protein coat was removed from spores of *A. flavithermus* and *Geobacillus* spp., there was no significant reduction in spore attachment to stainless steel. Surface proteins appear to be involved in the attachment of vegetative cells of *A. flavithermus* to stainless steel (Parkar *et al.*, 2001). The removal of surface proteins from a strain of *A. flavithermus* resulted in a 1-log reduction in its ability to attach to stainless steel (Palmer *et al.*, 2010). It was further shown that a mutant of *A. flavithermus* that exhibited a 10-fold reduction in its attachment ability when compared with the parental strain, was also less hydrophobic than the parental strain.

The hydrophobicity of both the spore and substratum surface may also play a role in the ability of a spore to attach to a surface (Wiencek *et al.*, 1991). Spores of *Geobacillus* attach in greater numbers to more hydrophobic surfaces (Seale, 2009). In some *Bacillus* studies it has been reported that the greater the hydrophobicity of a spore, the greater the attachment (Ronner *et al.*, 1990, Wiencek *et al.*, 1990, Faille *et al.*, 2002). The role of spore hydrophobicity of the obligate thermophilic bacilli is less clear. Parkar et al. (2001) found there was no correlation between hydrophobicity and the attachment of spores of thermophilic bacilli; whereas Seale *et al.* (2008) found that spores were mostly hydrophilic in nature at pH 6.8, and the most hydrophilic spore isolate with the smallest negative charge attached in the highest numbers to stainless steel surfaces. However, in the same study it was found that in the presence of milk these effects were negated.

EPS does not appear to be involved in the attachment process of thermophilic bacilli to surfaces as occurs in other bacterial species, as studies have shown that removal of EPS does not reduce attachment of these strains (Parkar *et al.*, 2001, Palmer, 2008).

### 1.4.2 Growth of attached populations

Following attachment, it has been speculated that thermophilic spores germinate and the vegetative cells then reproduce, forming a biofilm (Burgess *et al.*, 2009). However, it is also possible that a biofilm could be initiated from both spores and vegetative cells. Burgess et al. (2009) demonstrated that biofilms of *A. flavithermus* can be initiated from either spores or vegetative cells. Biofilms of *Geobacillus* spp. and *A. flavithermus* generated from vegetative cells inoculated into pasteurised skim milk appear to have similar growth rates and a maximum

cell density of approximately $10^6 - 10^7$ cells/cm$^2$, which is reached within 6 h (Flint *et al.*, 2001b, Parkar *et al.*, 2003, Burgess *et al.*, 2009). When a biofilm of *A. flavithermus* was grown over an extended period of 114 h, the cell density appeared to fluctuate over time between approximately $10^2$ and $10^7$ cells/cm$^2$ (Parkar *et al.*, 2003). In the same study, biofilms of thermophiles also formed when un-inoculated pasteurised skim milk was used that contained low levels (< 1 cell/ml) of natural thermophiles (whose identity is unknown). The cell density reached levels of greater than $10^6$ cells/cm$^2$ within 18 h. This growth rate is similar to the growth rate of thermophiles observed in dairy manufacturing plants.

### 1.4.3 Spore development within biofilms

The next stage is unclear, but it appears that vegetative cells within the biofilm sporulate as the biofilm reaches a steady-state. In the case of *A. flavithermus*, spore formation within a biofilm can occur within 4 h in a continuous flow reactor (Burgess et al., 2009). This differs from other studies on mesophilic bacilli, which have shown that spore formation occurs once the biofilm has matured (Branda *et al.*, 2001, Lindsay *et al.*, 2005, Faille *et al.*, 2014, Hayrapetyan *et al.*, 2016). The difference in sporulation rate may be related to a faster growth rate of the obligate thermophilic bacilli compared with the mesophilic bacilli or differences in using a static biofilm system (used for the mesophilic bacilli) compared with a continuous flow system. In *Bacillus subtilis* biofilms it has also been shown that multiple cell types occur within the biofilm, with spores localising to the top of the biofilm (Vlamakis *et al.*, 2008). In the case of *Geobacillus thermoglucosidans*, spores appear to be randomly dispersed within the biofilm (Zhao *et al.*, 2013).

As to be expected, temperature variation has a considerable effect on the formation of spores by thermophilic bacilli. For example, spores comprised as much as 10–50% of an 8 h *A. flavithermus* biofilm grown at 55 and 60 °C, while no spores were detected in a corresponding biofilm grown at 48 °C (Burgess et al., 2009).

## 1.5 Enumeration and identification of the thermophillic bacilli in milk powder product

### 1.5.1 Plate count methods used to enumerate thermophilic bacilli

Two broad approaches are used to enumerate thermophiles in food, namely a total thermophile plate count (TPC) and a thermophilic spore count (TSC). In the New Zealand dairy industry, the TPC involves transferring 1 ml of the sample or its decimal dilutions into a Petri dish, pour plating using milk plate count agar (MPCA) and incubating at 55 °C for 48 h. This method is based on the method described by Frank & Yousef (2004).

A TSC is carried out by heat-treating the sample at 100 °C for 30 min. This heat treatment inactivates vegetative cells and activates the spores, enabling them to germinate. The heat-treated sample is pour plated with MPCA supplemented with 0.2% starch and is incubated at 55 °C for 48 h. The starch is added as an aid for spore germination. The heat treatment used in this methodology is more severe than that in other published methods used to enumerate

thermophilic spores, such as 80 °C for 10 min, 80 °C for 20 min and 100 °C for 10 min (Murphy *et al.*, 1999, McGuiggan *et al.*, 2002, Rueckert *et al.*, 2004, Rueckert *et al.*, 2005a, Coorevits *et al.*, 2008). The use of a higher temperature selects for those spores that may survive additional heat treatments used for further processing during dairy manufacture.

A different method is used to enumerate highly heat-resistant spores in milk powder that is destined for UHT or retort treatment (ISO/TS 27265:2009). This method involves heat treating the sample or its decimal dilution at 106 °C for 30 min. The development of this method was based on work carried out in our laboratories, which demonstrated that a higher treatment temperature of 106−108 °C destroyed spores of *A. flavithermus* so that only heat-resistant spores of *Geobacillus* could be enumerated (Hill, 2004).

### 1.5.2 Alternative rapid methods

Plate count methods are labour intensive and time consuming. In the past, rapid test methods have been identified as being important for reducing labour costs, releasing product faster and, most importantly, giving greater control over the process. However, rapid methods generally require technical expertise and expensive reagents. Two rapid methods that were adapted for the enumeration of the thermophilic bacilli in milk powder used flow cytometry and real-time polymerase chain reaction (PCR).

The BactiFlow™ flow cytometer was used to count viable bacterial cells in milk powder based on esterase activity (Flint *et al.*, 2006). However, it was identified that it was important to develop a method to enumerate thermophiles only. The new method was still based on esterase activity. However, it was adapted by including a 55 °C incubation step to select for thermophiles (Flint *et al.*, 2007). The method showed good correlation with a TPC. However, the BactiFlow™ did not always correlate well with a TPC when it was used routinely in a manufacturing context. In addition, technical expertise was required for the BactiFlow™ protocols and the detection limit was not sensitive enough for some milk powder products.

A real-time PCR assay was adapted to enumerate total viable vegetative cells and spores of the thermophilic bacilli in milk powder (Rueckert *et al.*, 2005a, Rueckert *et al.*, 2005b). Originally, the assay targeted the 16S rRNA gene. However, it also amplified DNA from the mesophile *Bacillus megaterium*, but to a lesser degree compared with *A. flavithermus* and *B. licheniformis*. No *Geobacillus* strains were included in this study. The 16S rDNA gene can have a variable copy number in the genome, which may influence enumeration determined by real-time PCR. This led to the development of a similar assay targeting the *spo0A* gene (Rueckert *et al.*, 2006). The assay not only was successful at amplifying DNA from a variety of thermophilic bacilli, but also gave a positive result for other bacilli, including *B. cereus* and *B. smithii*. Although the assay was rapid (≤ 1 h), real-time PCR assays are costly compared with plate counts, require technical expertise and, in the case of the assay described by Rueckert et al. (2006), are not sensitive or specific enough.

### 1.5.3 Characterization and identification of milk powder isolates

Traditionally, biochemical tests have been used in the dairy industry to identify bacterial contaminants, including the thermophilic bacilli. These tests have provided important information on phenotypic characteristics, such as the ability to utilize lactose; however, these tests are time consuming, can be difficult to interpret and do not have the taxonomic resolution required for the thermophilic bacilli. In the late 1990s and early 2000s, PCR-based identification techniques, such as partial 16S rRNA sequencing, group-specific PCR and randomly amplified polymorphic DNA (RAPD) profiling, were developed to identify and sub-type thermophilic isolates from milk powder. Results from this type of work have suggested that strains of *A. flavithermus* may have been described previously as *G. stearothermophilus* before the advent of molecular techniques (Ronimus *et al.*, 1997, Flint *et al.*, 2001a, Ronimus *et al.*, 2003). 16S rRNA gene sequencing and group-specific PCR is useful as an initial identification tool, but it provides no information on strain variation within a manufacturing plant. In addition, some *Geobacillus* species (including *G. kaustophilus*, *Geobacillus thermoleovorans* and *G. stearothermophilus*) cannot be differentiated based on partial 16S rRNA gene sequences. Group specific PCR involves the use of primers specific for a species or genus. Primers were developed to identify *A. flavithermus* and members of the *Geobacillus* genus but it is only used as an initial screening tool. RAPD profile analysis was the first method used to show that multiple strains of thermophilic bacilli could be isolated from milk powder. This method is quick and easy; however, it is not always reproducible and can be difficult to interpret because of the large number of bands.

More recently other molecular-based methods, such as bar-coded 16S amplicon sequencing, multi-locus variable high resolution melting analysis (MLV-HRMA) and a microarray analysis have been used for typing bacilli from dairy products (Seale *et al.*, 2012, Chauhan *et al.*, 2013, Zhao *et al.*, 2013, Chauhan *et al.*, 2014). MLV-HRMA typing has demonstrated that multiple strains of *Geobacillus* can co-exist within the same manufacturing plant. Other methods have been used for identifying and typing *Geobacillus* species from environments other than dairy processing. In particular, the focus has been on evaluating other house-keeping genes, such as *recN*, *recA*, *rpoB*, *gyrB*, *parE* and *spo0A* that can be used as alternatives to the 16S rRNA gene (De Clerck *et al.*, 2004b, Zeigler, 2005, Kuisiene *et al.*, 2009, Weng *et al.*, 2009). However, in general the taxonomic resolution is still poor. Comparative genomics is likely to become the way of the future for identifying and characterising thermophilic bacilli. Several dairy strains, including *A. flavithermus*, *G. thermoglucosidasius* and *B. licheniformus*, have had their genomes sequenced, but with the exception of *B. licheniformis* no comparative analysis has been carried out. Genome sequencing has begun to be used for characterising species of the *Geobacillus*, in particular for the identification of novel functions. However, it has not been applied to dairy strains of *Geobacillus* (Zeigler, 2014, Brumm *et al.*, 2015a, Brumm *et al.*, 2015b, Studholme, 2015).

## 1.6 Conclusions

The thermophilic bacilli, especially *A. flavithermus* and *Geobacillus* spp., are an important group of bacterial contaminants in the dairy industry, particularly in milk powder production. Improvements are needed around the identification of these organisms, particularly in relation to those strains of *Geobacillus* that produce highly heat-resistant spores. In addition, a greater understanding of the structure of thermophilic biofilms within the context of the milk processing environment, and their link with spore formation, is needed to develop better control measures.

## 1.7 Objectives of this study

This study aimed to address three areas. Firstly, to begin to understand the diverse nature of strains isolated from one milk powder manufacturing plant from both a phenotypic and genotypic perspective. Secondly, to resolve the *Geobacillus* taxonomy using phylogenomic tools, with a particular focus on where the dairy strains fit into the *Geobacillus* taxonomy. Finally, to gain a genomic insight into how strains of *G. stearothermophilus* grow in dairy manufacturing plants, especially around biofilm formation and sporulation.

The specific objectives of this study were as follows:

(1) Isolate and characterise thermophilic bacilli from the surface of a milk powder manufacturing plant.

(2) Screen strains of *G. stearothermophilus* for their ability to form biofilms and produce spores.

(3) Compare genome sequences of dairy strains of *G. stearothermophilus* with other strains of *Geobacillus* to identify differences that may provide evidence of adaptation to a dairy environment.

(4) Establish the phylogenomic relationship of the dairy strains of *G. stearothermophilus* with other members of the *Geobacillus* genus.

(5) Carry out a comparative genomics analysis to determine the potential molecular mechanisms of biofilm formation and sporulation of *Geobacillus* spp.

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

## STATEMENT OF CONTRIBUTION
## TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Sara Burgess

**Name/Title of Principal Supervisor:** Steve Flint

**Name of Published Research Output and full reference:**

Burgess S A, Lindsay D, Flint S H (2010). Thermophilic bacilli and their importance in dairy processing. International Journal of Food Microbiology. 144: 215 – 225.

**In which Chapter is the Published Work:** 1

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  This paper was the work of the candidate who researched the information and wrote the paper with the guidance and editing by the two co-authors.

Sara Burgess
Digitally signed by Sara Burgess
DN: cn=Sara Burgess, o=Massey University,
ou=School of Food and Nutrition,

Candidate's Signature

02/3/16

Date

s.h.flint@massey.ac.nz
Digitally signed by s.h.flint@massey.ac.nz
DN: cn=s.h.flint@massey.ac.nz
Date: 2016.02.23 15:32:19 +13'00'

Principal Supervisor's signature

23/2/16

Date

GRS Version 3– 16 September 2011

Chapter 2

# 2 Isolation, identification and characterisation of strains of thermophilic bacilli from a milk powder manufacturing plant

This chapter is partially based on the publication by Burgess *et al.* (2014a). The literature cited in this chapter was up-to-date as of Febuary 2014.

## 2.1 Introduction

Biofilms are the main source of bacterial contamination of the final products in dairy manufacturing plants. One of the most common groups of biofilm-forming organisms in dairy manufacture is the thermophilic bacilli (Burgess *et al.*, 2010). This group of bacteria is able to grow in sections of manufacturing plants, such as preheating and evaporation sections of milk powder plants, in which the temperatures are suitable for their growth (40−65 °C) (Murphy *et al.*, 1999, Scott *et al.*, 2007). The thermophilic bacilli can be divided into facultative and obligate thermophiles. Facultative thermophiles, in the dairy processing context, belong to the *Bacillus* genus and include species, such as *B. coagulans*, *B. licheniformis, B. subtilis* and *B. pumilus*, which are able to grow at both mesophilic and thermophilic temperatures. However, it is the obligate thermophiles, such as *Geobacillus stearothermophilus* and *Anoxybacillus flavithermus*, which tend to grow to high numbers in milk powder manufacturing plants (Scott *et al.*, 2007). More recently *Geobacillus glucosidasius* has also been identified as a contaminant of milk powder manufacturing plants (Zhao *et al.*, 2013). Although these microorganisms are not pathogenic, their growth results in high bacterial numbers and their presence can be interpreted as an indicator of poor plant hygiene. In addition, they have the potential to spoil product through enzyme and acid production (Chopra & Mathur, 1984, Chen *et al.*, 2003, Chen *et al.*, 2004). Spores of *G. stearothermophilus* are able to survive ultra-high-temperature (UHT) (134 – 145 °C for 1 – 10 s) treatment (Hill, 2004). Many studies have focussed on the identification of the predominant thermophilic species in milk powder (Ronimus *et al.*, 1997, Flint *et al.*, 2001a, Ronimus *et al.*, 2003, Scott *et al.*, 2007), but few studies have looked at differences between strains of thermophilic bacilli within one manufacturing plant.

The most commonly used method to identify thermophilic bacilli in dairy products is partial 16S rRNA gene sequencing. However, the 16S rRNA gene sequence of some species of *Geobacillus* is so similar that this is not a reliable method of differentiating *Geobacillus* species. For example, the 16S rRNA gene of *G. kaustophilus*, *G. thermoleovorans* and *G. stearothermophilus* has greater than 99% identity (Weng *et al.*, 2009). Randomly amplified

polymorphic DNA (RAPD) profiling, has been used for typing strains of thermophilic bacilli that have been isolated from milk powder but not from different locations within a single manufacturing plant. Ronimus *et al*. (2003) used RAPD profiling for classifying strains in milk powder and found one strain of *B. subtilis*, two strains of *B. licheniformis*, three strains of *A. flavithermus* and only one strain of *G. stearothermophilus*. However, RAPD profiling is not always reproducible and can be difficult to interpret because of the large number of bands (often up to 20) as well as the presence of weak bands caused by mismatches between the primer and the DNA template. Other approaches, such as sequence analysis of various house-keeping genes, biochemical testing, fatty acid profiling, whole cell protein profiling and restriction fragment length polymorphism, have been used for identifying and characterizing *Geobacillus* species (Flint *et al.*, 2001a, Nazina *et al.*, 2001, Marchant *et al.*, 2002, Ronimus *et al.*, 2003, Zeigler, 2005, Yuan *et al.*, 2012). More recently, Seale *et al.* (2012) identified 16 different strains of *Geobacillus* species isolated from milk powder MLVA-HRM. They found that up to three different strains could co-exist in a milk powder manufacturing plant.

In this study, we evaluated whether there were other physiological characteristics that could differentiate strains that were isolated from the same milk powder processing plant. These included biochemical assays, fatty acid profiling and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). A milk powder manufacturing plant was selected that was known to have issues with thermophile growth. The production of milk powder involves three main steps: (1) heat-treatment of the milk, (2) evaporation and (3) drying. It is during the first two steps that thermophile growth can occur. This study focussed on the identification and characterisation of thermophilic bacilli from the evaporation section as it is easily accessible for taking samples.

## 2.2 Materials and methods

### 2.2.1 Operation of the evaporators

Thermophilic bacterial isolates were obtained from the evaporator section of a New Zealand milk powder manufacturing plant with permission from the manufacturing plant manager. The milk powder manufacturing process comprised of a preheat section which contained a dual plate heat exchanger (PHE), tubular heat exchanger, direct contact heater (DCH) and a direct steam injector (DSI) (Figure 2.1). Following preheating, the milk passed through the evaporator section into a concentrate tank before drying. The evaporator section of the plant contained two evaporators. Each evaporator had two effects: one five-pass mechanical vapour recompression (MVR) falling film effect and a one-pass thermal vapour recompression (TVR) effect. Prior to evaporation, the incoming milk was split and each stream passed through one of the evaporators, after which the concentrated milk from each stream was merged before drying.

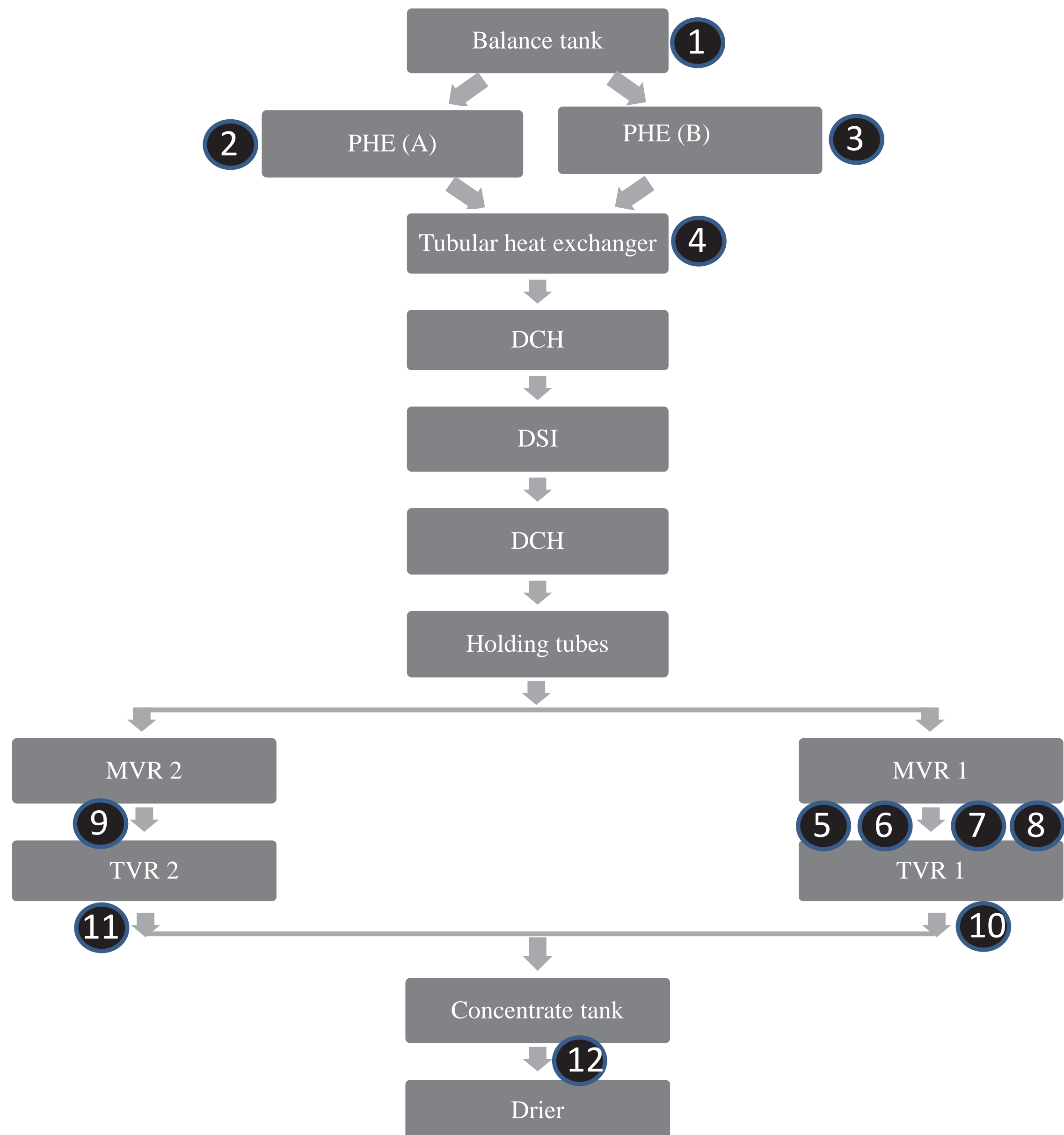



**Figure 2.1. Schematic diagram of the milk powder processing plant used for this study.** Numbers 1 – 12 indicate the points at which milk samples were taken. The samples from MVR 1 were taken from sample points as the milk was exiting each evaporator pass, where sample 5 is from pass 1, sample 6 from pass 2, sample 7 from pass 3 and sample 8 from pass 4. Only one sample point was available from MVR 2 which was as the milk was exiting pass 2.

### 2.2.2 Sampling regime

At the start and end of the manufacturing cycle, milk samples were taken from 12 different rubber septum sampling points (Figure 2.1), into sterile vacuum tubes (Vacutte grenier labortechnik, Thermofisher, Auckland, New Zealand). On completion of the manufacturing cycle, the evaporators were flushed with water, opened and ten different areas (Figure 2.2 and Table 2.1) within the evaporators were swabbed using 3M™ Sponge-Stick with neutralizing buffer (3M New Zealand Ltd., Auckland, New Zealand). The top of each pass on top of the

distribution plate (A, B and C of Figure 2.2) and the calandria bottom (D of Figure 2.2) were swabbed within the two MVR effects. The calandria bottoms only were swabbed in the two TVR effects (Table 2.1). The swabs were held at room temperature overnight and then spread directly onto tryptic soya agar (TSA) (BBL, Becton Dickinson, Cockeysville, MD, USA). Colonies (3−4) from each TSA plate were re-streaked on to TSA. For those plates that did not have individual colonies, a scraping was taken using a 10 μl loop and was streaked for single colonies on to TSA. The plates were sealed in a plastic bag to prevent drying, and incubated for 1–3 days at 55 °C. A single colony, for each re-streaked isolate, was transferred into 10 ml of tryptic soya broth and incubated at 55°C for 6–10 h until the bacteria were growing in approximately mid-exponential phase. An aliquot (1 ml) of each culture was frozen (at −80°C) with approximately 10% glycerol. In total, 33 isolates from the surface of the evaporators were obtained and archived.

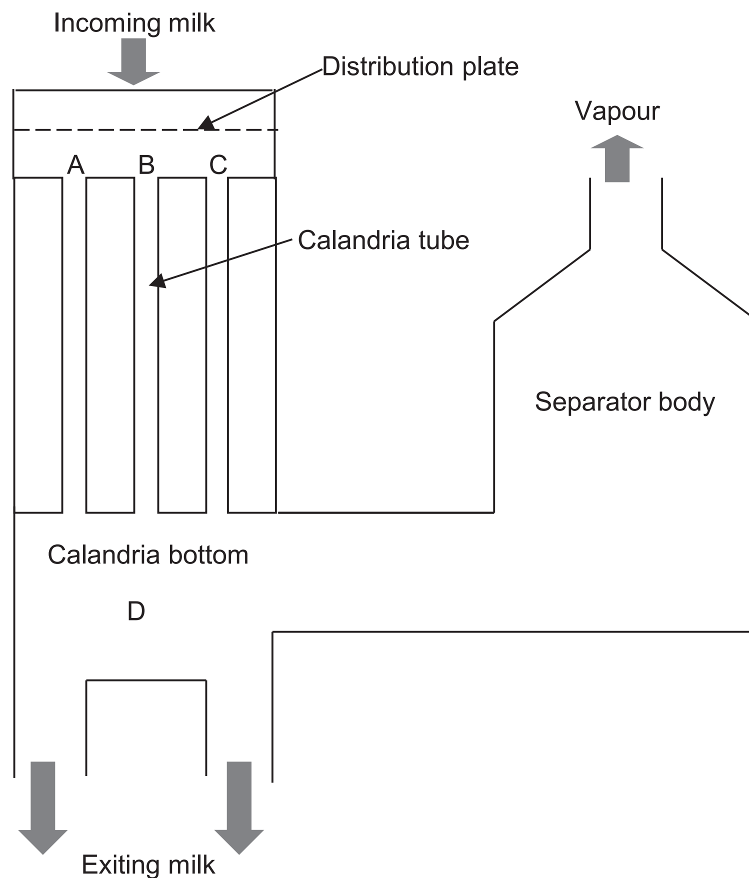



**Figure 2.2. Schematic diagram of one MVR effect from an evaporator.** The calandria tubes are approximately 6 m in length. A, B, C and D refer to the sample locations listed in Table 2.1.

### 2.2.3 Identification of isolates

#### *Primers*

The 16S rDNA-derived primer Levo (Flint *et al.*, 2001a) was used in combination with the universal primer Y1 (Young *et al.*, 1991) to identify those thermophile isolates that were *Geobacillus*. The Levo primer is genus specific for the *Geobacillus* group. The presence of a PCR product, approximately 450 bp in length, identifies the isolate as a *Geobacillus* species. The genus-specific PCR procedure for *Geobacillus* species was performed on the isolates as described by Flint et al., (2001a). To identify the unknown thermophiles, the 16S-derived universal primers 27F (5'-AGA GTT TGA TCC TGG CTC AG) and 16SR1 (5'-GGA CTA CCA GGG TAT CTA ATC) were used to amplify an approximately 800 bp fragment (Edwards *et al.*, 1989). For identification to the species level of ten *Geobacillus* isolates, which were selected based on sample location, the universal primers 27F and 1492R (5'-TAC GGC TAC CTT GTT ACG ACT T) (Edwards *et al.*, 1989) were used.

#### *DNA extraction*

The isolates were streak plated onto TSA and incubated overnight at 55°C. A crude DNA extraction was carried out by resuspending approximately 5−10 colonies in 200 µl of water, boiling for 10 min and freezing at −80°C for 20 min. The suspension was spun at 5,000 x *g* for 5 min and the supernatant was used as the PCR template.

#### *PCR reaction and gel electrophoresis*

The PCR reaction mix was made using a MasterAmp™ Tfl PCR kit: 528 µl of sterile Milli-Q water, 50 µl of 20x Tfl PCR buffer, 100 µl of 25 mmol $l^{-1}$ $MgCl_2$, 200 µl of dNTP mix containing 1.25 mmol $l^{-1}$ of each dNTP, 25 µl of 10 µmol $l^{-1}$ or each primer and 12 µl of Tfl DNA polymerase. Each PCR reaction was performed in a total volume of 50 µl containing 49 µl of PCR mix and 1 µl of template. The PCR reactions were performed in a Techne 512 thermal cycler using the following protocol: 94°C for 5 min, followed by 30 cycles of 94 °C for 30 s, 55 °C for 45 s and 72 °C for 1.5 min, followed by one cycle of 72 °C for 5 min. The products were analyzed by running on a 2% agarose gel (containing 10µl of GelRed™ (Biotium inc., Hayward, CA, USA) for 100ml of molten agarose) in 1x TBE buffer (10x TBE composition: 216g Tris, 110g Boric acid, 14.9g EDTA, made up to 2L with Milli-Q water, pH was adjusted to 8.35 with 6M HCl or 6M NaOH) for approximately 2 h at 80V. The gel was viewed under UV light. Those isolates that were not identified as *Geobacillus* species using the *Geobacillus* specific primers were identified using partial 16S rDNA sequencing, and the PCR products were sequenced as described below.

#### *Sequencing*

The PCR products were purified using DNA Clean and Concentrator™-5 (Zymo Research Corporation, Irvine, CA, USA) and sequenced. Sequence comparison was carried out against the GenBank database using the BLASTN algorithm.

### 2.2.4 Biochemical characterisation

Ten *Geobacillus* isolates were inoculated into the Microgen® Bacillus-ID (Microgen Bioproducts Ltd, Camberley, UK), the BBL™ Crystal (Becton, Dickinson and Company, Franklin Lakes, NJ, USA) and the API 50CHB® (bioMérieux, Marcy l'Etoile, France) identification kits, incubated at 55°C and read at 24 and 48 h. Additional tests, including tyrosine decomposition, casein hydrolysis, gelatin liquefaction and phenylalanine deamination, were all prepared according to the methods described by Barrow and Felham (2004).

### 2.2.5 Analyses of fatty acid methyl esters (FAMEs)

All of the strains were grown on TSA for 18–24 h and then scraped from the surface and freeze-dried as lyophilized cells were required for the extraction of the fatty acids. Lipids were extracted from the lyophilized cells by vortex mixing three times with fresh portions of chloroform:methanol 2:1 v/v. After mixing, the sample was left standing for 5 min, mixed again and centrifuged (8000 x *g*), and the solvent was transferred to a clean vial. The combined solvent was evaporated at 50 °C under nitrogen and the extracted material was weighed. The solids were again taken up in the solvent mixture and centrifuged to sediment any fines. An aliquot of the supernatant was transferred to a clean Reacti-Vial (1 ml capacity, Thermo-Fisher Scientific), the solvent was removed as before and the final weight of the test portion was calculated.

Fatty acids of neutral lipids were converted to the corresponding FAMEs using 0.1 mol $l^{-1}$ sodium methoxide at 40°C. After 5 min, 20 % boron trifluoride in methanol was added to methylate polar lipids. The mixture was allowed to stand for 10 min at ambient temperature before adding n-hexane (250 µl). After a further 2 min, the solution was neutralized to prevent hydrolysis of the methyl esters.

Where necessary, the FAME solutions were diluted with n-hexane to achieve an appropriate column loading. The methyl esters were separated by capillary gas liquid chromatography (Shimadzu model GC-17A) using on-column injection on a polar column (ZB-FFAP, Phenomenex). Fatty acids were identified by comparison of retention times with those of milk fat FAME and cod liver oil FAME samples. Relative amounts of fatty acids in the sample were calculated by area normalization using theoretical response factors (Ackman & Sipos, 1964).

### 2.2.6 Analysis by MALDI-TOF MS

Streak plates of the isolates on TSA were prepared, and proteins from a colony on each plate were extracted using the protein extraction method as outlined in the manufacturer's instructions (Bruker Daltonik GmbH, Bremen, Germany). Briefly, this method involves resuspension of colonies in 70% ethanol, followed by centrifugation and resuspension of the pellet in equal volumes of 70% formic acid and acetonitrile. The resulting supernatant from this mixture was added to the MALDI target plate in duplicate, air dried, overlaid with 1 µl of HCCA matrix (Bruker Daltonik GmbH) (cyano-4-hydroxycinnamic acid suspended in 50% acetonitrile and 2.5% trifluoroacetic acid), air dried again and then processed in the microflex LT mass spectrometer (Bruker Daltonik GmbH) using the Flex Control program. The resulting spectra

were analyzed using the Bruker Daltonics MALDI Biotyper 3.0 Real Time Classification program. To identify the isolates, the standard Bruker database, provided by the manufacturer, was used which contained 486 isolates belonging to the Bacillales order, of which four were *Geobacillus* spp. In addition, this database was supplemented with several reference cultures of *Geobacillus* and *Anoxybacillus* species, including *A. flavithermus* ATCC 21510, *A. flavithermus* DSM 2641, *A. flavithermus* CM (Burgess *et al.*, 2009), *G. kaustophilus* ATCC 8005, *G. stearothermophilus* DSM 5934 and *G. thermoleovorans* DSM 5366, giving a total of 10 strains of thermophilic bacilli used for comparison. Isolates were clustered using a minimum spanning dendrogram, which was generated using the Bruker Daltonics MALDI Biotyper 3.0 Offline Classification software.

## 2.3 Results

The aim of this work was to determine whether strains of *G. stearothermophilus*, isolated from the same manufacturing plant, had different physiological characteristics. The strains used in this study were isolated from the surface of ten different locations in the evaporator section of a milk powder manufacturing plant.

### 2.3.1 Isolation of strains

A milk powder manufacturing plant was selected that was known to have an issue with thermophile growth. During milk evaporation (prior to the evaporator opening), milk samples were taken at the start and end of a manufacturing run from a range of sample points (Figure 2.3). These samples were plated for thermophiles to determine whether the evaporator was a potential site for growth during the manufacturing run. From the tubular heat exchanger onwards there was an increase in thermophile numbers, with a > 1 log difference in counts between the start and the end of the manufacturing run (Figure 2.3).

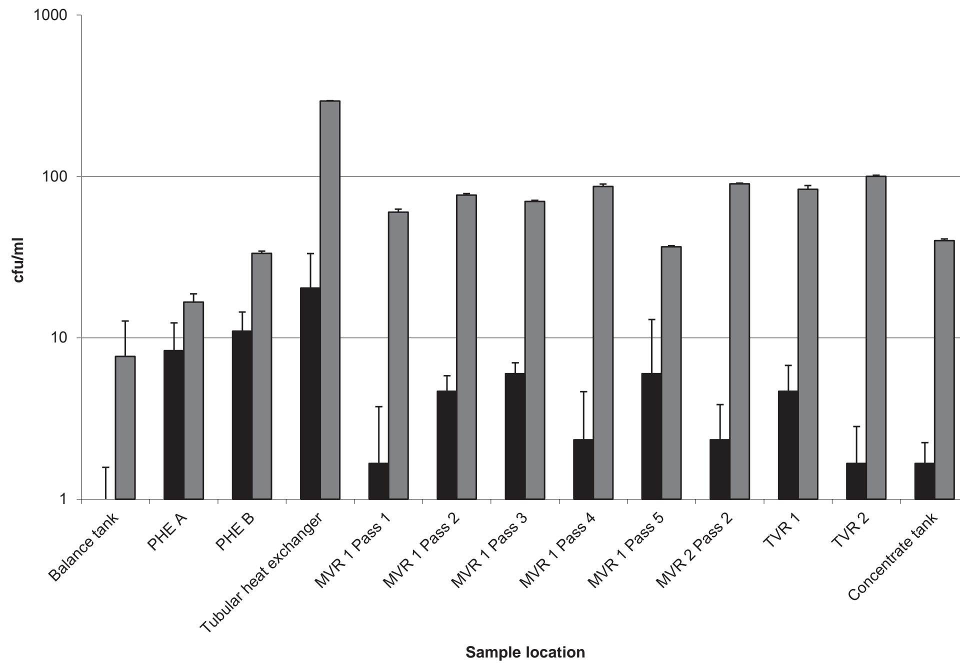



**Figure 2.3. Thermophile plate counts (cfu/ml) of milk samples taken at the start (black bars) and end (grey bars) of a nutrional milk powder processing cycle.**

At the end of a manufacturing run, the evaporator section of the manufacturing plant was rinsed with water and then swabbed in ten different locations. This was the only area of the manufacturing plant that was able to be opened. A minimum of three thermophile isolates were obtained from each of the ten locations, resulting in a total of 33 isolates. Using genus-specific and partial 16S rDNA sequencing, 29 of these isolates were identified as *Geobacillus* species, three as *Anoxybacillus* species and one as *B. licheniformis*. Table 2.1 shows a complete list of the isolates, their origin in the evaporator and identification.

**Table 2.1. Identification of strains isolated from the evaporator**

| **Sampling point**[a] | **Location in evaporator** | **Isolate** | **Identification** |
|---|---|---|---|
| A1 | MVR Evaporator 1, pass 1, top of the calandria tubes | **Isolate 1 (P2)** | *Anoxybacillus* |
| | | Isolate 2 | *Anoxybacillus* |
| | | **Isolate 3 (P3)** | *G. stearothermophilus* |
| | | Isolate 4 | *Geobacillus* |
| B1 | MVR Evaporator 1, pass 2, top of the calandria tubes | **Isolate 1 (G1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| | | Isolate 4 | *Anoxybacillus* |
| C1 | MVR Evaporator 1, pass 4, top of the calandria tubes | **Isolate 1 (H1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| D1 | Bottom of MVR evaporator 1 | **Isolate 1 (A1)** | *G.stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| A2 | MVR Evaporator 2, pass 1, top of the calandria tubes | **Isolate 1 (I1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| B2 | MVR Evaporator 2, pass 2, top of the calandria tubes | **Isolate 1 (J1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *B. licheniformis* |
| C2 | MVR Evaporator 2, pass 4, top of the calandria tubes | **Isolate 1 (K1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| D2 | Bottom of MVR evaporator 2 | **Isolate 1 (B1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| E | Top of TVR evaporator 1 | **Isolate 1 (D1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |
| | | Isolate 4 | |
| F | Top of TVR evaporator 2 | **Isolate 1 (C1)** | *G. stearothermophilus* |
| | | Isolate 2 | *Geobacillus* |
| | | Isolate 3 | *Geobacillus* |

* Those isolates in bold were selected for further characterisation.

[a] Refer to Figure 2.2 for location of each sample point in the evaporator.

Ten *Geobacillus* strains (one isolate from each location within the evaporator) were selected for further characterisation. To confirm which species of *Geobacillus* was present, a 1350 bp sequence of the 16S rDNA gene was sequenced and compared with reference strains of *Geobacillus* in the GenBank database. The ten *Geobacillus* strains were identified as *G. stearothermophilus.*

### 2.3.2 Biochemical characterisation

Biochemical assays were carried out to further characterise the *G. stearothermohilus* strains (Table 2.2). The dairy isolates were also compared with the type strain *G. stearothermophilus* ATCC 12980, based on data from Claus and Berkely (1986) and Nazina et al., (2001). All ten dairy strains were positive for the methyl red test and negative for casein hydrolysis, gelatin liquefaction, and starch hydrolysis. In addition, all ten strains were negative for tyrosine decomposition and phenylalanine deamination. Acid was produced from cellobiose, lactose, fructose, mannose, melibiose, saccharose, melezitose, salicin and sucrose, based on the API 50CHB identification kit. None of the dairy strains were able to produce acid from mannitol, adonitol, D-ribose or sorbitol. Acid production from amidon, galactose, melezitose, glucose, raffinose, gentiobiose, glycogen, inositol, inulin, rhamnose, D-turanose and xylose was variable between the *G. stearothermophilus* strains (see highlighted cells in Table 2.2). None of the strains was able to utilize citrate or produce indole. All of the strains were able to hydrolyse esculin and reduce nitrate. The Voges−Proskauer reaction was negative for all isolates. Importantly, several strains could be distinguished from each other based on different biochemical reactions. For example, isolates A1, D1 and P3 could be distinguished based on the amidon, trehalose, melezitose, L-pyroglutamic acid and *p*-nitrophenyl-β-D-glucoside biochemical reactions.

**Table 2.2. Biochemical reactions of thermophile *G. stearothermophilus* isolates based on the Microgen Bacillus-ID, BBL Crystal and API 50CHB identification kits**

| Biochemical reaction | A1 | B1 | C1 | D1 | G1 | H1 | I1 | J1 | K1 | P3 | ATCC 12980[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cellubiose | + | + | + | + | + | + | + | + | + | + | – |
| Lactose | + | + | + | + | + | + | + | + | + | + | – |
| Fructose | + | + | + | + | + | + | + | + | + | + | ND |
| Mannose | + | + | + | + | + | + | + | + | + | + | ND |
| Melibiose | + | + | + | + | + | + | + | + | + | + | ND |
| Saccharose | + | + | + | + | + | + | + | + | + | + | ND |
| Melezitose | + | + | + | + | + | + | + | + | + | + | ND |
| Salicin | + | + | + | + | + | + | + | + | + | + | ND |
| Sucrose | + | + | + | + | + | + | + | + | + | + | ND |
| Amidon | – | – | + | + | – | – | + | + | + | – | ND |
| Trehalose | + | + | + | – | Var* | + | + | Var | Var | + | ND |
| Arabinose | Var | Var | – | – | – | – | – | – | – | Var | Var |
| Galactose | Var | + | + | + | + | Var | + | + | + | + | – |
| Gentiobiose | – | – | + | – | – | – | + | – | – | – | ND |
| Glycogen | – | – | – | – | – | + | – | – | – | – | ND |
| Inositol | – | Var | – | – | – | – | – | – | – | – | – |
| Inulin | – | – | – | – | Var | – | – | – | – | – | ND |
| Melezitose | – | – | + | + | + | – | + | + | + | – | ND |
| Methyl-D-glucoside | – | – | – | – | – | – | – | Var | – | – | ND |
| Rhamnose | – | – | – | – | – | – | Var | – | – | – | – |
| D-Turanose | – | – | + | – | – | – | + | – | – | – | ND |
| Xylose | – | – | – | – | – | – | – | – | – | Var | – |
| Mannitol | – | – | – | – | – | – | – | – | – | – | ND |
| Adonitol | – | – | – | – | – | – | – | – | – | – | ND |
| D-ribose | – | – | – | – | – | – | – | – | – | – | ND |
| Sorbitol | – | – | – | – | – | – | – | – | – | – | – |
| L-Pyroglutamic acid | – | – | – | + | – | – | – | + | + | – | ND |
| *p*-Nitrophenyl-β-D-glucoside | + | + | + | – | + | + | + | + | + | + | ND |
| *p*-Nitrophenyl-β-D-cellobioside | Var | + | – | – | – | Var | – | – | – | – | ND |

* Var refers to variable results.

[a] Data obtained from Claus and Berkely (1986) and Nazina et al., (2001).

### 2.3.3 Analyses of FAMEs

Fatty acid profiling of the ten *G. stearothermophilus* isolates and one reference strain, *G. stearothermophilus* ATCC 12980, showed that the main fatty acids were iso-fatty acids as well as C16:0 (Table 2.3). The 11 strains could be separated into five groups based on the percentage composition of six main fatty acids: i15:0, a15:0, i16:0, 16:0, i17:0 and a17:0. Figure 2.4 illustrates the percentage composition of these six fatty acids from one strain of each of the five representative groups. Group 1 contained *G. stearothermophilus* ATCC 12980, group 2 contained isolates J1 and K1, group 3 contained isolates P3, I1, H1 and B1, group 4 contained isolate D1 and group 5 contained isolates A1, C1 and G1.

**Table 2.3 FAME percentage composition[*] of the thermophile *G. stearothermophilus* isolates compared with the type strain *G. stearothermophilus* ATCC 12980**

| | ATCC 12980 | A1 | B1 | C1 | D1 | G1 | H1 | I1 | J1 | K1 | P3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10:1 | | 0.1 | | | | 0.1 | | | | | |
| a13:0 | | 0.1 | 0.2 | 0.2 | 0.1 | 0.5 | | 0.1 | | | |
| i14:0 | 0.2 | 2.1 | 2.2 | 2.3 | 0.5 | 2.0 | 2.4 | 1.9 | 0.1 | 0.1 | 2.3 |
| 14:0 | 0.5 | 3.3 | 4.7 | 3.3 | 2.3 | 2.9 | 4.7 | 4.4 | 0.3 | 0.3 | 4.5 |
| 14:1 | 0.2 | 0.3 | | 0.2 | 0.1 | | 0.1 | 0.1 | | | 0.3 |
| **i15:0** | **22.1** | **22.6** | **12.3** | **18.1** | **23.3** | **20.2** | **11.4** | **12.4** | **46.6** | **45.0** | **12.5** |
| **a15:0** | **6.2** | **14.7** | **6.2** | **9.7** | **5.7** | **15.3** | **4.9** | **4.8** | **12.1** | **11.9** | **5.1** |
| 15:0 | 0.2 | 0.9 | 1.6 | 0.9 | 0.7 | 2.0 | 2.4 | 2.0 | 0.1 | 0.1 | 1.9 |
| **16:0 Br** | **6.7** | **15.7** | **18.8** | **18.3** | **5.6** | **13.9** | **22.0** | **19.3** | **2.3** | **2.4** | **20.7** |
| **i16:0** | **7.2** | **18.8** | **32.9** | **21.2** | **24.8** | **16.7** | **33.6** | **36.9** | **5.3** | **5.8** | **33.0** |
| 16:1 | 0.3 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.4 | 0.1 | | 0.6 |
| **i17:0** | **19.7** | **4.5** | **3.3** | **4.4** | **4.6** | **3.6** | **3.6** | **4.6** | **13.4** | **13.7** | **3.9** |
| **a17:0** | **28.8** | **10.3** | **5.7** | **9.2** | **6** | **10.3** | **6.3** | **6.0** | **17.9** | **19.1** | **6.3** |
| 17:0 | 0.3 | 0.3 | 1.4 | 1.4 | 0.4 | 3.2 | 0.5 | 0.9 | | | 0.7 |
| 18:0 | 4.3 | 3.1 | 7.1 | 8.2 | 23.4 | 7.1 | 5.5 | 4.9 | 0.3 | 0.3 | 4.7 |
| 18:1 | 0.3 | 0.4 | 0.4 | 0.4 | 0.7 | 0.4 | 0.1 | 0.2 | 0.5 | 0.5 | 0.4 |
| 18:2 | 0.4 | 0.2 | 0.1 | 0.1 | | | | | 0.1 | 0.1 | 0.1 |
| Other | 2.1 | 1.8 | 1.4 | 1.2 | 1.4 | 1.3 | 1.4 | 1.1 | 0.8 | 0.6 | 2.6 |

[*] Compositions are a percentage of the total amount of fatty acids. The main fatty acids are in bold.

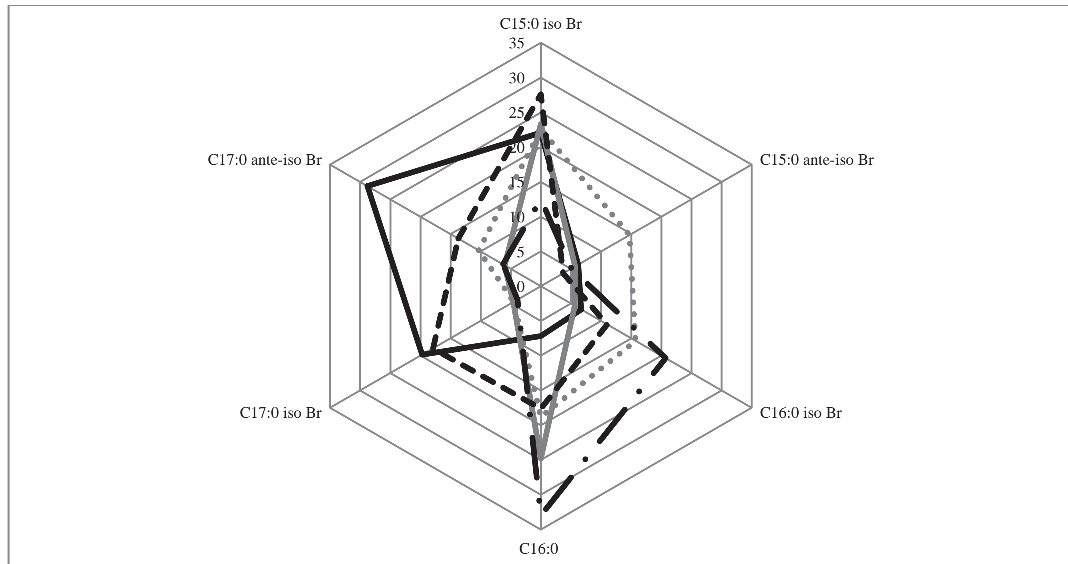



**Figure 2.4. Variation in the percentage of six fatty acids across five *Geobacillus stearothermophilus* strains.** A solid black line represents *G. stearothermophilus* ATCC 12980, grey dots represent strain A1, a solid grey line represents strain D1, black dashes and dots represent strain P3 and black dashes represent strain K1.

### 2.3.4 MALDI-TOF MS analysis

MALDI-TOF MS was used to group the ten *G. stearothermophilus* isolates based on the spectra produced from protein extracts. Based on the database provided by Bruker (Daltonik, GmbH, Bremen, Germany) and supplemented with several thermophilic type cultures, all ten isolates were identified as *G. stearothermophilus,* with P2 confirmed as *Anoxybacillus* (Figure 2.5). Eight of the isolates clustered together, separately to the reference strains. Isolate B1 clustered with the type strain *G. stearothermophilus* ATCC 12980 and isolate G1 did not cluster with any of the other *G. stearothermophilus* strains (Figure 2.5).

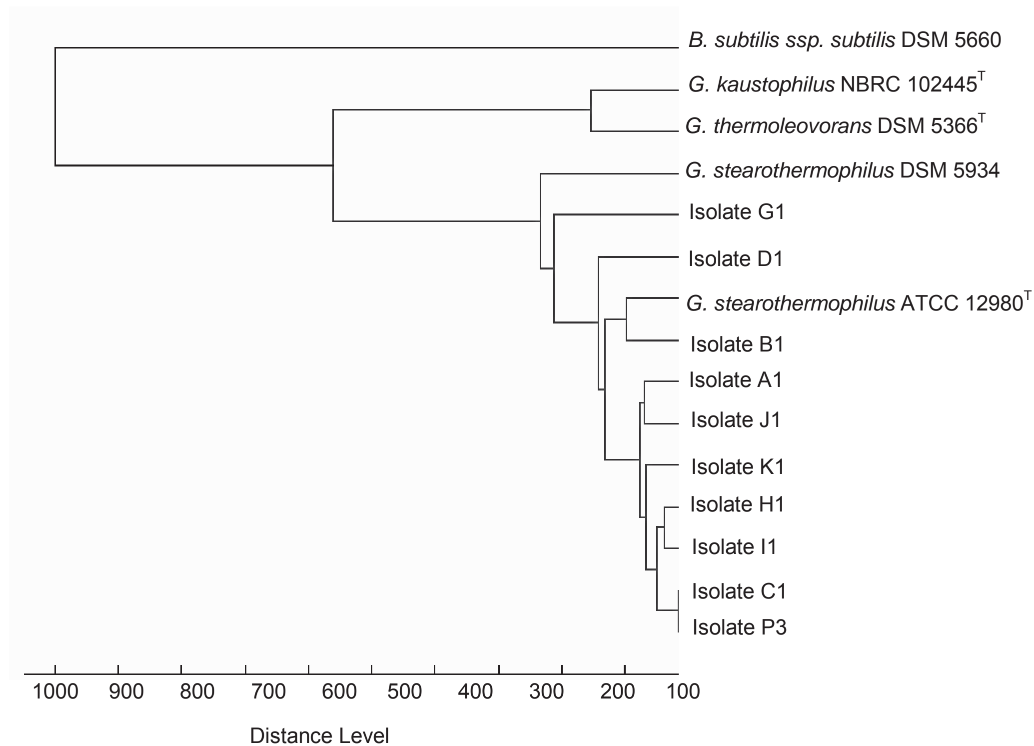



**Figure 2.5. Dendrogram based on MALDI-TOF spectra of the *G. stearothermophilus* isolates.**

## 2.4 Discussion

This study assessed the location of thermophile contamination and evaluated whether strains of *G. stearothermophilus*, isolated from the same milk powder manufacturing plant, had different physiological characterisitics. Isolation of the strains from the surface of the evaporators, rather than the milk, represented the true origin of biofilm formation.

The milk powder plant selected for the study was known to have issues with thermophile growth. To determine where the main source of thermophile contamination was, milk samples were taken at the start and end of the manufacturing run from thirteen sample points throughout the process. Overall, there was little growth, with starting CFUs for most of the sample sites being less than ten and a < 2 log difference between the start and end of run. The final CFU in the concentrate tank was less than 100 CFU/ml. This was probably due to the short manufacturing run of 12 h, whereas standard milk powder manufacturing runs are generally 16 – 20 h. In a study by Murphy et al., (1999), thermophile numbers were > $10^5$ cfu/g after 12 h. However, the starting numbers in that study were very high, approximately $10^3$ cfu/g. Scott et al., (2007) found thermophilic spore numbers increased after 10 -12 h , but they did not monitor the total thermophile counts. It is generally understood in the New Zealand dairy industry that thermophile growth occurs after approximately 10 – 12 h. Here, several locations were analyzed along the plant to the concentrate tank and observed little growth in the parts before the tubular heat exchanger. The colony counts at the end of the manufacturing run remained relatively constant from the first sample point of the evaporator and onwards. This suggests that the main

growth point in the system was the tubular heat exchanger and that growth carried through to the other parts of the plant. To confirm this, samples would need to be taken over multiple manufacturing runs. It is also possible that any vegetative cells originating from growth in the THE may get killed from the high temperature in the DSI, before going into the evaporator where more bacterial growth could occur. To verify this scenario a sample point would need to be introduced between the DSI and evaporator.

It was hypothesized that if the THE was the source of contamination, one would predict that strains isolated from various parts of the evaporator would be physiologically very similar indicating that contamination flowed through the plant (as opposed to arising independently at the different sites).To test this hypothesis, thirty-three strains were isolated from ten locations on the evaporator where the majority of CFU were observed in the milk. These were identified as predominantly *Geobacillus* or *Anoxybacillus*, and one isolate was identified as *B. licheniformis*. This snapshot agrees with other studies that have demonstrated that *G. stearothermophilus* and *A. flavithermus* are the predominant species found in milk powder product and milk powder manufacturing plants (Flint *et al.*, 2001a, Ronimus *et al.*, 2003, Scott *et al.*, 2007, Seale *et al.*, 2012). Milk powder manufacturing plants can be dominated by one thermophilic species or a mix of both (Hill & Smythe, 2012). In this study it was *G. stearothermophilus* which was the dominant species found in the evaporator section. *G. stearothermophilus* has previously been preferentially found in the evaporator section compared with the pre-heat section (Scott *et al.*, 2007).

The identity of ten *G. stearothermophilus* strains were confirmed by 16S rDNA sequencing and selected for in-depth further analyses. The strains were not distinguishable based on their 16S rDNA sequence. This is unsurprising given the 16S rRNA gene is sequence is well conserved between *Geobacillus* spp. Other house-keeping genes have been used for identifying *Geobacillus* spp., but given these also rely on one gene being analyzed giving poor resolution (De Clerck *et al.*, 2004b, Zeigler, 2005, Kuisiene *et al.*, 2009, Weng *et al.*, 2009). A range of methods have also been used for typing *Geobacillus* strains, but these methods also have their disadvantages. For example, RAPDs have poor reproducibility. In contrast, those methods that are more accurate are generally time-consuming (e.g.MLV-HRMA). In this study a different approach was taken. Firstly, biochemical assays and fatty acid profiling were used to identify phenotypic differences between the strains. Secondly, MALDI-TOF was used to confirm the identification of the strains and type them.

The biochemical assays and fatty acid profiling demonstrated that there were differences between the type strain of *G. stearothermophilus* and the selected dairy strains as well as differences between the dairy strains. All ten strains were positive for the methyl red test and negative for casein hydrolysis, gelatin liquefaction, and starch hydrolysis. This agrees with other strains of *G. stearothermophilus* (Claus & Berkeley, 1986, Nazina *et al.*, 2001). Acid was produced from cellobiose and lactose. In contrast, the type strain of *G. stearotherothermophilus* cannot produce acid from cellobiose or lactose (Claus & Berkeley, 1986, Nazina *et al.*, 2001). It

is not surprising that the dairy strains can produce acid from lactose; given this is the predominant sugar in milk, whereas the type strain ATCC 12980 is not able to grow in milk. Importantly, many of the ten isolates analyzed showed a different biochemical profile suggesting these did not all proliferate from the same contamination source or that multiple strains were proliferating in the same location.

The predominant fatty acids were iso-fatty acids, which agreed with the findings of Nazina et al., (2001). However, differences were seen in the amounts of iso-15:0, anteiso-15:0, iso-16:0, 16:0, iso-17:0 and anteiso-17:0. None of the strains grouped with the type strain *G. stearothermophilus* ATCC 12980. Fatty acid composition can differ at different stages of growth. However, preliminary experiments carried out demonstrated that fatty composition from spread-plates grown for 48 h was reproducible. In contrast, fatty acid composition from broth cultures was not reproducible (data not shown). Similar to what was observed with the biochemical profiling, the isolated strains displayed different fatty acid composition compared with each other and the type strain. While the strains were able to be clustered into five groups based on fatty acid composition, the groupings did not correspond with any similarities between strains in the biochemical profiles.

Recently, MALDI-TOF has been shown to be a simple method for the identification and typing of numerous bacterial species (van Baar, 2000, Fernandez-No *et al.*, 2013). To the author's knowledge, this is the first study that has used it for the identification of *G. stearothermophilus*. MALDI-TOF is a simple rapid method that could be used to identify and group strains of *G. stearothermophilus* with similar characteristics. In this work, we found that the majority of the dairy *G. stearothermophilus* isolates grouped together. Notably, the type strain ATCC 12980 grouped with the dairy strain B1 using MALDI-TOF; whereas, comparison of fatty acid composition grouped ATCC 12980 separately. This highlights the importance of using a variety of characterization methods that evaluate different components of the bacteria in order to make the best distinctions between similar strains. In summary, MALDI-TOF was useful in grouping different *Geobacillus* species, but there was not very good sensitivity at distinguishing between the majority of the strains that were isolated.

## 2.5 Conclusions

In conclusion, *G. stearothermophilus* was found to be the dominant thermophilic species isolated from the surface of the evaporator section of a milk powder processing plant. A selection of phenotypic methods was used to differentiate the *G. stearothermophilus* strains. The methods used were chosen to measure different parameters in order to provide a variety of data for comparison. Based on the 16S rDNA sequencing coupled with MALDI-TOF analysis, the majority of the strains isolated were *G. stearothermophilus* and not other *Geobacillus* species. The biochemical assays and fatty acid profiling were able to distinguish between these *G. stearothermophilus* dairy isolates and suggest that there were differences between the dairy

strains and also with the type strain *G. stearothermophilus* ATCC 12980. These differences point to various sources of contamination throughout the evaporator. Alternatively, there could be one source of diverse contamination early on that might be detected with a larger pool of isolates that would allow grouping of the isolates based on the profiling methods used. Assuming the variability observed between the isolates from different sites is representative, then it will be important to determine how this variability could influence the growth and control of *G. stearothermohilus* biofilms in the dairy industry.

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

**STATEMENT OF CONTRIBUTION**
**TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Sara Burgess

**Name/Title of Principal Supervisor:** Steve Flint

**Name of Published Research Output and full reference:**

Burgess S A, Flint S H and Lindsay D (2014). Characterization of thermophilic bacilli from a milk powder processing plant. Journal of Applied Microbiology. 11: 350-359.

**In which Chapter is the Published Work:** 2 & 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  The practical work and writing of the paper were done by the candidate, with guidance and editing by the co-authors.

Sara Burgess (Digitally signed by Sara Burgess; Date: 2016.03.02)
Candidate's Signature

02/03/2016
Date

s.h.flint@massey.ac.nz (Digitally signed by s.h.flint@massey.ac.nz; DN: cn=s.h.flint@massey.ac.nz; Date: 2016.02.23 20:31:19 +13'00')
Principal Supervisor's signature

23/02/2016
Date

Chapter 3

# 3 *In vitro* biofilm formation by *Geobacillus stearothermophilus* on stainless steel

The literature cited in this chapter was up-to-date as of 19 Febuary 2014. This chapter is partially based on the publication by Burgess *et al.*, (2014).

## 3.1 Introduction

The thermophilic bacilli, such as *Anoxybacillus flavithermus* and *Geobacillus* spp., are frequently associated with the contamination of milk powder (Murphy *et al.*, 1999, Flint *et al.*, 2001a, Rueckert *et al.*, 2004, Scott *et al.*, 2007, Zhao *et al.*, 2013, Burgess *et al.*, 2014a). They are generally not considered pathogenic. However, their growth in milk powder manufacturing plants may lead to enzyme and acid production causing off-flavours in the final product (Chopra & Mathur, 1984, Chen *et al.*, 2003, Chen *et al.*, 2004). In addition, they are used as hygiene indicators of a milk powder manufacturing plant because of their ability to form biofilms.

The development of a biofilm of thermophilic bacilli within a milk powder manufacturing plant is still not well understood. Laboratory studies indicate that both vegetative cells and spores can attach to the surface of stainless steel and initiate a biofilm (Seale *et al.*, 2008, Burgess *et al.*, 2009). It is believed that spores have a greater role in the initiation of a biofilm compared with vegetative cells. Spores have a greater propensity for attachment to sufaces compared with vegetative cells (Parkar *et al.*, 2001). In addition, the origin of thermophilic bacilli in milk powder manufacturing plants is believed to be spores in raw milk. It is suggested that the spores survive pasteurisation and attach to surfaces, downstream in the milk powder manufacturing process, in conditions which are optimal for their growth.

Cell surface proteins, hydrophobicity, cell surface charge, a pre-conditioned surface and exopolysaccharide production have been reported to be involved in the attachment of bacteria to a surface (Ronner *et al.*, 1990, Parkar *et al.*, 2001, Faille *et al.*, 2002, Faille *et al.*, 2007, Karunakaran & Biggs, 2011, Somerton *et al.*, 2013). In the case of thermophilic bacilli, attachment appears to involve various factors and may differ between strains. It appears hydrophobicity and cell surface charge may be involved in the attachment of vegetative cells but not spores. Palmer *et al.*, (2010) found that a mutant of *A. flavithermus*, which demonstrated a 10-fold reduction in its ability to adhere compared with its parental strain, also had a lower net-negative surface charge and was less hydrophobic than the wild type. Seale et al., (2008) compared the ability of *Geobacillus* strains to attach. The strain which showed the greatest

numbers of spores attaching to a surface also had the smallest net negative charge and was the most hydrophilic. However, Parkar et al., (2001) found there was no association between hydrobicity and spore adhesion.

Pre-conditioning of a surface may also increase attachment of some strains of thermophilic bacilli. Somerton et al., (2013) found that pre-conditioning the bacteria with cations can increase the ability of vegetative cells to attach to a stainless steel surface for some strains of *A. flavithermus* and *G. stearothermophilus*. Surface proteins may be involved in the attachment of vegetative cells of some strains of thermophilic bacilli (Parkar *et al.*, 2001, Palmer *et al.*, 2010), but had less effect on the attachment of spores compared with the vegetative cells (Parkar *et al.*, 2001). When spores have attached to a surface under the right conditions (i.e.with the correct temperatures for growth and in the presence of nutrients) they will germinate.

Within 6–8 hours, the vegetative cells of thermophilic bacilli will form a steady-state biofilm within a laboratory continuous flow reactor (Flint *et al.*, 2001b, Burgess *et al.*, 2009). It appears that the vegetative cells within the biofilm sporulate, as the biofilm reaches a steady-state within 8 h (Burgess, *et al.*, 2009). In the case of *Geobacillus thermoglucosidasius*, spores appear to be randomly dispersed within the biofilm (Zhao *et al.*, 2013). Sporulation within biofilms of thermophilic bacilli appears to be faster than for mesophilic bacilli. This may be related to the faster growth rate of thermophilic bacilli compared with the mesophiles. In *Bacillus subtilis* and *Bacillus cereus* spore formation occurs in a biofilm after the biofilm has matured (Branda *et al.*, 2001, Lindsay *et al.*, 2005). In addition, *B. subtilis* has been shown to form multiple cell types within a biofilm where different cell types localise to particular areas within the biofilm (Vlamakis *et al.*, 2008). For example, spores tend to localise at the top of the biofilm. Spore formation within biofilms of some of the mesophilic bacilli has been well studied but there are few studies on the thermophilic bacilli.

In this study, strains of *G. stearothermophilus* isolated from the surface of an evaporator in a milk powder manufacturing plant were tested for their ability to form biofilms and produce spores.

## 3.2 Materials and methods

### 3.2.1 Source of strains

The bacterial strains used in this study are listed in Table 3.1. Refer to Chapter 2, Section 2.2.1 for how the strains from this study were isolated from a milk powder manufacturing plant.

**Table 3.1. Bacterial strains used in this study.**

| Strain name | Source | Reference |
|---|---|---|
| *G. stearothermophilus* A1 | Bottom of Evaporator 1 | This study |
| *G. stearothermophilus* P3 | Evaporator 1, Pass 1 distribution plate | This study |
| *G. stearothermophilus* D1 | TVR1 top of calandria tubes | This study |
| *G. stearothermophilus* G1 | Evaporator 1, Pass 2 top of calandria tubes | This study |
| *G. stearothermophilus* I1 | Evaporator 2, Pass 1 top of calandria tubes | This study |
| *G. stearothermophilus* H1 | Evaporator 1, Pass 4 top of calandria tubes | This study |
| *G. stearothermophilus* B1 | Bottom of Evaporator 2 | This study |
| *G. stearothermophilus* J1 | Evaporator 2, pass 2 top of calandria tubes | This study |
| *G. stearothermophilus* K1 | Evaporator 2, pass 4 top of calandria tubes | This study |
| *G. stearothermophilus* C1 | TVR2 top of calandria tubes | This study |
| *A. flavithermus* P2 | Evaporator 1, Pass 1 distribution plate | This study |
| *A. flavithermus* CM | Milk powder | Burgess *et al.* (2009) and Ronimus *et al.* (1997) |

### 3.2.2 Media and culture preparation

For the routine preparation of cultures all of the strains were streak plated onto trypticase soy agar (TSA; BBL, Becton, Dickinson and Company) and incubated at 55 °C for 28 – 48 h. A single colony was restreaked onto TSA and incubated for a further 18 h at 55°C. For the crystal violet assay (Section 3.2.5) and the 80-h biofilm assay (Section 3.2.6) the growth was scraped off the plate and resuspended in 0.1 % peptone to an optical density (OD) of 0.1 at 600 nm.

For the 16-h biofilm assay (Section 3.2.3), biofilms were grown in reconstituted skim milk (RSM; 100 g of skim milk powder to 910 ml of water and autoclaved at 115°C for 15 min). For the 80-h biofilm assay modified RSM was used, the composition of which was: 20 % (v/v) RSM, 70 % (v/v) 0.1 % peptone (Fort Richard Laboratories Ltd., Auckland, New Zealand) and 100 mM MOPS (3-(N-Morpholino) propanesulfonic acid; Calbiochem, Merck KGaA, Darmstadt, Germany).

### 3.2.3 16-h biofilm and spore production assay

As described in Burgess *et al.*, (2014a) a 6-well plastic plate assay containing stainless steel coupons was used to screen the 11 selected isolates for their ability to form biofilms and produce spores. A 16-h incubation time was used to simulate a normal milk powder production run. For each isolate, 10 ml of sterile reconstituted skim milk (RSM) and four sterile 1 $cm^2$ stainless steel coupons with a grade 2B finish were placed into the well of a BD Falcon polystyrene six-well flat bottom microtitre plate (BD Biosciences, Auckland, New Zealand). . Each well was inoculated with a loop-full (using a 10 µl loop) of culture that had been revived on a TSA plate at 55 °C for 24 h. The inoculated microtitre plate was covered and left at room temperature for 45 min to allow attachment of the cells to the stainless steel coupons. Two coupons were removed for the 0-h samples, before incubating the microtitre plate at 55 °C with shaking at 100 rev $min^{-1}$.

After 8 h of incubation, the RSM was removed from each well and replaced with 8 ml of sterile RSM. After a further 8 h of incubation at 55 °C, the remaining coupons were removed and used as the 16-h sample. Total thermophile and thermophilic spore counts were performed on the cells attached to the coupons as described in Section 3.2.4.

### 3.2.4 Plate counts of attached cells

These methods were used to enumerate the number of bacterial cells attached to a stainless steel coupon used in the 16-h (Section 3.2.3) and 80-h (Section 3.2.6) biofilm assays.

#### *Removal of attached cells*

To prepare a suspension of the attached cells, a coupon was removed from the microtitre plate well and gently rinsed by dipping the coupon two times in 10 ml of 0.1 % peptone water. It was transferred to a sterile plastic tube containing 5 g of sterile glass beads and 10 ml of 0.1% peptone water and then vortex mixed for 1 min (Burgess *et al.*, 2014b).

#### *Droplet plate count*

Total thermophile and thermophilic spore counts were performed on the planktonic cells and the cells removed from the coupons using the droplet plate technique (Lindsay & Von Holy, 1999). In brief, serial 10-fold dilutions of the attached cell suspension were prepared in 0.1 % peptone. Pre-poured plates of milk plate count agar (MPCA; Oxoid Ltd., Adelaide, Australia) were divided into four and 50 µl of each dilution were plated onto one segment of the plate. This was carried out in duplicate. Thermophilic aerobic spores were similarly counted after a heat treatment of the cell suspension at 100 ºC for 30 min and enumerated on MPCA plus 0.2 % starch (MPCA+S) using the droplet plate method. A 1 ml spread plate on either MPCA or MPCA+S was also included for a lower detection limit of 1 log cfu/ ml or cfu/$cm^2$. The plates were incubated at 55 ºC for 48 h.

### 3.2.5 Crystal violet microtitre plate assay

To study the formation of biofilms on a polystyrene surface, a crystal violet assay was used for three selected strains of *G. stearothermophilus*: A1, P3 and D1. Bacterial suspensions were

prepared as described in Section 3.2.2. For each strain, 50 µl of the re-suspended culture were inoculated into the well of a 6-well microtitre plate containing 5 ml of the modified RSM medium. A control well was also set-up containing the modified RSM medium with no inoculum. After 8 h of incubation at 55 °C, the medium was removed from each well of the microtitre plate and replaced with 5 ml of fresh modified RSM. After a further 8 h of incubation at 55 °C, the medium was removed from each well and the plate was washed three times with 0.1 % peptone. The plates were stained with 0.1 % crystal violet, incubated at room temperature for 20 min, rinsed again three times with 0.1 % peptone, dried and photographed. To measure the $OD_{590nm,}$ the biofilm was dissolved in 5 ml of 33 % (v/v) acetic acid and a 100 µl aliquot transferred to a 96-well Nunc™ MicroWell™ plate wuth Nunclon™ Delta Surface (Thermo Fisher Scientific New Zealand Ltd, North Shore City, New Zealand). The OD was measured using a microtire plate reader (Spectrostar Nano, BMG Labtech, Germany). The OD values were normalized by using the formula BFI = B/C, where B is the OD of the test well, C is the OD of the control well and BFI is the biofilm forming index.

### 3.2.6 80-h biofilm and spore production assay

Mature biofilms on a stainless steel surface for the three selected strains of *G. stearothermophilus*: A1, P3 and D1 were compared. An 80 h time period was used to enable biofilm maturation and to simulate an area of the milk powder manufacturing plant that was not cleaned properly. Bacterial suspensions were prepared as described in Section 3.2.2. For each strain, 50 ml of the modified RSM was inoculated with a 500 µl of the bacterial suspension. For each time-point, 8 ml of the inoculated modified RSM and two sterile 1 $cm^2$ stainless steel coupons with a grade 2B finish were placed into the well of a six-well plate. The inoculated plate was left at room temperature for 45 min to allow attachment of the cells to the stainless steel coupons. Two coupons were removed for the 0-h samples, before incubating the plates at 55 °C. Two stainless steel coupons were removed at each time-point (8, 16, 32, 40, 56, 64 and 80 h) and the attached cells removed and enumerated as described in Section 3.2.4. After removal of the coupon the modified RSM medium was replaced with fresh medium. The pH of the spent medium was measured before disposal.

### 3.2.7 Scanning electron microscopy (SEM)

Biofilms of three selected *G. stearothermophilus* isolates on stainless steel coupons were prepared for SEM as described by Lindsay and von Holy (1999). Briefly, surfaces were fixed overnight in 4 % aqueous glutaraldehyde. They were then dehydrated in an ethanol series (30, 40, 50, 60, 70, 80, 90, 96 and 100 %) for 10 min at each concentration at room temperature. The coupons were then air dried, coated with carbon and gold/palladium and viewed using an FEI Quanta 200 Environmental scanning electron microscope.

## 3.3 Results

### 3.3.1 16-h biofilm and spore production assay

At the end of a manufacturing run, an evaporator was rinsed with water and then swabbed in ten different locations. Ten *Geobacillus* strains (one isolate from each location within the evaporator) and one *Anoxybacillus* strain were selected and screened for their ability to form biofilms and produce spores over a 16-h incubation period. In addition, *A. flavithermus* CM was included as a positive control as this strain forms biofilms and spores well (Burgess *et al.*, 2009). The incubation time of 16 h was based on the approximate length of time for a milk powder manufacturing run. Figure 3.1 illustrates the total number of bacterial cells and the number of spores attached to a coupon for each isolate. The strains varied in their ability to form biofilms and produce spores in a milk environment.

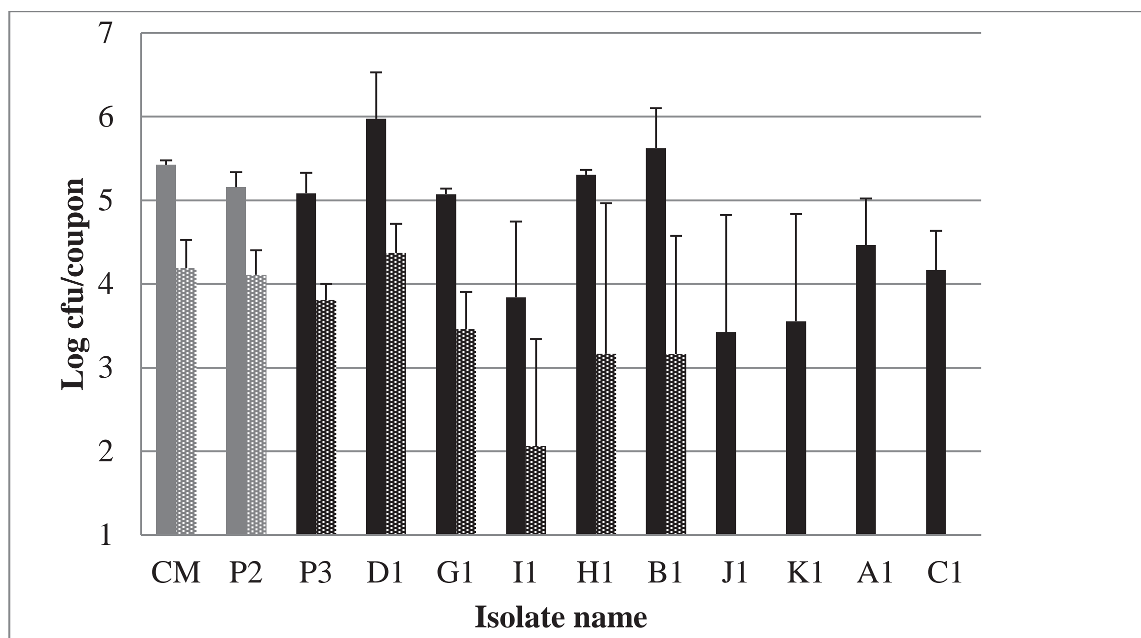



**Figure 3.1. Total thermophile counts (solid filled bars) and thermophilic spore counts (pattern filled bars) attached to the surface of a stainless steel coupon after 16 h of incubation at 55°C in RSM.** Black bars represent *Anoxybacillus* strains and grey bars represent *G. stearothermophilus* strains. Each bar represents the mean and standard deviation of triplicate experiments.

Three strains were selected for further analysis: A1, D1 and P3. Two of these (strains P3 and D1) formed a biofilm and produced spores and one (strain A1) formed biofilms but did not produce spores under the culture conditions used. SEM was carried out on the biofilms of the three selected strains to determine whether the biofilm morphology could account for differences in the cell counts determined by the microtitre plate assay (Figure 3.2). Surprisingly, the two strains that formed biofilms and produced spores (D1 and P3) had different biofilm morphologies. D1 tended to form a flat biofilm community, whereas P3 formed a three-dimensional structure. Spores were visible in the SEM image of D1, which supports the results from the biofilm and spore production assay. It was also noted that isolate D1 exhibited swollen tips at the end of some of the cells, potentially indicative of spore formation and qualitatively corresponding to the higher percentage spore production seen for this strain. While the P3 biofilm was robust (Figure 3.2A) no spores were easily visible. Strain A1 formed a flat biofilm similar to D1, however, no spore-like structures were visible (Figure 3.2C), which qualitatively corresponded to the lack of spores recorded in Figure 3.2.

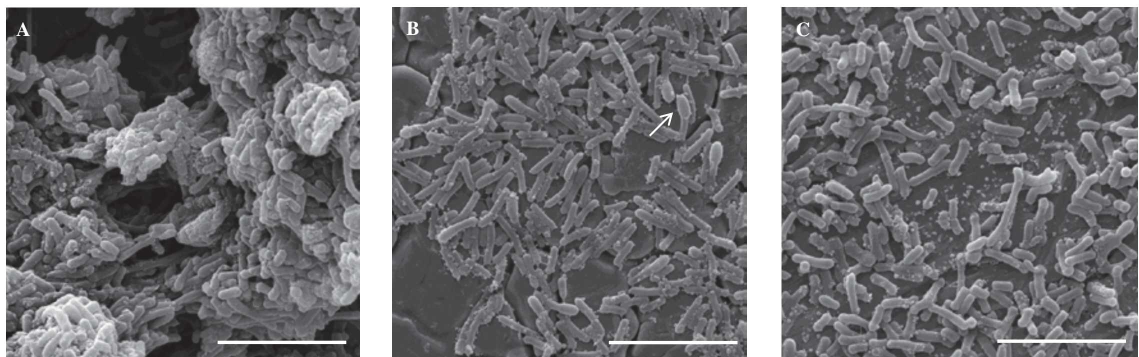



**Figure 3.2. SEM images of three selected *G. stearothermophilus* strains: P3 (A), D1 (B) and A1 (C); grown on the surface of stainless steel in RSM.** The arrow (B) indicates the location of one of the spores. Scale bars, 10 µm.

### 3.3.2 Crystal violet microtitre plate assay

A crystal violet assay was used to quantify the total amount of biofilm produced by the three selected strains and compare differences. After 16 h of growth on the surface of a microtitre plate, the total amount of biofilm was compared between the three strains by calculating the biofilm forming index as described in Section 3.2.5. The total amount of biofilm was less for strain D1 compared with strains A1 and P3 (Figure 3.3). To determine if the decrease in CV staining was due to a general decrease in growth, total cells present in the medium of each well was quantified by measuring CFU. As seen in Figure 3.4, the planktonic cell count was similar for the three strains suggesting that the decrease in biofilm for strain D1 relative to the other strains is not due to a general growth defect in this medium. Visually D1 also formed a different type of biofilm with less uniform CV staining across the well (Figure 3.5).

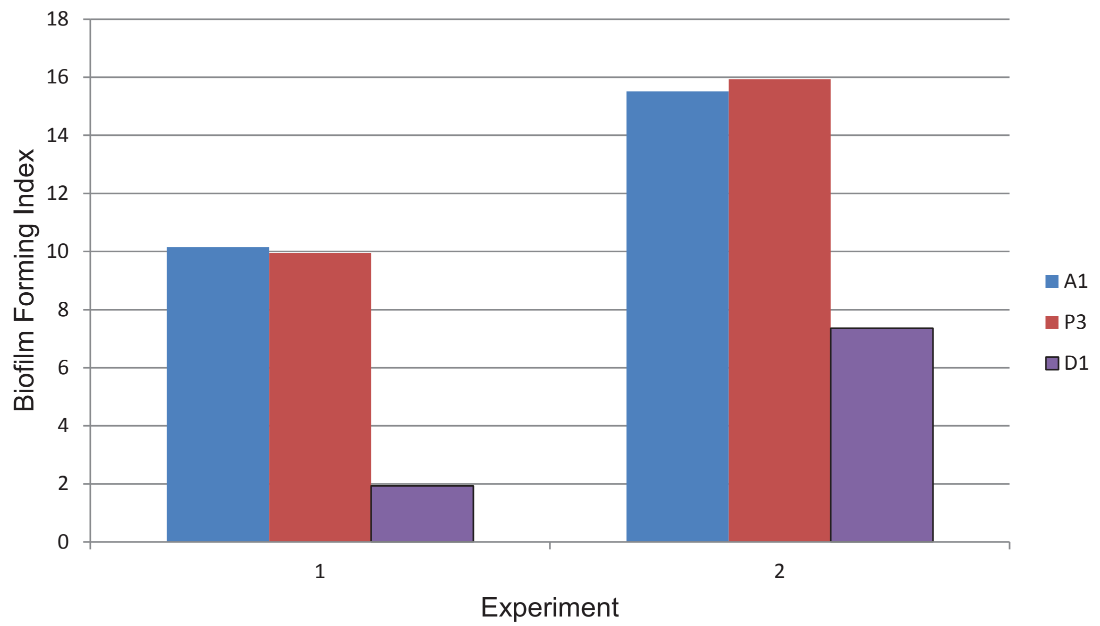



**Figure 3.3. Biofilm forming index based on the $OD_{590nm}$ obtained with the CV assay for the three selected strains of *G. stearothermophilus*.** Blue bars represent strain A1, red bars represent P3 and purple bars represent D1.

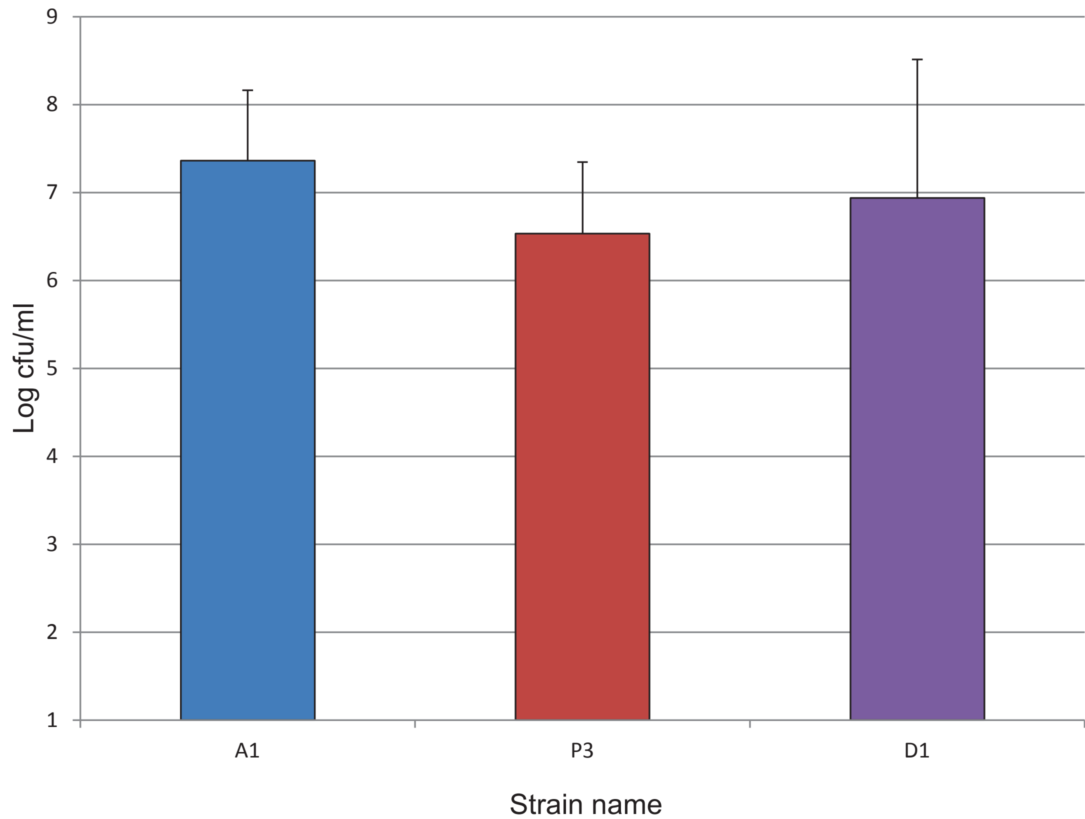



**Figure 3.4. Planktonic cell count of three selected strains of *G. stearothermophilus* after 16-h growth in modified milk medium with the CV assay.** The error bars are the mean of duplicate plates from one representative experiment.

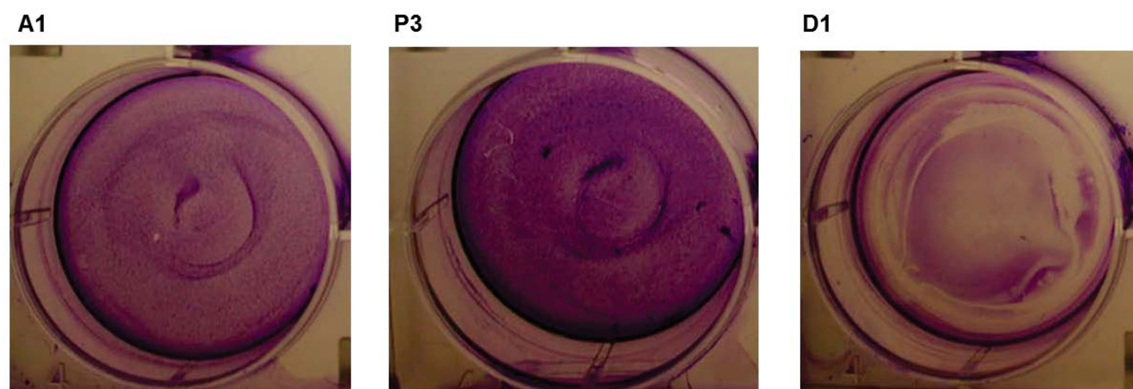



**Figure 3.5. Biofilm formation visualised by CV staining of three selected strains of *G. stearothermophilus* on the surface of a 6-well microtitre plate, grown in modified milk medium for 18 h.**

### 3.3.3 80-h biofilm and spore production assay

To determine whether biofilms of the strains A1, D1 and P3 differed over time in their ability to produce spores, the three strains were grown on stainless steel coupons in a milk based medium over 80-h. The medium was replaced every 8 – 18 h to prevent coagulation of the milk. Figure 3.6 illustrates the total number of bacterial cells attached to a stainless steel coupon for each isolate from one representative experiment. The experiment was repeated two more times with similar results (data not shown), although thermophilic spore counts were only obtained for two experiments.

All of the strains formed biofilms within 8 – 18 h. The three strains reached a similar maximum total thermophile count of approximately 6.5 – 7 log cfu/$cm^2$. The total thermophile count dropped substantially for A1 and to a lesser extent for P3 at 56 h and 80 h. There was also a large pH drop for both strains A1 and P3, over each 18 h incubation period before the medium was replaced (Figure 3.7). For example, the largest pH drop for A1 and P3 was 1.71 and 1.54 respectively, compared with 0.79 for D1.The pH goes back to approximately 6.6 with replacement of the medium.

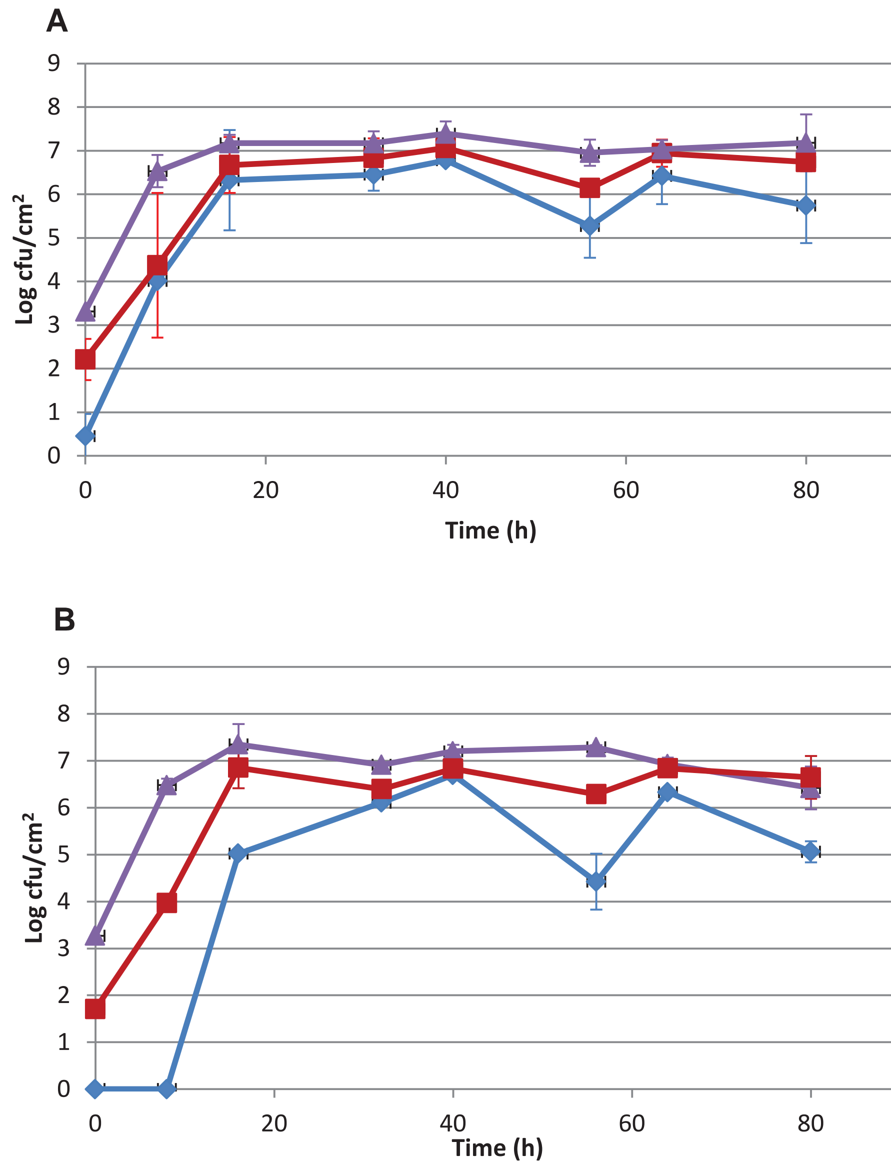



**Figure 3.6. Total thermophile counts (A) and thermophilic spore counts (B) of three selected *G. stearothermophilus* strains attached to the surface of a stainless steel coupon over an incubation period of 80 h at 55°C in 1/5 diluted RSM based medium**. Blue diamonds represent strain A1, red squares represent P3 and purple triangles represent D1. The error bars are the mean of duplicate plates from one representative experiment. The experiment was carried out in triplicate for the total thermophile counts and in duplicate for the spore counts with similar results.

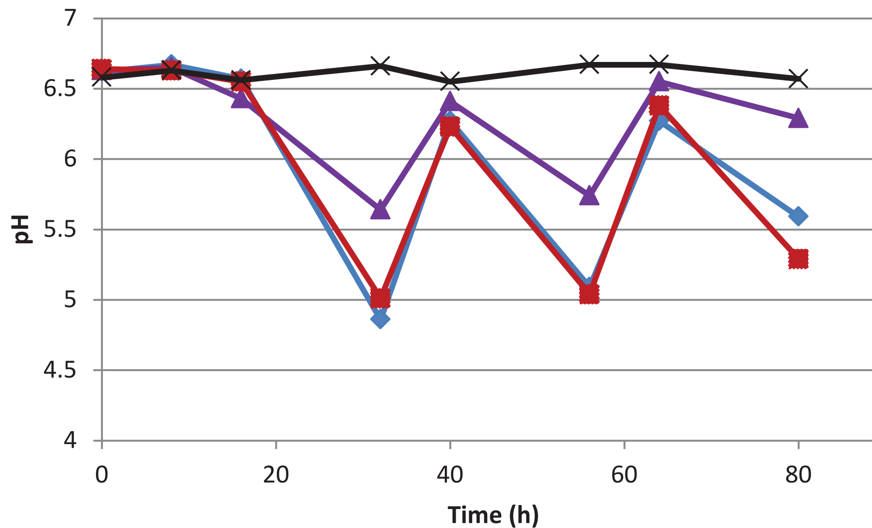



**Figure 3.7. pH readings of the medium, before it was replaced throughout an 80 h incubation period, used for growing biofilms of three selected strains of *G. stearothermophilus.*** After each time point the medium was replaced. Blue diamonds represent strain A1, red squares represent P3 and purple triangles represent D1, a black line with an X symbol represents the control (milk medium only).

The spore numbers differed between the three strains (Figure 3.6b) with A1 having the lowest numbers of spores once it had reached its maximum total cell count. However, A1 had a higher proportion of spores within the biofilm; whereas, the percentage of spores within the biofilm of P3 did not go above 1% (Table 3.2). For all three strains the percentage of spores within the biofilm decreases when the time between sampling and changing the media is 8 h, but increases when that time period is 16 h.

**Table 3.2. Percentage of the total thermophile counts that were spores attached to the surface of a stainless steel coupon for the three selected *G. stearothermophilus* strains.**

| Time | A1 | P3 | D1 |
|---|---|---|---|
| **0 h** | 0 | 0 | 0 |
| **8 h** | 5.8 | 0.024 | 34 |
| **16 h** | 0.20 | 0.0041 | 12 |
| **32 h** | 1.9 | 0.12 | 68 |
| **40 h** | 1.7 | 0.045 | 7.1 |
| **56 h** | 50 | 1.0 | 100 |
| **64 h** | 31 | 0.14 | 4.5 |
| **80 h** | 100 | 0.38 | 1.9 |

Values were obtained from one representative experiment.

SEM was used to observe the morphology of the biofilms over time (Figure 3.8). D1 appeared to have flatter structure, spread over the surface and it did not form a thick dense layer until 80 h. On the other hand, P3 appeared to form discrete microcolonies within 8 h and by 32 h both P3 and A1 had formed a thick dense layer so that the surface of the stainless steel coupon could no longer be seen and it was difficult to see individual bacterial cells. A1 had also formed a thick dense layer and the cells were very hard to see, whereas the bacterial cells of D1 and parts of the stainless steel surface could still be seen. By 32 h the bacterial cells of both D1 and A1 were elongated (Figure 3.9). It was difficult to see whether this also occurred with P3 because the surface was coated with “foulant” by 32 h. Spore formation could be seen within 8 h for all three strains (Figure 3.10). However, A1 appeared to have less cells with swollen tips compared with D1 and the tips of P3 were more swollen compared with A1.

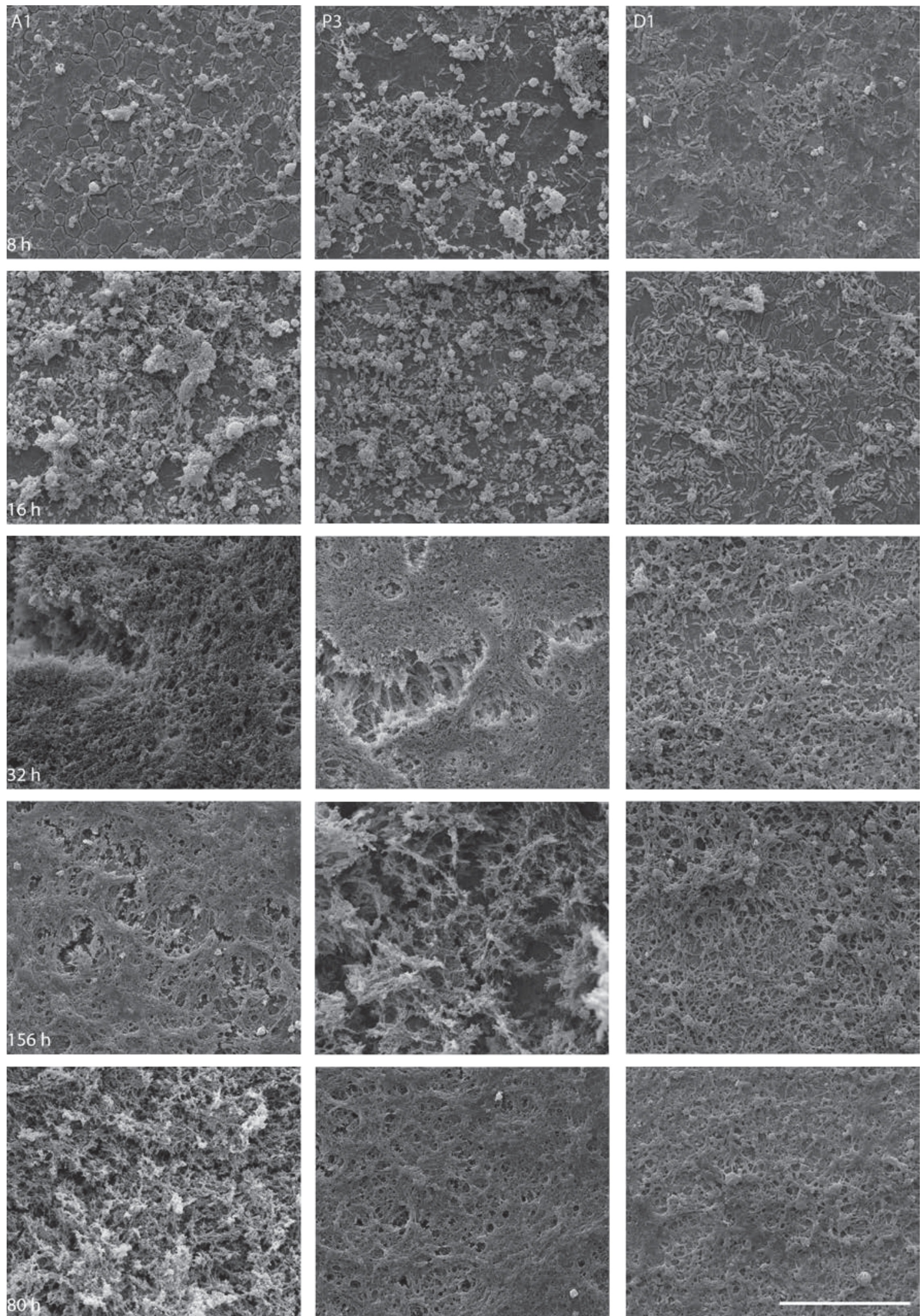



**Figure 3.8. SEM images with a magnification of 1300x of three selected *G. stearothermophilus* strains: A1, P3 and D1; grown on the surface of stainless steel in 1/5 diluted RSM based medium over an 80 h incubation period at 55°C.** Scale bar, 50 µm.

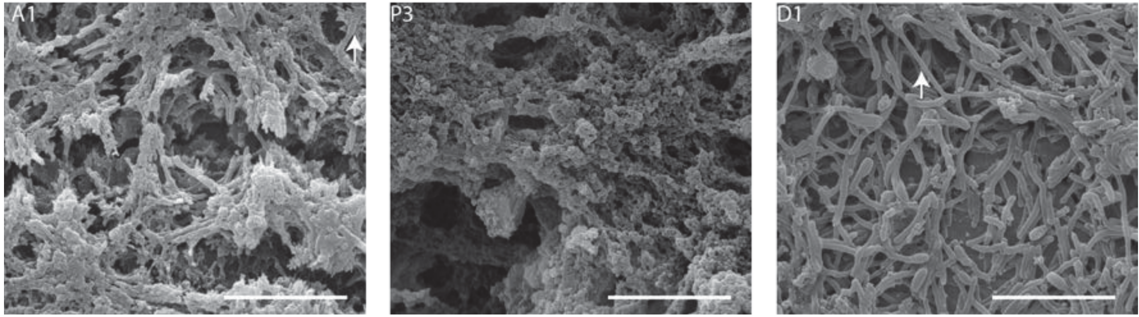



**Figure 3.9. SEM images with a magnification of 5000x of biofilms after 56 h of incubation of the three selected *G. stearothermophilus* strains strains: A1, P3 and D1; grown on the surface of stainless steel.** The arrows indicated elongated cells. Scale bars, 10µm.

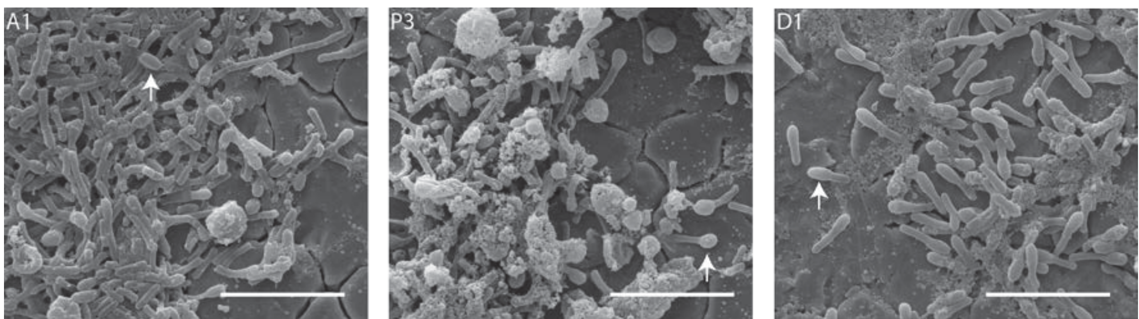



**Figure 3.10. SEM images with a magnification of 5000x showing spore formation within biofilms of the three selected *G. stearothermophilus* strains: A1 P3 and D1; after 8 h of incubation.** The arrows indicate locations of spore formation. Scale bars, 10µm.

## 3.4 Discussion

A stainless steel coupon assay was used to compare the ability of ten strains of *G. stearothermophilus* to form biofilms and produce spores. The strains varied in their ability to form biofilms and produce spores in a milk environment, which supported the results from the study carried out by Flint et al. (2001b). However, in this present study, some of the strains analyzed apparently produced no spores under the culture conditions used, whereas, in the study by Flint *et al.* (2001b), all six strains produced spores. This difference in results may be due to a different biofilm system and a different enumeration method or because fewer strains were analyzed. Flint *et al.* (2001b) used a flow-through reactor system and enumerated the attached cells using impedance. In their study only four of the six strains originated from milk powder. Variation in the ability to form biofilms has been observed for different strains of many bacterial species (Head & Yu, 2004, Rode *et al.*, 2007, Hayrapetyan *et al.*, 2015). Differences in the ability for these strains of *G. stearothermophilus* to form biofilms could have implications in the cleaning of milk powder processing plants in that some strains may be easier to control than others. As discussed in Chapter 2 it also provides further evidence that these strains are

different, meaning that there may be multiple contamination sources or that multiple strains are proliferating in the same location.

Three strains: A1, P3 and D1, were selected for further analysis. A crystal violet assay demonstrated that all three strains appeared to prefer forming biofilms on the bottom of the well rather than at the air/liquid interface. This differs from the results of Zhao *et al*., (2013), which found that thermophilic bacilli preferentially form biofilms on a solid surface at the liquid/air interface. However, a milk powder manufacturing plant is an enclosed system; therefore, it is unlikely there would be regions with an air/liquid interface at the correct temperature for growth of the thermophilic bacilli.

A period of 80 h was used to compare mature biofilms of the three strains A1, P3 and D1. This could also represent an area of manufacturing plant that was not getting cleaned properly. Despite the cells receiving fresh nutrients every 8 – 18 h the total count of the biofilm did not increase once a count of approximately 7 log cfu/cm$^2$ was reached. This observation has previously been noted in dairy strains of *G. stearothermophilus* and *A. flavithermus* (Flint *et al.*, 2001b, Parkar *et al.*, 2003, Burgess *et al.*, 2009); however, in the case of *G. stearothermophilus* only over shorter incubation times. Once the biofilm reached a cell count of approximately 7 log cfu/cm$^2$ there was little change in total cell numbers for the strains D1 as opposed to A1, which had a large drop in cell numbers at 56 and 80 h. This may indicate that A1 had large drops in cell numbers because of sloughing. ‘Footprints’, i.e. areas with larges chunks of cells missing, were not observed in the SEM images for A1, but only a small area was viewed.There was also a large pH drop at time points 56 and 80 h, but given there was a similar pH drop at 32 h with no drop in cell numbers it seems unlikely that this caused the drop in cell numbers for A1. One study carried out on mature biofilms of the thermophilic species *A. flavithermus* found that cell numbers had a cyclic trend where cell numbers repeatably dropped and increased (Parkar *et al.*, 2003). Their explanation for this trend was biofilm sloughing. From a manufacturing plant perspective a strain that is more susceptible to biofilm sloughing into the process fluid will have a greater impact on product quality.

All three strains produced spores. Surprisingly A1 did produce spores within 8 – 16 h, which differed from the 16-h biofilm assay. This may have been because a more dilute milk medium was used, providing less nutrients and therefore promoting spore formation. A1 produced fewer spores compared with P3 and D1, but the proportion of cells that were spores was higher compared with P3. Given that the strain A1 appears to take longer to form a biofilm and produce spores it would be easier to control its growth than that of the other two strains. It is also less of a risk to product quality as the vegetative cells are less likely to survive in a powdered product. When there was a longer period of time before the medium was replaced there was a greater percentage of spores within the biofilm for all three strains. This could indicate that spore formation is more likely to occur in areas of milk powder manufacturing plants where the supply of fresh nutrients is lower. For example, within fouling layers underneath the orifice plates in evaporators.

SEM was used to visualise the biofilm morphology after 16 h of growth in milk as well as throughout an 80 h incubation period in diluted milk medium. In both types of media there were differences in morphology between the three strains, particularly between D1 and P3. Differences in the morphology of biofilms and the composition of those biofilms has been observed for different strains of many bacterial species (Kirisits *et al.*, 2005, Allegrucci *et al.*, 2006, Vlamakis & Kolter, 2011, Chen *et al.*, 2013). For example, in natural isolates of *B. subtilis*, large variation has been observed in biofilm formation between strains (Chen *et al.*, 2013). In that study, the ability to form a robust biofilm correlated with a strain's ability to control disease when used as a biocontrol agent on plants. In this present study, D1 tended to form a flat biofilm community, which has previously been described in similar process biofilms (Burgess *et al.*, 2010) and is probably less susceptible to biofilm sloughing, whereas within 16 h P3 formed a three-dimensional structure that is more commonly seen by other biofilm-forming micro-organisms (Costerton *et al.*, 1994, Branda *et al.*, 2001). Under turbulent flow, biofilms have been shown to be compact rather than open three-dimensional structures (Stoodley *et al.*, 2002); therefore we would not expect the strain P3 to form this type of structure in turbulent areas.

A thick layer of foulant formed on the mature biofilms of both A1 and P3, which was not seen with D1. The foulant was probably caused by the milk coagulating with the lower pH. By 32 h, the bacterial cells of both D1 and A1 were elongated. This is also a characteristic of biofilm formation for *B. subtilis* (Branda *et al.*, 2001). The morphology of biofilms of thermophilic bacilli has not been well studied. In *B. subtilis*, biofilms that form at the air/liquid interface and on the surface of agar plates produce spores at the surface of the biofilm (Branda *et al.*, 2001, Vlamakis *et al.*, 2008). Given a genetic system has not been developed for dairy strains of thermophilic bacilli it is difficult to track spore formation within the biofilm using fluorescent-based or colorometric reporters. However, based on SEM images it appears spore formation occurs within the biofilm. Fluorescence microscopy carried out on *G. thermogluosidans* has also shown that spore production occurs randomly within the biofilm (Zhao *et al.*, 2013).

## 3.5 Conclusions

The strains used in this study were isolated from the surface of ten different locations in the evaporator section of a milk powder manufacturing plant. Isolation of the strains from the surface of the evaporators represented the true origin of biofilm formation. A 16-h stainless steel coupon assay was used to evaluate these strains for their ability to form biofilms and produce spores. Three strains were selected for further study. A longer incubation period was used to look at mature biofilms of the selected strains.

This study has shown that the isolates from the same milk powder manufacturing plant varied in their ability to form biofilms and produce spores, both in numbers and in the morphology of the biofilm. The variability in the biofilm structure of *G. stearothermophilus* dairy isolates has not been previously reported. The importance of how this variability could influence the control of *G. stearothermophilus* biofilms in the dairy industry needs to be determined.

Chapter 4

# 4 Genome sequencing of dairy strains of *Geobacillus stearothermophilus*

## 4.1 Introduction

There is extensive interest in the *Geobacillus* genus for biotechnological purposes such as for bioremediation, the production of thermostable enzymes and biofuels (Feng *et al.*, 2007, Bhalla *et al.*, 2013, Boonmark *et al.*, 2013, Wiegand *et al.*, 2013). With the development of next generation sequencing it has become affordable to sequence multiple bacterial genomes at relatively low cost. To date over 35 *Geobacillus* genomes have been sequenced, mainly to identify genes that could be used in different biotechnological applications.

*Geobacillus* spp. are also common spoilage organisms in food manufacturing plants and products. The genomes of several strains isolated from food manufacturing plants and final product have been sequenced. These include *G. stearothermophilus* ATCC 7953 originally isolated from cream-style corn (Donk, 1920), *G. kaustophilus* NBRC 102445 originally isolated from pasteurized milk and *G. glucosidasius* TNO-09 isolated from casein fouling in a dairy manufacturing plant (Zhao *et al.*, 2012). However, for this study it is the species *G. stearothermophilus* and its growth in milk powder manufacturing plants that is of interest.

Milk is a complex medium for these bacteria to grow in. The main sugar present is lactose. To date the species *G. stearothermophilus* (with the exception of the dairy strains) is regarded as a species that does not produce acid from lactose (Claus & Berkeley, 1986, Nazina *et al.*, 2001). The ability of dairy strains of *G. stearothermophilus* to utilise lactose (as shown in Chapter 2) is likely to be a result of niche adaptation through modification of their genomes.

Comparative genomics has revealed the extent to which bacteria are able to adapt to a specific niche (Bolotin *et al.*, 2004, Wiedenbeck & Cohan, 2011, Loman & Pallen, 2015), often through the loss of gene functionality and gene acquisition by horizontal gene transfer. As an example, both of these mechanisms can be seen with the adaptation of *Streptococcus thermophilus* starter strains to a dairy environment. In the study by Bolotin *et al.* (2004) it was found that in starter strains of *S. thermophilus,* loss of gene functionality occurred with 10 % of the genome containing pseudogenes as opposed to < 5 % in other streptococcal genomes. Many of these pseudogenes were related to carbohydrate uptake mechanisms that were not required in a milk environment. Pseudogenes generally occur as a result of mutations, rendering the gene inactive (Jacq *et al.*, 1977, Goodhead & Darby, 2015). Bolotin *et al.* (2004) also found that starter strains of *S. thermophilus* had acquired genes, such as *lacS,* which encodes a lactose transporter. The LacS permease is one of two different transport mechanisms (Figure 4.1),

which can exist in Lactic Acid Bacteria (LAB), that are used in the production of fermented dairy foods (Price *et al.*, 2012). After lactose is transported via the LacS permease, it is hydrolysed into glucose and galactose, with the galactose being feed into the Leloir pathway. Some strains of *S. thermophilus* are not able to utlise the galactose component and only utilise the glucose (Bogaard *et al.*, 2004). The second mechanism is encoded by the *lacABCDEF* operon, where lactose is transported into the cell by a phosphotransferase system (PTS) and is then utilised via the Tagatose pathway. In both systems, the glucose component is fed into the glycolysis pathway. Some LAB strains have both the lactose permease and lactose PTS systems (Qinglong *et al.*, 2015). Other lactose transportation systems include the lactose-$H^+$ (encoded by the *lacZYA* operon), found for example in enteric bacteria and the ABC protein dependent pathway, which has only been described in *Agrobacterium radiobacter* (Williams *et al.*, 1992). In terms of dairy *Geobacillus*, and to the author's knowledge there have been no lactose transport systems identified in this organism.

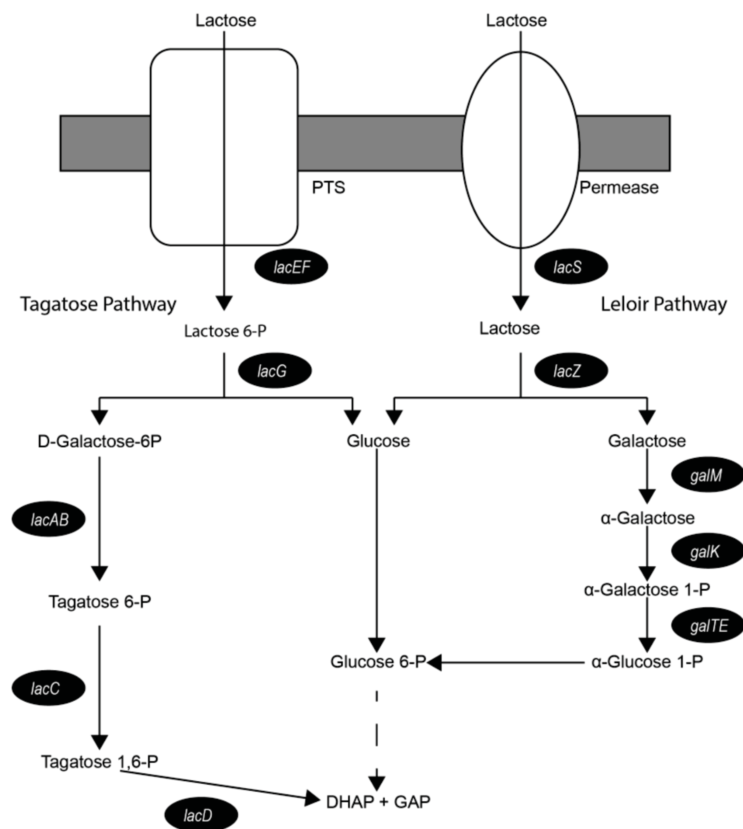



**Figure 4.1 Lactose utilisation in LAB.** Lactose is utilized by either the Tagatose or Leloir pathways, with the glucose component being fed into glycolysis. The genes involved in the Tagatose pathway are found on the *lacABCDEFG* operon and those involved with the Leloir pathway are found on the *galRKTEM-lacSZ* clusters.

In this study, the genomes of three strains of *G. stearothermophilus* A1, P3 and D1 were sequenced and compared with other database strains of *Geobacillus*. The aim of this study was to generate draft genome sequences and to determine whether there were differences in the

genomes of the newly sequenced dairy strains compared with other strains of *Geobacillus* that may provide evidence of adaptation by the three strains A1, P3 and D1 to a dairy environment.

## 4.2 Materials and methods

### 4.2.1 Genome sequencing of three *G. stearothermophilus* strains isolated from a milk powder plant

Three dairy strains of *G. stearothermophilus:* A1, P3 and D1, were selected for genome sequencing based on the results from Chapter 3. They were isolated from three different locations in the evaporator section of a milk powder manufacturing plant as described in Section 2.2.2.

### 4.2.2 Genomic DNA extraction

Cultures were streaked onto TSA and incubated overnight at 55 °C. One colony was transferred into 4 ml of TSB and incubated for 6 – 8 h until visibly cloudy. One ml of the broth culture was then transfered into 100 ml of TSB and incubated for 3 – 5 h until visibly cloudy. The cells were harvested by centrifugation at 5000 *g* for 10 min and washed twice in 4 ml of buffer A (50 mM NaCl, 30 mM Tris (pH8.0), 5 mM EDTA) at 5000 *g*, resuspended in buffer B (25 % sucrose, 50 mM Tris (pH 8.0), 1mM EDTA, 20 mg/ml (in water) lysozyme (Sigma Aldrich), 20 µg/ml (in water) mutanolysin (Sigma Aldrich) and incubated at 37 °C for 45 min. Following incubation, 0.5 ml of 250 mM EDTA (pH 8.0) was added to the mixture and incubated at room temperature for 5 min before adding 200 µl of 20 % (w/v) sodium dodecyl sulphate (SDS) and incubating at 65 °C for 60 – 90 min until the solution was clear. For digestion of protein 10 µl of 20 mg/ml proteinase K (Roche Applied Biosciences) was added to the lysate and incubated at 65 °C for 15 min. The mixture was diluted 1:1 with sterile $dH_2O$, and transferred to Phase Lock Gel heavy tubes (5 PRIME, Inc.). Equal volume of phenol/chloroform/isoamyl alcohol (IAA) (25:24:1 v/v) (Ultrapure™, Sigma Aldrich) was added to the tube, mixed by inverting the tube, incubated at room temperature for 5 min and centrifuged at 10 000 *g* for 10 min. The aqueous phase was removed and transferred to a clean phaselock Gel heavy tube. The phenol/chloroform/isoamyl alcohol step was repeated until there was no apparent layer at the interface. The aqueous phase was transferred to a clean eppendorf and RNase (Roche Applied Science) was added to a final concentration of 100 µg/ml and the mixture incubated at 37 °C for 30 min. The phenol/chloroform/isoamyl alcohol step was repeated and the aqueous phase was transferred to a clean eppendorf tube. An equal volume of chloroform/IAA (24:1 v/v) was added to the aqueous phase, mixed by inverting the tube, incubated at room temperature for 5 min and centrifuged at 13 000 *g* for 5 min. The aqueous phase was transferred to a clean eppendorf and the DNA was precipitated by adding 2 volumes of cold (4 °C) 95 % (v/v) ethanol and storing overnight at 4 °C. The DNA was pelleted by centrifugation at 13 000 *g* for 10 min at 4 °C. The pellet was washed twice in cold (4 °C) 70 % (v/v) ethanol, dried for approximately 30 min at room temperature and resuspended in 50 µl of 10 mM Tris HCl (pH 8.0).

To determine the purity and amount of DNA, gel electrophoresis, spectrophotometry and fluorometry was used. Gel electrophoresis was used to check the integrity of the DNA. 10 µl of

diluted (1/10 and 1/100 in Tris HCl pH 8.0) DNA extract was mixed with approximately 1 µl loading dye (composition: 0.5 g bromophenol blue, 0.25 g xylene cyanol, 5 ml glycerol, 0.186 g EDTA, made up to 10 ml with 1x TBE) and loaded onto a 0.8 % agarose gel containing 1:10, 000 GelRed™ Nucleic Acid Stain (10 000x in water, Biotium Inc.) in 1x TBE buffer (10x TBE composition: 216 g Tris, 11.0g Boric acid, 14.9 g EDTA made up to 2 L with Milli-Q water, pH 8.35). The gel was run at 80 V for approximately 2 h and viewed using the UVI Essential V2 (Uvitec, Total Lab Systems). Distinct bands, with no smearing indicated that the DNA extraction was successful with no degradation. Spectrophotometry was used to give an indication of the amount of DNA and its purity. Two µl of the DNA extract was loaded onto a nanophotometer (Implen, Total Lab Systems) and the absorbance was measured at wavelengths 260 nm and 280 nm. The Qubit® fluorometer (Life technologies™) was used to quantify the amount of DNA. The DNA was prepared using the dsDNA BR Assay (Qubit®, Life technologies) as per the manufacturers' instructions and loaded into the Qubit® fluorometer.

#### 4.2.3 Next Generation sequencing and assembly

Next generation Illumina sequencing was performed on an Illumina MiSeq™ sequencer by New Zealand Genomics Ltd (NZGL) (Massey University, Palmerston North, New Zealand). Aliquots (1.5 µg) of genomic DNA were used to prepare an Illumina TruSeq™ library as per the manufacturers' instructions. The three strains were run on one lane with 250 bp paired-end reads. An adapter-ligated library of the bacteriophage PhiX was used as a positive control. The raw sequence data was mapped against against the Phix genome and any hits removed. Three different quality control (QC) checks were carried out on the sequence data, these included SolexaQA (https://solexqa.sourceforge.net/), fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and fastQscreen (http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/) . In addition, NZGL carried out checks for potential contamination, which involved checking for *E. coli*, PhiX, yeast, Illumina adapters and cloning vectors. The reads were assembled into contigs using the Velvet assembler (v. 1.2.10) (Zerbino & Birney, 2008) using a range of k-mer values from k45 to k249. A total of 104 assemblies were generated for each genome. The assembly with the greatest number of residues was selected for further analysis. In addition, Contig Integrater for Sequence Assembly (CISA, v. 2013.08.22) was used for merging sets of contigs from the various Velvet assemblies in an attempt to generate a more accurate and less fragmented assembly. CISA combines multiple assemblies to generate one amalgamated assembly (Lin & Liao, 2013). This takes advantage of using the different parameters and features incorporated into each individual assembler, or in the case of this present study only one assembler was used (Velvet), but CISA enabled assemblies generated using different kmer lengths to be merged. The assembly with the greatest number of residues was selected for further analysis.

#### 4.2.4 Genome sequences

The genome sequences of 32 *Geobacillus* strains were downloaded from the NCBI (National Center for Biotechnology Information) RefSeq and WGS (Whole Genome Shotgun) project databases to compare with the three newly sequenced dairy strains (Table 4.1).

**Table 4.1. Bacterial genomes[a] used in this study**

| Abbreviated name | Full name | Reference | Source | GenBank accession no. |
|---|---|---|---|---|
| A1 | *G. stearothermophilus* A1 | This study | Milk powder manufacturing plant | LNDS01 |
| P3 | *G. stearothermophilus* P3 | This study | Milk powder manufacturing plant | LNDT01 |
| D1 | *G. stearothermophilus* D1 | This study | Milk powder manufacturing plant | LNDU01 |
| GcCIC | *G. caldoxylosilyticus* CIC9 | Unpublished | Hot spring, Indonesia | AMR001 |
| Gc10776 | *G. caldoxylosilyticus* NBRC10776[T] | Unpublished | Soil, Melbourne, Australia | BAWO01 |
| GicG1w | *Geobacillus icigianus* G1w1[T] | Bryanskaya *et al.*, (2014), Bryanskaya *et al.* (2015) | Hot spring, Valley of Geysers, Russia | JPYA01 |
| GkHTA | *G. kaustophilus* HTA426 | Takami *et al.*, (2004b) | Deep sea sediment, Mariana Trench | NC_006510 |
| GkGbl | *G. kaustophilus* Gblys | Doi *et al.*, (2013) | Phage ØOH2 integrated into the genome of *G. kaustophilus* NBRC 102445 | BASG01 |
| Gk102 | *G. kaustophilus* NBRC 102445[T] | Unpublished | Pasteurised milk | BBJV01 |
| Gs10 | *G. stearothermohilus* 10 (DSM 13240 or BGSC 9A21) | Unpublished | Hot springs, Yellowstone park | CP008934[b] |
| Gs795 | *G. stearothermophilus* ATCC7953 | Unpublished | Cream-style corn | JALS01 |
| Gs362 | *G. stearothermophilus* NUB3621 | Blanchard *et al.*, (2014) | Soil | AOTZ01 |
| GtdNG8 | *G. thermodenitrificans* NG80-2 | Feng *et al.*, (2007) | Oil reservoir, China | NC_009328 |
| Gtd465 | *G. thermodenitrificans* DSM 465[T] | Yao *et al.*, (2013) | Sugar beet juice, Austria | AYKT01 |
| GtgC56 | *G. thermoglucosidasius* C56YS93 | Unpublished | Obsidian Hot spring in Yellowstone Park | NC_015660 |
| GtgTNO | *G. thermoglucosidasius* TNO-09 | Zhao *et al.*, (2012) | Casein fouling, dairy factory | AJJN01 |
| Gtg107 | *G. thermoglucosidasius* NBRC 107763[T] | Unpublished | Soil, Shimogamo, Kyoto, Japan | BAWP01 |
| GtlCCB | *G. thermoleovorans* CCB_US3_UF5 | Unpublished | Hot spring, Malaysia | NC_016593 |
| GtlB23 | *G. thermoleovorans* B23 | Boonmark *et al.*, (2013) | Petroleum reservoir, Niigata, Japan | BATY01 |
| GsuPSS | *G. subterraneus* PSS2 | Unpublished | Dead trees, Puhimau thermal area, Hawaii | JQMN01 |
| GthcGS1 | *G. thermocatenulatus* GS-1 | Zheng *et al.* (2014) | Oil well, China | JFHZ01 |
| GvPSS | *G. vulcani* PSS1 | Unpublished | Dead trees, Puhimau thermal area, Hawaii | JPOI01 |
| GeoWCH | *Geobacillus* sp. WCH70 | Unpublished | Compost, Middleton, Wisconsin | NC_012793 |
| GeoMC61 | *Geobacillus* sp. Y412MC61 | (Brumm *et al.*, | Obsidian Hot spring in | NC_013411 |

| | | 2015b) | Yellowstone Park | |
|---|---|---|---|---|
| GeoMC52 | *Geobacillus* sp. Y412MC52 | Brumm *et al.* (2015a) | Obsidian Hot spring in Yellowstone Park | NC_014915 |
| GeoC56 | *Geobacillus* sp. C56-T3 | (Brumm *et al.*, 2015b) | Sandy's Spring West, Nevada | NC_014206 |
| GeoMC1 | *Geobacillus* sp. Y4.1MC1 | Unpublished | Hot spring in Yellowstone Park | NC_014650 |
| GeoGHH | *Geobacillus* sp. GHH01 | Wiegand *et al.*, (2013) | Botanischer Garden, University of Hamburg | NC_020210 |
| GeoJF8 | Geobacillus sp. JF8 | Shintani *et al.*, (2014) | Bark compost, Okayama, Japan | NC_022080 |
| GeoG11 | *Geobacillus* sp. G11MC16 | (Brumm *et al.*, 2015b) | Compost, Wisconsin | ABVH01 |
| GeoFW2 | *Geobacillus sp.* FW23 | Pore et al., (2014) | Oil well, Gujrat, India | JGCJ01 |
| GeoCAMR1 | *Geobacillus sp.* CAMR12739 | De Maayer *et al.*, (2014) | Iceland | JHUR01 |
| GeoCAMR5 | *Geobacillus sp.* CAMR5420 | De Maayer *et al.*, (2014) | N/A | JHUS01 |
| GeoWSU | *Geobacillus* sp. WSUCF1 | Bhalla *et al.*, (2013) | Soil from a compost facility, Washington State University | ATCO01 |
| GeoMAS | *Geobacillus* sp. MAS1 | Siddiqui *et al.*, (2014) | Hot spring, Pakistan | AYSF01 |
| GeoA8 | *Geobacillus* sp. A8 | Unpublished | Deep mine fissure water | AUXP01 |
| Bs168 | *B. subtilis* 168 | Barbe *et al.* (2009) | Laboratory strain | NC_000964 |
| Bsm347 | *Bacillus smithii* 7_3_47FAA | Unpublished | Inflamed tissue from a patient with Crohn's disease | ACWF01 |
| Sau832 | *Staphylococcus aureus* NCTC 8325 | Gillaspy *et al.* (2006), Herbert *et al.* (2010) | Clinial isolate (from a Sepsis patient) | NC_007795 |
| Sth183 | *S. thermophilus* LMG 18311 | Bolotin *et al.* (2004) | Yoghurt isolate | NC_006448 |

[a] Genome sequences that were available as of November 2014

[b] When the analyses were carried out this genome sequence was not available, so the genome was downloaded from the website http://www.genome.ou.edu/bstearo.html .

### 4.2.5 Gene prediction and annotation

Two methods were used to annotate the three dairy strains of *G. stearothermophilus*: the RAST pipeline (Aziz *et al.*, 2008) and Prokka (v.1.10) (Seemann, 2014). RAST annotation was carried out by uploading the genomes to the RAST server (http://rast.nmpdr.org/). Prokka annotation was integrated into an in-house script[1], which was run using the options described in Table 4.2.

[1] CD: Scripts/Prokka_work3SaraB.pl

**Table 4.2. Options for the in-house Prokka script**

| Option name | Parameter |
|---|---|
| - gram | 'pos' or 'neg' depending on the Gram status of the bacteria |
| - project | Project name e.g. Geobacillus, which is also declared in the project definition of the script |
| - extensionPre | Extension of input files e.g. fa, fasta, fas |
| - extensionPost | Extension of output files e.g. fna[a] |
| - mode[b] | '1' (for Prokka annotation) or '2' (for rMLST) |
| - species | Prokka requires a genus name e.g. Geobacillus |
| - inList[c] | Name of file containing the list of all of the genomes to be processed e.g. inList_Geobacillus |

[a] This is kept the same as the input file if no concatenation of contigs is required prior to analysis.

[b] Mode 1 must be carried out before mode 2 can be initiated

[c] A four column tab delimited text file must be created that lists all of the genomes for analysis. The columns are: name of input file | genus | species | name of output file.

The script was made up of two modes. The overall process of the first mode is illustrated in Figure 4.2, which included annotation of each genome sequence using Prokka (v. 1.10), finding the rRNA genes using RNAmmer (Lagesen *et al.*, 2007) and determining general features such as G+ C content, the number of contigs and total length of each genome sequence. All of the genomes analyzed were parsed in the same manner before any further analysis was carried out. The GBK files of all the genomes can be found on the CD[2]. These are labelled as per the GenBank accession number described in Table 4.1. The second mode carried out an rMLST analysis, as described in Chapter 5 Section 5.2. The outputs of the in-house Prokka script were stored as tables in a MySQL database (v. 5.5.47) (Table 4.3).

[2] CD: GenomeSequences/

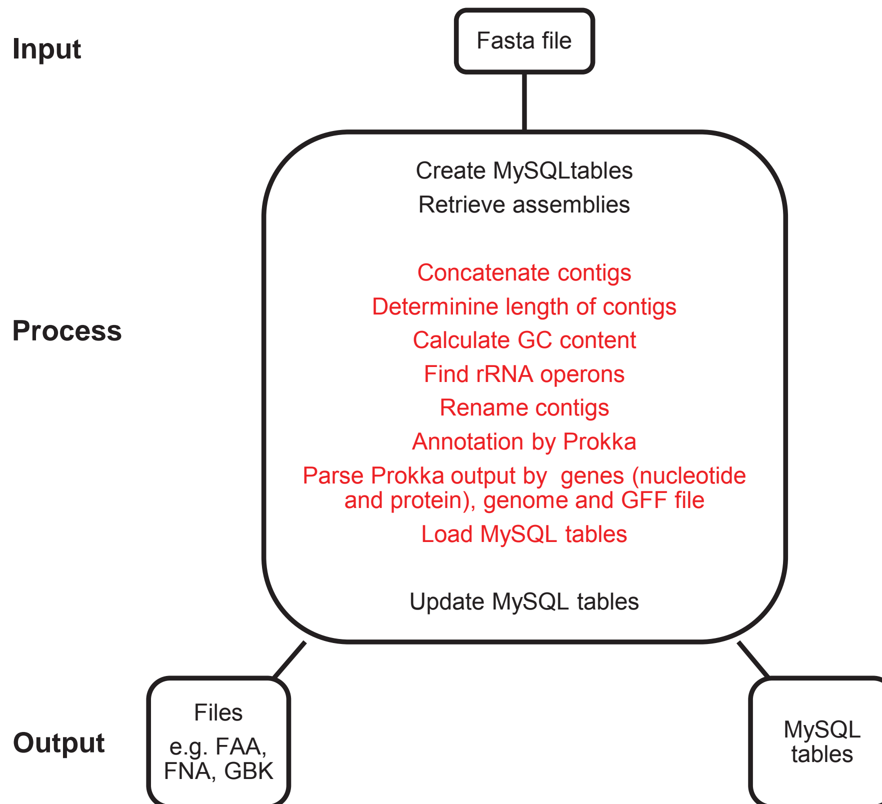



**Figure 4.2. Overall process of mode 1 from the in-house Prokka script for annotating genomes**

**Table 4.3. Description of the MySQL tables generated from the in-house Prokka script**

| Table name | Description |
| --- | --- |
| Geobacillus_genome | Contains the concatenated genome sequence of each strain analyzed |
| Geobacillus_gene | Information on each annotated gene, e.g. includes locus tag, location on the genome, gene sequence and predicted protein sequence. |
| Geobacillus_annotation | Description of each gene, e.g. includes predicted protein function as well as gene name, Pfam numbers and Uniprot numbers if known. |
| Geobacillus_alleles | Allocated allele number for each gene used in the rMLST analysis (mode 2) |
| Geobacillus_statistics | Statistics on each genome, e.g. contains GC content, number of contigs, bases, coding sequences, rRNAs and tRNAs |

### 4.2.6 Prediction of rRNA genes and generation of consensus gene sequences

The 5S rRNA genes were predicted using the web based RNAmmer 1.2 server (Lagesen *et al.*, 2007). The diversity of the 5S rRNA genes was calculated by counting the number of mismatches and dividing by the sequence length. To confirm whether fragments of the 16S and 23S rRNA genes would make a full-length 16S and 23S rRNA gene, a BLASTN database was prepared, as described in Appendix 1, using the 16S rRNA from *Geobacillus* sp. C56-T3 (GenBank Accession NC_014206, GeneID:9235013) and the 23S rRNA from *G. stearothermopilus* (GenBank Accession: X01387) as reference sequences. In order to maximise the chances of getting a full-length 16S and 23S RNA gene,the CISA assemblies were used as

the query instead of the Velvet assemblies. The output file (see Figure A1.1 for an example) was sorted by bit score and used to select the minimum number of contigs with the highest bit score needed to assemble a full-length 16S or 23S rRNA gene. The selected contigs were assembled using SeqMan Pro™, v. 10.1.1 (DNASTAR, Inc., Madison, WI, USA) and a consensus sequence was generated.

To determine any variability within the 16S and 23S rRNA genes for each of the three strains A1, P3 and D1, the consensus sequence was used as a reference for mapping the original sequence reads using Bowtie2, v. 2.1.0 (Langmead & Salzberg, 2012) as described in Appendix 2. Single nucleotide polymorphisms (SNPs) with a variant frequency greater than 8 % were replaced with an ambiguious nucleotide in the consensus sequence. A value of 8% was choosen as assuming there were ten rRNA operons then the variant frequency of a SNP across all of the operons should be greater than or equal to 10%. To account for any error a slightly lower value of 8% was taken. The variant frequency of each SNP was calculated by importing the reference sequence and the mapped reads (i.e the SAM file generated as per Appendix 2) into Geneious (v. 7.0.4), and using the “Find variations/SNPs” function.

### 4.2.7 Assigning predicted proteins to Clusters of Orthologous Groups (COGs)

The COG database was set-up to assign all known proteins in the complete genomes of microorganisms (prokayrotes, eukaryotes and archea) to orthologous groups (Tatusov *et al.*, 2000). COGnitor is used to assign new protein sequences the already established COGs. This is carried out by comparing the new sequence to the COG database by carrying out multiple PSI-BLASTs. In this present study, protein sequences, as predicted by Prokka, were assigned to a COG using the updated COGnitor software tool (Galperin *et al.*, 2015), which was integrated into an in-house perl script[3].

### 4.2.8 Identification of differences in gene presence within the *Geobacillus* genus

To identify genes that were present in the the dairy strains and not present in the other *Geobacillus* spp. genomes, and *vice versa*, two approaches were used; OrthoMCL and bidirectional best hit (BDBH) clustering. The same methods were used for identifying genes that were absent or present only in the entire *G. stearothermophilus* taxon.

#### *OrthoMCL clustering*

OrthoMCL uses a combination of BLASTP and a Markov clustering (MCL) algorithm for grouping genes into orthologous groups (Li *et al.*, 2003). OrthoMCL (v. 2.0.9) was integrated into an in-house script[4], which was run using the options described in Table 4.4. The overall process of this script is illustrated in Figure 4.3. Mode 5 was not used for this study. The initial inputs for this process are the “_gene” and “_annotation” MySQL tables that were generated by Prokka (Section 4.2.5) and the outputs were entered into new MySQL tables. The relationship between all of the MySQL tables is illustrated in Figure 4.3.

[3] CD:Scripts/COGnitorParse2.pl
[4] CD: Scripts/clusterByDB_ProkkaSB.pl

**Table 4.4. Options for the in-house OrthoMCL script**

| Option name | Parameter |
|---|---|
| - mode[a] | 1, 2, 3, 4, 5 |
| - lengthRange | Length range of orthologous gene members e.g. ‘20’ |
| - lengthType | ‘fraction’ or ‘value’ |
| - project | Project name e.g. Geobacillus[b] |
| - database | The name of the MySQL database being used e.g. ‘orthomcl’ |

[a] The modes must be run in order. Refer to Figure 4.3 for a description of the modes.

[b] The same project names need to be used as for the customized Prokka script.

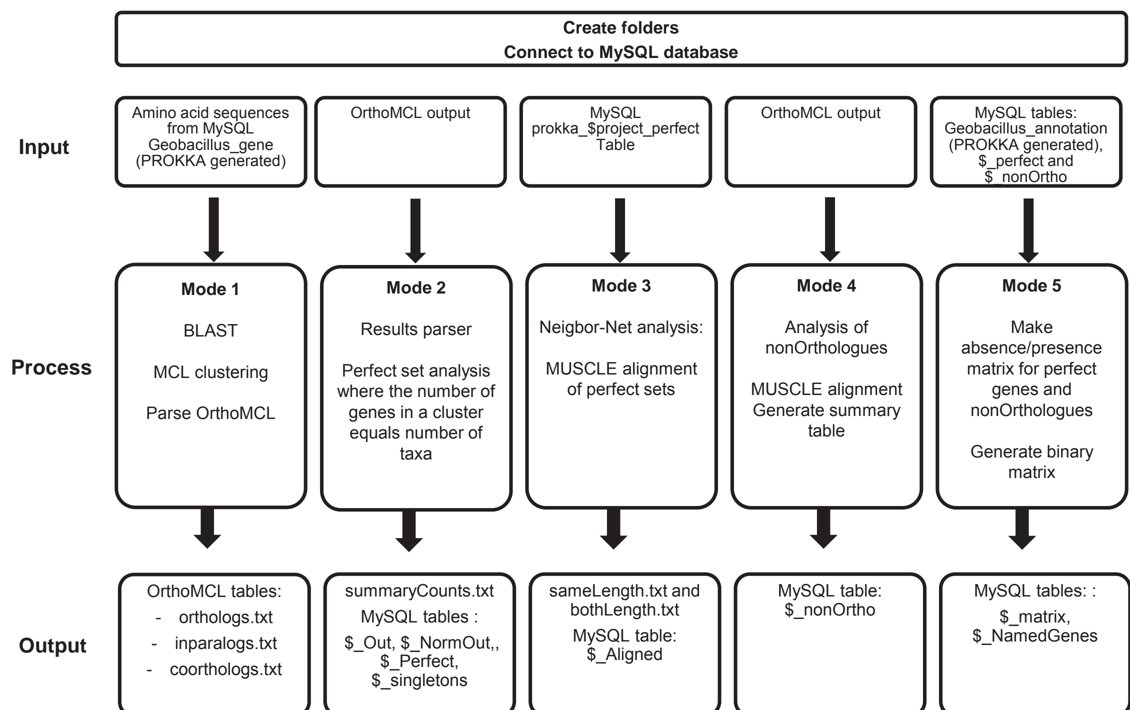



**Figure 4.3. Overall process of the in-house OrthoMCL script**

**Table 4.5. Description of the MySQL tables generated by the in-house OrthoMCL script**

| Name of table[a] | Description |
|---|---|
| **$_Out** | For each cluster it gives a cluster ID, number of genes and number of taxa within that cluster, list of gene locus tags |
| **$_NormOut** | This is a normalized version of the "_Out" table. The same information is contained as above but each locus tag is listed as a separate row along with the predicted protein length |
| **$_Perfect** | Lists the cluster IDs that only contain perfect sets, along with the locus tag and predicted protein length of each member |
| **$_singletons** | Lists all of the locus tags of all of the genes that do not sit within a cluster |
| **$_Aligned** | Lists the aligned sequences within each perfect cluster |
| **$_nonOrtho** | Lists the number of genes from each genome that sits within a cluster that is not perfect |
| **$_NamedGenes** | Gives the name of the gene from each genome that sits within a perfect cluster |

[a] Where $ refers to "Prokka_[name of project]"

Using the code described in Appendix 3, two of the MySQL tables: "Prokka_Geobacillus_Perfect" and "Prokka_Geobacillus_nonOrtho", were used to identify clusters that contained only the three dairy strains of *G. stearothermophilus* and not any of the other *Geobacillus* strains or vice versa. The Perfect table contains those clusters with one gene per taxon. The nonOrtho table contains any clusters with paralogues, so the number of genes does not equal the number of taxa. For those clusters that only contained the three dairy strains of *G. stearothermophilus*, the closest orthologue of the gene from A1, was determined by carrying out a BLASTX against the protein component of the RefSeq (Reference Sequence) database. This database was built by the National Center for Biotechnology Information (NCBI) and contains a collection of non-redundant protein sequences. Any genes that had another *Geobacillus* gene as a best hit were removed from the list.

***BDBH clustering***

The BDBH algorithm using the program GET_HOMOLOGUES (v. 2014112) (Contreras-Moreira & Vinuesa, 2013) was used as an additional tool for confirming gene absence or presence. The BDBH algorithm was run twice: firstly, using A1 as the reference to identify genes only present in the dairy strains and secondly, using *G. thermoglucosidasius* C56T3 as the reference to identify genes absent in the dairy strains or *G. stearothermophilus* taxon. Both runs used 80 % as the cut-off for coverage (option –C) and a percentage identity (option –S). For each run, a pangenome matrix was created; using the pangenome_matrix.pl script and then this was interrogated using the parse_pangeome_matrix.pl script. Both scripts come with the GET_HOMOLOGUES program and more details on how to run these scripts can be found in the GET_HOMOLOGUES manual (http://161.111.227.80/compbio/soft/manual_get_homologues.pdf) .

### 4.2.9 Identification of genes important for lactose metabolism

To identify whether each of the *Geobacillus* strains has the potential to metabolise lactose, KAAS (KEGG automatic annotation server) (http://www.genome.jp/tools/kaas/) was used to assign translated amino acid sequences to KEGG orthology groups. Each genome (in FASTA file format) was uploaded to KAAS using the representative genes data set for prokaryotes and BBDH as the method used to assign orthologues. The KEGG galactose metabolism pathway was used to identify genes absent or present that are required for lactose metabolism via the Leloir or Tagatose pathways (as described in Figure 4.1). The resulting KO (KEGG orthology) assignment files can be found on the CD[5].

### 4.2.10 CRISPR arrays

CRISPR arrays were analyzed as an additional tool for identifying differences between the three dairys strains, A1, P3 and D1. CRISPR arrays were determined by uploading the fasta file of each genome sequence into CRISPRfinder (http://crispr.u-psud.fr/Server/) (Grissa *et al.*, 2007). The potential target of each spacer sequence was identified using CRISPRTarget (http://brownlabtools.otago.ac.nz/CRISPR_WEB/crispr_analysis.html). Each contig (in fasta file format) containg a CRISPR array was uploaded into CRISPRDetect (https://brownlabtools.otago.ac.nz/CRISPRDetect/predict_crispr_array.html) and submitted using the default settings. In the results window each CRISPR array was selected and exported to CRISPRTarget (Biswas *et al.*, 2013). CRISPRTarget uses BLASTN to identify potential targets of the CRISPR RNA, where the spacer sequence is the query and a BLASTN is carried out against selected databases. In this study, the databases selected were Genbank-Phage and RefSeq-Plasmid. The default settings were used for the BLASTN. CRISPRTarget scores each BLAST hit by scoring each nucleotide match with (+1), each mismatch with (-1) and each gap (-10). Unique spacer sequences were identified using CRISPRtionary (http://crispr.u-psud.fr/CRISPRcompar/Dict/Dict.php). Two mismatches were allowed between spacers of the same allele. Annotation of the *cas* genes was carried out using a combination of BLASTX and Prokka.

[5] CD: Chapter4/Results/KOAssignments/

## 4.3 Results

Three strains of *G. stearothermophilus*, namely A1, P3 and D1 were selected for further analysis as described in Chapter 3. These strains originated from three different locations in the evaporator section of a milk powder manufacturing plant. P3 originated from the top of the first evaporator (MVR1), A1 from the bottom of the first evaporator and D1 from the top of the second evaporator (TVR1) (refer to Chapter 2, Table 2.1 for sampling location details).

### 4.3.1 Next generation sequencing and assembly outputs

Next generation sequencing was carried out on the three strains A1, P3 and D1, The coverage for each of the genomes sequenced was approximately 300x (Table 4.6), with the quality of the reads being reasonable as shown in Appendix 4.

**Table 4.6. Coverage for each G*. stearothermophilus* genome sequenced**

| Strain | Number of reads | Average nucleotide coverage |
|---|---|---|
| **A1** | 3 749 346 | 311x |
| **P3** | 3 517 967 | 291x |
| **D1** | 3 841 631 | 323x |

Despite the high coverage, the assembled genomes were very fragmented with strain P3 having the largest number of contigs of 449 (Table 4.7). This appeared to be because these strains contained a large number of repetitive regions greater than 150 bp in length. Other genomes of *Geobacillus* have been fully sequenced that also contain a larger number of repeats; for example, *G. kaustophilus* HTA426 (Figure 4.4).

**Table 4.7. Assembly statistics**

| | A1 | P3 | D1 |
|---|---|---|---|
| **Kmer value used for assembly** | 221 | 209 | 221 |
| **Number of contigs** | 434 | 449 | 398 |
| **Number of contigs > 500 bp** | 328 | 328 | 295 |
| **Mean length of contig*** | 9037 bp | 9056 bp | 9914 bp |
| **Longest contig** | 111887 bp | 100485 bp | 99417 bp |

* Only those sequences greater than 500 bp were used to calculate the mean.

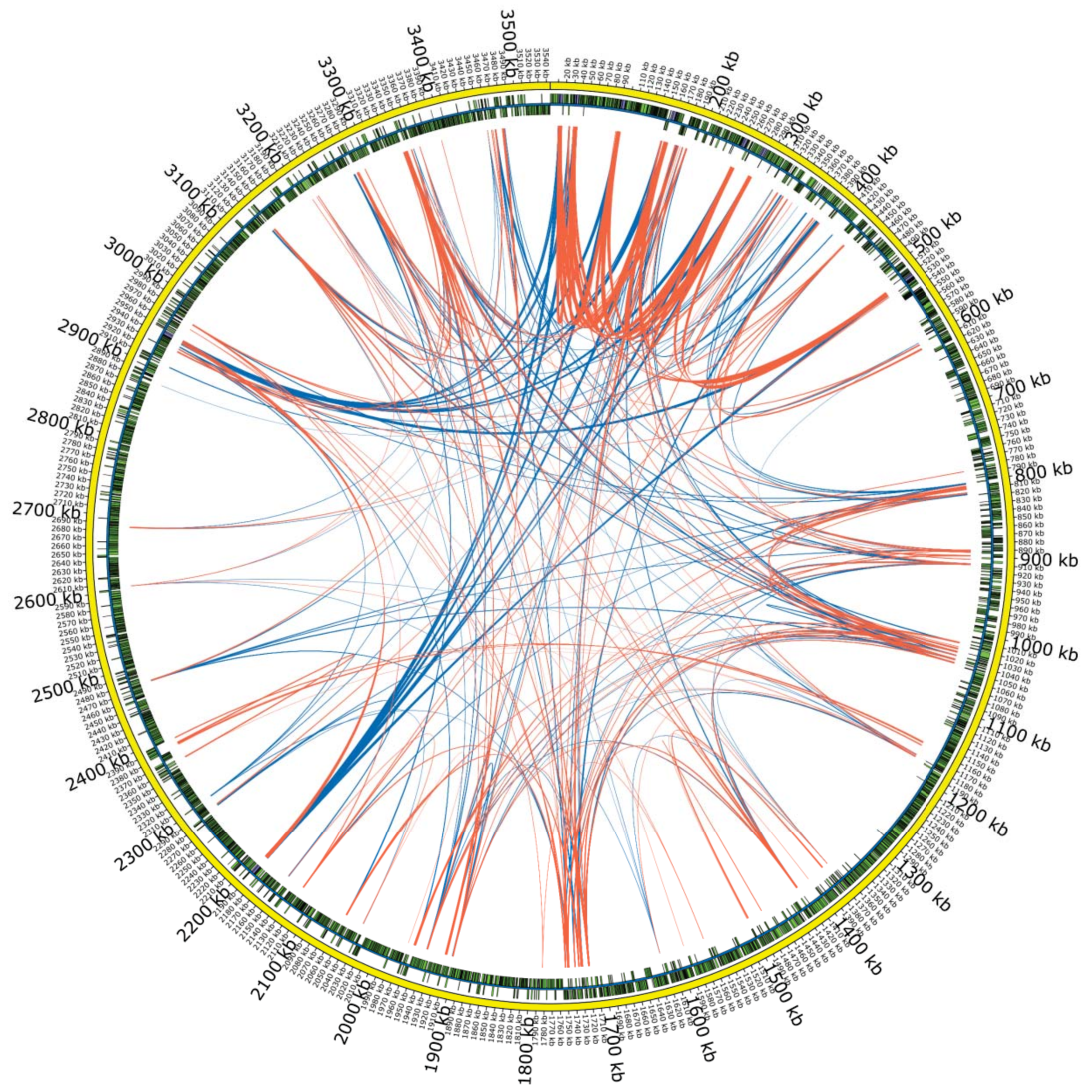



**Figure 4.4. Reputer analysis showing repetative regions greater than or equal to 150bp in the genome of *G. kaustophilus* HTA426.** Different types of repeats are shown in the following colours: forward, red; reverse, dark green; complement, black; palindromic (reverse complement), blue.

### 4.3.2 General genomic features

The genome sizes of the three strains were all approximately 3 Mb (Table 4.8). This was slightly less than most *Geobacillus* strains previously sequenced (approximately 3.5Mb), with the exception of *G. stearothermophilus* ATCC 7593, which has a genome size of approximately 2.8 Mb. Other general features are listed in Table 4.8. General features of the published *Geobacillus* genomes can be found in Appendix 5. The newly sequenced dairy strains have a similar number of coding sequences and of these 67 % were allocated a biological function.

**Table 4.8. General features of the draft genomes for the sequenced *G. stearothermophilus* strains A1, D1 and P3.**

| | A1 | P3 | D1 |
|---|---|---|---|
| **Size (bp)** | 3 012 236 | 3 022 732 | 2 971 009 |
| **Number of coding sequences** | 3169 | 3200 | 3171 |
| **GC content** | 52.02 % | 52.01 % | 52.2 % |
| **Number of pseudogenes** | 234 | 238 | 226 |
| **Number of CRISPR arrays (number of repeats)** | 7 (72, 66, 51, 38, 12, 12 and 5) | 7 (66, 59, 51, 38, 12 12 and 5) | 3 (24, 8 and 5) |
| **Number of tRNAs** | 93 | 93 | 89 |
| **Prophages** | 3 | 4 | 6 |

### 4.3.3 Assembly and analysis of the rRNA genes

To determine whether the rRNA operons were present, each draft genome was uploaded into RNAmmer. The 5S rRNA genes were detected in all of the strains and two fragmented 23S rRNA genes were detected in strain D1, but not the other two strains (refer to the CD[6] for the RNAmmer output containing the 5S rRNA gene sequences). The 16S rRNA genes were not found in any of the three genomes.

The fragmented genome sequences meant that the rRNA operons were also fragmented. When the identified 5S rRNA genes were located on the draft annotated genomes, using Geneious (v. 7.0.4), there was no full-length 23S rRNA gene located on the same contigs. Small fragments of the 16S rRNA genes were found alongside some of the fragmented 23S rRNA genes. This was concerning given RNA operons are fundamental to a bacterial genome and they are traditionally used for studying phylogenies. This section describes how the consensus rRNA gene sequences were determined.

#### *5S rRNA genes*

Using RNAmmer it was determined that each of the three dairy strains, A1, P3 and D1, had 10 copies each of the 5S rRNA gene, all of which were 116bp in length (refer to the CD[6] for the RNAmmer output containing the 5S rRNA gene sequences). The program Geneious (v. 7.0.4) was used to compare the sequences of the ten 5S rRNA gene copies and calculate the variant frequencies (Table 4.9 and Figure 4.5).

[6] CD: Chapter4/Results/RNAmmer_Outputs/

**Table 4.9. Variant frequency of polymorphisms in the 5S rRNA gene sequences of the *G. stearothermohilus* dairy strains.**

| Nucleotide position | Change | Variant Frequency (%) | | |
|---|---|---|---|---|
| | | **A1** | **P3** | **D1** |
| 6 | G/A | 90/10 | 90/10 | 90/10 |
| 13 | C/T | - | - | 90/10 |
| 88 | C/T | 90/10 | 90/10 | - |
| 95 | A/G | 90/10 | 90/10 | 90/10 |
| 107 | C/T | 80/20 | 80/20 | 90/10 |
| 108 | T/C | 90/10 | 90/10 | 90/10 |
| 116 | A/G/T | 70/20/10 | 70/20/10 | 70/20/10 |

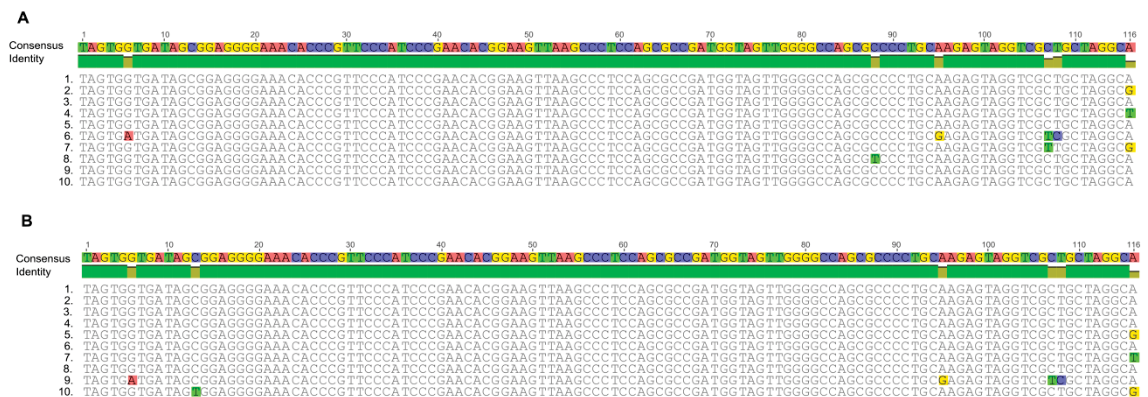



**Figure 4.5. Alignment of the 5S rRNA copies.** (**A**) A1 and (**B**) D1.

In D1 one of the 5S rRNA genes is either a lone copy or part of a split operon, with the surrounding genes of this 5S rRNA gene being the same as those surrounding the 5S rRNA gene from a split operon in the genome of *G. kaustophilus* HTA426 (Figure 4.6). In this case the 5S rRNA gene is separated from the 23S rRNA gene by 398 853 bp.

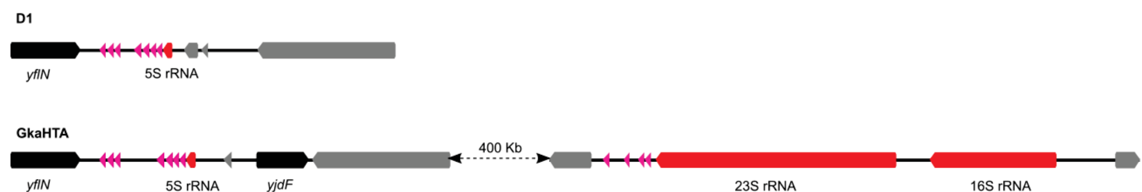



**Figure 4.6. Genetic context of a 5S rRNA gene from strain D1, which sits by itself, compared with the split rRNA operon found in *G. kaustophilus* HTA426.** The rRNA operon is coloured in red, tRNAs in pink, genes encoding hypothetical proteins are in grey and other genes are in black.

The three dairy strains had a 5S rRNA gene diversity of 5.2% (Table 4.10). This is similar to the strain *G. kaustophilus* HTA426.

**Table 4.10. Diversity of the 5S rRNA genes.**

| Strain | Copy number | Sequence length | Percentage diversity (%) |
|---|---|---|---|
| ***G. stearothermophilus*** **A1** | 10 | 116 bp | 5.2 |
| ***G. stearothermophilus*** **P3** | 10 | 116 bp | 5.2 |
| ***G. stearothermophilus*** **D1** | 10 | 116 bp | 5.2 |
| ***Geobacillus*** **sp C56-T3**[a] | 9 | 117 bp | 8.6 |
| ***G. kaustophilus*** **HTA426**[a] | 9 | 116 bp | 4.9 |
| ***G. thermodenitrificans*** **NG80-2**[a] | 10 | 116 bp | 5.5 |
| ***B. subtilis*** **168**[a] | 10 | 118 bp | 6.9 |

[a] Data taken from Pei *et al.* (2012)

### *16S and 23S rRNA genes*

No full-length copy of any of the 16S or the 23S rRNA genes could be found in the final assemblies for the strains A1 and P3, nor was a full-length copy of a 16S rRNA gene found for strain D1; therefore, the draft assemblies were also uploaded into RNAmmer to determine whether a full-length copy of either of these genes could be found. One full length 16S rRNA and one full length 23S rRNA gene was found in each genome using the highest k-mer value of 249, but no 5S rRNA genes were detected. When the k-mer value was decreased no full-length copies of the 16S and 23S rRNA genes were detected, but the 5S rRNA genes were detected.

Given at least 8 copies of the 16S and 23S rRNA genes were still missing, a mapping approach was undertaken to determine a consensus sequence for both the 16S and 23S rRNA genes. The original sequence reads were mapped against the full-length sequence of the two ribosomal genes detected using RNAmmer as described in Section 4.2.3. The consensus 16S and 23S rRNA gene sequences are available on the CD[7].

To determine the diversity of the 16S and 23S rRNA genes and to give an indication of how many rRNA operons were present in each of the newly sequenced strains, the variant frequency of each SNP, insertion and deletion was calculated (Table 4.11 and 4.12). The variant frequency gives an indication of how many ribosomal operons would contain a given SNP in a given gene. For example, the SNP at position 462 in the 16S rRNA gene of D1 (highlighted in grey in Table 4.11) had a variant frequency of 88.9/10.7 %, which would probably equate to 9 out of 10 copies of this gene having nucleotide C and 1 out of 10 copies having nucleotide T.

---

[7] CD: Chapter4/Results/16S_23S_Consensus/

**Table 4.11. Variant frequency of polymorphisms in the 16S rRNA gene sequences of the *G. stearothermophilus* dairy strains.**

| Nt position | Change | Polymorphism type | Variant frequency (%) | | | Coverage | | |
|---|---|---|---|---|---|---|---|---|
| | | | A1 | P3 | D1 | A1 | P3 | D1 |
| 36 | C/T | SNP | 89.1/10.4 | 91.1/8.7 | - | 4036 | 2167 | |
| 67 | G/A | SNP | 74.2/25.5 | 76.3/23.4 | 76.4/23.5 | 4211 | 2705 | 3731 |
| 69 | T/C | SNP | 74.1/25.5 | 76.3/23.4 | 76.3/23.5 | 4225 | 2727 | 3758 |
| 70 | T/G/C | SNP | 74.1/25.7/< 1 | 76.5/23.4/<1 | 57.6/23.8/18.6 | 4227 | 2742 | 3774 |
| 71 | G/A | SNP | 74.3/25.3 | 76.5/23.2 | 76.3/23.6 | 4225 | 2755 | 3790 |
| 72 | G/A | SNP | 90.0/9.8 | 85.4/14.3 | - | 4236 | 2773 | - |
| 73 | G/A | SNP | - | - | 81.7/18. | - | - | 3822 |
| 80 | Plus T | Insertion | 23.5 | 22.3 | - | 4245 | 2850 | - |
| 80 | C/T | SNP | - | - | 58.9/40.9 | - | - | 3908 |
| 81 | T/C | SNP | - | - | 58.8/41.0 | - | - | 3918 |
| 83 | G/T | SNP | - | - | 76.1/23.7 | - | - | 3952 |
| 84 | Minus A | Deletion | 23.3 | 21.9 | - | 4264 | 2921 | - |
| 84 | A/G | SNP | - | - | 76.0/23.7 | - | - | 3982 |
| 87 | C/T | SNP | 74.4/25.0 | 76.5/23.1 | 59.3/40.4 | 4265 | 2962 | 4020 |
| 183 | A/G | SNP | 88.0/11.2 | 89.8/9.6 | - | 4596 | 4256 | - |
| 196 | C/T | SNP | 75.4/24.1 | 74.3/25. | - | 4461 | 4229 | - |
| 451 | A/G | SNP | - | - | 88.4/11.0 | - | - | 5185 |
| 462 | C/T | SNP | 87.5/11.9 | 86.0/13.5 | 88.9/10.7 | 4046 | 3769 | 5216 |
| 482 | G/A | SNP | 87.1/12.1 | 85.9/13.6 | 88.5/11.2 | 4113 | 3800 | 5359 |
| 751 | C/T | SNP | 76.9/22.5 | 78.6/20.4 | 89.7/9.5 | 5316 | 4976 | 6502 |
| 1175 | A/G | SNP | 88.5/10.9 | 87.5/11.8 | - | 5379 | 5148 | - |
| 1428 | C/T | SNP | 87.6/11.8 | 88.0/11.7 | - | 3553 | 3389 | - |
| 1459 | G/T | SNP | 80.7/19.1 | 77.3/22.3 | 67.8/31.9 | 2942 | 2794 | 3144 |
| 1460 | C/T | SNP | 80.2/19.3 | 77.3/22.4 | 67.4/32.2 | 2915 | 2767 | 3093 |
| 1462 | A/C | SNP | 80.4/19.2 | 77.4/22.2 | 67.4/32.2 | 2872 | 2733 | 3042 |

**Table 4.12. Variant frequency of polymorphisms in the 23S rRNA gene sequences of the *G. stearothermophilus* dairy strains.**

| Nt position | Change | Polymorphism type | Variant Frequency (%) | | | Coverage | | |
|---|---|---|---|---|---|---|---|---|
| | | | A1 | P3 | D1 | A1 | P3 | D1 |
| 154 | G/T | SNP | - | - | 90.6/9.1 | - | - | 5577 |
| 455 | A/T | SNP | - | - | 82.7/10.2 | - | - | 8050 |
| 639 | G/A | SNP | - | - | 83.1/16.7 | | | 8286 |
| 690 | C/T | SNP | 58.7/40.5 | 56.6/42.9 | - | 7069 | 6786 | - |
| 892 | G/A | SNP | - | - | 81.7/18.1 | - | - | 8850 |
| 995 | G/A | SNP | 78.2/21.5 | 78.4/21.2 | - | 7524 | 6958 | - |
| 1085 | T/C | SNP | 78.2/21.5 | 77.1/22.5 | - | 6977 | 6754 | - |
| 1460 | G/A | SNP | 89.1/10.3 | 88.0/11.7 | - | 5548 | 5261 | - |
| 2157 | G/A | SNP | 87.4/12.1 | 86.7/12.6 | - | 5134 | 4843 | - |
| 2237 | T/G | SNP | - | - | 89.0/10.8 | - | - | 5858 |
| 2239 | G/T | SNP | - | - | 83.0/16.5 | - | - | 5846 |
| 2654 | C/T | SNP | - | - | 88.5/10.9 | - | - | 6673 |
| 2674 | G/A | SNP | 88.3/11.1 | 88.0/11.4 | 87.1/12.4 | 5522 | 5260 | 6715 |

The number of polymorphisms was used to calculate the diversity of both the 16S and 23S rRNA genes (Table 4.13 and Table 4.14). The diversity of the 16S rRNA gene was quite high, but was similar to that of *G. kaustophilus* HTA426. The diversity of the 23S rRNA gene was a lot lower. As with the variant frequencies A1 and P3 have the same diversity whereas D1 was slightly lower for the 16S rRNA gene and slightly higher for the 23S rRNA gene. The rRNA operon copy number could be 9 or 10 based on both variant frequency data and the number of 5S rRNA genes. It is possible that one of the 5S rRNA genes is an orphan gene. Orphan genes have been found in Bacilli e.g. *Bacillus megaterium*, *B. clausii* and *B. halodurans*. However, in these cases the diversity of the 5S rRNA genes was > 3% (Pei *et al.*, 2012).

**Table 4.13. Diversity of the 16S rRNA genes.**

| Strain | Copy number | Sequence length | Percentage diversity (%) |
|---|---|---|---|
| ***G. stearothermophilus*** **A1** | 9 or 10[a] | 1547 | 1.23 |
| ***G. stearothermophilus*** **P3** | 9 or 10[a] | 1547 | 1.23 |
| ***G. stearothermophilus*** **D1** | 9 or 10[a] | 1547 | 1.16 |
| ***G. kaustophilus*** **HTA426** | 9 | 1553 | 0.77[b] |
| ***G. thermodenitrificans*** **NG80-2** | 10 | 1551 | 1.22[b] |
| **Mean diversity[c]** | | | 0.55[b] |

[a] Estimated copy number

[b] Data taken from Pei *et al.* (2010)

[c] Mean diversity of all of the bacteria analyzed by Pei *et al.* (2010)

**Table 4.14. Diversity of the 23S rRNA genes.**

| Strain | Sequence length | Number of polymorphisms | Percentage diversity (%) |
|---|---|---|---|
| ***G. stearothermophilus*** **A1** | 2926 | 6 | 0.21 |
| ***G. stearothermophilus*** **P3** | 2926 | 6 | 0.21 |
| ***G. stearothermophilus*** **D1** | 2926 | 8 | 0.27 |
| ***G. kaustophilus*** **HTA426** | | | 0.41[a] |
| ***Bacillus*** **species** | | | 0.17 – 0.92[a] |
| **Mean diversity** | | | 0.40[a] |

[a] Data taken from Pei *et al.* (2009)

[b] Mean diversity of all of the bacteria analyzed by Pei *et al.* (2009)

### 4.3.4 Gene prediction and annotation

Both RAST and Prokka were used for gene prediction and annotation. RAST is easier to use in that it is web based (Aziz *et al.*, 2008), whereas Prokka software is run from the command line. Prokka is faster than RAST at annotating one genome (approximately 10 min compared with overnight) and Prokka has the advantage of generating multiple file types (e.g. GBK and SQN files) (Seemann, 2014). Prokka generated files were used for further analyses. This was in part because Prokka was faster and generated multiple file types, but also because in general the annotations appeared to be more reliable. These annotations were used in Chapter 6 as one way of identifying putative genes involved in biofilm formation and sporulation.

Two examples of the resulting annotations, using strain P3, from using these two methods are described below. The first example, (Figure 4.7) shows a region of the genome that encodes for a kinase and modulator that may be involved in biofilm formation and the annotations are described in Table 4.15. The RAST annotation gave the correct description but the protein

names (EpsC and EpsD) were incorrect. In addition, there appear to be no studies describing an EpsX protein. The annotations were verified by carrying out a BLASTX against *B. subtilis* 168.

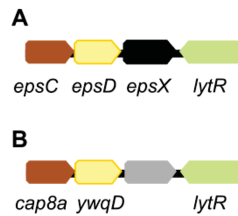



**Figure 4.7. Gene organisation of a region of the genome for *G.stearothermophilus* strain P3 containing putative biofilm genes. A.** RAST annotation**. B.** Prokka annotation**.** Colours represent the encoded function of each gene as annotated by RAST or Prokka. Grey represents a hypothetical protein.

**Table 4.15. Comparison of annotation descriptions between RAST and Prokka for a region of the genome of strain P3 containing putative biofilm genes.**

| Position on genome | RAST description | Prokka description | BLASTX result using *B. subtilis* 168 as a reference | |
|---|---|---|---|---|
| | | | **Top hit** | **Evalue** |
| 1,082,894-1,083,637 | Tyrosine-protein kinase transmembrane modulator EpsC CDS | Capsular polysaccharide type 8 biosynthesis protein Cap8A | *ywqC* (BSU36260)<br>*epsA* (BSU34370) | 7e-84<br>8e-52 |
| 1,083,627-1,084,325 | Tyrosine-protein kinase EpsD | Tyrosine-protein kinase YwqD | *epsB* (BSU34360)<br>*ptkA* (previously known as *ywq*D) (BSU36250 | 1e-77<br>1e-74 |
| 1,084,384-1,085,169 | EPSX protein | Hypothetical protein | No hit | |
| 1,086,142-1,085,201 | Cell envelope-associated transcriptional attenuator LytR-CpsA-Psr, subfamily F2 | Transcriptional regulator LytR 1 | *lytR* (BSU35650) | 3e-127 |

The second example, comparing the two different methods of annotation, shows a region of the genome that contains a CRISPR array (Figure 4.8). The RAST pipeline did not pick up the repeat region; instead it identified regions within the CRISPR array as hypothetical proteins (Table 4.16).

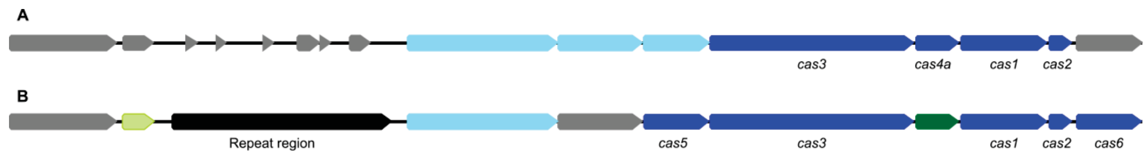



**Figure 4.8. Gene organisation of a region of the genome for *G.stearothermophilus* strain P3 containing a CRISPR array. A.** RAST annotation**, B.** Prokka annotation**.** Colours represent the encoded function of each gene as annotated by RAST or Prokka as follows: grey, hypothetical protein; light green, flagellar assembly protein; black, repeat region; light blue, CRISPR associated genes; dark blue *cas* genes; and dark green, member of the PD-(D/E)XK nuclease superfamily.

**Table 4.16. Comparison of annotation descriptions between RAST and Prokka for a region of the genome of strain P3 containing a CRISPR array.**

| Position | RAST description | Prokka description |
|---|---|---|
| 1-1221 | Hypothetical protein | Hypothetical protein |
| 1289-1654 | Hypothetical protein | Flagellar assembly protein H |
| 1851-4347 | ------------------------- | Repeat region |
| 2004-2141 | Hypothetical protein | ------------------------- |
| 2352-2465 | Hypothetical protein | ------------------------- |
| 2890-3009 | Hypothetical protein | ------------------------- |
| 3268-3534 | Hypothetical protein | ------------------------- |
| 3531-3659 | Hypothetical protein | ------------------------- |
| 3868-4104 | Hypothetical protein | ------------------------- |
| 4527-6242 | CRISPR-asscociated protein, TM1802 family | CRISPR-asscociated protein, TM1802 |
| 6244-7203 | CRISPR-asscociated protein, TM1801 family | Hypothetical protein |
| 7216-7959 | CRISPR-asscociated protein, TM1800 family | CRISPR-asscociated protein Cas5 |
| 7961-10291 | CRISPR-asscociated helicase Cas3 | CRISPR-asscociated nuclease/helicase Cas3 |
| 10301-10810 | CRISPR-asscociated RecB family exonuclease Cas4a | PD-(D/E)XK nuclease superfamily |
| 10813-11814 | CRISPR-asscociated protein Cas1 | CRISPR-asscociated endonuclease Cas1 |
| 11825-12088 | CRISPR-asscociated protein Cas2 | CRISPR-asscociated endoribonuclease Cas2 |
| 12129-12875 | Hypothetical protein | CRISPR-asscociated protein Cas6 |

### 4.3.5 COG for assigning protein function

The COG system was developed to assign all known proteins in the complete genomes of microorganisms to orthologous groups. Each of these orthologous groups can also be assigned to a general category, termed COG functional categories, of which there are 26. To give an overview of how the proportion of genes assigned to a COG functional category might differ across the *Geobacillus* genus a heat-map was generated (Figure 4.9). The COG functional category which showed the most variation between the genomes was category X, which contains the mobilome. The importance of the mobilome in genome plasticity and adaptation of a *Geobacillus* strain to a new environment has not bee previously studied. However, one would hypothesize that the mobilome would be important for a *G. stearothermophilus* isolate to adapt from a soil or silage environment to a dairy environment. In addition, one would expect strains with a high mobilome content to have fewer mechanisms (such as the CRISPR-Cas system) against invading DNA.

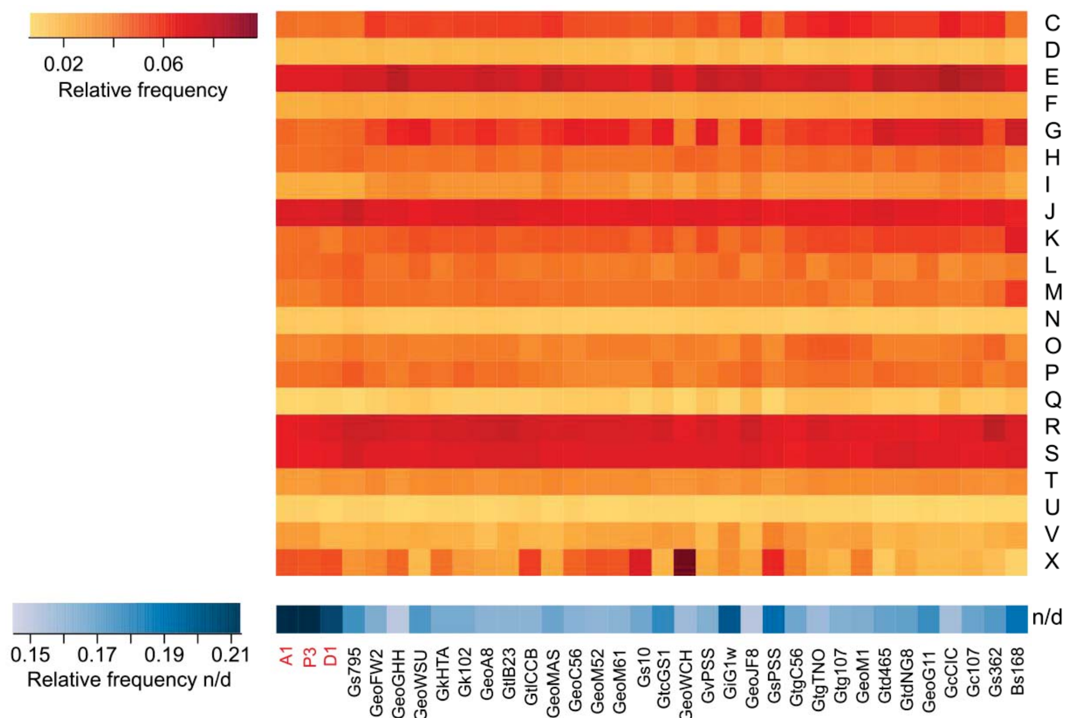



**Figure 4.9. Proportion of coding sequences assigned to a protein function for the three *G.stearothermophilus* dairy strains (coloured in red) compared with other strains of *Geobacillus* spp. and *B. subtilis.*** The letters refer to the COG functional categories: (C) energy production and conversion, (D) cell division and chromosome partitioning, (E) amino acid transport and metabolism, (F) nucleotide transport and metabolism, (G) carbohydrate transport and metabolism, (H) coenzyme metabolism, (I) lipid metabolism, (J) translation, ribosomal structure and biogenesis, (K) transcription, (L) DNA replication, recombination and repair, (M) cell envelope biogenesis outer membrane, (N) cell motility, (O) post-translation modification, protein turnover, chaperones, (P), inorganic ion transport and metabolism, (Q) secondary structure, (R) general function prediction only, (S) function unknown, (T) signal transduction, (U) intracellular trafficking, secetion and vesicular transport, (V) defense mechanisms, (X)

mobilome: prophages and transposons. In addition, a large number of protein sequences were not able to be determined (n/d).

The absence and presence of predicted proteins within each COG as well as variation within the *Geobacillus* genus was also analyzed; where 2147 COGs did not contain any proteins from the *Geobacillus* genus[8], 433 COGs had predicted protein(s) present from the *Geobacillus* genus with no variation between the genomes analyzed[6] and 2052 COGs were variable across the *Geobacillus* genus[9]. For the Prokka generated locus tags of a gene encoding a protein within a particular COG refer to the CD[10].

### 4.3.6 Differentiation of the *G. stearothermophilus* dairy strains

Three methods: BDBH clustering using the program GET_HOMOLOGUES, OrthoMCL clustering and a COG approach, were used to identify differences between the three newly sequenced *G. stearothermophilus* strains, and the other *Geobacillus* strains. Both the BDBH and OrthoMCL clustering produced similar results. Using the BDBH method there were 115 COGs that only contained the three *G. stearothermophilus* dairy strains and of these 75 % contained genes encoding for hypothetical proteins and 5 % for transposases. Sixteen of the COGs contained genes with homologues in other *Geobacillus* species, identified using BLASTX. Table 4.17 describes the seven remaining genes that had assigned biological function. Of particular interest were *lacA*, *lacB* and *lacC*. These will be described in more detail in Section 4.3.7. There were no COGs that were unique to the three dairy strains.

Using both the BDBH and OrthoMCL clustering it was found that genes present in the other *Geobacillus* genomes were missing from the three dairy strains of *G. stearothermophilus* (Table 4.18). A COG approach was also used to identify potentially defective biological functions by searching for COGs that did not contain any genes from the three dairy strains of *G. stearothermophilus*, but were present in the other *Geobacillus* genomes. These strains did not contain predicted proteins in six COGs (Table 4.19).

---

[8] CD: Chapter4/Results/COG/COG_GeobacillusPresentAbsent.xlsx
[9] CD: Chapter4/Results/COG/COG_GeobacillusVariable.xlsx
[10] CD: Chapter4/Results/COG/COG_LocusTag

**Table 4.17. Selected genes only found in *G. stearothermophilus* A1, P3 and D1**

| Prokka Annotation | | | Predicted function/role(s) | Pfam | | BLASTX | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Locus tag** | **Gene name** | **Gene description** | | **Pfam number** | **Pfam family, description** | **Best homologue found in*:** | **Description** | **% ID** | **% query cover** | **E-value** | **Bit score** |
| **A1_00879, P3_03064, D1_00064** | *lacR* | Lactose phosphotransferase system repressor | Involved in metabolism of lactose and galactose via Tagatose pathway. | PF08220<br>PF00455 | DeoR-like helix-turn-helix domain<br>DeoR C terminal sensor domain | *Bacillus smithii* | DeoR family transcriptional regulator | 89 % | 97 % | 1e-129 | 375 |
| **A1_00880, P3_03065, D1_00065** | *lacA* | Galactose-6-phosphate isomerase subunit LacA | | PF02502 | Ribose/Galactose Isomerase | *Bacillus smithii* | Galactose-6-phosphate isomerase | 96 % | 99 % | 1e-94 | 281 |
| **A1_00881, P3_03066, D1_00066** | *lacB* | Galactose-6-phosphate isomerase subunit LacB | | PF02502 | Ribose/Galactose Isomerase | *Bacillus smithii* | Galactose-6-phosphate isomerase | 95 % | 99 % | 5e-91 | 276 |
| **A1_00883, P3_03068, D1_00068** | *lacC* | Tagatose-6-phosphate kinase | | PF00294 | pfkB family carbohydrate kinase | *Bacillus smithii* | Tagatose-6-phosphate kinase | 95 % | 99 % | 0.0 | 617 |
| **A1_00229, P3_00246, D1_02257** | | PIN domain | Involved in nutritional stress response | PF01850 | PIN domain | *A. flavithermus* | Nucleic acid-binding protein | 97 % | 98 % | 1e-62 | 201 |
| **A1_01659, P3_03135, D1_02950** | *ftsH* | ATP-dependent zinc metalloprotease FtsH | Involved in heat shock or a stress response | PF00004 | ATPase family associated with various cellular activities | *Butyrivibrio fibrisolvens* | Hypothetical | 57 % | 99 % | 0.0 | 578 |
| **A1_00955** | | Phage tail protein | | | | *Anoxybacillus sp. ATCC*<br>*Geobacillus* phage GBSV1 | Hypothetical protein<br>Hypothetical protein | 98 %<br>62 % | 99 %<br>97 % | 4e-176<br>1e-94 | 502<br>290 |

* A BLASTX was carried out using the A1 nucleotide sequence as the query.

**Table 4.18. Selected genes absent in *G. stearothermophilus* A1, P3 and D1**

| | Prokka annotation | | Pfam | | | |
|---|---|---|---|---|---|---|
| **Locus tag of *G. kaustophilus* HTA** | **Gene name** | **Gene description** | **Pfam number** | **Pfam description** | **K0 number** | **Predicted function** |
| **NC_006510_02146** | *ywpJ* | Putative phosphatase | PF08282 | Haloacid dehydrogenase superfamily | K07024 | Uknown |
| **NC_006510_00756** | *ygaZ* | Inner membrane protein | PF03591 | AzlC | N/A* | Unknown |
| **NC_006510_03460** | *nuoM* | NADH-quinone oxidoreductase | PF00361 | NADH dehydrogenase | K00342 | Subunit of NADH dehydrogenase required for respiration |
| **NC_006510_03469** | *ndhC* | NAD(P)-H-quinone oxidoreductase subunit 3 | PF00507 | NADH dehydrogenase | K00330 | Subunit of NADH dehydrogenase required for respiration |
| **NC_006510_01148** | N/A* | Nitrate monooxygenase | PF03060 | Nitronate monooxygenase | K02371 | Fatty acid biosynthesis |
| **NC_006510_00751** | *lcfB* | Long-chain-fatty-acid CoA ligase | PF00501 | AMP-binding enzyme | K01897 | Fatty acid biosynthesis |
| **NC_006510_00429** | *xdhA* | Xanthine dehydrogenase molybdenum-binding subunit | PF01315 & PF02738 | Aldehyde oxidase and xanthine dehydrogenase a/b hammerhead domain & Molybdopterin-binding domain of aldehyde dehydrogenase | K00087 | Molybdate metabolism |
| **NC_006510_00328** | *rebO* | Flavin-dependent L-tryptophan oxidase RebO precursor | PF01593 | Flavin containing amine oxidoreductase | N/A* | Unknown |

* Not applicable as there was no assigned gene name or K0 number.

**Table 4.19. COGs absent in *G. stearothermophilus* A1, P3 and D1 and present in the other *Geobacillus* spp.**

| COG | Functional Category | COG description[a] |
|---|---|---|
| **COG0838** | C | NADH:ubiquinone oxidoreductase subunit 3 (chain A) |
| **COG1008** | C | NADH:ubiquinone oxidoreductase subunit 4 (chain M) |
| **COG1231** | E | Monoamine oxidase |
| **COG1296** | E | Predicted branched-chain amino acid permease (azaleucine resistance) |
| **COG1529** | C | CO or xanthine dehydrogenase, Mo-binding subunit |
| **COG4377** | S | Uncharacterized membrane protein YhfC |

[a] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

In addition to the recently sequenced *G. stearothermohilus* strains, there are two other *Geobacillus* strains that have been isolated from a dairy environment: *G. kaustophilus* 102445 and *G. thermoglucosidasius* TNO-09. Using the same approaches as described above, it was determined that there were no genes that were unique to the five strains isolated from a dairy environment, compared with those found in the other *Geobacillus* strains.

The genomes of the four *G. stearothermophilus* strains (A1, P3, D1, ATCC 7953) have smaller genomes compared with the other *Geobacillus* genomes. To determine which genes have been lost the same approach as described above was used. Sixty eight clusters did not contain these four *G. stearothermophilus* strains. Table 4.20 describes those genes that have an assigned biological function, using *G. kaustophilus* HTA as a reference. The alternative COG approach identified fifteen COGs that did not contain predicted proteins from any of the four *G. stearothermophilus* strains (Table 4.21).

**Table 4.20. Selected genes absent in *G. stearothermophilus* A1, P3, D1 or ATCC 7953 but present in the other *Geobacillus* spp**

| Prokka annotation | | | | Pfam | | | |
|---|---|---|---|---|---|---|---|
| **Locus tag** | **Gene name[a,b]** | **Gene description** | ***B. subtilis* homologue[a]** | **Pfam number[c]** | **Pfam description** | **K0 number[a]** | **Predicted function** |
| NC_006510_00241 | *aprX* | Serine protease AprX | *aprX* | PF00082 | Subtilisin family | K17734 | Protein degradation during stationary phase |
| NC_006510_00415 | *liar 1* | Transcriptional regulatory protein LiaR | N/A | PF00196 | LuxR-type DNA-binding HTH[a] domain | N/A | DNA binding transcriptional regulator |
| | | | | PF10114 | PocR ligand binding domain for sensing small hydrocarbon derivatives | | |
| NC_006510_00416 | *rsbV 1* | Anti-sigma-B factor antagonist | N/A | PF01740 | STAS (Sulphate transporter and antisigma antagonist) domain | N/A | Antaganist of an anti-sigma factor |
| NC_006510_00446 | *lutA 1* | Lactate utilization protein A | lutA | PF02754 | Cysteine-rich domain | K18928 | Lactate utilization |
| NC_006510_00447 | *lutR 1* | HTH-type transcriptional regulator LutR | lutR | PF077729 | FadR C-terminal domain | K05799 | |
| | | | | PF00392 | GntR-like transcription factor | | |
| NC_006510_00448 | *lutP* | L-lactate permease | lutP | PF02652 | L-lactate permease | K03303 | |
| NC_006510_00495 | *acdA 1* | Acyl-CoA dehydrogenase | N/A | | Acyl-CoA dehydrogenase domains: | | Metabolism of aromatic hydrocarbons or fatty acid degradation |
| | | | | PF00441 | C-terminal | | |
| | | | | PF02770 | Middle | | |
| | | | | PF02771 | N-terminal | | |
| NC_006510_00496 | *bbsF* | Succinyl-CoA:(R)-benzylsuccinate CoA-transferasesubunit BbsF | N/A | PF02525 | Co-A-transferase family III | K18703 | |
| NC_006510_00637 | *dcuS* | Sensor histidine kinase DcuS | dctS | | | K11691 | C4-dicarboxylate sensing |
| NC_006510_00638 | *dctP* | C4-dicarboxylate-binding periplasmic protein | dctB | | | K11688 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | precursor | | | | | |
| NC_006510_00752 | N/A | ABC transporter substrate binding protein | N/A | PF04392 | ABC transporter substrate binding protein | K01989 | Branched chain amino acid transport |
| NC_006510_00753 | N/A | Branched-chain amino acid transport system / permease component | N/A | PF02653 | Branched-chain amino acid transport system / permease component | | |
| NC_006510_01113 | *kduD* | 2-dehydro-3-deoxy-D-gluconate 5-dehydrogenase | N/A | PF00106 | Short chain dehydrogenase | N/A | Utilzation of glycosaminoglycans |
| NC_006510_01394 | N/A | Long-chain-fatty-acid--CoA ligase FadD13 | N/A | PF00501<br>PF13193 | AMP-binding enzyme<br>AMP-binding enzyme C-terminal domain | N/A | Metabolism of hydrocarbons or fatty acid degradation |
| NC_006510_01395 | *mmgC 2* | Acyl-CoA dehydrogenase | N/A | PF02771<br>PF02770<br>PF00441<br>PF08028 | Acyl-CoA dehydrogenase domains | K00257 | |
| NC_006510_01396 | *bbsE* | Succinyl-CoA:(R)-benzylsuccinate CoA-transferasesubunit BbsE | N/A | PF02515 | CoA-transferase family III | K01796 | |
| NC_006510_01397 | *N/A* | Long-chain-fatty-acid--CoA ligase | N/A | PF00501<br>PF13193 | AMP-binding enzyme<br>AMP-binding enzyme C-terminal domain | K00666 | |
| NC_006510_01399 | *fadA 1* | Putative acyltransferase Rv0859 | N/A | PF00108<br>PF02803 | Thiolase<br>Thiolase, C-terminal domain | K00626 | |
| NC_006510_01506 | *tpa* | Taurine--pyruvate aminotransferase | *yhxA* | PF00202 | Amino transferase class III | N/A | Unknown |
| NC_006510_01608 | Hgd | 2-(hydroxymethyl)glutarate dehydrogenase | N/A | <br>PF03446<br>PF14833 | NAD binding domains of:<br>6-phosphogluconate dehydrogenase<br>NADP-dependent 3-hydroxyisobutyrate | K18121 | Unknown |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | dehydrogenase | | |
| NC_006510_01611 | N/A | Putative FAD-linked oxidoreductase Rv2280 | N/A | PF01565<br>PF02913 | FAD binding domains<br>FAD linked oxidases, C-terminal domain | K00104 | Glyoxylate pathway, conversion of glyoxylate to malate and malate utilization |
| NC_006510_01612 | *glcB* | Malate synthase G | N/A | | | K01638 | |
| NC_006510_01613 | N/A | NAD-dependent malic enzyme | *ytsJ* | | | K00027 | |
| NC_006510_01733 | *fabG 5* | 3-oxoacyl-[acyl-carrier-protein] reductase FabG | N/A | PF00106 | Short chain dehydrogenase | | Unknown |
| NC_006510_01734 | | bifunctional enoyl-CoA hydratase/phosphate acetyltransferase | N/A | PF13452 | N-terminal half of MaoC dehydratase | | |
| NC_006510_01735 | | bifunctional enoyl-CoA hydratase/phosphate acetyltransferase | N/A | PF01575 | MaoC like domain | | |
| NC_006510_02003 | *fadB 1* | putative enoyl-CoA hydratase | *yhaR* | PF00378 | Enoyl-CoA hydratase/isomerase | | Unknown |
| NC_006510_02029 | *gntT 1* | High-affinity gluconate transporter | *gntP* | PF02447 | GntP family permease | K06157 | Gluconate uptake and utilization |
| NC_006510_02031 | *xylB 2* | Xylulose kinase | *gntK* | PF00370 | FGGY family of carbohydrate kinases, N-terminal domain | K00851 | |
| NC_006510_02479 | *rocR 2* | Arginine utilization regulatory protein RocR | *bkdR* | | | | Regulation of branched-chain amino acid utilization |
| NC_006510_02784 | *acrA* | Acryloyl-CoA reductase electron transfer subunitbeta | etfA | PF01012<br>PF00766 | Electron transfer flavoprotein domain<br>Electron transfer flavoprotein FAD-binding domain | K03522 | Fatty acid degradation, |
| NC_006510_02798 | *marR* | Multiple antibiotic resistance protein MarR | N/A | | | | Regulation of genes involved in antibiotic resistance |
| NC_006510_02900 | *acsA 4* | Acetyl-coenzyme A synthetase | *acsA* | PF00501 | AMP-binding enzyme<br>AMP-binding enzyme C- | K01895 | Utilization of acetate |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | PF13193 | terminal domain | | |
| NC_006510_02901 | *acuA* | Acetoin utilization protein AcuA | *acuA* | No domains predicted | | K04766 | Control of AcsA activity |
| NC_006510_02902 | *guaB 2* | Inosine-5'-monophosphate dehydrogenase | *acuB* | PF00571 | CBS domain | K04767 | Unknown |
| | | | | PF01842 | ACT domain | | |
| NC_006510_03118 | *fadA 3* | 3-ketoacyl-CoA thiolase | fadA | PF00108 | Thiolase, N-terminal domain | K00632 | Fatty acid degradation |
| | | | | PF02803 | Thiolase, C-terminal domain | | |
| NC_006510_03316 | *lacF 1* | Lactose transport system permease protein LacF | N/A | PF00528 | Binding-protein-dependent transport system inner membrane component | K05814 | Components of an ABC type sugar transport system |
| NC_006510_03320 | *alsB* | D-allose-binding periplasmic protein precursor | N/A | PF13407 | Periplasmic binding protein domain | | |
| NC_006510_03325 | N/A | putative FAD-linked oxidoreductase Rv2280 | N/A | PF01565 | FAD binding domain | K00102 | Unknown |
| | | | | PF02913 | FAD-oxidase | | |
| NC_006510_03462 | *nuoK* | NADH-quinone oxidoreductase subunit K | N/A | PF00361 | Proton-conducting membrane transporter | K05576 | Respiratory complex I Proton-pumping NADH |
| NC_006510_03463 | *nuoJ* | NADH-quinone oxidoreductase subunit J | N/A | PF00499 | NADH-ubiquinone/plastoquinone oxidoreductase chain 6 | K05578 | |
| NC_006510_03465 | *nuoH* | NADH-quinone oxidoreductase subunit H | N/A | PF00146 | NADH dehydrogenase | K00337 | |
| NC_006510_03466 | *ndhH* | NAD(P)H-quinone oxidoreductase subunit H | N/A | PF00346 | Respiratory-chain NADH dehydrogenase, 49 Kd subunit | K00333 | |
| NC_006510_03467 | *nuoC* | NADH-quinone oxidoreductase subunit C/D | N/A | PF00329 | Respiratory-chain NADH dehydrogenase, 30 Kd subunit | K00332 | |
| NC_006510_03509 | *mmgB 3* | putative 3-hydroxybutyryl-CoA dehydrogenase | *mmgB* | | 3-hydroxyacyl-CoA dehydrogenase: | K00074 | Fatty acid degradation |
| | | | | PF02737 | NAD binding domain | | |
| | | | | | C-terminal domain | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | PF00725 | | |
| NC_006510_03510 | *mmgA* | Acetyl-CoA acetyltransferase | *mmgA* | | Thiolase: | K00626 |
| | | | | PF00108 | N-terminal domain | |
| | | | | PF02803 | C-terminal domain | |
| NC_006510_03511 | *lutA 4* | Lactate utilization protein A | *fadF* | PF13183 | 4Fe-4S dicluster domain | N/A |
| | | | | PF02754 | Cysteine rich domain | |

[a] A number after a gene means there was more than one ORF assigned this gene name

[b] N/A represents “not applicable” as there was no assigned gene name, homologue or K0 number

[c] Multiple Pfam numbers indicate there were multiple domains in the predicted protein sequence

**Table 4.21. COGs absent in *G. stearothermophilus* and present in the other *Geobacillus* spp.**

| COG | Functional Category | COG description[a] |
|---|---|---|
| **COG0161** | H | Adenosylmethionine-8-amino-7-oxononanoate aminotransferase |
| **COG0247** | C | Fe-S oxidoreductase |
| **COG1007** | C | NADH:ubiquinone oxidoreductase subunit 2 (chain N) |
| **COG1055** | P | Na+/H+ antiporter NhaD or related arsenite permease |
| **COG1143** | C | Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I) |
| **COG1250** | I | 3-hydroxyacyl-CoA dehydrogenase |
| **COG1620** | C | L-lactate permease |
| **COG1638** | G | TRAP-type C4-dicarboxylate transport system, periplasmic component |
| **COG2135** | O | Putative SOS response-associated peptidase YedK |
| **COG2225** | C | Malate synthase |
| **COG2427** | S | Uncharacterized conserved protein YjgD, DUF1641 family |
| **COG2610** | GR | H+/gluconate symporter or related permease |
| **COG2972** | T | Sensor histidine kinase YesM |
| **COG2984** | R | ABC-type uncharacterized transport system, periplasmic component |
| **COG4120** | R | ABC-type uncharacterized transport system, permease component |

[a] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

### 4.3.7 Lactose metabolism

To gain a genomic insight into whether the *lac* cluster of genes are the most likely candidates for giving the dairy strains the ability to metabolise lactose the KEGG database was used. All of the *Geobacillus* genomes were analyzed using the KEGG annotation server (KAAS) and the KEGG orthology (KO) groups assigned for each gene were used to determine if any of the *Geobacillus* strains could be capable of metabolising lactose. Figure 4.10 illustrates the absence or presence of genes that may be involved in the metabolism of lactose for each of the *Geobacillus* strains. Of all the *Geobacillus* strains the newly sequenced dairy strains contained the greatest number of genes possibly involved in lactose metabolism and as previously mentioned they were the only three strains that contained *lacA*, *lacB* and *lacC* (Figure 4.11). However, they were missing genes assigned to the orthology group K01220 containing a galactosidase (i.e. *lacG*) required for splitting lactose into galactose and glucose (EC 3.2.1.85). Instead of a galactosidase they had a gene encoding for a glucosidase, annotated as *bglC*. However, the two enzymes LacG and BglC are closely related and in *Lactococcus lactis* it has been shown that a glucosidase enzyme can act as a galactosidase under certain conditions (Hall, 2001, Aleksandrzak-Piekarczyk *et al.*, 2005).

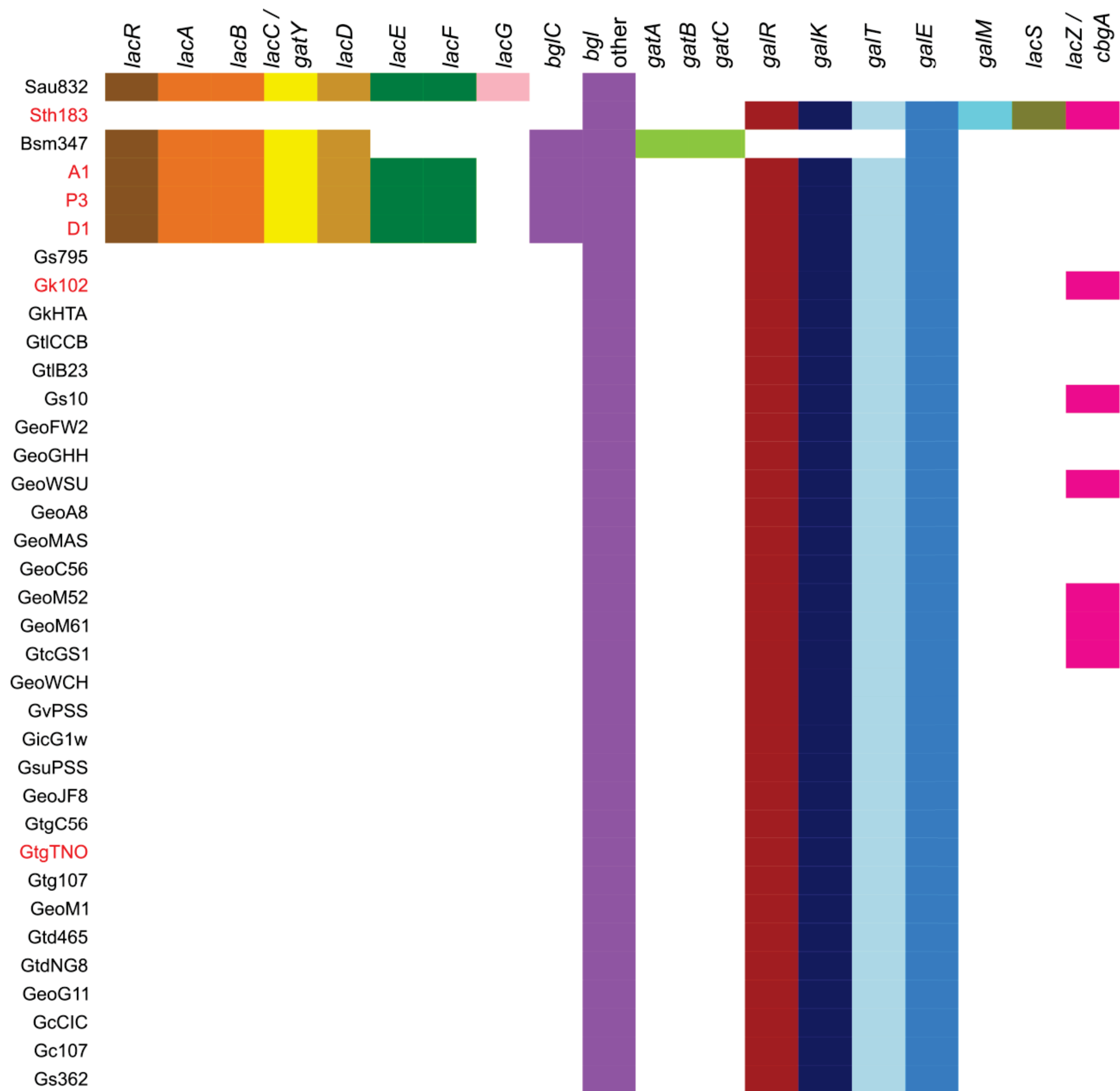



**Figure 4.10. The conservation of selected genes involved in the utilisation of lactose and galactose.** *S. aureus* and *S. thermophilus* were used as references. Those *Geobacillus* strains written in red were originally isolated from a dairy environment.

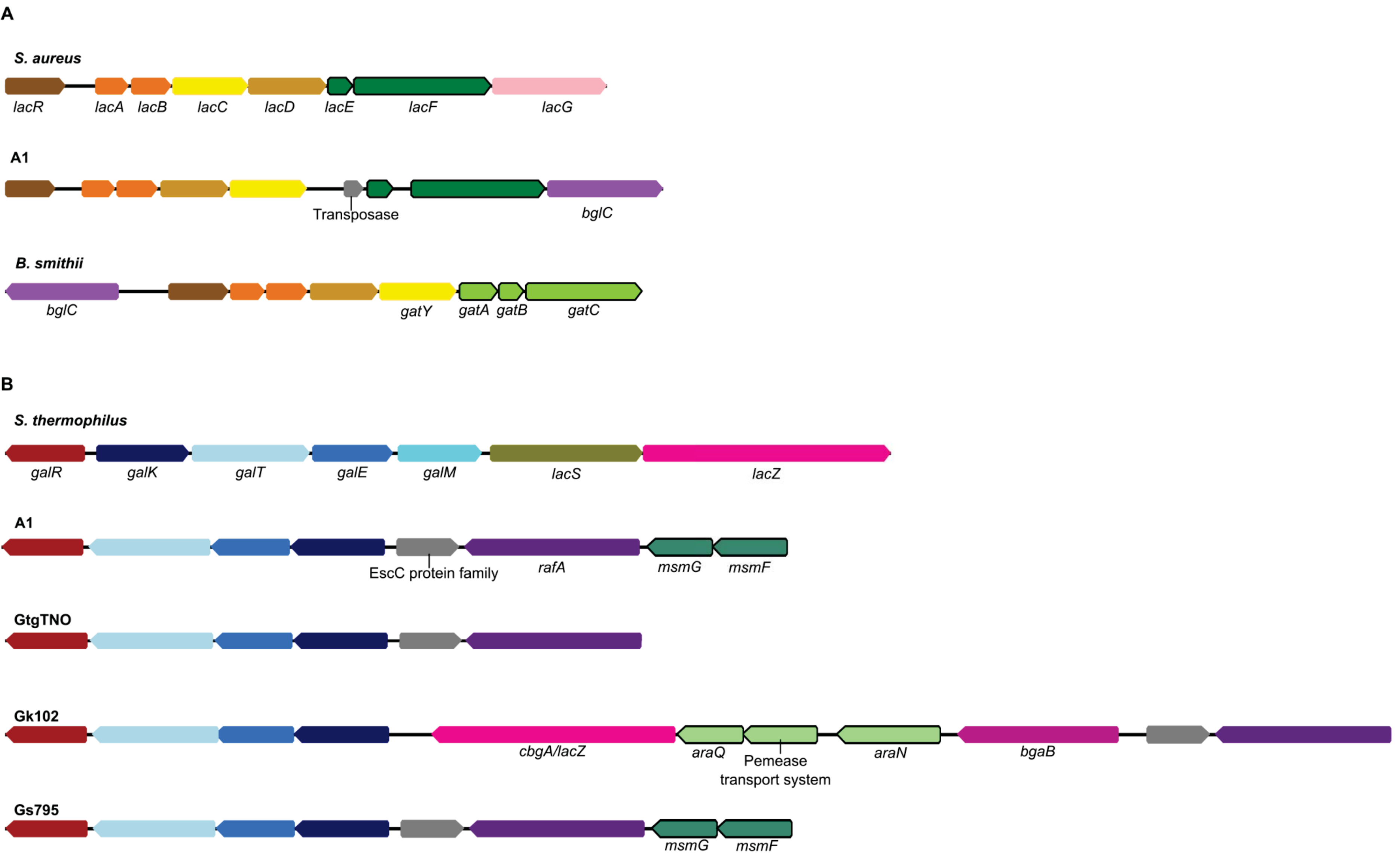

A
S. aureus
lacR
lacA
lacB
lacC
lacD
lacE
lacF
lacG
A1
Transposase
bglC
B. smithii
bglC
gatY
gatA
gatB
gatC
B
S. thermophilus
galR
galK
galT
galE
galM
lacS
lacZ
A1
EscC protein family
rafA
msmG
msmF
GtgTNO
Gk102
cbgA/lacZ
araQ
Pemease
transport system
araN
bgaB
Gs795
msmG
msmF

**Figure 4.11. Comparison of the organisation of the *lac* genes.** Annotations in both diagrams are based on the assigned KEGG KO for each gene. Colours represent those genes belonging that to the same KO group and/or KEGG enzyme entry.  **A.** The *lac* operon in *S. aureus* and the putative *lac* operons in strain A1 as well as *B. smithii* (which showed the highest similarity to the putative A1 *lacA*, *lacB* and *lacC* genes as described in Table 4.17). The gene organisation of the putative *lac* operon in strains P3 and D1 was syntenic with that of A1.  The *gatABC* operon encodes a galactitol transport system and *gatY* a component of the of the GatYZ tagatose aldolase as described by (Van der Heiden *et al*., 2013). GatY and LacD both belong to the same enzyme group (EC 4.1.2.40). **B.** The *gal-lac* operon in *S. aureus* and the putative *gal* operon plus downstream genes in selected *Geobacillus* strains of dairy origin as well as *G. stearothermophilus* ATTCC 7953, the closest relative of strain A1, which was originally isolated from creamed corn. Those genes outlined in black encode for proteins of various transport systems.

To evaluate whether the *lac* cluster might have been acquired by horizontal gene transfer, a BLASTX search was carried out on the various components of the A1 *lac* cluster and the GC content compared. As determined in Section 4.3.7 the closest homologues of *lacA-B-C* were found in *B. smithii*. However, the closest homologues of *lacE-F* and *bglC* were found to be in *B. cereus* (Table 4.22). Homologues of *lacE-F* and *bglC* were also found in other species both Gram positive and negative. The average GC content of the *lacA-B-C* and *gatY* cluster in strain A1 is 38.9 % and *B. smithii* is 38.5 % compared with the overall genome GC content of 52.0 % and 40.7 % respectively. The GC content of the A1 contig, from which the *lac* cluster originates, is also very low with a value of 39.5 %.

**Table 4.22. BLASTX of *G. stearothermophilus* A1 *lac* cluster components**

| Gene | Species homologue found in | Coverage | Identity | Evalue |
|---|---|---|---|---|
| ***lacA-B-C*** | *B. smithii* | 95 – 96 % | 99 % | Refer to Table 4.17 |
| ***lacA-B*** | *Clostridium* spp. | 96 - 98 % | 70 – 75 % | 1e-61 – 1e-75 |
| ***gatY*** | *B. smithii* | 99 % | 96 % | 0 |
| ***lacF*** | *B. cereus* | 95 % | 70 % | 2e-45 |
| | *Enterococcus* spp. | 95 % | 59 – 63 % | 1e-39 – 1e-41 |
| ***lacE*** | *B. cereus* | 99 % | 75 % | 0 |
| | *Clostridium* spp. | 99 % | 67 – 69 % | 0 |
| | *Streptococcus uberis* | 99 % | 65 % | 0 |
| ***bglC*** | *B. cerueus* | 99 % | 79 % | 0 |
| | *Listeria* spp. | 99 % | 71 % | 0 |

### 4.3.8 Differences between the three dairy strains

In Chapter 2 it was demonstrated that the three dairy strains A1, P3 and D1 were phenotypically different. To determine whether their genomic content could account for any of these differences the COG approach was used as in Section 4.3.8; where 199 COGs were identified with variation across the three strains[11] and of this variation, 83 % was between D1, and the other two strains A1 and P3. Twenty-six COGs contained variation between A1 and P3.

[11] CD: Chapter4/Results/COG_GstDairyStrains_Variable.xlsx

### 4.3.9 CRISPR arrays

The CRISPR (Clustered regularly-interspaced short palindromic repeats) - Cas system is an important defense system against foreign DNA, such as phage and plasmids. A CRISPR array consists of short repeat sequences interspersed by unique sequences of DNA called spacers. These arise from incorporation of short fragments of foreign DNA into the CRISPR array. The *cas* genes, which generally neighbour a CRISPR array, encode for proteins responsible for incorporating new spacers, recognizing foreign DNA and stimulating a process equivalent to an immune response interaction of a strain with phage. As of August 2014, approximately 45 % of bacterial genomes contained CRISPR-Cas systems, according to the CRISPRfinder database (Grissa *et al.*, 2007). CRISPR-Cas systems have been studied for a variety of reasons including for typing and providing an historical record of phage exposure (Horvath & Barrangou, 2010, Amitai & Sorek, 2016). In the *Geobacillus* genus CRISPR arrays were found in all of the complete genomes that were available in the CRISPRFinder database (Table 4.23).

**Table 4.23. CRISPR arrays found in the *Geobacillus* genus.**

| Strain | Number of CRISPR |
|---|---|
| ***G. kaustophilus* HTA426** | 4 |
| ***Geobacillus* sp. C56-T3** | 4 |
| ***Geobacillus* sp. GHH01** | 5 |
| ***Geobacillus* sp. JF8** | 2 |
| ***Geobacillus* sp. WCH70** | 6 |
| ***Geobacillus* sp. Y4.1MC1** | 5 |
| ***Geobacillus* sp. Y412MC52** | 6 |
| ***Geobacillus* sp. Y412MC61** | 6 |
| ***G. thermodenitrificans* NG80-2** | 3 |
| ***G. thermoglucosidans* C56-YS93** | 6 |
| ***G. thermoleovorans* CCB_US3_UF5** | 2 |

The spacer sequences can be very different between two near identical strains (Tyson & Banfield, 2008). In this present study, an analysis of the CRISPR arrays was carried out to look for differences between the three strains A1, P3 and D1. Four different CRISPR types were found based on the sequence of the repeat region (Table 4.24 and Figure 4.12). Two of these types were found in other *Geobacillus* spp. In total, seven CRISPR arrays were found in the genomes of A1 and P3, whereas only three were found in the genome of D1. CRISPR associated genes were associated with three of the arrays in A1 and P3, but only one array in D1 (Figure 4.12A). In addition, for each of the strains, CRISPR associated genes were also found on one contig that did not contain a CRISPR array (Figure 4.12B).The strains A1 and P3 could be differentiated based on CRISPR 3A. The CRISPR 3A array of A1 contained a larger number of spacers, which were not present in any of the other CRISPR arrays (Figure 4.13).

The origin of most of the spacers remains unknown, with 80% of the spacers having no match with a known sequence in the phage and plasmid databases[12]. Five of the spacers had an identitical match to the strain's own genome.

CRISPRDetect was used to predict the direction of each CRISPR and its putative subtype (Biswas *et al.*, 2014). Using CRISPRDetect the subtype prediction was initially based on whether the repeat sequences have previously been classified. For those CRISPR arrays that were associated with *cas* genes a subtype prediction was also carried out based on the gene content according to the classifications by Makarova *et al.*, (2015). One of the CRISPRs is also likely to belong to the subtype III-B based on the *cas* genes that are located in a contig by themselves.

[12] CD: Chapter4/Results/Spacers.xlsx

A
Type 1
D-aminopeptidase
Type 2
A1, P3
2A
A1, P3
2B
cas8b1
cas7
cas5
cas3
cas4
cas1
cas2
cas6
2C
D1
2D
flagella assembly protein
Type 3
A1
cas6
cas8b2
cas7
cas5
cas3
cas4
cas1
cas2
P3
Type 4
A1, P3
cas3
cas5
cas8c
cas7
cas4
cas1
cas2
4A
D1
4B
B
A1, P3
csm6
csx1
cas7/cmr1
cas10
cas5/cmr3
cas7/cmr4
cmr5
cas7/cmr6
D1
cas3HD
cas5
cas8c

**Figure 4.12. Organisation of the CRISPR arrays and associated genes in the dairy strains A1, P3 and D1.** Homologous genes are represented by the same colour. White filled genes are not known CRISPR associated genes. Dark grey genes encode hypothetical proteins and light grey transposases. Red brackets indicate the end of a contig. CRISPR associated genes were annotated according to the system described by Makarova *et al.* (2015). (**A**) The four types of CRISPR arrays represented by red diamonds (repeats) and black squares (spacers). The number of diamonds and squares are not representative of the actual number of repeats and spacers. Refer to Table 4.24 for a description of each CRISPR array. (**B**) The CRISPR associated genes located on a contig without a CRISPR array.

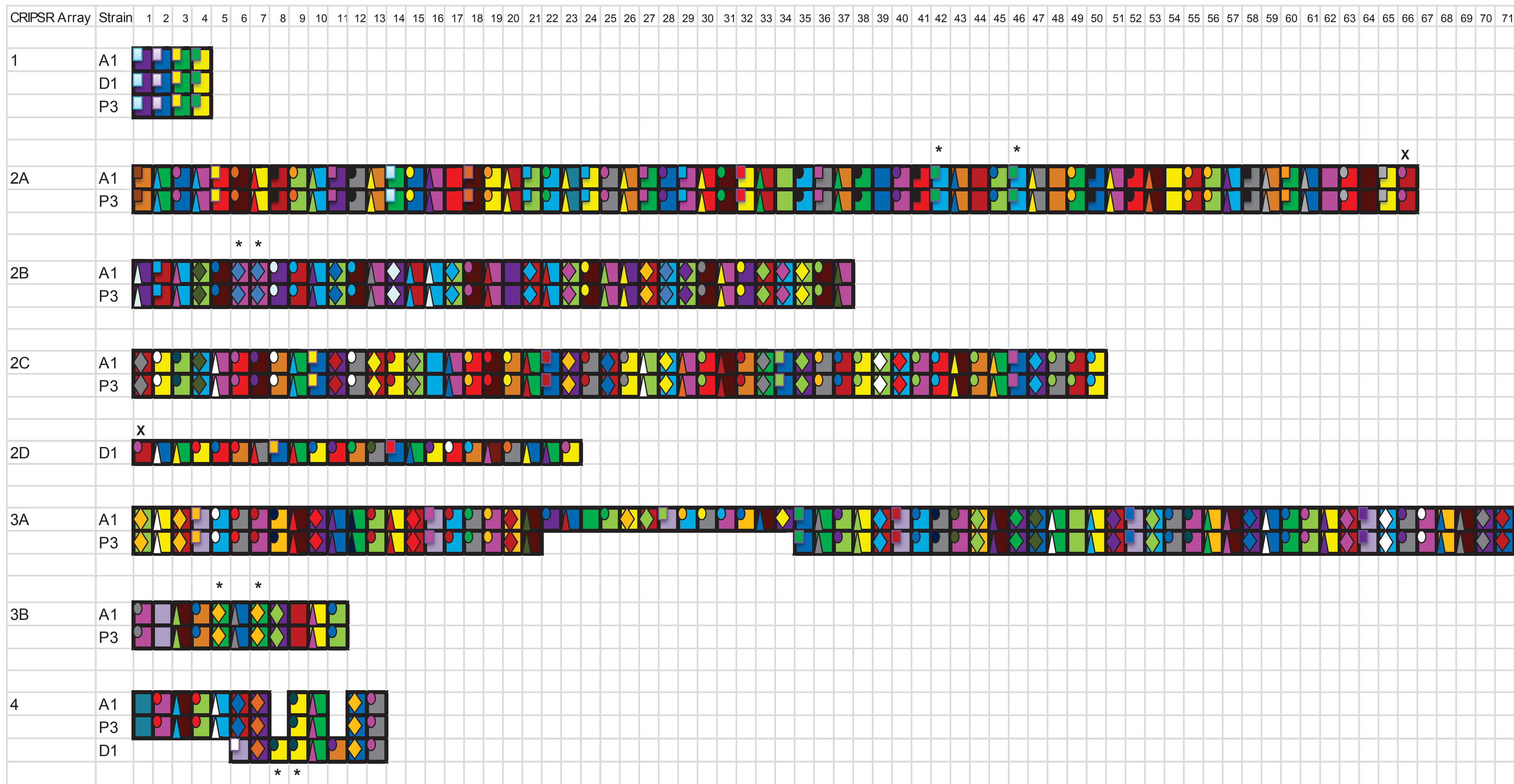



**Figure 4.13. Organisation of the spacers in each CRISPR array.** For simplicity the repeat regions have been removed. Each spacer with a unique sequence is characterized by a combination of a different colour and symbol. The * illustrates where an allele is repeated within a CRISPR array and the X designates where a spacer is the same as that found in another array type. The spacer sequences between the three strains have been aligned and missing spacers are represented by empty squares.

**Table 4.24. Characteristics of the CRISPR arrays**

| | Repeat consensus | Repeat length (bp) | Putative Array family | Number of spacers | Position on draft genome | | | Presence in other *Geobacillus* genome |
|---|---|---|---|---|---|---|---|---|
| | | | | | A1 | P3 | D1 | |
| **1** | GTTTTTATCGTACCTATGAGGGATTGAAACC | 31 | I-B[a] | 4 | 403806-404102 | 944410-944706 | 649597-649893 | Geo JF8 |
| **2** | GTTTTTATCGTACCTATGAGGGATTGAAAC | 30 | | | | | | GkHTA, GeoM61, GeoM52, GtICCB, GeoGHH |
| **A** | | | I-B[a] | 65 (A1 & P3) | 3061237-3056882 | 87729-83374 | | |
| **B** | | | I-B[a,b] | 37 (A1 & P3) | 2743477-2745973 | 412587-415083 | | |
| **C** | | | I-B[a,b] | 50 (A1 & P3) | 2754684-2758049 | 423794-427159 | | |
| **D** | | | I-B[a] | 23 (D1) | | | 864719-866279 | |
| **3** | GTTTTATCTGAACGGAGTGGGATTTAAAG | 29 | | | | | | |
| **A** | | | I-B[b] | 71 (A1) / 58 (P3) | 77244-72598 | 1805853-1802056 | | |
| **B** | | | Unknown | 11 (A1 & P3) | 1071454-1070710 | 398202-398946 | | |
| **4** | GTTGCACCCGGCTATTAAGCCGGGTGAGGA TTGAAACA | 38 | I-C[b] | 11 (A1 & P3) / 8 (D1) | 107709-106879 | 1139516-1138686 | 528815-529356 | |

[a] As determined by CRISPRDetect

[b] As determined by gene content using the classifications by Makarova *et al.* (2015).

## 4.4 Discussion

Thermophilic bacilli, particularly *Geobacillus* species and *A. flavithermus*, are important contaminants of milk powder. Initial entry into a dairy manufacturing process is believed to occur in low numbers via the raw milk, in the endospore form. Endospores of these bacteria may then attach to the surfaces of equipment such as plate heat exchangers and evaporators where they probably germinate and form a biofilm. The resulting biofilm sheds both vegetative cells and endospores, resulting in high bacterial numbers in the final product (Murphy *et al.*, 1999, Scott *et al.*, 2007). In a laboratory setting, several studies have shown how both vegetative cells and spores can attach to a surface and subsequently form a biofilm (Flint *et al.*, 2001b, Burgess *et al.*, 2009, Zhao *et al.*, 2013, Burgess *et al.*, 2014a, Somerton *et al.*, 2015). However, the question remains, what is it about these strains that enable them to adapt to a dairy manufacturing environment compared with other strains of the same species? To provide a genomic insight into this question the genomes of three strains (A1, P3, D1) of *G. stearothermophilus,* which were isolated from the surface of an evaporator, were sequenced using next generation sequencing. In June 2015, over 36 genomes of *Geobacillus* had been sequenced. This is the first study that has sequenced the genomes of *G. stearothermophilus* originating from a milk processing environment.

### 4.4.1 Fragmentation of genomes

This study used the Illumina MiSeq for next generation sequencing of the three dairy strains of *G. stearothermophilus*. The Illumina MiSeq belongs to a class of sequencing technologies sometimes referred to as second generation sequencing. This class of sequencing has resulted in high throughput sequencing for relatively low cost. However, the short read length (generally ranging between 150 – 250 bp) makes genome assembly difficult compared with traditional Sanger sequencing (also referred to as first generation sequencing). This can be seen in this present study where the genomes were very fragmented ranging between 398 – 449 contigs. This is the highest number of contigs out of all of the *Geobacillus* genomes analyzed in this study, with the other genomes ranging from a complete genome to 346 contigs for *Geobacillus* sp. WSUCF1.

Genome fragmentation is caused by repetitive DNA sequences (Koren & Phillippy, 2015). The most obvious repeat region found in all bacteria is the rRNA gene operon, which also contains repetitive regions between each rRNA gene. In this study fragmentation caused by the multiple rRNA gene operons was evident with no full-length rRNA gene operon being present in the genome. However, fragmentation due to the rRNA gene operon would only account for approximately 30 fragments, depending on the number of rRNA gene operons. Other repeat regions range from small DNA repeats < 400 bp (e.g. miniature inverted-repeat transposable elements and species specific repeat sequences) to larger mobile elements (e.g. transposons, insertion sequences and genomic islands) and less common are repeats larger in size than the rRNA gene operon (5 – 7 kb), such as phage associated inserts (Delihas, 2011, Ricker *et al.*, 2012, Koren *et al.*, 2013).

Many DNA repeat regions are mobile elements and are often associated with transposases. The dairy strains had a higher mobilome content (Section 4.3.5), compared with many of the other *Geobacillus* genomes analysed, which may have contributed to these genomes containing a higher number of repeats.

There were many transposases identified in the three dairy strains and it is likely that they would be associated with repeat regions that caused the fragmentation. *Geobacillus kaustophilus* HTA426 was one of the *Geobacillus* strains most closely related to the three dairy strains with a complete genome and also contained many repeat regions. Analysis of this genome may provide insight into what type of repeat regions could also be present in the three dairy strains.

To determine the nature of the repeat regions in the newly sequenced dairy strains, a third generation sequencing technique, such as PacBio RS sequencing (Eid *et al.*, 2009), would need to be carried out on one of the dairy strains. PacBio RS sequencing was the first commercially available single-molecule sequencing technology. This third generation sequencing technology generates long read lengths > 14 kb, which not surprisingly improves the assembly process compared with second generation technologies such as the MiSeq, often resulting in assembling a complete genome (Koren *et al.*, 2013, Miyamoto *et al.*, 2014). The exception in the study by Koren *et al.* (2013) was an *E. coli* genome that contained repeats > 7kb in length. However, since this paper was published the read-length of PacBio sequencing has improved.

According to Ricker *et al.* (2012) leaving genomes in their draft form may mean that important data is missed, particularly in relation to how a strain has evolved and adapted to a specific environment. DNA repeats have a range of functions such as gene regulation and acquisition of genes required for growth in a specific niche as well as as pathogenicity and resistance (Delihas, 2011). In the study by Ricker *et al.* (2012) draft assemblies were compared with full genomes. They found that repetitive regions, often from mobile elements, caused fragmentation of the draft genomes. In the species *Cupriavidus metallidurans,* some of these repetitive regions originated from genomic islands. Some of these genomic islands contained genes involved in the ability of *C. metallidurans* to persist in environments with heavy metals. In the draft genome of this species, most of these genomic islands would have been fragmented; therefore, information on how this strain had adapted to an environment containing heavy metals would have been missed. In two *S. thermophilus* strains used as starters in dairy products, it was found that insertion sequence (IS) elements were associated with genes important in milk adaptation (Bolotin *et al.*, 2004). In the case of the draft genomes of the three dairy strains, it is possible that there may be information missing on how these strains have adapted to a dairy environment.

### 4.4.2 Diversity of the rRNA operons

None of the genome assemblies generated in this study contained a full-length copy of the rRNA gene operon. As already discussed, this is not suprising given the short read lengths (250 bp) and the repetitive nature of the rRNA gene operon. What was surprising was the potential

copy number of rRNA gene operons and their diverse nature. It is unknown how many rRNA operons are present in the newly sequenced *G. stearothermophilus* strains, but based on the number of 5S rRNA genes and the variant frequency ratios of the 16S and 23S rRNA genes, it is likely to be 9 or 10. This approximate number is similar to the copy number in those *Geobacillus* strains with complete genomes, which range from 8 – 10 (Sun *et al.*, 2013). Prokaryotic rRNA gene operon copy numbers range from 1 – 15 with an average number of 3.61 (Sun *et al.*, 2013).

Unsurprisingly the shortest rRNA gene, the 5S rDNA, has the most diversity (Pei *et al.*, 2012). In the same study it was found that those strains with significant diversity (> 3 %) could be caused by five different mechanisms. Two of these mechanisms: "split ribosomal operon", and "5S-23S spacer length", have been found in *Geobacillus* species. A third mechanism "orphan 5S rRNA gene" has also been found in *Bacillus* species. One of these three mechanisms could be a possible cause of the high diversity in the newly sequenced *G. stearothermohilus* strains. Full-length sequences of each of the rRNA operons would be needed to determine this.

Inevitably, the diversity of the 16S rRNA gene is the most well-studied of the three rRNA genes because of its taxonomic importance (Coenye & Vandamme, 2003, Acinas *et al.*, 2004, Pei *et al.*, 2010, Sun *et al.*, 2013). The largest of these studies found that 12.5 % of the genomes analyzed had > 1 % 16S rRNA gene diversity. Traditionally strains with > 3 % difference in full-length 16S rRNA gene sequences were regarded as different species (Tindall *et al.*, 2010), but more recently it was proposed that this be decreased to 1.3 % (Kim *et al.*, 2014). Given the high similarity between some species like *G. thermoleovorans* and *G. stearothermophilus*, significant 16S rRNA gene diversity could affect the taxonomic positioning of some strains. In the case of the *Geobacillus* genus, to date, only *G. thermodenitrificans* (Sun *et al.*, 2013), and the three newly sequenced *G. stearothermophilus* strains from this present study, have been found to have a 16S rRNA gene diversity of > 1 %.

The 23S rRNA gene is the least diverse of the three rRNA genes and, in the case of the three three dairy strains, was at the lower end of the diversity scale compared with other bacteria that have been analyzed (Pei *et al.*, 2009). In the same study, the diversity of the 23S rRNA gene was only analyzed in one other strain of *Geobacillus*: *G. kaustophilus* HTA426. This strain has a higher diversity compared with the three dairy strains (0.41 % compared with 0.21 % for A1 and P3 and 0.27 % for D1), but it was still lower than the mean diversity of those bacterial species analyzed in the study by Pei *et al.* (2009).

### 4.4.3 Differentiation of the *G. stearothermophilus* taxon

The genomes of the four *G. stearothermophilus* strains analyzed in this present study were all approximately 2.8 - 3 Mb compared with 3.5 Mb for the other *Geobacillus* genomes. This suggests that gene loss occurred when the *G. stearothermophilus* taxon diverged from the common ancestor of the *Geobacillus* genus. Many of the genes that are absent from the *G. stearothermophilus* genomes can be related to its phenotypic traits. For example, lactate was not able to be utilized by both the dairy strains of *G. stearothermophilus* (refer to Chapter 2) and

the type strain of *G. stearothermophilus* (Claus & Berkeley, 1986, Nazina *et al.*, 2001). In this study it was found that some of the *lut* genes, which are required for lactate utilization (Chai *et al.*, 2009, Chiu *et al.*, 2014), were missing from the four genomes of *G. stearothermophilus*. The presence or absence of genes can also be used to predict a phenotype. For example, *gntP* and *gntK*, which encode for a gluconate permease and gluconokinase respectively and form part of the *gnt* operon are required for gluconate utilization (Fujita *et al.*, 1986). These two genes were not present in the four *G. stearothermophilus* strains, but were present in the other strains of *Geobacillus*. This suggests that the *G. stearothermophilus* strains cannot utilize gluconate. To the author's knowledge this phenotype has not been tested in strains of *G. stearothermophilus*.

A description of *G. stearothermophilus* has not been republished since 1986 by Claus & Berkeley (1986); therefore, Logan *et al.* (2009) advise that this description is likely to have encompassed a variety of thermophilic bacilli strains that would now be regarded as separated taxa. Given the lack of a recent description of *G. stearothermophilus,* it is difficult to make comparisons between phenotype and genotype.

The *G. stearothermophilus* genomes analyzed in this study were all derived from a food product or manufacturing plant, creamed corn in the case of *G. stearothermophilus* ATCC 7953, and a milk powder manufacturing plant in the case of the three dairy strains. The loss of genes involved in carbohydrate metabolism may be because they were no longer needed in such a specific niche. Given few strains of *G. stearothermophilus* have been analyzed in terms of phenotype and genotype it is unknown whether this is a phenomenon specific to these four strains or the *G. stearothermophilus* taxon as a whole.

### 4.4.4 Niche adaptation of the dairy strains of *G. stearothermophilus*

The adaptation of bacteria to a dairy environment has been studied in starter strains such as *S. thermophilus* and *Lactococcus lactis* (Price *et al.*, 2012). In *S. thermophilus* it was found that dairy adaptation was mainly a result of the loss of gene functionality, particularly in relation to carbohydrate metabolism, with a higher proportion of pseudogenes found in the genomes of dairy isolated *S. thermophilus* compared with pathogenic streptococci. In this present study, the number of pseudogenes was obtained by using the NCBI prokaryotic genome annotation pipeline. However, given the genomes were in their draft format, and some of the pseudogenes were located at the end of a contig, it is unknown whether this is a true number of pseudogenes; therefore, no comparison was made with other strains of *Geobacillus*. There did appear to be some complete gene loss compared with other strains of *Geobacillus*; however, none of these appeared to be related to carbohydrate metabolism.

In *L. lactis*, genes that appear to be related to dairy adaptation, such as the lactose operon (*lacRABCDFEGX*), proteololysis (*prtP* and *prtM*) as well as magnesium and manganese transport reside only on plasmids (Siezen *et al.*, 2005) . In addition, these genes are generally only found on plasmids from strains isolated from dairy environments not other environments such as soil (Siezen *et al.*, 2011). The dairy strains of *G. stearothermophilus* also appeared to have acquired a lactose operon. Given that the GC content of this lactose operon was lower

than the average GC content of the genome, it is likely that this operon was acquired by horizontal gene transfer. However, as discussed in Section 4.4.1 the complete genome would be needed to gain more information as to how this acquisition occurred.

Interestingly, the other two strains of dairy origin that were analyzed, *G. kaustophilus* and *G. glucosidasius* TNO-09, did not appear to have acquired a lactose operon. Although, some (< 10 %) strains of *G. kaustophilus* and *G. thermoglucosidasius* have been shown to utilize lactose (White *et al.*, 1994), it is not known whether this is the case for these particular strains. It is possible that lactose could be transported into the cell using a non-specific PTS transport system. However, once in the cell the lactose needs to be split into glucose and galactose using a galactosidase. A galactosidase-encoding gene was found in the *G. kaustophilus* genome but not the *G. thermoglucosidasius* genome. Like the dairy strains of *G. stearothermophilus* a more detailed analysis using the complete genomes would need to be carried out to determine whether there is any evidence of dairy adaptation.

#### 4.4.5 Differences between the dairy strains

Phenotypic differences were seen between the three strains A1, P3 and D1, when attributes such as the fatty acid profile, metabolic activity and ability to form biofilms and spores were compared (Refer to Chapter 2 and 3). Fatty acid and biochemical profiling showed that strain D1 was different to the other two strains P3 and A1. This same trend was also seen at a genome level. When a COG approach was used for identifying differences between the three strains, more differences were seen between D1 and the other two strains A1 and P3, compared with differences between A1 and P3. From a phylogenomics perspective, differences in the three strains will be evaluated in Chapter 5.

The COG approach could not be used to determine a link between phenotype and genotype. This was mainly because the COGs were generally generic and a gene specific approach was not used. A more useful approach could be using PhenoLink (Bayjanov *et al.*, 2012), which is a web based tool for linking phenotype to genotype. This would also be a useful tool for identifying differences between the dairy strains and other strains of *G. stearothermophilus*, to enable further study of the adaptation of *G. stearothermophilus* to a specific environment. To be more useful a wider range of phenotypic tests would need to be carried out on a range of *G. stearothermophilus* strains from different environments. PhenoLink has been used for linking phenotype to genotype and evaluating differences between strains of *Lactobacillus paracasei* as well as *Lactobacillus plantarum* (Bayjanov *et al.*, 2012, Smokvina *et al.*, 2013). In both of these studies over 30 strains were evaluated.

Differences in a phenotype between strains of a particular bacterial species can also be attributed to mutations in a gene or the promoter region. For example, although some *S. thermophilus* strains are not able to utilize galactose (Gal -), they still have a *galKTE* operon and when compared with *S. thermophilus* strains that were Gal + it was found that there were several single base pair mutations in the promoter region (Bogaard *et al.*, 2004). Another example is where *B. subtilis* 168 has lost its ability to form biofilms compared with the near

genetically identical strain NCIB3610, which forms strong biofilms. This difference has been attributed to mutations in four different genes (McLoon *et al.*, 2011b). In these examples, linking the phenotype to genotype was more difficult to establish as opposed to when a phenotype originates from a missing gene. A genome alignment program such as Mauve (Darling *et al.*, 2004), could be used for identifying sequence differences between the three strains, A1, P3 and D1; however, this still provides no information on whether phenotypic differences could be attributed to the sequence differences. As will be discussed in Chapter 6 a transcriptomics study is needed to identify which genes are upregulated in response to different conditions.

#### 4.4.6 CRISPR arrays

Analysis of the CRISPR arrays was also carried out to identify differences between the three dairy strains of *G. stearothermophilus*. CRISPR arrays can be different between closely related strains (Horvath *et al.*, 2008, Shariat *et al.*, 2015). In this present study, the CRISPR arrays in strain D1 were different overall to the other two strains P3 and A1. One CRISPR was different between A1 and P3, with additional spacers found in one of the A1 CRISPRs. Strain A1 originates from the top and P3 from the bottom of the same evaporator; whereas D1 originates from a different evaporator. CRISPRs could provide an an evolutionary history of an organism's exposure to extrachromosomal DNA such as plasmids and phage (Horvath & Barrangou, 2010), given the spacers generally originate from mobile elements. Given, the dairy strains had a higher mobilome content than many of the other Geobacillus strains, one would expect them to contain fewer mechanisms against invading DNA. However, this was not the case; with two of the dairy strains (A1 and P3) containing six CRISPR arrays each. *Geobacillus* sp. WCH (originally isolated from compost) also had six CRISPR arrays as well as the highest mobilome content of all of the *Geobacillus* genomes analyzed.

In this present study, the potential origin of most of the spacers could not be identified, as the spacers showed little or no homology with known sequences. For example, in strain A1 only 19 % of the spacers showed some homology to known sequences, such as with some known *Geobacillus* phages. Some of the spacers also showed some homology to phage specific for other bacteria that may also be present in milk, such as *S. aureus* and *Lactobacillus helveticus*. The method used for the identification of spacers is only as good as the database. For example, CRISPRs as well as phage specific for the cheese starter, *S. thermophilus*, have been well studied. Biswas *et al.* (2013) was able to identify the origin of 75 % of the spacers from a strain of *S. thermophilus*; whereas, in the plant pathogen *Pectobacterium carotovorum*, which has not had its CRISPR system so well characterized, the origin of only 46 % of spacers were identified. Raw milk has a very diverse microbial population, so presumably bacterial strains that originate from a milk environment have been exposed to a lot of different phages and plasmids that have not previously been characterized. In a study of the human pathogen, *Porphyromonas gingivalis*, which is probably also exposed to a diverse microbial environment (in the mouth and gut) and has had few studies carried out on its CRISPR system, < 20 % of its spacers show any sequence similarity to known sequences (Burmistrz *et al.*, 2015).

A small number of spacers matched chromosomal DNA from the host, which suggests self-targeting. Several recent studies have shown that sometimes the spacers originate from targets within the chromosome (Burmistrz *et al*., 2015, Dy *et al*., 2013, Shariat *et al*., 2015, Stern *et al*. 2010). Shariat *et al.* (2015) found that self-targeting spacers were often associated with a variant of the direct repeat. The authors suggest that this mutation probably prevents the CRISPR-Cas system from stimulating an immune response. Stern *et al*. (2010) examined CRISPR arrays from 330 organisms and of the spacers they examined, 0.4 % were self-targeting, with most being deemed inactive. The authors proposed several mechanisms for protection of the cell from autoimmunity, including loss of the *cas* genes, a mutated leader sequence (which is required for incorporation of new spacers), mutated flanking repeat regions, and a lack of or a mutated protospacer adjacent motif. In this present study, the self-targeting spacers were only associated with one type of CRISPR array. However, it was not examined in enough detail to determine whether it was likely to be active or not. Active self-targeting spacers have been shown to be deadly to a cell (Vercoe *et al.*, 2013).

CRISPR arrays have been used for typing different bacteria (Horvath *et al.*, 2008, Shariat *et al.*, 2013, Smokvina *et al.*, 2013, Shariat & Dudley, 2014). However, there is evidence that CRISPRs are not a good measure of phylogeny as they evolve differently to the core genome (Horvath *et al.*, 2009). In this present study, the CRISPR arrays were not used as a phylogenetic tool, but rather as a means for separating strains. CRISPR arrays were found in other *Geobacillus* spp. but to the author's knowledge no analysis has been carried out to determine the type and diversity of CRISPR arrays in *Geobacillus* spp.

CRISPR arrays have been classified into five types and sixteen sub-types based on the components of the *cas* loci (Makarova *et al.*, 2015). The dairy strains did not appear to contain any unusual features that have not been previously described, and for those CRISPRs that were associated with a *cas* locus, they could be classified using the system described by Makarova *et al.* (2015). In the analysis carried out by Makarova *et al.* (2015), some *Geobacillus* genomes were included and these were found to contain Types I-B, II-C, or III-B. In the present study, the three *G. stearothermophilus* dairy strains contained Types I-B, I-C and III-B. The Type III-B system is commonly found in archea as well as cyanobacteria and other extremophilic bacteria such as thermophiles (e.g. *Deinococcus* spp., *Caldicellulosiruptor* spp. and *Thermos thermophilus*), acid tolerant (e.g. *Alicyclobacillus* spp.) and salt tolerant bacteria (e.g. *Bacillus halodurans* and *Myxococcus fulvus*) (Makarova *et al.*, 2015). An unusual feature of the Type III-B system is that they often do not have the *cas1* and *cas2* genes, which encode proteins required for recognizing foreign DNA and integrating new spacers into the CRISPR array. This means if a CRISPR system does not have the *cas1* and *cas2* genes it relies on another Cas system to carry out this process (Makarova *et al.*, 2015). The Type III-B loci of the strains A1 and P3 were also missing the *cas1* and *cas2* genes.

CRISPRs generally contain repeats that fall into the size range of 20 – 50 bp (Sorek *et al.*, 2008). In this present study all of the repeats fell within this range. The number of spacers in

each array varied from four spacers to over 40 in three of the arrays of strains A1 and P3, with A1 containing the largest array consisting of 71 spacers. D1 contained a smaller number of arrays (three arrays as opposed to seven found in A1 and P3) and the largest array contained only 23 spacers. This suggests that strains P3 and particularly A1 have a wider protective capability against foreign DNA compared with strain D1.

## 4.5 Conclusions

Dairy adaptation in the *G. stearothermophilus* genomes of the three newly sequenced dairy strains was evident by the presence of a putative lactose operon. Further evidence of dairy adaptation may have been missed due to the fragmented nature of the draft genome sequences. Some gene loss appears to have occurred in the *G. stearothermophilus* taxon during evolution from the common ancestor of the *Geobacillus* genus. This was indicated by the *G. stearothermophilus* genomes having a smaller genome, compared with other members of the *Geobacillus* genus. In addition, clusters of genes were missing that were involved in various functions, such as carbohydrate metabolism. It seems unlikely that this gene loss is a result of the fragmented genomes; given other *Geobacillus* genomes are also very fragmented.

A study incorporating a wider range of *G. stearothermophilus* strains, including at least one complete genome is needed to fully understand how these newly sequenced strains have adapted to a dairy environment and the potential gene loss within the *G. stearothermophilus* taxon.To generate a complete genome, third generation sequencing, such as PacBio RS sequencing would need to be carried out.

Chapter 5

# 5 Comparative analysis of *Geobacillus* taxonomy using phylogenomic tools

## 5.1 Introduction

The *Geobacillus* genus contains Gram-positive rod shaped, spore-forming bacteria that have an optimum growth temperature of 55 – 65 ºC. Members of the *Geobacillus* genus were originally classified in Group 5 of the *Bacillus* genus (Ash *et al.*, 1991). In 2001, based on a combination of 16S rRNA sequence analysis, fatty acid composition and DNA-DNA hybridization (DDH), some members of Group 5 were reclassified into the new genus *Geobacillus,* with the word *Geobacillus* meaning "soil or earth small rod" (Nazina *et al.*, 2001). As of February 2016, there were twenty described *Geobacillus* species with validly published names, and of these four species: *G. debilis*, *G. gargensis*, *G. pallidus* and *G. tepidamans* have been reclassified as *Caldibacillus debilis*, *G. thermocatenulatus*, *Aeribacillus pallidus* and *Anoxybacillus tepidamans* respectively (Minana-Galbis *et al.*, 2010, Dinsdale *et al.*, 2011, Coorevits *et al.*, 2012) The remaining sixteen include *G. galactosidasius*, *G. kaustophilus*, *G. lituanicus*, *G. stearothermophilus*, *G. vulcani*, *G. jurassicus*, *G. thermodenitrificans*, *G. thermoglucosidasius*, *G. thermantarcticus*, *G. thermocatenulatus*, *G. subteraaneus*, *G. uzenensis*, *G. icigianus*, *G. thermoleovorans*, *G. toebii* and *G. caldoxylosilyticus* (Ahmad *et al.*, 2000, Fortina *et al.*, 2001, Nazina *et al.*, 2001, Sung *et al.*, 2002, Kuisiene *et al.*, 2004, Nazina *et al.*, 2004, Nazina *et al.*, 2005, Dinsdale *et al.*, 2011, Poli *et al.*, 2011, Coorevits *et al.*, 2012, Bryanskaya *et al.*, 2015). Of these, there are six species with genome sequences of the type strain.

In bacterial taxonomy, the two methods on which the most emphasis has been placed to demarcate a new bacterial species are 16S rRNA gene sequence analysis and DDH, alongside various phenotypic methods (Tindall *et al.*, 2010). However, in some cases, including the *Geobacillus* genus, the sequence similarity of the 16S rRNA is > 97 % between species (Rainey *et al.*, 1994, White *et al.*, 1994, Weng *et al.*, 2009, Coorevits *et al.*, 2012); therefore the identification of new *Geobacillus* species generally relies on other approaches, such as DDH.

In general DDH is also fraught as a method for the differentiation of bacterial species because of the lack of replication, reciprocation and calibration of the method with a reference strain of a known DDH value (Young, 2015, personal communication, University of York, UK). In the case of new *Geobacillus* species, there has been a lot of variation in DDH values between studies (Rainey *et al.*, 1994, White *et al.*, 1994, Nazina *et al.*, 2004, Weng *et al.*, 2009, Dinsdale *et al.*, 2011, Coorevits *et al.*, 2012), which has led to the taxonomic classification of some species of *Geobacillus* being uncertain. Recently, Dinsdale et al. (2011) showed that some of the

previously published species were in fact synonymous with current species and were no longer considered valid. For example, the described species *G. kaustophilus* (Priest *et al.*, 1988, White *et al.*, 1994, Nazina *et al.*, 2001)], *G. lituanicus* (Kuisiene *et al.*, 2004)and *G. vulcani* (Nazina *et al.*, 2004) were shown to be synonymous with *G. thermoleovorans*. In addition, the described species *G. gargensis* (Nazina *et al.*, 2004) was synonymous with *G. thermocatenulatus*. Most of the disagreement in assigning new species to the *Geobacillus* genus comes from the DDH similarity values distinguishing two species being different between studies.

Other house-keeping genes have been evaluated as alternatives to the 16S rRNA gene, such as *recN*, *recA*, *rpoB*, *gyrB*, *parE* as well as *spo0A,* for their use in identifying *Geobacillus* species, all with limited success (Zeigler, 2005, Kuisiene *et al.*, 2009, Weng *et al.*, 2009, Tourova *et al.*, 2010). Of the genes analyzed, *recN* appears to be the most reliable, with a higher taxonomic resolution compared with 16S rDNA (Zeigler, 2005). However, the taxonomic resolution between some species of *Geobacillus* is still poor (for example, between *G. subterraneus* and *G. uzenensis*).

In the era of next generation sequencing and especially the introduction of third generation (single molecule) sequencing, it is likely that DDH will become outdated. This is already apparent with the proposal to use comparative genomics approaches to demarcate new species with genomic DNA as the type material archived alongside live cultures (Chun & Rainey, 2014, Whitman, 2015).

To the author's knowledge, only one phylogenomic study has been carried on the *Geobacillus* genus, and this only included the *G. thermoleovorans* taxon (Studholme, 2015). In addition, none of the broader taxonomic studies on the *Geobacillus* genus have included *G. stearothermophilus* strains of dairy origin as part of their comparison. Traditionally both a genotypic and phenotypic analysis is carried out to identify a new species. However, the relationship between phenotype and genotype is not always straightforward. This can be seen with dairy strains of *G. stearothermophilus*, which show some different physiological characteristics such as their metabolism (biochemical assays) and fatty acid profile from the type strain *G. stearothermophilus* ATCC 12980 as described in Chapter 2.

The aims of this study were two-fold: firstly, to resolve relationships within the *Geobacillus* genus using three different phylogenomic approaches, and secondly to identify differences within the *G. stearothermophilus* group itself.

## 5.2 Materials and methods

### 5.2.1 Genome sequences

The three dairy strains of *G. stearothermophilus*: A1, P3 and D1 that have been newly sequenced as described in Chapter 4, were compared with the genome sequences of 32 other strains of *Geobacillus* as described in Chapter 4, Table 4.1 as well as two additional newly sequenced genomes; *G. stearothermophilus* ATCC 12980$^T$ (GenBank Accesion JYNW01) and *G. thermoglucosidasius* DSM 2542 (GenBank Accession CP012712), abbreviated to Gs129 and

Gtg254 respectively in this present study. These genomes were not published at the time when the analyses in Chapter 4 were carried out. All of the genomes were parsed and re-annotated using Prokka (v. 1.10) (Seemann, 2014), as described in Chapter 4, Section 4.2.2.

### 5.2.2 Phylogenetic tree of 16S rRNA sequences

Consensus 16S rRNA gene sequences for the three selected dairy strains were generated as described in Chapter 4, Section 4.2.3. These sequences are available under NCBI GenBank accession numbers KU248350 (A1), KU248348 (P3) and KU248349 (D1). The remaining 16S rRNA gene sequences from the other *Geobacillus* genomes were obtained using Prokka (v. 1.10). Only those strains with a full-length 16S rRNA gene sequence were used for the final analysis. All of the 16S rRNA gene sequences were imported into MEGA v. 6.06 (Tamura *et al.*, 2011), aligned using CLUSTALW, and a phylogenetic tree was generated using the maximum likelihood method with 1000 bootstrap replicates. *B. subtilis* 16S rRNA gene was used as the out-group. 16S rDNA percentage identities were calculated using Geneious (v. 7.1.9), and compared in a heat-map. The heat-map was generated using the `heatmap.2` function included in the `gplots` library of the statistics software package R (v3.2.0), visualized in Rstudio (v0.98.1103).

### 5.2.3 Ribosomal multi-locus sequence typing (rMLST)

rMLST was carried out based on the method described by Jolley *et al.* (2012). In the current study, the rMLST analysis was performed using an in-house Perl script. The 53 *rps* genes in each species were retrieved based on a combination of the Prokka annotations followed by a BLAST search. For each genome, the 53 *rps* genes were individually aligned using MUSCLE (v. 3.8.31) (Edgar, 2004), and then subsequently concatenated. The aligned fasta file was imported into MEGA v. 6.06 (Tamura *et al.*, 2011), and a phylogenetic tree was generated using the maximum likelihood method with 1000 bootstrap replicates.

### 5.2.4 Average nucleotide identity (ANI) and Tetra nucleotide frequency calculations

The ANI between two genomes has been proposed as an *in-silico* method to replace DDH (Richter & Rossello-Mora, 2009). The ANI can be calculated using a BLAST approach or by using the program MUMmer. The BLAST approach (called ANIb) was proposed by Goris *et al.* (2007) and involves dividing the genome up into 1020 nucleotide fragments and carrying out a BLAST of each fragment against the subject genome; whereas, the MUMmer approach (called ANIm) involves alignment of the entire genome. This study used the default parameters in the JSpecies software package (v. 1.2.1) to calculate both the ANIb and ANIm between each pair of *Geobacillus* genomes. In addition, the tetranucleotide frequency was calculated using the program TETRA (Teeling *et al.*, 2004b), which is incorporated into JSpecies. Tetranucleotide frequency has been proposed as an additional method to complement ANI for determining genome similarities (Teeling *et al.*, 2004b, Richter & Rossello-Mora, 2009). The tetranucleotide

frequency involves determining the number of times each of the 256 possible tetranucleotides repeats itself within a genome (Teeling *et al.*, 2004a).

The ANIm values were used to compare the relationship of the *Geobacillus* genomes by generating a heat-map. The heat-map was generated using the `heatmap.2` function included in the `gplots` library of the statistics software package R (v. 3.2.0), visualized in Rstudio (v. 0.98.1103).

### 5.2.5 Core genome comparisons

The program OrthoMCL (v. 2.0.9) (Li *et al.*, 2003) was used for determining the core genome as described in Chapter 4, Section 4.2.5. Comparison of the core genome was based on predicted amino acid sequences from "perfect sets" of orthologuous gene clusters (i.e., for a given gene, there were no paralogues identified within a genome), as previously described (Biggs *et al.*, 2011). The length range of the amino acid sequences within a cluster did not vary by more than 20 %. This value allows some variation, without being too flexible, in the length of the protein amongst all cluster members. Variation in predicted protein length may occur for example, from the actual gene starting at a different start codon from that of the predicted annotation. The core genes were aligned individually using MUSCLE (v. 3.8.31) (Edgar, 2004) and concatenated. The Neighbor-Net algorithm (Huson & Bryant, 2006) in SplitsTree (v. 4.13.1) was used to generate a Neighbor-Net with the aligned sequences.

### 5.2.6 Identification of strains based on the 16S rRNA gene

The Seqmatch tool (http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp) in the ribosomal database project (RDP) (Cole *et al.*, 2014), was used to confirm the identity of those *Geobacillus* species with conflicting names or where there was no type strain within a species group. A seqmatch score was given for each submission. As described on the RDP website the Seqmatch score was determined by taking the number of unique 7-base oligomers shared between the two sequences and dividing it by the lowest number of unique 7-base oligomers from one of the sequences; therefore, for two identical sequences the score would be 1.0.

## 5.3 Results

To gain an understanding of how the newly sequenced *G. stearothermophilus* strains isolated from dairy manufacture are related to other *Geobacillus* species, the traditional method of 16S rRNA sequence analysis was carried out, as well as three different phylogenomic approaches including rMLST, ANI and a comparison of the core genomes. In addition, these methods were evaluated for their ability to resolve the *Geobacillus* taxonomy.

### 5.3.1 16S rRNA gene sequencing and rMLST

Initially a traditional approach was followed, by analysing the similarities of the 16S rRNA gene sequences (Figure 5.1A), which was compared with rMLST (Figure 5.1B). For the 16S rRNA maximum Likelihood phylogeny three strains of *Geobacillus*: *G. thermodenitrificans* DSM 465, *G. thermocatenulatus* GS-1 and *Geobacillus* MAS1, were not included as there were no full 16S

rRNA gene sequences available. The dairy strains A1, P3 and D1 fell within one clade along with *G. stearothermophilus* ATCC 7953 and ATCC 12980. 16S rRNA gene sequencing was not able to distinguish between some clades (with bootstrap values less than 70 %). In practice, bootstrap values should be higher as seen with the rMLST tree. 16S rDNA percentage identities were also calculated[13] and the values compared using a heat-map (Figure 5.2).

16S rDNA sequence analysis was compared with rMLST to show that it is possible to resolve differences between some taxa of *Geobacillus* when a multi-locus approach is taken. rMLST compared the sequences of 53 house-keeping genes (Figure 5.1B). rMLST is more useful for identifying closely related taxa, given a larger number of nucleotides are analyzed (approximately 21kb compared with 1.5kb) and because there is more variability between some of the ribosomal subunit genes. As with the 16S rRNA, the dairy strains A1, P3 and D1 formed a unique clade with the *G. stearothermophilus* strains ATCC 7953 and ATCC 12980. rMLST clustered the *Geobacillus* strains into six clades, with other strains falling as individuals outside any clusters. There were also inconsistencies between the two methods. For example, using 16S rRNA gene sequence analysis, *G. vulcani* PSS1 clustered within the *G. thermoleovorans* clade, whereas using rMLST, this isolate appeared as an individual isolate.

[13] CD: Chapter5/Results/Geobacillus16S%ID.xlsx

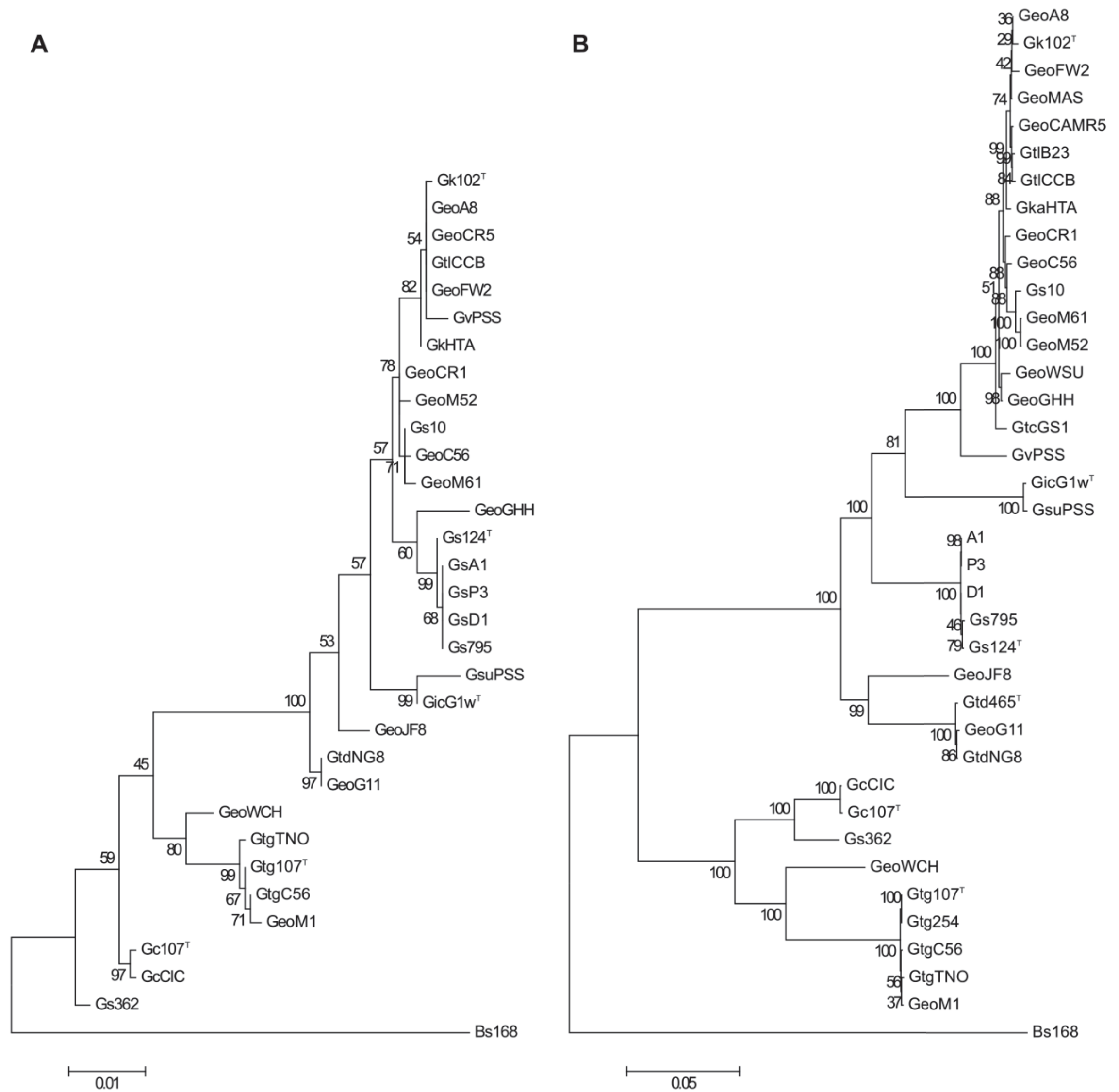



**Figure 5.1. Un-rooted maximum likelihood phylogenetic trees for the *Geobacillus* genus. A.** Phylogeny based on full-length sequences of 16S rRNA genes. **B**. rMLST using 53 concatenated house-keeping genes. Bootstrap values (expressed as percentages based on 1000 replicons) are indicated at the major branching points.

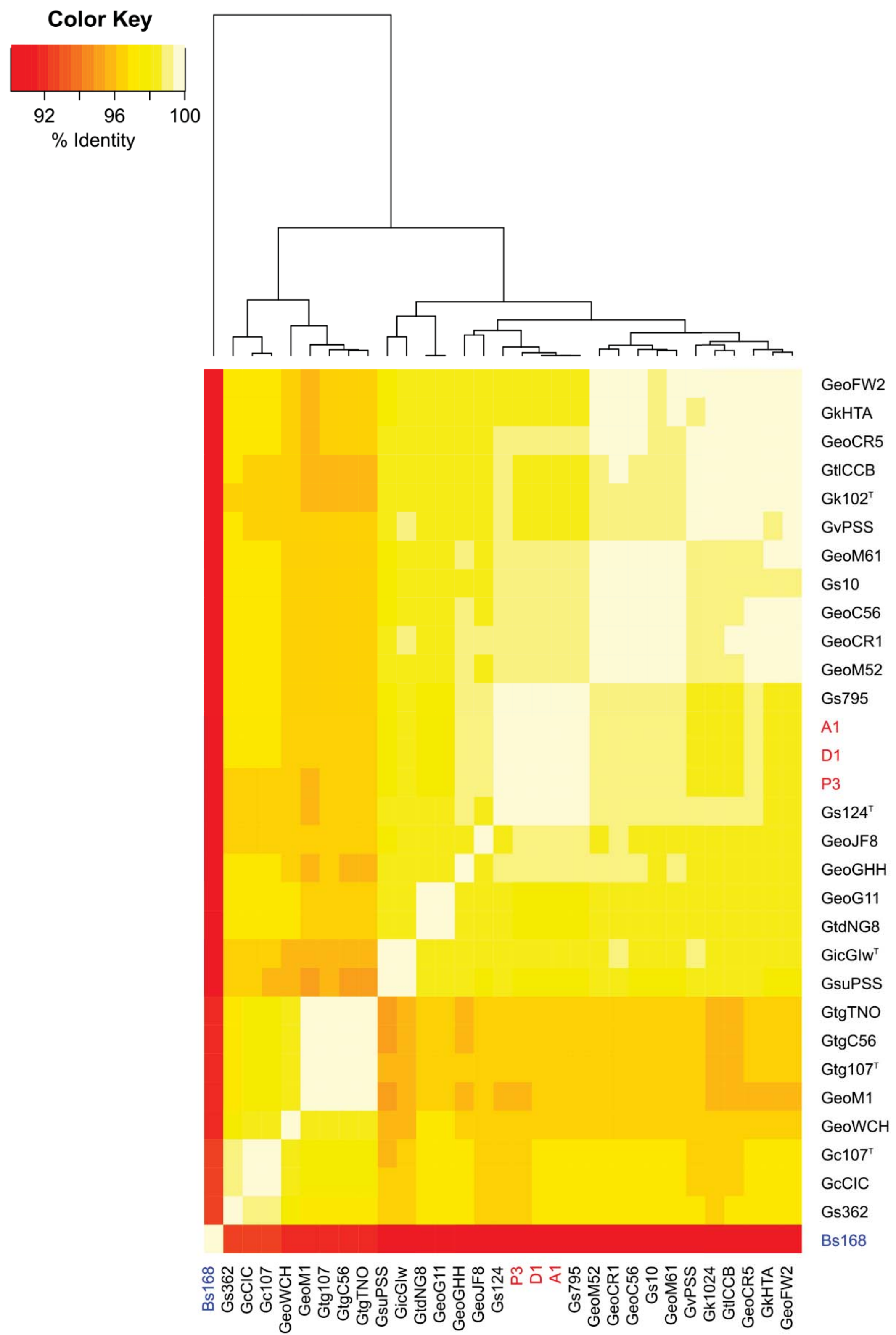



**Figure 5.2. Heat-map comparison of 16S rRNA gene percentage identities.**

### 5.3.2 ANI and TETRA calculations

To examine the taxonomic classification of *Geobacillus* spp., the ANI and tetranucleotide frequencies were calculated for all of the sequenced genomes of the *Geobacillus* genus[14], where those values highlighted in green are considered to be members of the same taxon (> 95 % for ANIb and ANIm and > 0.99 for TETRA) and those values highlighted in red are considered to be different taxa.

In some cases the ANIb values did not support the ANIm values. For example, the ANIb values classified *Geobacillus* WSUCF1 as a different taxon from five strains including *Geobacillus* A8 with an ANIb value of 89.34 %. In contrast, the ANIm values classified these same strains as belonging to the same taxon with the ANIm value between *Geobacillus* WSUCF1 and *Geobacillus* A8 being 96.11 %. In addition, despite repeated attempts, including using smaller groups, the ANIb calculation was not successful for some of the pairs of *Geobacillus* strains (represented by NaN in the spread-sheet). It did not appear to be because the strains were too different to compare as there were some values that were quite low, for example, between those strains classified as *G. glucosidasius* and *G. stearothermophilus* ATCC7953.

TETRA (tetranucleotide) calculations were carried out to support the ANI values. However, results showed that the TETRA values did not group the *G. stearothermophilus* strains (ATCC 12980$^T$, ATCC 7953, A1, D1 and P3), as a separate taxon from those classified as *G. thermoleovorans*[2].

The ANIm frequencies were visualized using a heat-map (Figure 5.3). The heat-map was non-symmetrical as a result of greater differences between some of the ANIm values. No correlation (using the data in the Excel spread-sheet[2]) could be obtained between a larger difference in ANIm value and the number of contigs. However, when the difference in contig number was zero (this was only the case when both genomes being analyzed were complete genomes; i.e., one contig) the difference between the two ANIm values was less than 0.26 %. When the difference between two ANIm values is greater than 0.5 % around the 95 % threshold it could potentially place ambiguity around the taxonomic position of a strain. However, this was not seen in this present study where the difference in two ANIm values between two members of the same taxon was always less than 0.5 %, so that there were defined demarcations between taxa (as designated by a red box in Figure 5.3). The newly sequenced dairy strains and *G. stearothermophilus* strains ATCC 7953 and ATCC 12980 had ANIm values > 95 % grouping them within the same taxon.

---

[14] CD: Chapter5/Results/GeobacillusJspecies.xlsx

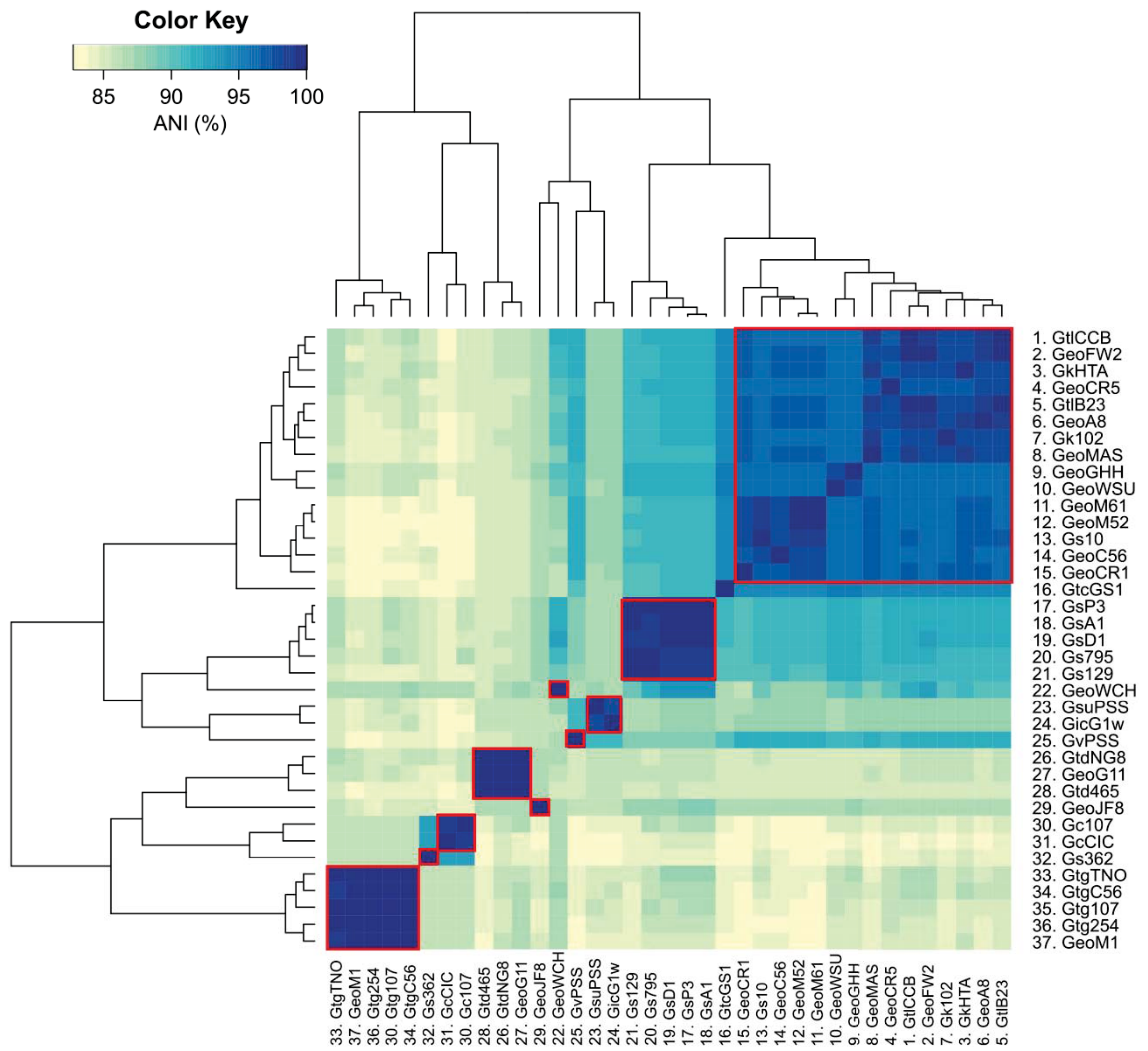



**Figure 5.3. Heat map comparison of the ANIm values.** Some strains were placed in a different order for the recipricol pairwise comparison by the dendrogram option using the `heatmap.2` function in R. Those ANI values greater than 95 %, grouping the strains within the same species, were enclosed by a red box.

### 5.3.3 Core genome comparisons

Another potential substitute for DDH is to use core genome comparisons. In this study, the core genome was defined using the program OrthoMCL. Each orthologous group contained only one gene per genome to be defined as a member of a gene cluster. In addition, to be included in the core genome, the length range of the amino acid sequences within each cluster did not vary by more than 20 %. A Neighbor-Net was then generated using the concatenated sequence of those orthologous genes. Two analyses were carried out for all of the *Geobacillus* genomes, using those translated amino acid sequences that had no difference in sequence length (Appendix 6) and those that were within a length range of 20 % (Figure 5.4A). For the subsequent analyses a length range of 20 % was used given there was little difference in clusters between the two analyses.

Core genome comparisons separated the *Geobacillus* genus (Group A, Table 5.1) into similar clusters to those defined from the rMLST and ANI analyses. To analyze the relationship of the *G. stearothermophilus* taxon more closely, comparison of the core genome was carried out on three smaller groups (Group B, C and D, Table 5.1), of *Geobacillus* taxa. Although strains A1, P3, D1, ATCC 7953 and ATCC 12980 belong to the same species, Figure 5.4D shows that based on the core genome, A1 and P3 are the same, whereas the remaining three strains are different.

**Table 5.1. Number of genes and amino acids used in the OrthoMCL clustering.**

| Group[a] | Members of group | Same length and same sequence | | Same length and different sequence | | Both length[b] and different sequence | | Total core genes |
|---|---|---|---|---|---|---|---|---|
| | | Genes | Amino acids | Genes | Amino acids | Genes[c] | Amino acids | |
| **A** | All *Geobacillus* spp. | 1 | 116 | 103 | 15 658 | 638 | 146 875 | 639 |
| **B** | *G. thermoleovorans*[d], *G. stearothermophilus*[e], *G. icigianus*[f] and *G. thermodenitrificans*[g] groups as well as *G. thermocatenulatus* and *G.vulcani* | 9 | 990 | 186 | 32 000 | 725 | 164 902 | 727 |
| **C** | *G. thermoleovorans*[d], *G. stearothermophilus*[e] groups as well as *G. thermocatenulatus* and *G.vulcani* | 9 | 990 | 276 | 49 838 | 799 | 184 151 | 808 |
| **D** | *G. stearothermophilus*[e] group. | 958 | 197 978 | 1 003 | 329 507 | 1 247 | 406 988 | 2 205 |

[a] Refer also to Figure 5.4

[b] To be in an orthologous cluster, genes had to have a length range of 20% across all cluster members, and only one member per strain

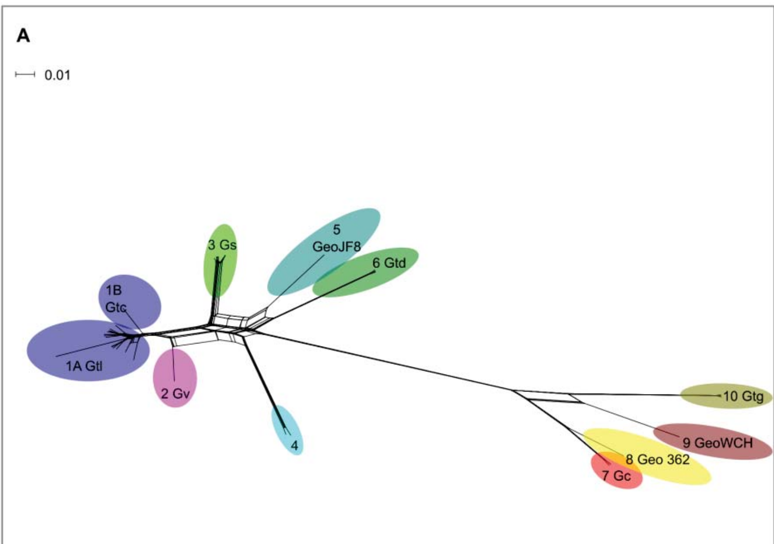
[c] This number of genes was used in the Neighbor Net analysis

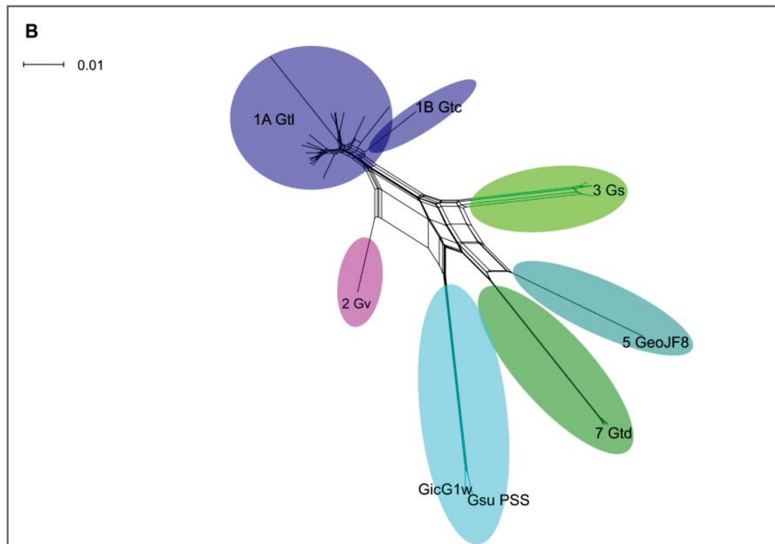
[d] Includes GkHTA, Gk102445, GkGbl, GthlCCB, GthlB23, GeoWSU, GeoGHH, GeoC56, Gs10, GeoMC61, GeoM52, GeoCAMR1, GeoCAMR5, GeoFW2, GeoA8 and GeoMAS

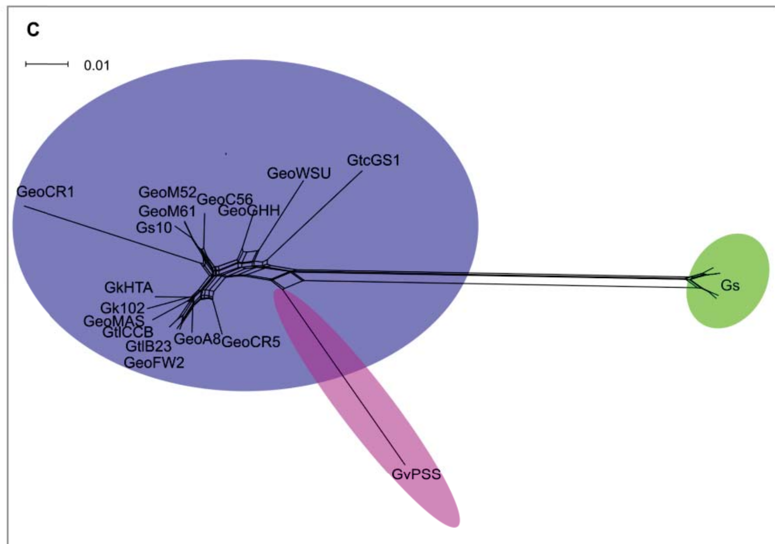
[e] Includes Gs795, Gs129, A1, P3 and D1

[f] Includes GsuPSS and GeoG1w

[g] Includes GthdNG8, Gthd465 and GeoG11


A
0.01
1A Gtl
1B
Gtc
2 Gv
3 Gs
4
5
GeoJF8
6 Gtd
7 Gc
8 Geo 362
9 GeoWCH
10 Gtg


B
0.01
1A Gtl
1B Gtc
2 Gv
3 Gs
5 GeoJF8
7 Gtd
GicG1w
Gsu PSS


C
0.01
GeoCR1
GeoM52
GeoC56
GeoM61
GeoGHH
Gs10
GeoWSU
GtcGS1
GkHTA
Gk102
GeoMAS
GtlCCB
GtlB23
GeoFW2
GeoA8
GeoCR5
GvPSS
Gs

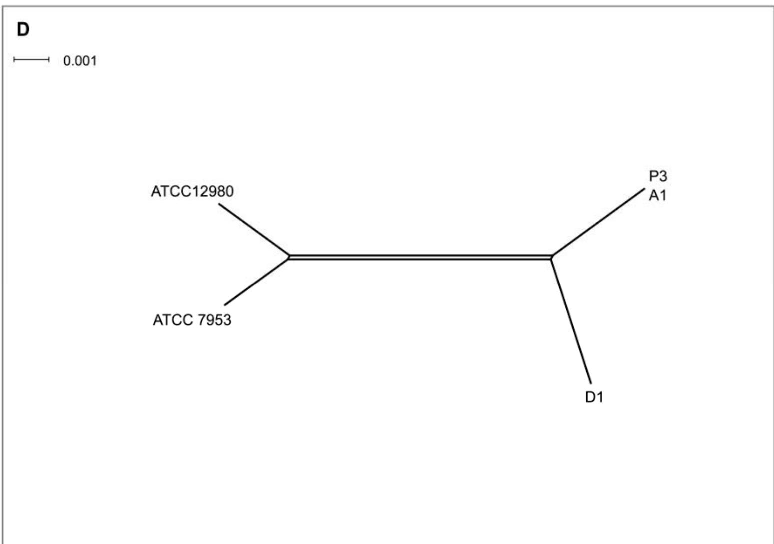

D
0.001
ATCC12980
ATCC 7953
P3
A1
D1

**Figure 5.4 Core genome sequence comparisons.** The phylogenetic networks were generated using the Neighbor-Net algorithm in SplitsTree (v4.13.1). The orthologous groups were defined using the program OrthoMCL (v. 2.0.9) and the analyses were based on those genes that have orthologous gene members with a length range less than or equal to 20 %. The first analysis included all of the *Geobacillus* genomes (**A)** where *G. thermoleovorans* group 1A refers to strains GkGbl, GthlB23, GthlCCB, GeoCAMR5, GeoFW2, GeoMAS, Gk102445, GeoA8 and GkHTA; *G. thermoleovorans* group 1B refers to GeoMC52, GeoMC61, Gs10 and GeoC56; *G. thermoleovorans* group 1C refers to GeoGHH and GeoWSU; Group 4 refers to GsuPSS and GicG1w; *G. stearothermophilus* group 5 includes Gs7953, A1, P3 and D1; *G thermodenitrificans* group 7 includes Gthd465, GthdNG8 and GeoG11; *G. caldoxylosilyticus* group 8 includds GcCIC and Gc10776; *G. thermoglucosidans* group 11 includes Gthg10776, GthgC56, GthgTNO and GeoMC1. Each subsequent analysis (**B**, **C** and **D**) includes fewer genomes as described in Table 5.1.

### 5.3.4 Classification of *Geobacillus* species

Some of the strains used in this study did not have a species name and sat in a cluster by themselves, or did not contain a type strain within their cluster. In addition, one cluster contained two species names (*G. subterraneus and G. icigianus*). To determine whether they were related to any of the validly published type strains, the RDP was used to determine the best match (Table 5.2). *Geobacillus icigianus* was the only *Geobacillus* species, which did not have its type strain included in the RDP. *G. vulcani* PSS was the only strain analyzed that gave a significant match to a type strain (highlighted in grey in Table 5.2). The identities of the other strains were unable to be confirmed.

**Table 5.2. Identification of *Geobacillus* strains using the RDP.**

| | Top identify in RDP database | Acccession number | Seqmatch score[b] |
|---|---|---|---|
| ***G. subterraneus* PSS1** | *G. thermocatenulatus* | AY608935 | 0.915 |
| ***G. vulcani* PSS** | *G. vulcani*[a] | AJ293805 | 0.991 |
| | *G. thermoleovorans* | ZZ26923 | 0.969 |
| ***G. iciguinius* G1w** | *G. vulcani*[a] | AJ293805 | 0.934 |
| | *G. thermocatenulatus* | AY608935 | 0.931 |
| ***Geobacillus* JF8** | *G. thermodenitrificans* | AY608961 | 0.855 |
| ***Geobacillus* WCH** | *G. toebii* | AF326278 | 0.961 |
| ***G. stearothermophilus* NUB3621** | *G. caldoxylosilyticus* | AF067651 | 0.928 |

[a] The top two hits were included given *G. vulcani* given in the study by Dinsdale *et al.* (2011) *G. vulcani* was deemed to be synonymous with *G. thermoleovorans*, although *G. vulcani* is a validly published name according to the List of prokaryotic names with standing in nomenclature (LPSN) bacterio.net (http://www.bacterio.net/geobacillus.html).

[b] Number of unique 7-base oligomers shared between the two sequences divided by the lowest number of unique 7-base oligomers from one of the sequences

Based on the results from this study it is clear that some of the *Geobacillus* strains, which have had their genome sequences published, have either not been identified, or have been mis-identified (Table 5.3), as found in previous studies (Dinsdale *et al.*, 2011, Coorevits *et al.*, 2012). Using a combination of rMLST, core genome comparisons and ANIm values, the *Geobacillus* genus could be divided into ten species (Figure 5.4) based only on those species with genome sequences available. The taxonomic position of *G. thermocatenulatus* is less clear. It borders

on the ANIm demarcation threshold from genomes in the *G. thermoleovorans* group (94.1 – 95.0 %) and it is not clearly delineated from this same group based on the core genome comparisons.

**Table 5.3. Identity of *Geobacillus* strains.**

| Given species name | Strain names | Species name determined by this study |
|---|---|---|
| *Geobacillus kaustophilus* | NBRC 102445 and HTA426, | *Geobacillus thermoleovorans* |
| *Geobacillus* sp. | Y412MC61, Y412MC52, C56-T3, GHH01, FW23, CAMR12739, CAMR5420, MAS1, A8 and WSUCF1, | *Geobacillus thermoleovorans* |
| *Geobacillus stearothermophilus* | 10 | *Geobacillus thermoleovorans* |
| *Geobacillus stearothermophilus* | NUB3621 | Unknown, possible new species |
| *Geobacillus* sp. | Y4.1MC1 | *Geobacillus thermoglucosidans* |
| *Geobacillus* sp. | G11MC16 | *Geobacillus thermodenitrificans* |
| *Geobacillus sp.* | WCH70 | Unknown, possibly *G. toebii* |
| *Geobacillus sp.* | JF8 | Unknown, possible new species |

## 5.4 Discussion

This study carried out a comparative analysis of whole genome sequences to determine differences between the three dairy strains and how they fit within the taxonomy of the *Geobacillus* genus. In addition, a number of genomic approaches were taken for resolving the taxonomy of the *Geobacillus* genus as much as possible, based on the genome sequences available. Genome sequences are available for eleven of the fifteen described *Geobacillus* species with validly published names, which were included in this study.

Based on the methods used in this study, the three newly sequenced *Geobacillus* strain from a dairy origin clustered within the *G. stearothermophilus* group along with *G. stearothermophilus* ATCC 7953 and the type strain ATCC 12984. The results from this chapter agree with the work described in Chapter 4 that showed that that strain D1 is different to strains A1 and P3. The work also further confirmed that A1 and P3 were similar to each other.

In bacterial taxonomy the two methods on which the most emphasis has been placed on demarcating a new bacterial species are 16S rRNA gene sequence analysis and DDH, alongside various phenotypic methods. Generally isolates with < 97 % identity for the 16S rRNA gene are regarded as separate species. More recently, it was proposed that this threshold for demarcating species should be increased to 98.65 % (Kim *et al.*, 2014). In reality, setting a threshold based on 16S rRNA gene similarity, let alone such a specific number, does not work,

as can be seen in this present study where such a threshold would not separate some taxa. For example, the 16S rRNA ranged between 98.4 – 99.0 % similarity between strains of *G. stearothermophilus* and *G. thermoleovorans*. However, these are clearly two different taxa based on the various genomic approaches used in this study. In addition, it is clear that variation within a species and between closely related species can be missed using traditional Sanger sequencing methods, such as sequence analysis of the 16S rRNA genes. However, in this genomics era, comparison of the 16S rRNA gene may still have a role to play in the initial identification of a strain, using databases such as RDP, as was used in this study.

In this present study, 16S rDNA sequence analysis was compared with rMLST to show that it is possible to resolve differences between some strains of *Geobacillus* when a multi-locus approach is taken. Not surprisingly, rMLST had a higher resolution and better support from the bootstrap values compared with 16S rDNA sequence analysis. 16S rDNA has been criticized for having insufficient resolution to identify and compare the relationships between *Geobacillus* spp. Tindall *et al.* (2010) recommends that 16S rRNA gene analysis only be used as a preliminary screen and that other methods should also be used, such as sequence analysis of other genes and DDH. Surprisingly, a genomics based approach was not proposed, although the authors do state that a genomics based method such as ANI may replace DDH in the near future. A number of house-keeping genes have been analyzed in *Geobacillus* spp. as an alternative to 16S rRNA gene analysis, with little success (Zeigler, 2005, Kuisiene *et al.*, 2009, Weng *et al.*, 2009, Tourova *et al.*, 2010). Two issues were encountered with these approaches including: (1) inability to differentiate between more closely related species or (2) no universal primers available that could amplify the selected gene(s) in all of the *Geobacillus* species analyzed. These are not issues faced by a comparative genomics approach such as rMLST. With the invention of next generation sequencing it has become affordable to sequence multiple genomes, which may lead to comparative genomics approaches such as rMLST being used more frequently for *Geobacillus* taxonomy.

Alternatively a cheaper option, requiring less bioinformatics input, would be to develop a MLST scheme for the *Geobacillus* genus. In this study, core genome comparisons were also carried out, which separated the *Geobacillus* genus into similar clusters to the rMLST and ANI analyses. Determination of the core genome may provide a basis for identifying 6 – 7 genes that would be suitable for an MLST scheme for *Geobacillus* spp. A core genome comparison of *Geobacillus* spp. was carried out by Studholme (2015); however, the analysis only included those genome sequences that make up the *G. thermoleovorans* and *G. thermocatnulatus* group. The groupings were similar to that found in this study for the OrthoMCL clustering of group C (Figure 5.4C), providing evidence that core genome comparisons are comparable between research groups. Studholme (2015), did not describe the method used for determining the core genome.

When identifying new *Geobacillus* species DDH association values have been relied on because of the high 16S rRNA gene similarity, where < 70 % for a DDH value is defined as

different species (Tindall *et al.*, 2010). However, in the case of new *Geobacillus* species there has been a lot variation in DDH values between studies. For example, Nazina *et al.* (2004) found *G. vulcani* had DDH values < 70 % when hybridized with a number of other type strains of *Geobacillus* including *G. thermoleovorans*. In contrast Dinsdale *et al.* (2011) found *G. vulcani* to be synonomous with *G. thermoleovorans* with a DDH value of 79.4 %. In the era of next generation sequencing and the introduction of third generation sequencing with single molecule sequencing it is immenent that DDH will become outdated. This is already apparent with the proposal of using a comparative genomics approach for demarcating new species and genomic DNA as the type material alongside or instead of live cultures (Chun & Rainey, 2014, Whitman, 2015). The most likely substitute for DDH is ANI (Tindall *et al.*, 2010, Chun & Rainey, 2014). It has been shown that an ANI value of < 95 – 96 % generally corresponds well with the thresholds of < 70 % for DDH and < 97 - 98 % for 16S rRNA gene identity for defining new species (Goris *et al.*, 2007, Kim *et al.*, 2014). The Genome Blast Distance Phylogeny tool and multi-locus sequence analysis have also been proposed as alternatives (Whitman, 2015).

In this present study, the Jspecies software package was used to calculate the ANI and tetranucleotide frequency. The ANIm values were found to be more consistent for demarcating species compared with the ANIb values. This may be because the ANIb relies on pre-processing of the genome sequence by dividing it into 1020 nucleotide fragments, as opposed to ANIm which involves alignment of the genome. Some of the draft genomes contain contigs with less than 1020 nucleotides in length, which may affect the ANIb process. The tetranucleotide frequencies did not support the ANI values for dermacating closely related taxa. This was also observed in a study comparing the genetic diversity of *Lactococcus lactis* (Cavanagh *et al.*, 2015). In the evaluation of the Jspecies package, Richter & Rossello-Mora (2009) found a number of cases where the ANI values were low and the tetranucleotide frequencies were high, but it was very rare for the opposite to occur. This present study also found that the reciprical ANIm calculations did not always emulate each other. This also occurred in the study by Cavanagh *et al.* (2015), but to a lesser extent. In this present study despite differences in the reciprocal values, there were still clear demarcations between taxa.

The use of ANI for defining new species is not without its problems (Richter & Rossello-Mora, 2009). Two key issues are that the genome sequences of many type strains are not available, and there are many strains that have been incorrectly identified to a given species. In the analysis of Richter and Rossello-Mora (2009), it was found that for those genomes with validly published names, only 45 % actually belonged to the same species as the type strain (as defined by other means such as DDH). As of 31 July 2013 there were 10,546 validly published bacterial species names, but only 14.9 % of these had genome sequences available for the type strain (Chun & Rainey, 2014). This issue has arisen within the *Geobacillus* genus where, in defining the new species *G. icigianus,* Bryanskaya *et al.* (2015) carried out an ANI analysis with three other genome sequences with only three of these strains being a validly published type strains. Another issue faced when using ANI is that it takes into account the entire genome including accessory elements such as the mobilome. This may explain why the placing of the

genome *Geobacillus* WCH70 results in a unsymmetrical placing of this taxon on the ANIm heat-map (Figure 5.3) as it showed the closest similarity to the *G. stearothermophilus* strains A1, P3 and D1 (approximately 93 %) as opposed to the other two strains of *G. stearothermophilus*, ATCC 7953 and ATCC 2461 (92 % and 89 % respectively). When a core-genome comparison was carried out, *Geobacillus* WCH70 was more closely related to the *G. thermoglucosidasius* and *G. caldoxylosilyticus* taxa.

A core genome approach is useful for determining phylogenomic relationships. In this present study, a Neighbor-Net was generated for making core genome comparisons. An advantage of using a Neighbor Net, as opposed to a branching phylogenetic tree, is that it can show any ambigious signal as to the taxonomic relationship between strains (Huson & Bryant, 2006). Ambiguous signal can arise from events such as gene duplication, gene transfer, different rates of mutation and recombination (Bryant & Moulton, 2004).

Based on this present study, there are currently ten *Geobacillus* taxa with genome sequences available. Of these, seven appear to have validly published names (*G. stearothermophilus*, *G. thermodenitrificans*, *G. thermoglucosidans*, *G. icigianus*, *G. thermoleovorans, G. vulcani and G. caldoxylosilyticus*). The remaining three had no species names and may be new taxa. However, to confirm this, the genome sequences for all of type strains need to be generated. Genome sequences are also available for the validly published *G. kaustophilus*, *G. subterraneus* and *G. thermocatenulatus.* This present study agrees with previous studies that have shown *G. kaustophilus* is synonymous with *G. thermoleovorans* (Sunna *et al.*, 1997, Coorevits *et al.*, 2012). In the case of *G. subterraneus,* this study showed it belonged to the same taxon as *G. icigianus.* Based on 16S rRNA analysis it appears the *G. subterraneus* strain PSS2 has been incorrectly identified. The taxonomic position of *G. thermocatenulatus* GS-1 is less clear. It borders on the ANIm demarcation threshold from genomes in the *G. thermoleovorans* group (94.1 – 95.0 %) and it is not clearly delineated from this same group based on the core genome comparisons. This may indicate that it is a subspecies of *G. thermoleovorans* rather than a separate species. Previous studies disagree on whether *G. thermocatenulatus* can be regarded as a separate species (Sunna *et al.*, 1997, Dinsdale *et al.*, 2011), and analysis of further *G. thermocatenulatus* strains is required to determine its taxonomic position.

## 5.5 Conclusions

A comparative genomics approach was taken to show how three newly sequenced *Geobacillus* isolates of dairy origin are related to other *Geobacillus* strains. Core genome sequence comparisons demonstrated that the strain *G. stearothermophilus* D1 was different to strains A1 and P3. Based on three different genomic approaches (rMLST, ANI and core genome comparisons), the *Geobacillus* genus could be divided into ten taxa for those *Geobacillus* strains that have genome sequences available. Comparison of the genome is able to resolve differences between species of the *Geobacillus* genus that cannot be resolved using the traditional approach of 16S rRNA gene sequence analysis. However, 16S rRNA gene sequencing can still provide a useful method as a preliminary screen for identifying newly

isolated strains. All three genomic approaches used in this study would be useful for classifying new species in the genus *Geobacillus*. However, although ANI was able to be used for demarcating taxa, it should not be used for determining phylogenetic relationships. For any genomic approach to become routine, all of the type strains would need to be sequenced first. Defining a threshold for the demarcation of new species can be difficult, even for some genomic approaches; therefore, it is recommended that a multi-faceted approach as well as multiple strains should be used.

Chapter 6

# 6 Genomic insights into biofilm formation and sporulation of *Geobacillus* spp.

## 6.1 Introduction

When bacteria are under stress they can respond in different ways. In dairy manufacturing environments bacteria face a multitude of stresses, including heat, chemicals and low water activity. Despite this harsh environment some bacterial species such as *Geobacillus stearothermophilus* are able to colonise manufacturing equipment. Biofilm formation and sporulation are two characteristics that make this species difficult to control in dairy manufacturing.

In the model organism, *B. subtilis,* distinct developmental processes can be activated to enable cell survival, resulting in a population of different cell types (Lopez *et al.*, 2009a, Murray *et al.*, 2009). Three of these processes, cannabolism, matrix production (for biofilm formation) and sporulation, are all governed by the master regulator Spo0A (Lopez & Kolter, 2010, Eduardo Gonzalez-Pastor, 2011, Higgins & Dworkin, 2012). Spo0A is activated via a phosphorelay, where kinases sense a signal and respond, ultimately resulting in transfer of a phosphoryl group to Spo0A via the phosphotransferases Spo0B and Spo0F (Figure 6.1).

Five orphan histidine kinases (HK): KinA, KinB, KinC, KinD and KinE, have the ability to phosphorylate Spo0A either directly or indirectly via the phosphorelay (Jiang *et al.*, 2000b). These HKs are all comprised of sensor domain(s) (at the N-terminal end) and an autokinase (at the C-terminal end) which transfers the phosphoryl group (Stephenson & Hoch, 2002). KinA is the main HK required for entry to sporulation *in vivo*, and to lesser extents KinB and KinC (Fujita & Losick, 2005). The fourth HK, KinD, has been described as a “checkpoint protein” in biofilms by sensing matrix completion and thus allowing sporulation to proceed (Aguilar *et al.*, 2010). The authors of this latter study suggest KinD acts as a phosphatase, resulting in the dephosphorylation of Spo0F~P or Spo0A~P and once matrix formation is complete this is sensed by KinD resulting in a change of role for KinD from phosphatase to kinase. The role of the fifth HK, KinE, *in vivo* is unknown.

The phosphorelay is regulated by a number of different mechanisms including control of HK activity by regulatory factors (Dartois *et al.*, 1997, Wang *et al.*, 1997, Burkholder *et al.*, 2001, Rowland *et al.*, 2004, Banse *et al.*, 2011), dephosphorylation of Spo0F~P and Spo0A~P by phosphatases (Perego, 2001, Diaz *et al.*, 2008, Diaz *et al.*, 2012) as well as regulation at the transcription level by SigE (Higgins & Dworkin, 2012). The amount of Spo0A~P determines

which genes are switched on. For example, low levels of Spo0A~P result in activation of genes involved in matrix production and high levels for sporulation (Fujita *et al.*, 2005).

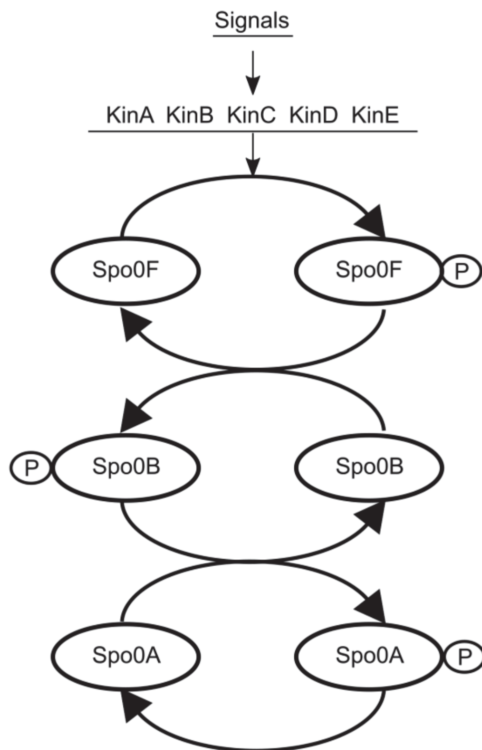



**Figure 6.1. Initiation of biofilm formation and sporulation through activation of Spo0A.** For a more detailed schematic the reader is referred to publications by Higgins & Dworkin (2012) and Tojo *et al.* (2013).

In the case of biofilm formation, Spo0A~P indirectly controls entry to the matrix pathway via the master regulator of matrix production SinR. Matrix production involves a number of different genes as summarized in Table 6.1. The regulatory network that directs matrix production as well as further descriptions of the genes involved to produce extracellular matrix was recently reviewed by Vlamakis *et al.* (2013).

**Table 6.1. Genes involved in production of the extracellular matrix in *B. subtilis*.** Adapted from Vlamakis *et al.* (2013).

| Gene or operon[a] | Role | Reference |
|---|---|---|
| ***epsABCDEFGHIJKLMNOP*** | Exopolysaccharide production | Branda *et al.* (2001), Branda *et al.* (2006) |
| ***tkmA – ptkA – ptpZ – ugd*** | Exopolysaccharide production and sporulation within a biofilm | Gao *et al.* (2015a), Gerwig *et al.* (2014), Kiley & Stanley-Wall (2010) |
| ***tapA – sipW – tasA*** | Protein component of matrix | Branda *et al.* (2006), Romero *et al.* (2010), Romero *et al.* (2011) |
| ***bslA*** | Formation of a hydrophyobic layer | Hobley *et al.* (2013), Kobayashi & Iwano (2012) |
| ***pgsABC*** | $\gamma$-polyglutamic acid production | Stanley & Lazazzera (2005) |
| ***pgcA*** | Nucleotide sugar production | Branda *et al.* (2004), Lazarevic *et al.* (2005) |
| ***gtaB*** | Nucleotide sugar production | Lazarevic *et al.* (2005) |

[a] Those genes separated by a hyphen are adjacent to each other and believed to be transcribed in an operon.

Once Spo0A~P levels are high enough, some of the sporulation genes are induced. Sporulation is a last resort for cell survival and results in the formation of a resistant dormant structure called an endospore (hereafter referred to as a spore). The sporulation process can be divided into seven stages from starvation (Stage 0) through to release of the spore (Stage VII), whereupon when the conditions are right, germination of the spore can occur. This present study will focus on the identification of genes involved in Stages I – VII. Germination will not be addressed further in this study. Upon intiation of sporulation, asymmetric division occurs separating the forespore from the mother cell. Assymetric division must be completed before a commitment is made to sporulate. Once committed this complex process involves the expression of over 500 genes that are regulated in a tightly controlled manner, overseen by four sporulation-specific sigma factors $\sigma^E$ , $\sigma^K$, $\sigma^F$and $\sigma^G$ (Eichenberger *et al.*, 2003, Wang *et al.*, 2006). The expression of these sigma factors and their respective regulons is compartmentalized in terms of both location and timing (Figure 6.2) (Eichenberger *et al.*, 2004). The sporulation-specific sigma factors are initially synthesized as inactive precursor sigma factors and once activated they initiate transcription of genes within their regulon.

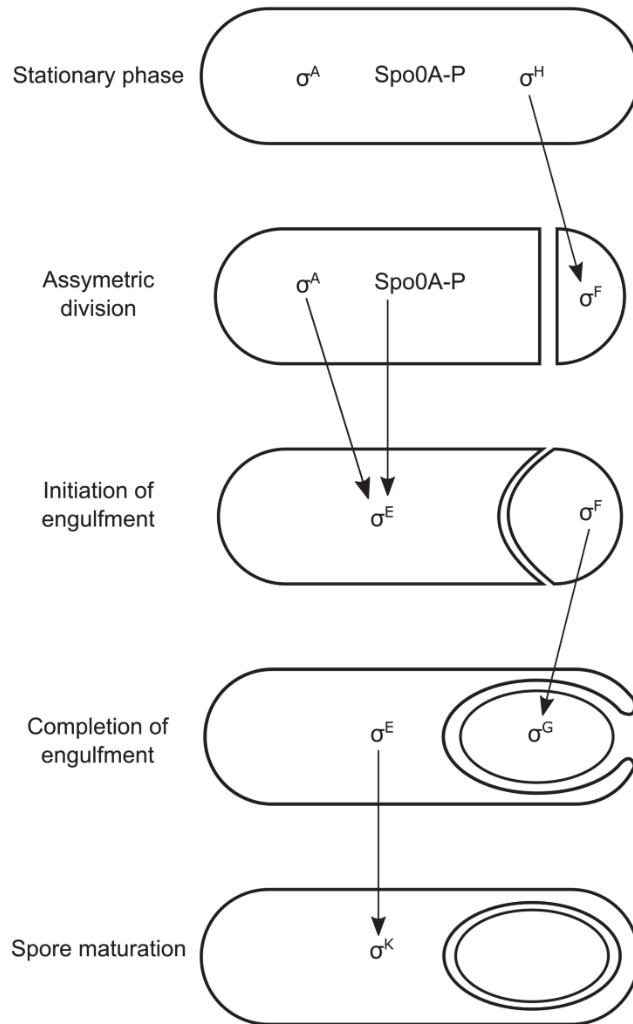



**Figure 6.2. Simplified schematic of the sporulation sigma factor regulatory network**. Gene expression during stationary phase is predominantly regulated by the sigma factor $\sigma^H$. Upon initiation of sporulation gene expression is under the control of the four sporulation specific sigma factors, $\sigma^E$ , $\sigma^K$, $\sigma^F$and $\sigma^G$, whose expression is compartmentalized in terms of both location (as depicted by the mother cell and the forespore) and timing. The master regulator, Spo0A as well as two other sigma factors, $\sigma^H$ and the main sigma factor for vegetative growth, $\sigma^A$, are involved with expression of the sporulation specific sigma factors as depicted by the arrows. For a more detailed schematic the reader is referred to the review by de Hoon *et al.* (2010).

Both biofilm formation and sporulation have been extensively studied in *B. subtilis* but poorly characterized in the Geobacillus genus or other members of the Bacillaceae family. *Geobacillus* spp. can be found in a wide variety of environments, both hot such as hot-springs, compost, and dairy manufacturing plants, as well as cold, such as soils and ocean sediments (Marchant *et al.*, 2002, Sung *et al.*, 2002, Rueckert *et al.*, 2004, Takami *et al.*, 2004a, Scott *et al.*, 2007, Poli *et al.*, 2011, Bryanskaya *et al.*, 2015). In their natural environment bacteria seldom grow in their planktonic state, instead they form biofilms. One would imagine *Geobacillus* spp. also form biofilms when growth conditons are favorable. Biofilms have only been studied in two species

isolated from dairy manufacturing plants, namely *G. stearothermophilus* and *G. glucosidasius* (Flint *et al.*, 2001b, Somerton *et al.*, 2013, Zhao *et al.*, 2013). The ability of *Geobacillus* spp. to sporulate enables them to survive the extreme and diverse environments they are found in. Although, sporulation has been shown to occur in *Geobacillus* spp. this process has not been characterized at the molecular level.

Many genomic studies have been carried out to identify sporulation-specific genes, but these studies have generally included few *Geobacillus* genomes (Onyenwoke *et al.*, 2004, Alcaraz *et al.*, 2010, de Hoon *et al.*, 2010, Galperin *et al.*, 2012). In these studies, homologues have been identified by carrying out either a bidirectional BLAST of protein sequences or using a Clusters of Orthologous Groups (COG) approach. In the case of a bidirectional BLAST, one sequence from a reference genome is used as the “query” and this is searched against a database containing the “subject” sequences (Altschul *et al.*, 1990, Wheeler & Bhagwat, 2007). The best match is then taken and searched against the reference. The algorithm can be adjusted by modifying parameters such as the coverage, percentage identity and E (Expect) value. The COG database was generated to assign all known proteins in the complete genomes of microorganisms (including prokayrotes, eukaryotes and archea) to orthologous groups (Tatusov *et al.*, 2000). At present there are over 4,500 COGs, each with a description of their function (Galperin *et al.*, 2015). The process of assigning a protein to a COG involves multiple PSI-BLASTS. The result is that a particular COG generally contains proteins all with similar domains, but potentially different intervening sequences, presumably all evolved from the same ancestor and therefore have the same function. This present study used a multi-faceted approach to identify genomic similarities between *Geobacillus* spp. and other species well characterized for biofilm formation and sporulation, particularly of *B. subtilis*. In particular, to determine whether the genome sequences of the three dairy strains A1, P3 and D1 can provide new insights into how these strains go about the processes of biofilm formation and sporulation.

## 6.2 Materials and methods

### 6.2.1 Genome sequences

The genome sequences of 32 *Geobacillus* strains were analyzed as described in Chapter 4 (Table 4.1, Section 4.2.1), including the three dairy strains of *G. stearothermophilus* A1, P3 and D1. Four *Geobacillus* spp. genomes: *Geobacillus* sp. WSUCF1 (Accession ATCO01), *Geobacillus* sp. CAMR1 (Accession JHUR01), *Geobacillus* sp. CAMR5 (Accession JHUS01) and *G. kaustophilus* Gblys (Accession BASG01), were omitted from this study. This was because of the high number of contigs in the first genome, lack of source information for the second two strains and in the case of *G. kaustophilus* Gblys this was a variation of another strain (*G. kaustophilus* HTA426), which was altered by integration of a phage. All of the *Geobacillus* genomes analyzed were parsed in the same manner as described in Chapter 4 (Section 4.2.5) before any further analysis was carried out. The GBK files of all these genomes can be found on the CD[15]. The *Geobacillus* genomes and their corresponding species group (as described in Chapter 5) are described in Table 6.2 and the reference genomes used are described inTable 6.3.

[15] CD:GenomeSequences/

**Table 6.2. *Geobacillus* genomes used in this study.**

| Abbreviated name | Full name | Species group[a] |
|---|---|---|
| A1 | *G. stearothermophilus* A1 | 3 |
| P3 | *G. stearothermophilus* P3 | 3 |
| D1 | *G. stearothermophilus* D1 | 3 |
| Gs795 | *G. stearothermophilus* ATCC7953 | 3 |
| GkHTA | *G. kaustophilus* HTA426 | 1A |
| Gk102 | *G. kaustophilus* NBRC 102445[T] | 1A |
| Gs10 | *G. stearothermophilus* 10 (DSM 13240) | 1A |
| GtlCCB | *G. thermoleovorans* CCB_US3_UF5 | 1A |
| GtlB23 | *G. thermoleovorans* B23 | 1A |
| GeoM61 | *Geobacillus* sp. Y412MC61 | 1A |
| GeoM52 | *Geobacillus* sp. Y412MC52 | 1A |
| GeoC56 | *Geobacillus* sp. C56-T3 | 1A |
| GeoGHH | *Geobacillus* sp. GHH01 | 1A |
| GeoFW2 | *Geobacillus* sp. FW23 | 1A |
| GeoMAS | *Geobacillus* sp. MAS1 | 1A |
| GeoA8 | *Geobacillus* sp. A8 | 1A |
| GthcGS1 | *G. thermocatenulatus* GS-1 | 1B |
| GvPSS | *G. vulcani* PSS1 | 2 |
| GicG1w | *G. icigianus* G1w1[T] | 4 |
| GsuPSS | *G. subterraneus* PSS2 | 4 |
| GeoJF8 | *Geobacillus* sp. JF8 | 5 |
| GtdNG8 | *G. thermodenitrificans* NG80-2 | 6 |
| Gtd465 | *G. thermodenitrificans* DSM 465[T] | 6 |
| GeoG11 | *Geobacillus* sp. G11MC16 | 6 |
| GcCIC | *G. caldoxylosilyticus* CIC9 | 7 |
| Gc10776 | *G. caldoxylosilyticus* NBRC10776[T] | 7 |
| Geo362 | *Geobacillus* sp. NUB3621 | 8 |
| GeoWCH | *Geobacillus* sp. WCH70 | 9 |
| GtgC56 | *G. thermoglucosidasius* C56YS93 | 10 |
| GtgTNO | *G. thermoglucosidasius* TNO-09 | 10 |
| Gtg107 | *G. thermoglucosidasius* NBRC 107763[T] | 10 |
| GeoM1 | *Geobacillus* sp. Y4.1MC1 | 10 |

[a] As defined in Chapter 5

**Table 6.3. Reference genomes used in this study.**

| Abbreviated name | Full name | Process analyzed | Reference | Source | GenBank accession no. |
|---|---|---|---|---|---|
| **Bs168** | *B. subtilis* 168 | Initiation, biofilm and sporulation | Barbe *et al.* (2009) | | NC_000964 |
| **Bc145** | *B. cereus* ATCC 14579 | Biofilm | Ivanova *et al.* (2003) | | NC_004722 |
| **Bc905** | *B. cereus* 905 | Biofilm | Gao *et al.* (2015b) | Wheat rhizosphere | KP076269 – KP076277 |
| **BanAme** | *Bacillus anthracis* Ames | Exosporium | Read *et al.* (2003) | Dead cow | NC_003997 |
| **Sau** | *Staphylococcus aureus* NCTC 8325 | Biofilm | Gillaspy *et al.* (2006) | | NC_007795 |
| **Sth** | *Streptococcus thermophilus* JIM8232 | Biofilm | Unpublished | Milk | NC_017581 |

### 6.2.2 Gene lists

The scope of this study included those genes involved with biofilm formation, sporulation and initiation of both of these processes via the phosphorelay. It did not include genes expressed during stationary phase under the control of the stationary phase sigma factor SigE, unless they were essential for sporulation. In addition, those genes involved with germination were not included unless they had an additional role in sporulation. The phosphorelay genes were compiled based on the list of genes in the phosphorelay category (http://subtiwiki.uni-goettingen.de/wiki/index.php/Phosphorelay) of *Subti*Wiki (Michna *et al.*, 2014). A list of these genes and relevant references can be found on the CD[16]. The list of biofilm genes[17] was compiled based on publications by Cairns *et al.* (2014) and Vlamakis *et al.* (2013) for *B. subtilis*, Gao *et al.* (2015b) for *B. cereus*, O'Gara (2007) for *S. aureus* and Couvigny *et al.* (2015) for *S. thermophilus*.

The *B. subtilis* sporulation genes[18] were initially compiled based on the lists of genes described by de Hoon *et al.* (2010) and Stragier & Losick (1996). This list was added to with those genes which have been described as being a member of one of the sporulation specific sigma factor regulons: $\sigma^E$, $\sigma^F$, $\sigma^G$ and $\sigma^K$ (Kuwana *et al.*, 2002, Eichenberger *et al.*, 2003, Feucht *et al.*, 2003, Eichenberger *et al.*, 2004, Wang *et al.*, 2006) as well as additional genes encoding for coat proteins (McKenney & Eichenberger, 2012). Given the number of genes in each regulon differed between studies only those genes that were described in at least two studies were included in this study, unless they were a member of an operon that was already included.

Each sporulation gene was grouped into a functional category based on its function or putative function in *B. subtilis.* These categories were partially based on those described by previous studies (Eichenberger *et al.*, 2004, de Hoon *et al.*, 2010, Galperin *et al.*, 2012). A description of all of the genes analyzed can be found in each of the spreadsheets[2,3,4]. The descriptions for those genes from *B. subtilis* originate from SporeWeb (Eijlander *et al.*, 2014) or *Subti*Wiki (http://subtiwiki.uni-goettingen.de/ ).

[16] CD:Chpt6/InitiationGenes.xlsx
[17] CD:Chpt6/BiofilmGenes.xlsx
[18] CD:Chpt6/SporulationGenes.xlsx

*B. subtilis* does not contain an exosporium, so as described in Table 6.3, *B. anthracis* was used as the reference. The genes encoding for exosporium proteins of *B. anthracis* are described in Table 6.4.

**Table 6.4. Genes encoding for exosporium proteins of *B. anthracis.***

| Gene name | Locus tag | Reference |
|---|---|---|
| ***bclA*** | BA_1222 | Sylvestre *et al.* (2002) |
| ***bclB/exsH*** | BA_2450 | Thompson *et al.* (2012), Todd *et al.* (2003) |
| ***betA/bclF*** | BA_3550 | Thompson *et al.* (2011) |
| ***bxpA*** | BA_2162 | Moody *et al.* (2010) |
| ***exsB*** | BA_2045 | Todd *et al.* (2003) |
| ***exsC*** | BA_2894 | Todd *et al.* (2003) |
| ***exsD*** | BA_2617 | Todd *et al.* (2003) |
| ***exsE*** | BA_1786 | Todd *et al.* (2003) |
| ***exsFA/bxpB*** | BA_1237 | Sylvestre *et al.* (2005) |
| ***exsFB*** | BA_2477 | Sylvestre *et al.* (2005) |
| ***exsG*** | BA_2150 | Todd *et al.* (2003) |
| ***exsK*** | BA_2554 | Redmond *et al.* (2004), Severson *et al.* (2009) |
| ***exsM/bxpC*** | BA_2332 | Fazzini *et al.* (2010) |
| ***exsY*** | BA_1234 | Johnson *et al.* (2006) |

### 6.2.3 Identification of homologues

#### *BDBH algorithm*

Homologues were initially identified using the BDBH algorithm in the program GET_HOMOLOGUES (v. 2014112) using the default E-value of 1 x $10^{-5}$ and coverage of 10% (Contreras-Moreira & Vinuesa, 2013). Some exceptions were made as described below. The GBK files generated by Prokka (as described in Chapter 4, Section 4.2.5) were used as the input, and the GenBank full file, which was downloaded from the NCBI website, was used for each of the reference genomes (described in Table 6.3), with the exception of *B. cereus* 905, as this has not had its genome sequenced. In this case the GBK file of each gene of interest was concatenated with the genome sequence of *B. cereus* ATCC 14579 in Geneious (v. 9.0.2) and exported as a GBK file, before running the GET_HOMOLOGUES program. A separate run was carried out for each reference genome. GET_HOMOLOGUES uses the protein sequences in the GBK file for identifying protein clusters with the BDBH algorithm. Refer to the CD[19] for the file output containing the location coordinates of each homologue. Exceptions to using this method were used in three situations: (1) where there was no protein sequence available in the reference (e.g. with split genes such as *spoIVCB* and *spoIIIC* which make up *sigK* in *B. subtilis* 168) (2) there was no annotated gene in the input file as was the case with *spoVM* or (3) there

[19] CD:Chapter6/BDBHHomologueCoordinates

were multiple gene copies. In the first two situations a TBLASTX database was prepared as described in Chapter 4, Section 4.2.7 and Appendix 1, example A1.2, using the *B. subtilis* genes (*spoIVCB*, *spoIIIC* and *spoVM*) as the reference. The BDBH algorithm includes inparalogues by default; therefore allowing multiple copies per genome. However, in the third situation this only appeared to be the case when multiple copies had identical E-values. In some cases the *Geobacillus* genomes had two copies of a homologue (including *spoVAC*, *spoVAD*, *spoVAEB*), but only the copy with the higher E value was identified as a homologue by the BDBH algorithm. In these cases the best match was confirmed for the second copy by carrying out a BLASTX.

Two additional approaches were taken to pick up any potential homologues that may have been missed using the BDBH algorithm as described below.

#### ***Clusters of Orthologous Groups (COG)***

As described in Section 4.2.7 the COGnitor software tool was used to assign predicted proteins in the *Geobacillus* genomes as well as *B. subtilis* 168 to pre-existing COGs. Where homologues of the *B. subtilis* genes analyzed were absent or variable (as defined by the BDBH algorithm) across the *Geobacillus* genus the COGnitor output[20] was used to identify other proteins in the same COG as the *B. subtilis* protein that could potentially carry out the same function as the “missing” homologue. The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015).

#### ***Prokka annotations***

This approach was used for identifying additional genes encoding for kinases (KinA, KinD, KinE), Rap, RsfA, SpoVAC, SpoVAD, SpoVAEB, SASPs (Small Acid-Soluable Proteins) as well as genes potentially involved in sporulation of *Geobacillus*, but with no homologues in *B. subtilis*. As described in Chapter 4 (Section 4.2.5), all of the *Geobacillus* genomes were annotated using Prokka as part of an in-house script. One of the outputs of this script was the MySQL table “Geobacillus_annotation” (as described in Section 4.2.5 and Figure 4.3). The “infoFact” column of this table, which contained descriptions of the gene product, was queried for the terms “%sporulat%”, “%spore”, “cot%”, “ssp%” as well as the protein names listed above.

[20] CD:Chapter4/Results/COG/COG_LocusTag.xlsx

#### *Heat-map generation*

Heat-maps were generated to visualize the number of homologues for a particular gene of *B. subtilis* as defined by the BDBH algorithm as well the number of proteins in a COG using the `heatmap.2` function included in the `gplots` library of the statistics software package R (v3.2.0), visualized in Rstudio (v0.98.1103). For each set of genes, a heat-map was originally generated using the BDBH data only to determine the order of strains (Appendix 7). In the final heat-map, which included both number of homologues and number of proteins in a COG, the dendrogram function was turned off.

### 6.2.4 Catergorization of histidine kinases

The HKs were categorized into different types based on their domain organization, which was determined using Pfam (Finn *et al.*, 2014) with the predicted amino acid sequences. Transmembrane (TM) regions were determined using the TM pred program (http://www.ch.embnet.org/software/TMPRED_form.html). Domain organization diagrams were drawn to scale using the program IBS (Illustrator of Biological Sequences, v. 1.0) (Liu *et al.*, 2015).

To determine if the *Geobacillus* HKs belonged to the same class of HK (orphan kinase class IIIB) as the *B. subtilis* KinA, KinB, KinC, KinD and KinE, the 14 amino acid sequence surrounding the histidine residue (the phosphorylation target) was compared as described by Fabret *et al.* (1999). The sequences were aligned using the default settings of ClustalW in Geneious (v. 9.0.2) imported into MEGA 6.06 (Tamura *et al.*, 2011) and a phylogenetic tree was generated using the maximum likelihood method with a boostrap value of 1000.

## 6.3 Results and discussion

*G. stearothermophilus* is one of the major contaminants of milk powder. Two characteristics that play an important role in its ability to colonise and survive the harsh dairy manufacturing environment are biofilm formation and sporulation. These characteristics are presumably also important for the growth and survival of *Geobacillus* spp. in a variety of environments. The genome analysis carried out in this study provides a basis for understanding the potential molecular mechanisms used by this genus for both biofilm formation and sporulation.

The model organism, *B. subtilis,* is the closest relative to *Geobacillus* in which both biofilm formation and sporulation has been well characterised. When the model organism *B. subtilis* is faced with starvation a cell can respond in different ways via a master regulator to ensure survival; this results in a population containing different cell types (Lopez & Kolter, 2010). The phosphorylation of one such master regulator, Spo0A, controls the activation of the matrix production and sporulation pathways. In this study a three-part approach was taken to determine whether the genes involved in these processes: activation of the phosphorelay, biofilm formation and sporulation are conserved in *Geobacillus* spp. i) Homologues of *B. subtilis* genes involved in biofilm formation and sporulation were identified by BLASTP using the BDBH algorithm in the program GET_HOMOLOGUES, ii) for those genes that were variable or absent

across the *Geobacillus* genus COG functional grouping was used for classification; and iii) annotations generated by Prokka were analyzed.

### 6.3.1 Activation of the phosphorelay

A key question behind this present study was do *Geobacillus* spp. initiate biofilm formation and sporulation in the same way as *B. subtilis*? In *B. subtilis* the phosphorelay is activated via multiple sensor HKs. The key components of the phosphorelay were conserved in the *Geobacillus* genus (Figure 6.3 and CD[21,22]). The *Geobacillus* genomes contained multiple sensor HKs, the two phosphotransferases Spo0F and Spo0B as well as the phosphorylation target Spo0A. Previous studies studies have also shown that these key components are conserved across the *Bacillus* genus as well as other members of the Bacillaceae family, including *Geobacillus* (Onyenwoke *et al.*, 2004, de Hoon *et al.*, 2010, Galperin *et al.*, 2012). This differs from the other main group of spore formers, the *Clostridiaceae* family, which only contain the HKs and the master regulator Spo0A (Paredes *et al.*, 2005, Galperin *et al.*, 2012, Pettit *et al.*, 2014, Al-Hinai *et al.*, 2015).

Despite the core components being similar there are differences between *B. subtilis* and the *Geobacillus* genus in the kinases and control of the phosphorelay, which may indicate that the pathways activated by Spo0A are different between these two groups of bacteria. Those products involved in controlling the phosphorelay through the levels of phosphorylated Spo0F and Spo0A were less well conserved. For those genes that were variable (*kinA*, *kinB* and *yisI*) or absent (*kinC*, *kinD*, *rapA*, *rapB* and *rapE*) COG functional grouping identified other proteins with similar functions, which could be possible homologues (Figure 6.3 and Table 6.5). The COG groups associated with each of those proteins encoded for by these genes are described in Table 6.5. As will be discussed in Section 6.3.5 these COG groups could not be relied on for the identification of homologues.

[21] CD:Chapter6/InitiationGenes.xlsx
[22] CD:Chapter6/BDBHHomologueCoordinates/InitiationHomologues/

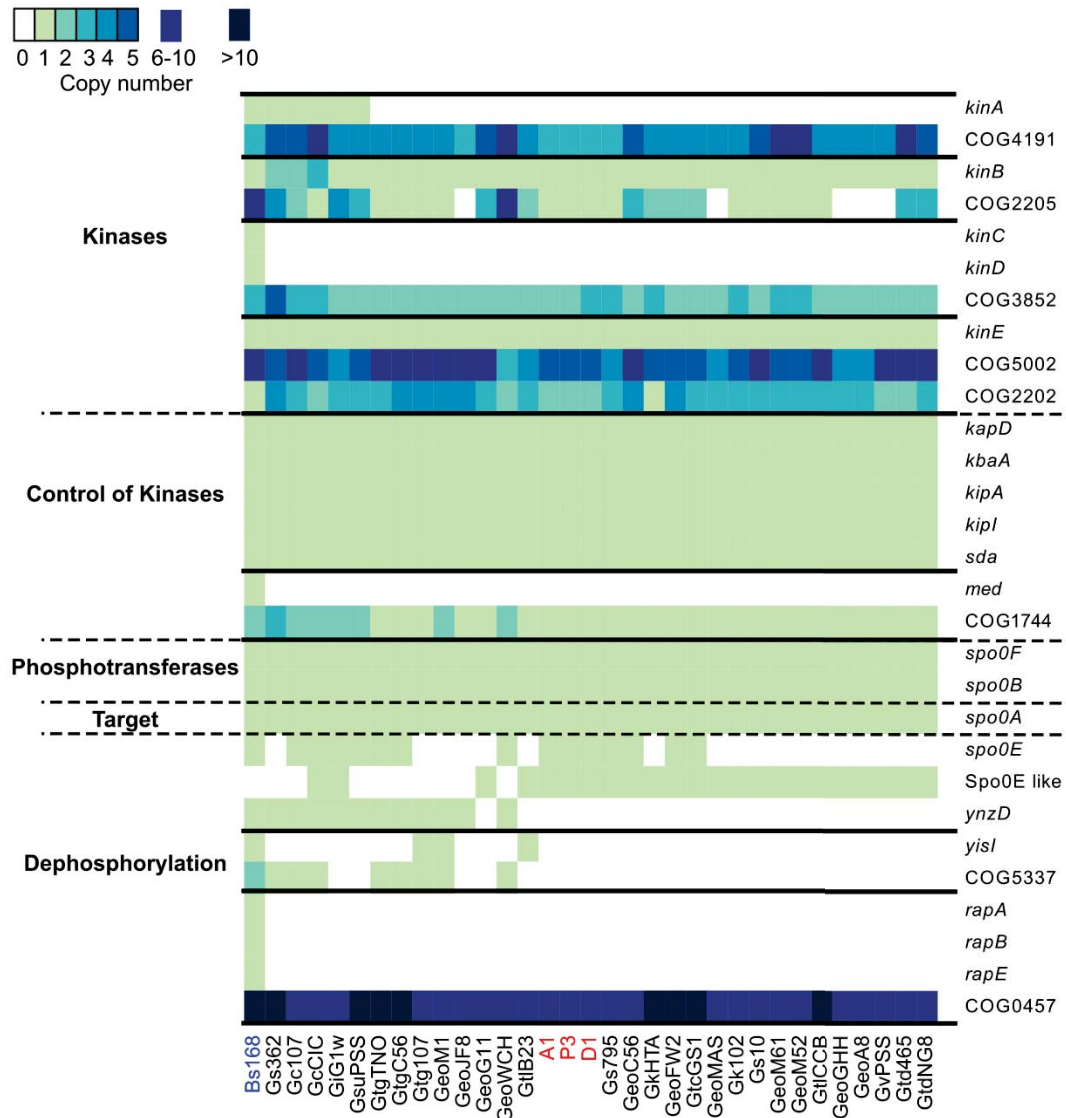



**Figure 6.3. Conservation of those genes involved in the phosphorelay and their equivalent COGs in *Geobacillus* spp.** This heat map shows 20 genes involved in the phosphorelay of *B. subtilis* and the presence/absence of homologues in the *Geobacillus* genus. Genes were categorized according to their role in the phosphorelay. Each category is separated by a dotted line. The presence of a homologue was determined by using the BDBH algorithm in the program GET_HOMOLOGIES using a cut-off of 10% for the coverage and an E value of $1x10^{-5}$. The COG associated with each gene(s) is in the row below and is separated from the next gene and its associated COG by a black line. The COG analysis was only included for those genes that were variable or absent in *Geobacillus* spp. as defined by the BDBH algorithm. COG5002 contained the *B. subtilis* KinE, and COG2202 contained the *Geobacillus* KinE homologues. No homologues (as defined by the BDBH algorithm) were found for the *B. subtilis* genes: *phrA, phrC, phrE* and *kapB* . The following proteins did not belong to a COG: PhrA, PhrC, PhrE, Spo0E, YnzD and KapB.

**Table 6.5. Description of each COG, which contained a protein involved in the *B. subtilis* phosphorelay.**

| COG[a, b] | Description[c] | Gene names | |
|---|---|---|---|
| | | Phosphorelay | Other[d] |
| **COG4191** | Signal transduction histidine kinase regulating C4-dicarboxylate transport system | *kinA* | *glnJ, ynzK* |
| **COG2205** | Potassium ion (K+) sensing histidine kinase KdpD | *kinB* | *bceS, cssS, ycbM, yigA, yvrG, yxdK* |
| **COG3852** | Signal transduction histidine kinase, nitrogen specific | *kinC, kinD* | *ypjB* |
| **COG5002** | Signal transduction histidine kinase | *kinE* | *resE, phoR, yvcQ, ywdI, walK* |
| **COG2202**[e] | PAS domain | n/a | *yxiF* |
| **COG1744** | Basic membrane lipoprotein Med, periplasmic binding protein (PBP1-ABC) superfamily | *med* | *yufN* |
| **COG0457** | Tetratricopeptide (TPR) repeat | *rapA, rapB, rapE, rapH* | *glpG, gutR, rapC, rapD, rapF, rapG, rapI, rapJ, rapK, yotC, ypiA, yrrB, ysoA, yvcD* |
| **COG5337** | Spore coat protein CotH | *yisI* | *cotH* |

[a] The following proteins did not belong to a COG: PhrA, PhrC, PhrE, Spo0E, YnzD, KapB

[b] Refer to CD: Chapter4/Results/COG/COG_LocusTag.xlsx for a list of locus tags within each COG

[c] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

[d] These genes listed are not known to be involved in the phosphorelay but they belong to the same COG.

[e] COG2202 contained the *Geobacillus* KinE homologues.

### *Histidine kinases*

In *B. subtilis* five sensor HKs (KinA, KinB, KinC, KinD and KinE) have the capability of transferring a phosphoryl group into the phosphorelay (Jiang *et al.*, 2000b). However, the role of KinE *in vivo* has not been established. Of the five *B. subtilis* HKs, KinE showed the most similarity to its homologues in *Geobacillus*. In addition, at least one homologue of KinB was found in each of the *Geobacillus* genomes (Figure 6.3). A homologue of KinA was only found in five of the *Geobacillus* genomes and no KinC or KinD homologues were found in any of the genomes. This present study found that the KinA and KinB homologues in *Geobacillus* belonged to the same COG as the *B. subtilis* KinA and KinB (COG4191 and COG2205 respectively). However, the KinE of *B. subtilis* and the putative *Geobacillus* KinE belonged to separate COGs (COG5002 and COG2202 respectively). KinC and KinD of *B. subtilis* belonged to the same COG (COG3852).

Using a combination of the Prokka annotations and the COG approach an additional 2 – 4 puative *kin* open reading frames (ORFs) were identified. This resulted in seven different types of putative HKs being identified across the Geobacillus genus and like the *B. subtilis kinA, kinB, kinC* and *kinD* genes these additional *kin* ORFs all appeared to be orphan genes as they were not associated with a response regulator. This differed from a previous study that found five different types of orphan HKs in one strain of *G. stearothermophilus* (Stephenson & Hoch, 2002). However, the study by Stephenson & Hoch (2002) did not state what strain they analyzed; therefore it is difficult to make comparisions between their study and this present study. Another member of the *Bacillaceae* family, *B. anthracis*, also has a larger number (compared with *B. subtilis*) of putative orphan HKs, with nine being identified through a bioinformatics approach (Brunsing *et al.*, 2005).

As an additional check, the domain organization of the putative HKs in seven of the *Geobacillus* genomes (A1, D1, GcCIC, Gs362, GtgTNO, GtdNG8, GkHTA and GsuPSS) representing six of the main *Geobacillus* species groups were compared with KinA – KinE of *B. subtilis.* As expected of a HK, they are composed of a sensor component (at the N-terminal end) and an autokinase, which transfers the phosphoryl group (at the C-terminal end). Seven distinct types were were identified in *Geobacillus* based on the sensor component of each HK (Figure 6.4). Three of these (Type 1, 2 and 3) were identified as homologues of *B. subtilis* KinA, KinB and KinE using the BDBH algorithm. The type 3 was most similar to its homologue KinE based on its amino acid similarity and the domain organization, containing four Per-Arnt-Sim[23] (PAS) domains. Type 2 was also quite similar to its homologue KinB containing multiple transmembrane domains. However, Type 1 was a transmembrane HK as opposed to *B. subtilis* KinA, which is cytoplasmic. Analysis of the signal sensing domains did not provide any information on the type signal that might be received by *Geobacillus* spp. in order to trigger

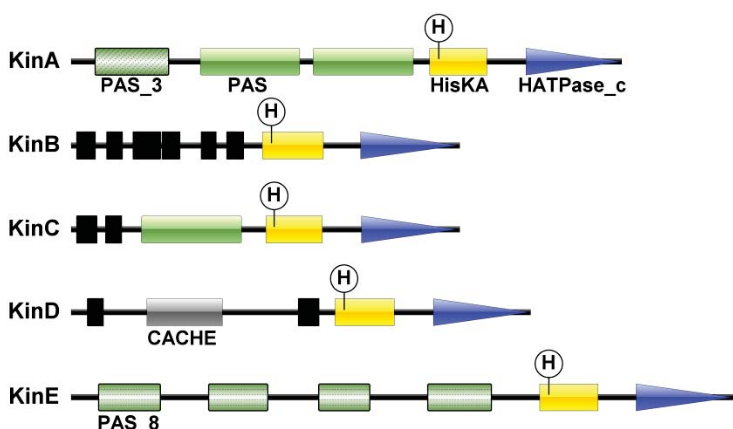
---

[23] The naming of the Per-Arnt-Sim domain originates from this domain being first identified in the three proteins: PERIOD, a circadian protein found in *Drosphila*; the mammalian aryl hydrocarbon nuclear translocator (ARNT) and the single-minded protein (SIM), a developmental regulator found in *Drosphila* (Huang *et al.*, 1993, Henry & Crosson, 2011).

biofilm formation or sporulation. This is not surprising given the sensor domains, such as the PAS, CACHE and HAMP domains are are commonly found in signalling proteins (Aravind & Ponting, 1999, Anantharaman & Aravind, 2000, Anantharaman *et al.*, 2001, Galperin, 2004). Some mystery still surrounds the role of the *B. subtilis* orphan HKs. It has been established that KinA is the major sporulation kinase and to a lesser extent KinB and KinC. However, more understanding is still needed around the mechanisms required to trigger downstream processes via these kinases, particularly sporulation. The signals which trigger matrix production are better understood, for example, oxygen depletion, changes in osmolarity as well as chemical signals such as sugars, free fatty acids and metal ions have all been shown to trigger matrix production via KinA, KinB, KinC or KinD (Lopez *et al.*, 2009c, Aguilar *et al.*, 2010, McLoon *et al.*, 2011a, Chen *et al.*, 2012, Beauregard *et al.*, 2013, Kolodkin-Gal *et al.*, 2013, Shemesh & Chai, 2013, Pasvolsky *et al.*, 2014, Grau *et al.*, 2015).

Over 35 sensor HKs are found in *B. subtilis*, all of which are composed of a sensor region and an autokinase (Fabret *et al.*, 1999). The *B. subtilis* KinA, KinB, KinC, KinD and KinE can be differentiated from other sensor HKs by the 13 amino acid sequence surrounding the phosphorylated histidine residue as well as not being associated with a response regulator on the chromosome. The identified *Geobacillus* HKs also met these two criteria; therefore there is likely to be some similarity in the role that they play. This short 13 amino acid sequence was analyzed in two of the *Geobacillus* species, A1 and *Geobacillus* sp. NUB3621 (Figure 6.5). These sequences all grouped with the *B. subtilis* KinA – E sequences. Sequences from PhoR (another kinase) were used as the out-group to confirm that the *Geobacillus* sequences clustered within the same group as the *B. subtilis* orphan sensor kinases and not the other kinases as described by Fabret *et al.* (1999).

**A**




**B**

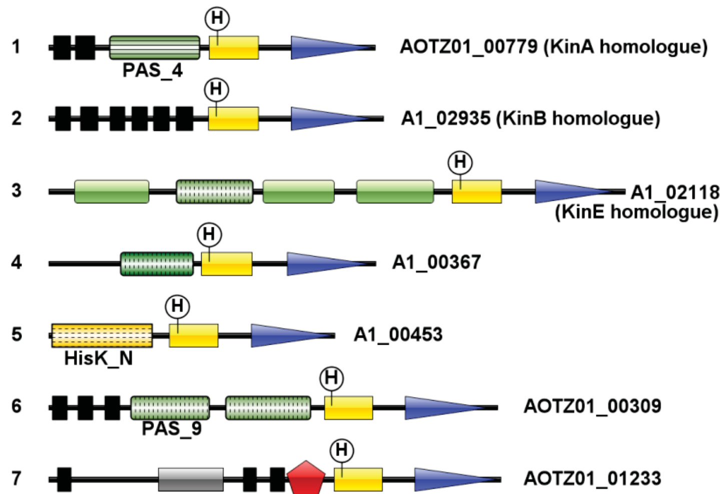



**Figure 6.4. Domain organization of the orphan HKs from (A) *B. subtilis* and (B) *Geobacillus* strains A1 (Types 2, 3, 4 and 5) and NUB3621 (Type 1, 6 and 7).**

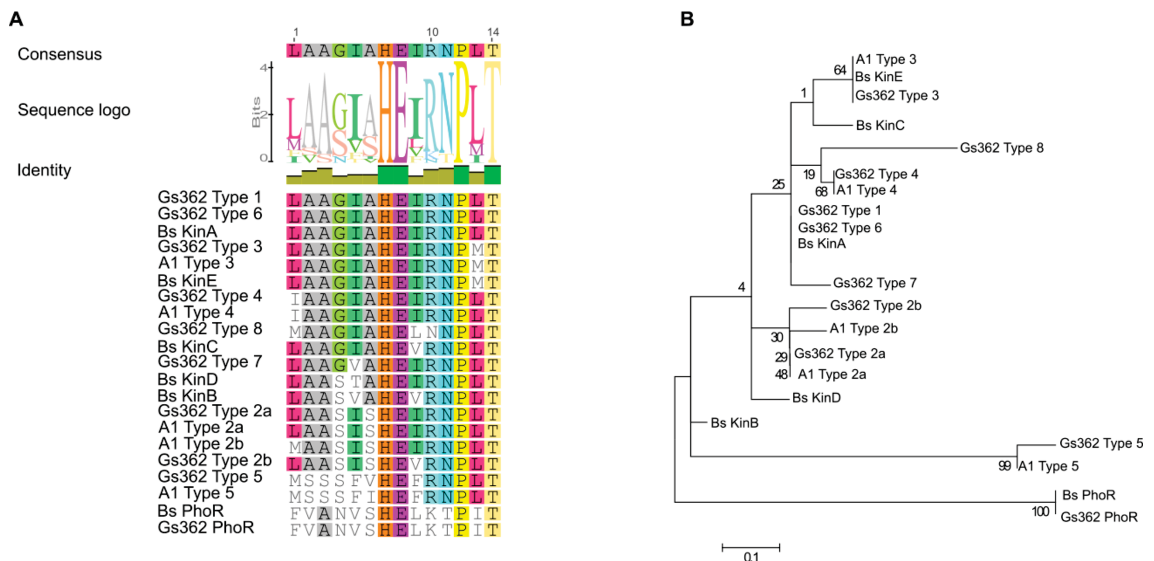



**Figure 6.5. Alignment of the sequence surrounding the phosphorylated histidine of orphan HKs from *B. subtilis* 168, *G. stearothermophilus* A1 and *Geobacillus* sp. NUB3621.**

Do orphan kinases trigger sporulation in *Geobacillus*? Given that orphan HKs are conserved across all endspore forming bacteria and have been shown to initiate sporulation in both *Bacillus* spp. and *Clostridium* spp., at least some of the orphan HKs are likely to be involved in triggering sporulation in *Geobacillus*. A novel HK was identified in *B. anthracis* that is not found in *B. subtilis* and it was shown to be both an inducer and inhibitor of sporulation (Brunsing *et al.*, 2005, Scaramozzino *et al.*, 2009). This HK is homologuous with one of the seven types of HKs found in the *Geobacillus* spp. analysed. In *B. subtilis* KinA is the main kinase responsible for sporulation (Fujita & Losick, 2005) and is a cytoplasmic kinase as opposed to KinB-D, which contains transmembrane regions. Although, some KinA homologues were found (as defined by

the BDBH algorithim) in *Geobacillus* spp. they were not cytoplasmic (i.e. they contained transmembrane regions), they only contained one PAS domain as opposed to three in *B. subtilis* and they did not neighbour *patA* and *pbhA*; therefore, they are unlikely to be true homologues of KinA (i.e. derived from the same ancestor).

In the absence of KinA, KinB is able to initiate sporulation (Trach & Hoch, 1993). All of the *Geobacillus* genomes contained at least one KinB homologue. Although they do not show synteny to the *B. subtilis kinB* as they do not neighbour *patB* and *kapB*, these homologues were more similar (as opposed to the KinA homologues) to KinB based on amino acid similarity and the type of domains each contained. In addition, the structure of one of these KinB homologues complexed to the kinase inhibitator Sda has been determined from the *G. stearothermophilus* strain 10, and like in *B. subtilis*, they showed that Sda is an inhibitor of the *G. stearothermohilus* KinB (Bick *et al.*, 2009). In the same study it was shown that the *G. stearothermophilus* Sda protein also inhibited phosphotransfer from KinB to Spo0F, highlightling another difference between the *B. subtilis* and *Geobacillus* phosphorelay.

KinC is also involved in sporulation, under certain genetic and growth conditions a *kinC* mutant can cause minor defects in sporulation, but it is not able to initiate sporulation when KinA and KinB are absent (Kobayashi *et al.*, 1995, Ledeaux & Grossman, 1995, Ledeaux *et al.*, 1995, Devi *et al.*, 2015). No homologues of KinC were found in the *Geobacillus* genomes, although, the KinA homologues also showed some similarity to KinC.

The role, if any, of the putative *Geobacillus* HKs in biofilm formation is less clear, based on this bioinformatics analysis. Both KinC and KinD are involved in triggering matrix production in B. subtilis. However, neither of these HKs have homologues in *B. subtilis* as defined by the BDBH algorithm. Under biofilm conditions KinC triggers matrix production by sensing disruptions to the membrane, although the mechanisms of disruption to the membrane and suggested signal are different (Lopez *et al.*, 2009c, Shemesh *et al.*, 2010). In KinD the CACHE domain has been shown to be responsible for sensing small molecules triggering matrix production (Chen *et al.*, 2012, Shemesh & Chai, 2013). The only *Geobacillus* HK which contained a CACHE domain was found in *Geobacillus* sp. 3621 and *G. caldoxylosilyticus,* although it was not a homologue of KinD and also contained a HAMP domain in the sensor region, both domains are commonly found in signalling proteins. It has also been suggested that KinD acts as both a phosphatase and a kinase and in its role as a phosphatase it inhibits the onset of sporulation under biofilm forming conditions (Aguilar *et al.*, 2010). Given there are no KinD homologues in *Geobacillus*, it is likely that regulatory switch between sporulation and biofilm formation is different compared with *B. subtilis*. Further evidence that no equivalent of KinD was present in *Geobacillus* was the lack of any gene encoding the lipoprotein Med, which is a KinD specific regulatory factor that is required for SpoA phosphorylation (Banse *et al.*, 2011). Other regulatory factors that act on KinA or KinB in an inhibitory manner (Sda and KipI) or as an effector (KapB and KbaA on KinB) had homologues in *Geobacillus*.

Are other developmental processes initiated via the *Geobacillus* kinases? Orphan HKs have been shown to be involved in other processes in other bacteria. In *B. subtilis* cannabalism is triggered via KinC (Lopez *et al.*, 2009b) and KinB has also been shown to trigger sliding through a K+ selective sequence (Grau *et al.*, 2015). In some *Clostridium* spp. genes involved in virulence and metabolism have been shown to be under the control of Spo0A; therefore, it is likely that this occurs via one of the orphan HKs (Paredes *et al.*, 2005, Pettit *et al.*, 2014). It is possible that some of the putative HKs found in *Geobacillus* could initiate other developmental processes, in addition to sporulation.

#### *Control of the Phosphorelay*

In *B. subtilis* the phosphorelay is tightly controlled. One way is through dephosphorylation of Spo0F~P and Spo0A~P. In *B. subtilis* Spo0F~P is dephosphorylated by associating with some of the Rap proteins. Galperin *et al.* (2012) proposed that it is Spo0F which desphosphorylates itself and this is activated by binding of the Rap protein to Spo0F. The authors of this same study state that because the Rap proteins are only composed of TPR domains they do not contain any enzymatic properties. *B. subtilis* contains eleven *rap* genes (Rap A – K), nine of which are associated with a peptide inhibitor (*phr* gene). RapA, B, E, H, and J all have the capability to cause dephosphorylation of Spo0F~P (Perego *et al.*, 1996, Jiang *et al.*, 2000a, Mirouze *et al.*, 2011, Parashar *et al.*, 2011, Diaz *et al.*, 2012). Most of the other Rap proteins play a role in regulating the other master regulators ComA and DegU, which are involved in activating pathways for other cell types in *B. subtilis* (Bongiorni *et al.*, 2005, Lopez *et al.*, 2009a, Murray *et al.*, 2009, Perego, 2013). There were no homologues of the Rap proteins in *Geobacillus* as defined by the BDBH algorithm. However, they did contain ORFs encoding proteins containing only TPR domains, which belonged to the same COG as the *B. subtilis* Rap proteins. None of these TPR containing genes were associated with peptide inhibitors, which suggests that *Geobacillus* may not have the same control over Spo0A phosphorylation as *B. subtilis*.

The dephosphorylation of Spo0A~P is controlled by a group of aspartyl-phosphate phosphatases. Spo0E is the most well characterized of these proteins. Although not all of the *Geobacillus* genomes have a homologue of Spo0E, most have at least one of the aspartyl-phosphate phosphatases Spo0E, YnzD or YisI. Furthermore, 20 of the genomes, including the three dairy strains A1, P3 and D1, have an additional ORF annotated as “Spo0E like sporulation regulatory protein”, which forms a separate homologous group. One genome *G. thermodenitrificans* NG80-2 did not contain any ORF encoding a Spo0E related phosphatase, instead there was a transposase in its place.

### 6.3.2 Biofilm formation

To determine what genes might be involved in biofilm formation of *Geobacillus* spp. the BDBH algorithm from GET_HOMOLOGUES was used to identify homologues of genes involved in biofilm formation from *B. subtilis*. Members of the *Geobacillus* genus contain some homologues

of genes required for biofilm formation in *B. subtilis*, such as as those that are involved in regulation and exopolysaccharide production (Figure 6.6, Table 6.6 and CD[24, 25]).

Many of the regulators identified are also involved in other cell survival responses such as competence, motility and sporulation. For example, homologues of *ymdB*, *degU* and *degS* were found in all of the *Geobacillus* spp. YmdB is a phosphodiesterase that controls the expression of over 800 genes, including genes involved in biofilm formation and sporulation (Diethmaier *et al.*, 2014). DegU and DegS are a two component system response regulator and sensor HK respectively, involved in the regulation of multiple cell types in *B. subtilis* and is conserved in many bacilli (Murray *et al.*, 2009). Similarly to Spo0A~P the level of DegU~P initiates the transcription of a subset of genes. For example, lower levels of DegU~P are involved with initiation of motility and higher levels are involved with initiating expression of *pgsB*, *bslA* and *yvcA*, which are involved in producing the extracellular matrix in some strains of *B. subtilis* (Stanley & Lazazzera, 2005, Murray *et al.*, 2009, Verhamme *et al.*, 2009).

The *eps* operon in *B. subtilis* is composed of 15 genes and is the main operon involved in producing the exopolysaccharide component of the biofilm matrix. Putative *eps* clusters were located in some of the *Geobacillus* genomes, two examples can be seen in *G. thermoleovorans* CCB and *Geobacillus* C56-T3 (Figure 6.7). This large cluster of putative *eps* genes was not located in the genomes of the dairy strains A1, P3 or D1. The putative *eps* cluster found in some of the *Geobacillus* genomes is also similar to a cluster of genes found in *B. cereus* (Figure 6.7 and Table 6.7). It has been shown that this putative *eps* cluster in *B. cereus* is not involved with matrix production in a pellicle or within a submerged biofilm (Gao *et al.*, 2015b). . Exopolysaccharide is found in a number of different bacterial structures such as the capsule, S-layer in the cell envelope and spore coat and cell envelope (Wang *et al.*, 2014, Geno *et al.*, 2015, Liszewski Zilla *et al.*, 2015, Zilla *et al.*, 2015). In addition, *lytR* which is clustered with these *eps* homologues in both *B. cereus* and *Geobacillus* belongs to a family of enzymes termed LCP proteins that are involved with attaching secondary cell wall polysaccharide to peptidoglycans in cell walls in *B. anthracis*; therefore it seems that those genes from the putative *eps* clusters of *Geobacillus* are likely to have another another role other than exopolysaccharide production in a biofilm. If this putative *eps* cluster is involved with exopolysaccharide production for another process it can be seen that this process probably varies between strains of *Geobacillus* spp. or in some cases may not occur. Another factor pointing towards a different role of this *eps* cluster is that the most well characterised *eps* gene in *B. subtilis, epsE*, only has homologues in some of the *Geobacillus* genomes. Where it is present it is found in a different region of the genome to the putative *eps* cluster. In *B. subtilis*, EpsE has a dual role, in that it acts a switch between motility and biofilm formation as well as functioning as an enzyme to produce sugar involved in the formation of the exopolysaccharide (Guttenplan *et al.*, 2010).

---

[24] CD:Chapter6/BiofilmGenes.xlsx

[25] CD:Chapter6/BDBHHomologueCoordinates/BiofilmHomologues/Bsubtilis/

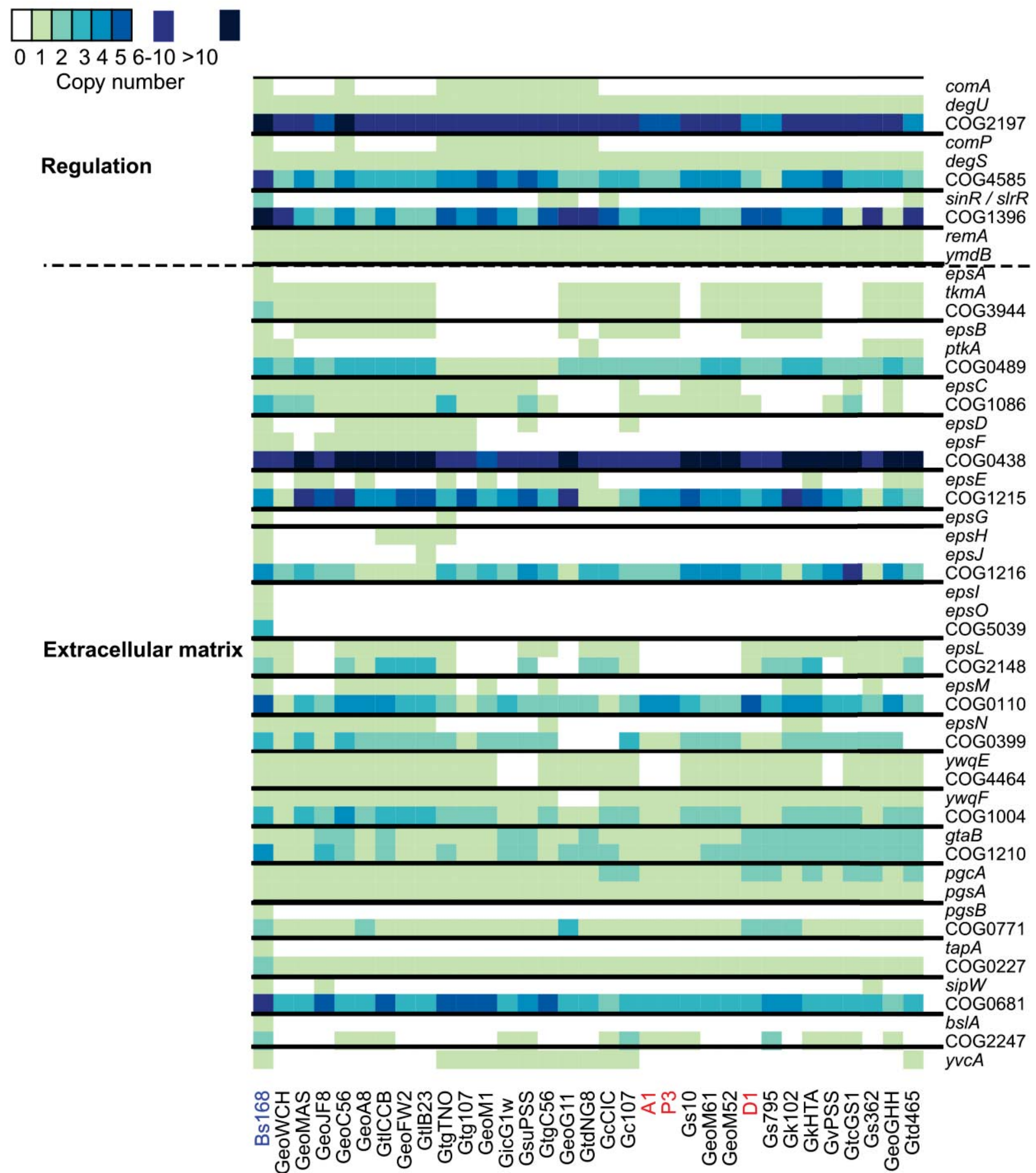



**Figure 6.6. Conservation of selected *B. subtilis* biofilm genes in *Geobacillus* spp. and their equivalent COGs.** This heat map shows 33 genes involved in biofilm formation of *B. subtilis* and the presence/absence of homologues in the *Geobacillus* genus. The presence of a homologue was determined by using the BDBH algorithm in the program GET_HOMOLOGIES using a cut-off of 10% for the coverage and an E value of $1x10^{-5}$. The COG associated with each gene(s) is in the row below and is separated from the next gene and its associated COG by a black line. The COG analysis was only included for those genes that were variable or absent in *Geobacillus* spp. as defined by the BDBH algorithm. No homologues (as defined by the BDBH algorithm) were found for the *B. subtilis* genes: *pgsB*, *pgsC*, *degQ, epsA, remB, sinI, slrA, tapA*, and *tasA*. The following proteins did not belong to a COG: PgsC, DegQ, TasA, RemB, and EpsG.

**Table 6.6. Description of each COG, which contained a protein involved in biofilm formation of *B. subtilis*.**

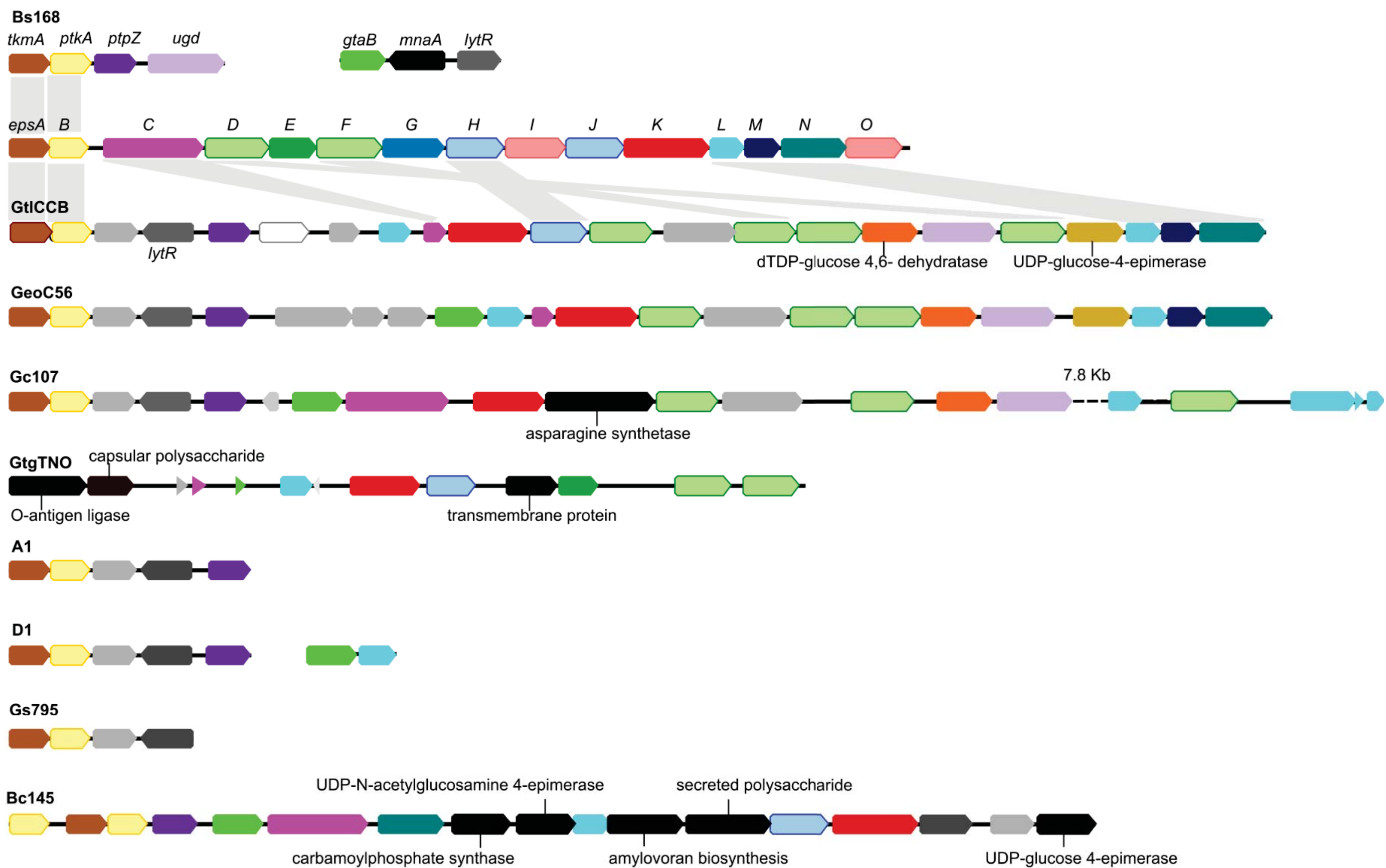
| COG[a,b] | Description | *B. subtilis* gene names | |
|---|---|---|---|
| | | **Biofilm** | **Other[c,d]** |
| **COG2197** | DNA-binding response regulator, NarL/FixJ family, contains REC and HTH domains | *comA, degU* | *gerE, desR, liaR, ydfI, yfiK, yhcZ, ylxL, yosQ, yvfU, yxjL* |
| **COG4585** | Signal transduction histidine kinase | *comP, degS* | *desK, liaS, ydfH, yfiJ, yvfT, yxjM* |
| **COG1396** | Transcriptional regulator, contains XRE-family HTH domain | *sinI, sinR, slrR, slrA* | *ansR, immR, rghRA, rghRB, xre, yazB, yobD, yopO, yqaG* |
| **COG3944** | Capsular polysaccharide biosynthesis protein | *epsA, ywqC* | *No* |
| **COG0489** | Chromosome partitioning ATPase, Mrp family, contains Fe-S cluster | *epsB, ptkA* | *salA* |
| **COG1086** | NDP-sugar epimerase | *epsC* | *yodU, ypqP* |
| **COG0438** | Glycosyltransferase involved in cell wall bisynthesis | *epsD, epsF* | ***cotSA**, tagE, tuaH, ypjH, yqqM, ytcC* |
| **COG1215** | Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine | *epsE* | *ggaB, tuaG, ydaM* |
| **COG1216** | Glycosyltransferase, GT2 family | *epsH, epsJ* | *yfnE, ywdF* |
| **COG5039** | Exopolysaccharide biosynthesis protein EpsI, predicted pyruvyl transferase | *epsI, epsO* | *yxaB* |
| **COG2244** | Membrane protein involved in the export of O-antigen and teichoic acid | *epsK* | ***spoVB**, tuaB, ykvU, yabM, ytgP* |
| **COG2148** | Sugar transferase involved in LPS biosynthesis | *epsL* | *tuaA* |
| **COG0110** | Acetyltransferase | *epsM* | *maa, **tgl**, yjdH, yvoF* |
| **COG0399** | dTDP-4-amino-4,6-dideoxygalactose transaminase | *epsN* | *ntdA, **spsC*** |
| **COG4464** | Tyrosine-protein phosphatase YwqE | *ywqE* | *No* |
| **COG1004** | UDP-glucose 6-dehydrogenase | *ywqF* | *ytcA, tuaD* |
| **COG1210** | UTP-glucose-1-phosphate uridylyltransferase | *gtaB* | *yngB, ytdA, yvzE* |
| **COG2843** | Poly-gamma-glutamate biosynthesis protein CapA/YwtB (capsule formation) | *pgsA* | *No* |
| **COG0771** | UDP-N-acetylmuramoylalanine-D-glutamate ligase | *pgsB* | *murB* |
| **COG0227** | Ribosomal protein L28 | *tapA* | *rpmB* |
| **COG0681** | Signal peptidase I | *sipW* | *sipS, sipT, sipU, sipV, yozO* |
| **COG2247** | Putative cell wall-binding domain | *bslA* | *lytB* |

[a] The following proteins did not belong to a COG: PgsC, DegQ, TasA, RemB, EpsG, YvcA

[b] Refer to CD: Chapter4/Results/COG/COG_LocusTag.xlsx for a list of locus tags within each COG

[c] The gene names in bold are involved with sporulation

[d] These genes listed are not known to be involved in the phosphorelay but they belong to the same COG.



Bs168
tkmA
ptkA
ptpZ
ugd
gtaB
mnaA
lytR
epsA
B
C
D
E
F
G
H
I
J
K
L
M
N
O
GtICCB
lytR
dTDP-glucose 4,6- dehydratase
UDP-glucose-4-epimerase
GeoC56
Gc107
7.8 Kb
asparagine synthetase
GtgTNO
capsular polysaccharide
O-antigen ligase
transmembrane protein
A1
D1
Gs795
Bc145
UDP-N-acetylglucosamine 4-epimerase
secreted polysaccharide
carbamoylphosphate synthase
amylovoran biosynthesis
UDP-glucose 4-epimerase

**Figure 6.7. Gene organisation of biofilm genes in *B. subtilis* and putative exopolysaccharide genes in selected *Geobacillus* strains and *B. cereus* ATCC 14579.** Colours represent those genes which encode for proteins within the same COG group. Only those genes in the same region as other putative exopolysaccharide genes are shown. The three clusters of *B. subtilis* genes are all in different regions of the genome. The genes from the strain D1 were found on two separate contigs. Grey lines between *B. subtilis* and *G. thermoleovorans* CCB denote homologues as defined by the BDBH algorithm.

**Table 6.7. Putative *eps* cluster in *B. cereus* and their homologues in *B. subtilis* and presence in *Geobacillus* spp.**

| Bc905 gene | Bc145 locus tag | Description[a] | Bs168 homologue | Presence in *Geobacillus*[b] |
|---|---|---|---|---|
| **B905_5465, B905_5463** | BC5279, BC5277 | Tyrosine-protein kinase | *ptkA* | Variable |
| **B905_5464** | BC5278 | Capsular polysaccharide biosynthesis protein | *tmkA* | Variable |
| **B905_5462** | BC5276 | Tyrosine-protein phosphatase | *ywqE* | Variable |
| **B905_5461** | BC5275 | UTP-glucose-1-phosphate uridyltransferase | *gtaB* | Variable |
| **B905_5460** | BC5274 | Polysaccharide biosynthesis protein | *epsC* | Variable |
| **B905_5459** | BC5273 | Pyridoxal phosphate-dependent aminotransferase | *spsC* | Variable |
| **Unknown** | BC5272 | Carbamoylphosphate synthase small subunit | Absent | Variable |
| **Unknown** | BC5271 | UDP-N-acetylglucosamine 4-epimerase | Absent | Gc107 only |
| **B905_5458** | BC5270 | Sugar transferase | *epsL* | Variable |
| **B905_5457** | Absent | Acetyltransferase | *epsM* | Variable |
| **B905_5456** | Absent | Glycosyltransferase | *epsD* | Variable |
| **B905_5455** | Absent | Poly(glycerol-phosphate) alpha-glucosyltransferase | Absent | Gc107 only |
| **B905_5454** | Absent | Hypothetical | Absent | Gc107 only |
| **B905_5453** | Absent | Poly(glycerol-phosphate) alpha-glucosyltransferase | Absent | Variable |
| **B905_5452** | Absent | Hypothetical | Absent | Absent |
| **B905_5451** | Absent | Teichuronic acid biosynthesis | *tuaB* | Geo362 only |
| **B905_5450** | Absent | UDP-glucose 6-dehydrogenase | Absent | Variable |
| **B905_5449** | Absent | UDP-glucose epimerase | *ytcB* | Variable |
| **Unknown** | BC5267 | Glycosyltransferase | *epsJ* | Variable |
| **B905_5448** | BC5265 | Transcriptional regulator LytR | *lytR* | Variable |
| **B905_5447** | BC5264 | Hypothetical | Absent | Variable |
| **B905_5446 & B905_5445** | BC5263 | UDP-glucose 4-epimerase | Absent | Present |

[a] Taken from Gao *et al.* (2015b)

[b] For those genes listed as variable refer to the CD[26] for the *Geobacillus* genomes that have a homologue and its coordinates.

[26] CD:Chapter6/BDBHHomologueCoordinates/BiofilmHomologues/Bcereus/

If *Geobacillus* spp. are able to produce exopolysaccharide in biofilms it seems likely that the molecular mechanisms to generate it as well as its composition would differ from *B. subtilis*. The nature of the polysaccharide component of the extracellular matrix has been shown to vary between strains and is also affected by the substrates available for growth. For example in the presence of sucrose it has been shown that *B. subtilis* can include levan via a levansucrose instead of producing exopolysaccharide via the *eps* operon (Dogsa 2013). One would deduce that in the case of the dairy strains of *G. stearothermophilus* where the formation of the biofilm occurs in milk, that the nature of any matrix would be quite different from other species of *Geobacillus*, particularly those found in natural environments such as hot springs. The dairy strains of *G. stearothermophilus* were clearly able to form biofilms as shown in Chapter 3; therefore other components other than or as well as exopolysaccharide, such as proteins or DNA, may also be responsible for formation of the matrix. In *B. subtilis* the protein component of the matrix is generated via the *tapA – sipW – tasA* operon. No homologues of *tasA* or *tapA* were found in *Geobacillus*. *B. cereus* contains a similar cluster of genes containing homologues of *sinI, sinR, tasA and sipW* in *B. subtilis*, and is also required for biofilm formation (Gao *et al.*, 2015b). In the presence of milk it was shown that cell bundling (also termed flocs or floating biofilms) of *B. subtilis* as well as *B. cereus* and *B. licheniformis* occurred as opposed to free floating cells in Lysogeny broth (Pasvolsky *et al.*, 2014). In the same study, TasA was essential for this bundling to occur and was triggered by butyric acid, a free fatty acid, via KinD. It is unknown whether bundling of *G. stearothermophilus* occurs in liquid cultures of skim milk. If it does occur, the molecular mechanisms would be different to those in *B. subtilis*.

An analysis was also carried out to see if there were homologues of genes that may be involved in biofilm formation of two other Gram positive species: *S. aureus* and *S. thermophilus*. The molecular mechanisms of biofilm formation in *S. aureus* are well characterized, although there are variations between strains. For PIA (polysaccharide intercellular adhesin) dependent biofilm formation the exopolysaccharide protein PIA is produced via the *icaADBC* operon (Archer *et al.*, 2011). Some strains can form biofilms via a PIA independent mechanism through the production of surface proteins. This pathway seems to be confined to MRSA strains and will not be discussed further (O'Neill *et al.*, 2007). This present study limited the identification of homologues to the *ica* operon. A homologue of *icaA* from *S. aureus* was found in 19 genomes of *Geobacillus* (A1, D1, P3, Gs795, GicG1w, Gk102, GkHTA, GtlCCB, GtlB23, GsuPSS, GvPSS, GtcGS1, GeoM52, GeoM61, Gs10, GeoA8, GeoC56, GeoJF8 and GeoFW2). In *S. aureus icaA* encodes a N-acetyl-D-glucosamine synthase. However, for optimal synthesis of PIA and attachment *icaDBC* is also required (O'Gara, 2007). Where an *icaA* homologue in *Geobacillus* was present, it it clustered with a HEAT (**H**untingtin, elongation factor **E**F3, protein phosphatase PP2**A**, kinase **T**OR1) repeat and a diguanylate cyclase, generally annotated as *pleD.* An example of the gene organisation of this cluster in *G. stearothermopohilus* A1 is shown in Figure 6.8. The strain D1 was the exception, where the *icaA* ORF was at the start of a contig.

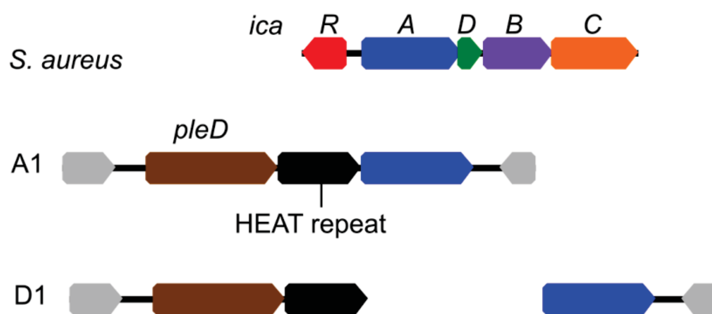



**Figure 6.8. Gene organisation of the *ica* operon in *S. aureus* and the homologue of *icaA* in the dairy strains *G. stearothermophilus* A1 and D1.** The D1 *icaA* homologue (coloured in blue) was on a separate contig to *pleD* (coloured in brown).

In other bacterial species, *pleD*, encodes a diguanylate cyclase responsible for the synthesis of c-di-GMP (Romling *et al.*, 2013). This is a universal bacterial activator and is involved in the regulation of a number processes for example, motlity, biofilm formation and virulence. In terms of biofilm formation c-di-GMP has been shown to be involved in the regulation of motile to biofilm state, as well as the production of the matrix component in a number of bacterial species including the production of cellulose in *Salmonella enterica*, Pel polysaccharide production in *Pseudomonas aeroginosa*, as well as PIA in some Gram negative species (Ahmad *et al.*, 2011, Romling *et al.*, 2013, Mika & Hengge, 2014, Ha & O'Toole, 2015). In *B. subtilis* c-di-GMP has been shown to affect swarming but not biofilm formation (Gao *et al.*, 2013).

*S. thermophilus* is one of the most commonly used starter strains in cheese and yoghurt. Although, many strains of *S. thermophilus* are not good biofilm formers, recently one strain strain of *S. thermophilus* (JIM8232) isolated from milk was found to be a good biofilm former and three genes were identified as being important for biofilm formation in this strain whereas in strains that were poor biofilm formers these genes were predominantly absent (Couvigny *et al.*, 2015). In this present study, no homologues of these genes were found in any of the *Geobacillus* genomes. However, the *S. thermophilus* gene locus tag STH8232_1361 is located in a putative *eps* cluster of 10 genes. Three of the genes from this cluster have homologues in *Geobacillus*, although not in the three dairy strains A1, P3 or D1.

In general, the overall process of biofilm formation and the main components of the extracellular matrix (exopolysaccharides, proteins and eDNA) is similar between species (Vlamakis *et al.*, 2013). However, the molecular mechanisms of biofilm formation vary between species and even between some strains of the same species (Kirisits *et al.*, 2005, Stanley & Lazazzera, 2005, O'Neill *et al.*, 2007, Ryder *et al.*, 2007). Not surprisingly, the approach taken in this present study of gaining a genomic insight into biofilm formation of *Geobacillus* spp. was limited.

### 6.3.3 Sporulation

#### *Distribution of sporulation genes*

The genes (387) were analyzed to determine how the sporulation process is conserved in *Geobacillus* compared with *B. subtilis*. 39% of these genes were conserved across the *Geobacillus* genus, 29% were absent and the remaining were variable as defined by the BDBH algorithm (CD[27],[28]). Each gene was grouped into a functional category based on their function or putative function in *B. subtilis* (Figure 6.9). The functional categories of transport, metabolism and cell wall biosynthesis were based on the products (or putative products) encoded for by the genes in these categories. Although, these genes are under control of one of the sporulation specific sigma factors, generally their actual role in sporulation is less clear; therefore, these genes were not analyzed any further. For those genes that were variable or absent and whose role in sporulation has been well characterized the additional approach of using COG functional grouping was taken to identify possible homologues. This will be discussed in more detail below, for each functional category.

Those genes involved in regulation and the initial stages of sporulation were well conserved; whereas structural genes particularly those involved in coat formation were less conserved. This is in agreement with previous studies which have found that the basic components of the sporulation pathway are well conserved amongst spore-forming members of the Bacillales as well as Clostridiales orders, with the regulatory genes being the most well conserved group (Onyenwoke *et al.*, 2004, de Hoon *et al.*, 2010, Galperin *et al.*, 2012, Galperin, 2013). Previous studies have identified a minimal set of sporulation genes that are conserved across all spore-forming *Bacilli* spp. and *Clostridia* spp. as well as a broader set that are *Bacilli*-specific sporulation genes (Paredes *et al.*, 2005, de Hoon *et al.*, 2010, Galperin *et al.*, 2012, Abecasis *et al.*, 2013, Al-Hinai *et al.*, 2015). Many of these conserved genes have been shown to be essential for sporulation of *B. subtilis* and are required in the initial stages of sporulation (Eichenberger *et al.*, 2003, Feucht *et al.*, 2003, Eichenberger *et al.*, 2004, Silvaggi *et al.*, 2004, Abecasis *et al.*, 2013). A small number of genes that have been described as essential for sporulation in *B. subtilis* were not found in some members of the *Geobacillus* genus, such as *stoA* (CD[11]). However, it is possible that there is a homologue of this gene but the criteria used by the BDBH method was too stringent and/or other members of the same COG group may fill the same function.

[27] CD:Chapter6/SporulationGenes.xlsx
[28] CD:Chapter6/BDBHHomologueCoordinates/SporulationHomologues/

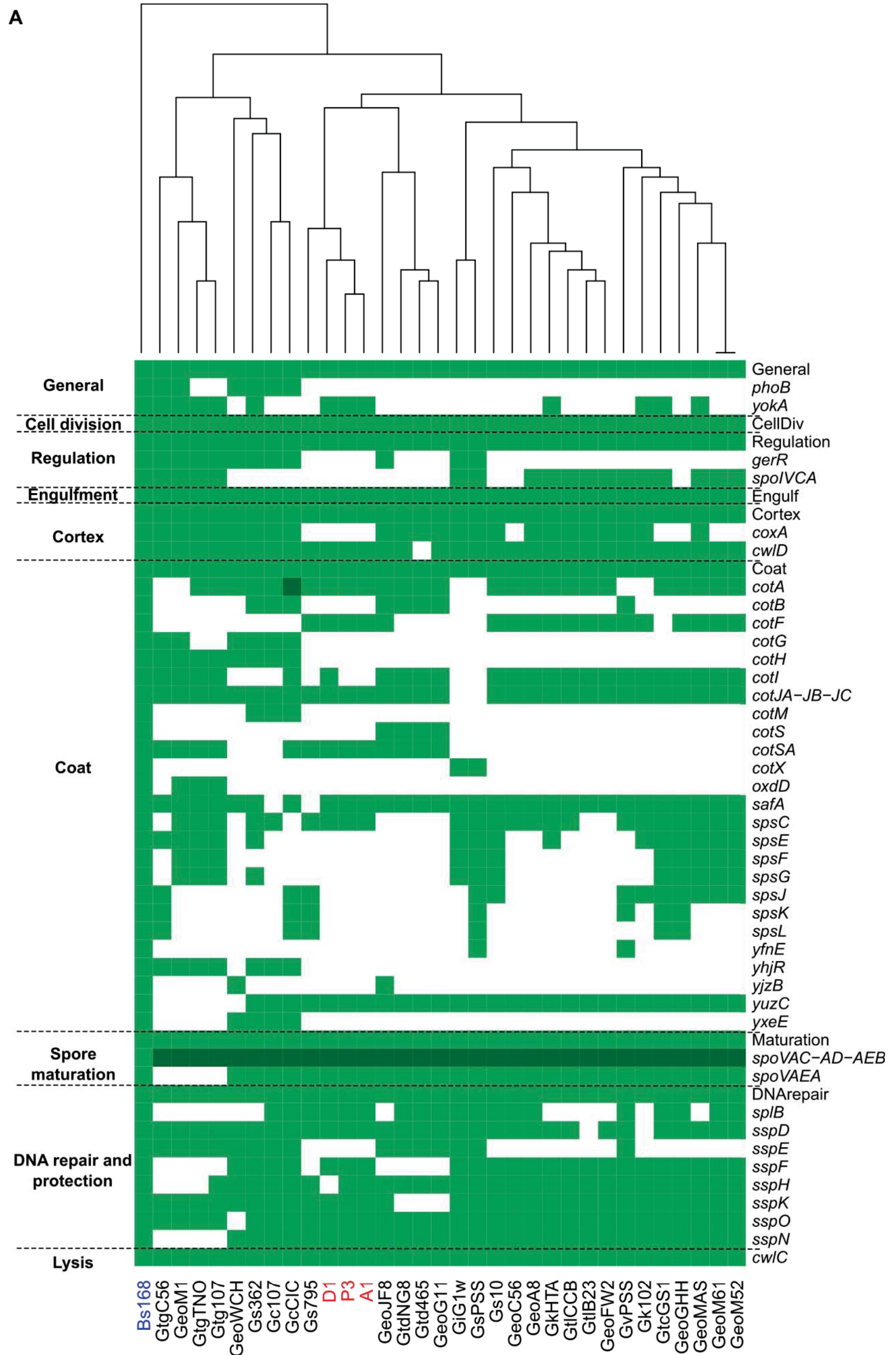

A
General
Cell division
Regulation
Engulfment
Cortex
Coat
Spore maturation
DNA repair and protection
Lysis
General
phoB
yokA
CellDiv
Regulation
gerR
spoIVCA
Engulf
Cortex
coxA
cwlD
Coat
cotA
cotB
cotF
cotG
cotH
cotI
cotJA−JB−JC
cotM
cotS
cotSA
cotX
oxdD
safA
spsC
spsE
spsF
spsG
spsJ
spsK
spsL
yfnE
yhjR
yjzB
yuzC
yxeE
Maturation
spoVAC−AD−AEB
spoVAEA
DNArepair
splB
sspD
sspE
sspF
sspH
sspK
sspO
sspN
cwlC
Bs168
GtgC56
GeoM1
GtgTNO
Gtg107
GeoWCH
Gs362
Gc107
GcCIC
Gs795
D1
P3
A1
GeoJF8
GtdNG8
Gtd465
GeoG11
GiG1w
GsPSS
Gs10
GeoC56
GeoA8
GkHTA
GtlCCB
GtlB23
GeoFW2
GvPSS
Gk102
GtcGS1
GeoGHH
GeoMAS
GeoM61
GeoM52

**Figure 6.9. Conservation of selected *Bacillus subtilis* sporulation genes in the *Geobacillus* genus.** This heat-map shows 281 genes involved in sporulation of *B. subtilis* that have homologues in at least one member of the *Geobacillus* genus. The presence of a homologue was determined by using the BDBH algorithm as described in Section 6.2.3. White represents no homologue being present, mid green when one homologue was present and dark green when two homologues were present. Genes were categorized according to function or putative function. Each functional category is separated by a dotted line and where the first line contains all of the genes that were were present across all of the genomes analyzed. **Functional caterogies** are: **(A) General** which contains *ftsY, mbl, obg, parA, parB, patA, spoVC* (*pth*), *spoVG, spoVIF, spoVS, yabP, ydcC, yhaL, yhaL, yhbH, ypjB, yqfD, ytaF, ytxC, yunB* as all present; **Cell division** where "CellDiv" contains *divIB, divIC, divIVA, ftsA, ftsH, ftsZ, racA, spoIIE, spoIIIE*; **Regulation** which contains *bofA, bofC, csfB, ctpB, gerE, lonB, rsfA, sigE, sigF, sigG, sigH, spoIIAA, spoIIAB, spoIIGA, spoIIID, spoIIR, spoIVB, spoIVFA, spoIVFB, spoVT*, and *subA* as all present;

**Engulfment** which contains *spoIID, spoIIIAA-B-C-D-E-F-G-H, spoIIIJ, spoIIM, spoIIP* and *spoIIQ* as all present; **Cortex** which contains *dacB, dacF, murB, murG, pbpG, pbpI, pdaA, sleB, spoIVA, spoVB, spoVD, spoVE, spoVIE, spoVIGA, spoVIGB, spoVR, yabQ, ypeB* and *yqfC* as all present; **Coat** which contains *cotD, cotE, cwlJ, gerQ, prkA, spoIVA, spoVID, spoVM, spsI, tgl, yaaH, yabG, yhaX, yhbB, yheC, yheD, ylbD, yppG, ysxE* and *yutH* as all present; **Spore maturation** where "Maturation" contains *spmAB*, *spoVAA-B-C-D-EB-F, spoVFA-B (dpaAB)* and *spoVK*; **DNA repair and protection** where "DNArepair" contains *mutTA, sspB, sspI* and *sspP*; **Lysis** and **(B) Metabolism** which contains *asnO*, *comER, mlpA, ydcA* and *yisZ*; **Cell wall biosynthesis** where "CellWall" contains *dapA, dapG* and *ldt*; **Transport** which contains *ytlC* and *ytlD*; and **Resistance** . Those genes present categorized as function **Unknown** included *ybaK, yckC, yhcV, yjbA, yjbE, yjcA, ylaM, ylaK, ylbC, ytvI, ylbE, ylbJ, ylxY, ymxH, ymfB, yozD, ypzA, yqhG, yqfX, yqfT, yrrS, yrrD, yrzE, ykvI, ytfI, ytcB, yuzA, yunC, yusN, yfhS, yhfW, yteA, yuiC, yqhH, yqhP, yqhQ, ytcB, ylmC, yphA, yqfQ, ywcA, yteV, ydfS* and *yloB*.

Genes absent were not shown in this diagram. They included: *proJ, proH, spoIISA, spoIISB, usd* and *ykvP* (General Sporulation); *spoIIB* (Engulfment); *ykvU, stoA* and *yoaR* (Cortex); *cotC, cotO, cotP, cotQ, cotR, cotT, cotU cotV, cotW, cotY, cotZ, cgeCDE, cgeAB, gerT, lipC, spsA, spsB, spsD, yeeK, yfnD, yfnFGH, yisY, yjqC, yknT, ymaG, yncD, yodU, yotN, ysnD, ytxO* and *yybI* (Coat); *splA, sspA, sspC, sspG, sspJ, sspL, sspM, ykoU* and *ykoV* (DNA repair and protection); *cwlH* and *nucB* (Mother cell lysis); *gdh, yitAB, yitD, yjmC, exuT, yqiQ, mmgED* and *ywjE* (Metabolism); *mdxF, citH* and *glcU* (Transport); *cypA, katX,* and *yhcM* (Resistance). Refer to the CD[29] for those genes absent in the Unknown category.

[29] CD:Chapter6/SporulationGenes.xlsx

### *Assymetric cell division*

Initiation of sporulation begins with asymmetric cell division, whereby formation of a septum splits the cell assymetrically. Some of the proteins involved in this process, such as DivIB, DivIC, FtsA and FtsZ are also required during vegetative cell growth (King *et al.*, 1999, Ben-Yehuda & Losick, 2002, Kemp *et al.*, 2002, Kobayashi *et al.*, 2003, Thompson *et al.*, 2006). The process also involves the sporulation specific proteins SpoIIE and RacA (Feucht *et al.*, 1996, Ben-Yehuda *et al.*, 2003, Carniol *et al.*, 2005). The septum forms across one of the chromosomes, which results in about 30% of the chromosome in the forespore, with the remaining part of the chromosome transferred across by the translocase SpoIIIE. The genes known to encode proteins involved in assymetric division are conserved in the *Geobacillus* genus (Figure 6.9).

### *Regulation*

Once a cell is committed to sporulate most of the genes involved are under control of one of the sporulation specific sigma factors: $\sigma^E$, $\sigma^F$, $\sigma^G$ and $\sigma^K$, all of which were conserved in the *Geobacillus* genus (Figure 6.10, Table 6.8). The function of many of the genes under the control of these sporulation specific sigma factors is unknown and many of these genes were absent or variable in *Geobacillus* (CD[30]).

[30] CD:Chapter6/SporulationGenes.xlsx

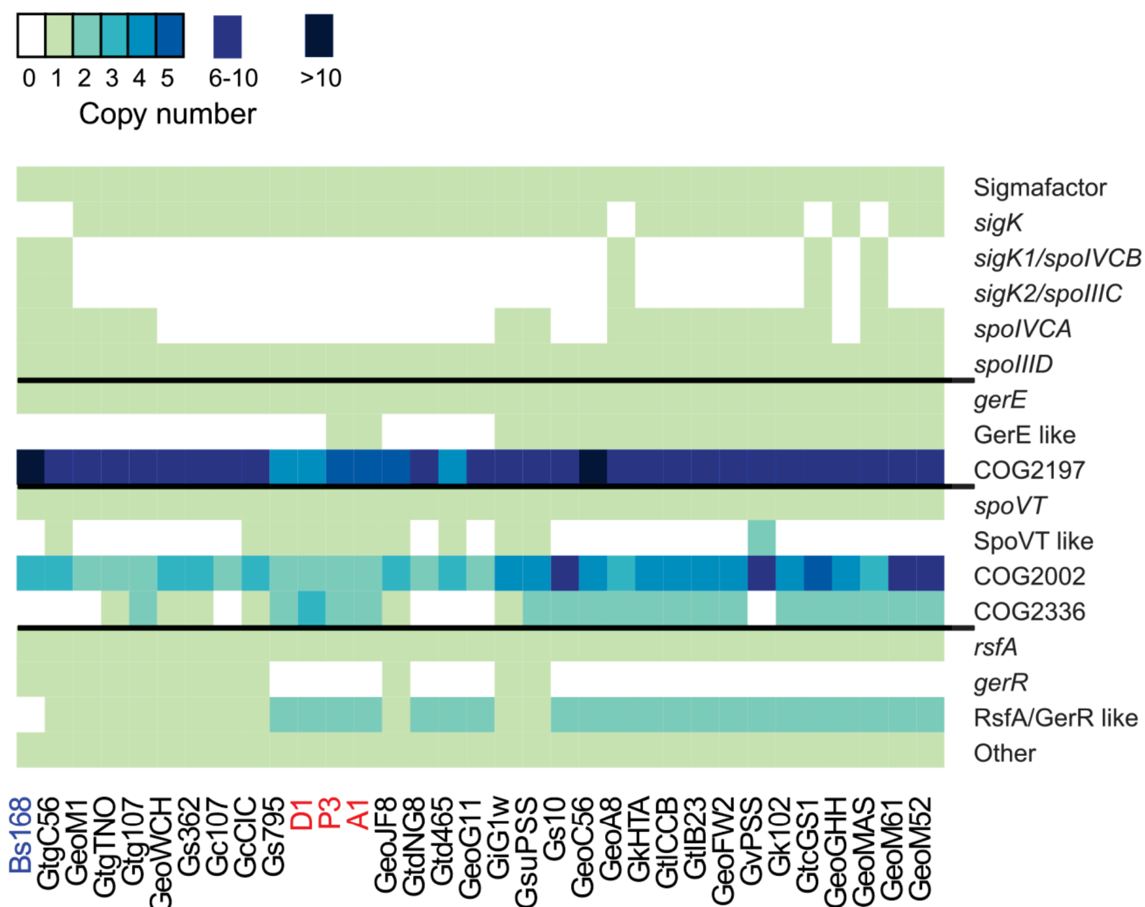



**Figure 6.10. Conservation of selected *B. subtilis* regulatory network genes in *Geobacillus* spp. and their equivalent COGs.** This heat-map shows 26 genes involved in the regulation of sporulation in *B. subtilis* and the presence/absence of homologues in the *Geobacillus* genus. Refer to Section 6.2.3 for determination of a homologue. The COG associated with each gene(s) is in the row below and is separated from the next gene and its associated COG by a black line. The COG analysis was only included for those genes that were variable or absent in the *Geobacillus* genus as defined by the BDBH algorithm. The row "Sigmafactor" refers to the genes *sigE*, *sigG* and *sigF*. The row *sigK* refers to a full-length *sigK*. The rows "SpoVT like", "GerE like" and "RsfA/GerR like refer to additional ORFs in some of the *Geobacillus* genomes that were annotated as *spoVT*, *gerE* and *rsfA* respectively.

**Table 6.8. Description of each COG, which contained a protein involved in *B. subtilis* spore regulation.**

| COG[a,b] | Description[c] | Genes in *B. subtilis* | |
|---|---|---|---|
| | | **Spore regulation** | **Other[d]** |
| **COG2002** | Bifunctional DNA-binding transcriptional regulator of stationary/sporulation/toxin gene expression | *spoVT* | *abrB*, *abh* |
| **COG2336[e]** | Antitoxin component of the MazEF toxin-antitoxin module | N/A | No members |
| **COG2197** | DNA-binding response regulator, NarL/FixJ family, contains REC and HTH domains | *gerE* | *comA*, *degU*, *desR*, *liaR*, *ydfI*, *yfiK*, *yhcZ*, *ylxL*, *yosQ*, *yvfU*, *yxjL* |

[a] The following proteins did not belong to a COG: GerR, RsfA

[b] Refer to CD:Chapter4/Results/COG/COG_LocusTag.xlsx for a list of locus tags within each COG

[c] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

[d] These genes listed are not known to be involved in the *B. subtilis* regulatory network of the mother-cell and forespore as described by de Hoon *et al.* (2010) but they belong to the same COG as one of the spore regulatory genes. The anomaly is AbrB, a transition state regulator, which represses activation of the phosphorelay.

[e] This COG contained the second SpoVT like protein of *G. stearotheromophilus* A1, P3, D1, ATCC7953 and *Geobacillus* sp. JF8.

In *B. subtilis*, to ensure each stage of sporulation is carried out at the right time the sporulation network is tightly controlled. One way this is achieved is through activation of each sigma factor, through proteolytic cleavage in the case of $\sigma^E$ and $\sigma^K$ , release from an anti-sigma factor complex in the case of $\sigma^F$ and less understood is the activation of $\sigma^G$, which is likely to occur via a signaling molecule through the SpoIIIAH-SpoIIQ channel (Errington, 2003, Higgins & Dworkin, 2012). The genes known to be required for activation of each of the sigma factors in *B. subtilis* are also conserved in the *Geobacillus* genus.

The stationary phase sigma factor, $\sigma^H$, initiates transcription of the first of the sporulation specific sigma factors, the forespore specific $\sigma^F$, as part of the *spoIIAA–spoIIAB–sigF* operon. This operon was conserved in the *Geobacillus* genus (Figure 6.18). $\sigma^F$ is not activated until assymetric division has been completed and the commitment has been made to sporulate. Approximately 50 genes are under the control of $\sigma^F$, including *sigG* (encoding the other forespore specific sporulation factor) and *spoIIR* (Wang *et al.*, 2006). Approximately 100 genes are in the forespore specific $\sigma^G$ regulon (Steil *et al.*, 2005, Wang *et al.*, 2006). The *spoIIIA* operon is required for activation of $\sigma^G$, which will be discussed further under “Engulfment”.

SpoIIR is required for activation of $\sigma^E$, a mother-cell specific sigma factor. *sigE* is in the same operon as *spoIIGA* and *sigG*. This operon was conserved within the *Geobacillus* genus. The only unknown was the strain D1 where *sigG* and *sigE* sat on a separate contig compared with *spoIIGA*. The $\sigma^E$ regulon is the largest of the four sporulation specific sigma factor regulons, with approximately 250 genes (Eichenberger *et al.*, 2003). Activation of $\sigma^E$ is required for engulfment.

The final sigma factor to be activated is $\sigma^K$. In *B. subtilis* 168, $\sigma^K$ is encoded on two genes: *spoIVCB* (hereafter called *sigK1*) encodes the N-terminal half and *spoIIIC* (hereafter called

*sigK2*) encodes the C-terminal half. The intervening DNA (approximately 48 Kb in length and called the *skin* element) is removed by the DNA recombinase SpoIVCA (Stragier *et al.*, 1989, Kunkel *et al.*, 1990). It was found that in *Geobacillus*, $\sigma^K$ was generally encoded by one gene, *sigK* (Figure 6.12). The exceptions were *G. thermoglucosidasius* C56YS93, *Geobacillus* sp. A8, *Geobacillus* MAS1 and *G.thermocatenulatus* GS-1. These four genomes all contained two genes annotated as *sigK* by Prokka; one encoding the N-terminal and the other the C-terminal portion. In *Geobacillus* A8 these two genes were alongside each other whereas in the other three genomes the intervening DNA ranged between 15.5 – 22.5 Kb (Table 6.9). The *skin* element was predominantly the same between the three *Geobacillus* genomes (Appendix 8), except for the absence of *spoIVCA* (the recombinase) neighbouring *sigK1* and the additional ORFs in *G. thermoglucosidasius* C56YS93. However, they differed from the *skin* element found in *B. subtilis* 168. Only two homologues (of *spoIVCA* and *yqaF*) were found in *Geobacillus* MAS1 and *G.thermocatenulatus* GS-1, and no homologues of the *B. subtilis skin* element genes were found in *G. thermoglucosidasius* C56YS93. The *skin* elements of the *Geobacillus* genomes were predominantly made up of genes encoding hypothetical proteins as well as some recombinases/resolvases and restriction enzymes.

The *skin* element in two of the *Geobacillus* (GeoMAS and GthcGS1) genomes contained a homologue of *spoIVCA* (Figure 6.11), whereas although *G. thermoglucosidasius* C56 contained recombinases within the skin element they did not neighbour *sigK1* and they were not identified as being the closest homologue of *spoIVCA*. *Geobacillus* sp. A8 also had a homologue of *spoIVCA*, but it was in another region of the genome (data not shown). Many of the *Geobacillus* genomes did not contain a homologue of *spoIVCA*, which is not surprising given most of them contain a full-length *sigK* gene.

**Table 6.9. Genomes containing a split *sigK* gene.**

| Genome | Length of *skin* element (Kb) | Neighbouring recombinase | Neighbouring recombinase is closest homologue of *spoIVCA* |
|---|---|---|---|
| **Bs 168** | 47.9 | | |
| **GeoMAS** | 15.5 | Yes | Yes |
| **GthcGS1** | 15.5 | Yes | Yes |
| **GtgC56** | 22.5 | Yes | No |
| **GeoA8** | NP* | No | No |

* Not present

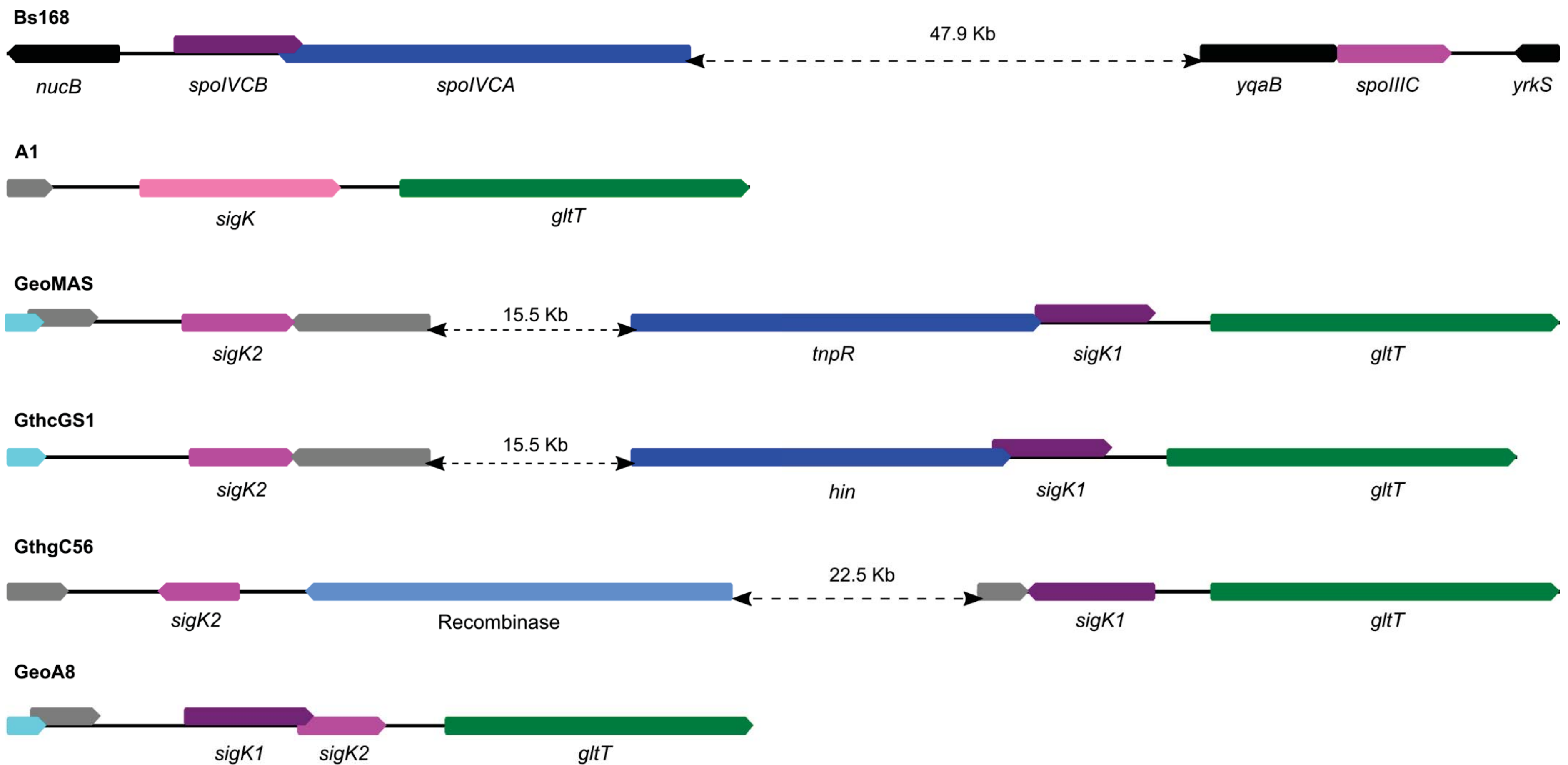



**Figure 6.11. Gene organisation of *sigK* in *B. subtilis* 168 and selected *Geobacillus* strains.** The remaining *Geobacillus* strains not shown in this diagram have a full-length *sigK* as represented in this diagram by the strain A1. Those genes coloured in grey encode hypothetical proteins and the genes in light blue represent a gene that encodes a putative sporulation protein YrzI as annotated by Prokka.

It was previously suggested that a split *sigK* was one way in which to prevent *sigK* expression from occurring too early, under conditions still suitable for growth (Oke & Losick, 1993). However, split *sigK* genes and the *spoIVCA* gene (encoding the DNA recombinase required for removing the *skin* element) are not found in all *B. subtilis* strains, for example *B. subtilis* subsp spizizenii W23 contains a full-length *sigK* and does not contain the *skin* element (Earl *et al.*, 2007, Zeigler, 2011). In addition, a full-length *sigK* gene appears to be more common in endospore formers than a split gene (Abe *et al.*, 2013). When the *skin* element is present it has been shown to vary in both size and composition between different species (Abe *et al.*, 2013).

Additional control of the sporulation network is provided by the transcription factors GerE, RsfA, SpoVT, SpoIIID and GerR by activating and/or repressing genes to give longer or shorter periods of gene expression (Eichenberger *et al.*, 2004, Wang *et al.*, 2006, Wang *et al.*, 2007, Cangiano *et al.*, 2010, de Hoon *et al.*, 2010). GerE, RsfA, SpoVT and SpoIIID all contained homologues in the *Geobacillus* genomes as defined by the BDBH algorithm, but a homologue of *gerR* was only found in eleven of the *Geobacillus* genomes analyzed. GerR acts as a repressor of some of the $\sigma^E$-dependent genes and an activator of some of the $\sigma^K$-dependant genes (Eichenberger *et al.*, 2004, Cangiano *et al.*, 2010). In *B. subtilis* GerR shows 52% similarity to RsfA, another member of the leucine zipper family of transcriptional regulators (Wu & Errington, 2000). RsfA is in the $\sigma^F$ regulon and acts as both an activator and repressor of genes within this regulon. All of the *Geobacillus* genomes contained three separate ORFs showing sequence similarity to both *rsfA* and *gerR*, as opposed to *B. subtilis* which only contains the two genes (*rsfA* and *gerR*). All three of the *rsfA/gerR* like ORFs in the *Geobacillus* genomes were annotated as *rsfA*, one of which was a homologue of *B. subtilis rsfA* as defined by the BDBH algorithm. The second was a similar match to both *gerR* and *rsfA* and the third was a homologue of *gerR* in eleven of the genomes. An example of how these three ORF were similar to both *rsfA* and *gerR* is outlined in Appendix 9.

Analysis of the Prokka annotations also identified two other additional putative transcription factors in some of the *Geobacillus* genomes, including the three dairy strains A1, P3 and D1. These were annotated as *spoVT* or *gerE*, labelled “SpoVT like” and “GerE like” respectively in the heat-map (Figure 6.10), with the closest match being *spoVT* and *gerE* respectively, in *B. subtilis* 168. GerE is in the $\sigma^K$ and SpoVT is in the $\sigma^G$ regulon with both of these transcription factors acting as activators and repressors in their respective regulons. The additional *gerE* ORF and most of the *spoVT* ORF, in *Geobacillus*, also belonged to the same COG as *B. subtilis gerE* (COG2197) and *spoVT* (COG2002). However, the *spoVT* like ORF from the *G. stearothermophilus* strains A1, P3, D1, ATCC7953 as well as *Geobacillus* sp. JF8 belonged to COG2336. This is probably because the translated amino acid sequence is shorter in these proteins belonging to COG2336 and only contained the MazE_Antitoxin domain, whereas *B. subtilis* SpoVT and its homologues in *Geobacillus* contained two domains: MazE_Antitoxin and SpoVT_C.

The presence of additional putative transcription factors involved in sporulation of *Geobacillus* spp. or other bacilli has not been identified in the other comparative genomics study by de Hoon *et al.* (2010). In that study *spoIIID* (a member of the $\sigma^E$ regulon) and *spoVT* were found in the genomes of all of the endospore producing species they analysed, whereas *gerR*, *gerE* and *rsfA* were not so well conserved. Their analysis included one strain of *Geobacillus* (*G. kaustophilus* HTA426) in which no homologues of *gerR* were found but homologues of *gerE* and *rsfA* were found. The authors of that study suggest that the conservation of each transcription factor relates to their significance, with RsfA and GerR controlling fewer genes. This does not explain the presence of additional transcription factors found in the *Geobacillus* genus in this present study. However, given no promoter search was carried out in this present study, it is unknown whether these additional putative transcription factors could be members of one of the sporulation-specific sigma factor regulons. In *B. subtilis*, GerR was also found to repress genes that were not members of the $\sigma^E$ regulon nor known to be involved with sporulation (Eichenberger *et al.*, 2004). To confirm the additional putative transcription factors found in *Geobacillus* play a role in sporulation a transcriptomics analysis would need to be carried out.

***Engulfment***

Once asymmetric division is completed, the cell can commit to sporulating and the process of engulfment begins. Engulfment involves synthesis of peptidoglycan, degradation of the septal membrane, and relocation of the mother-cell membrane around the forespore (Meyer *et al.*, 2010, Crawshaw *et al.*, 2014). Two protein complexes involved in this process are the SpoIIDMP complex (encoded by *spoIID*, *spoIIM* and *spoIIP*) and the SpoIIQ/SpoIIAH complex (encoded by *spoIIQ* and the *spoIIAA – AH* operon) (Broder & Pogliano, 2006, Chastanet & Losick, 2007).

SpoIID, SpoIIM and SpoIIP are hydrolases which interact with each other to degrade peptidoglycan in the septal membrane (Morlot *et al.*, 2010). These three proteins were all conserved in *Geobacillus*. In *B. subtilis* these are all encoded by monocistronic genes. In *Geobacillus* this appears to be the case for *spoIIP*, but in contrast it seems *spoIID* and *spoIIM* may each be a member of an operon. It appears *spoIID* is in the same operon as *spoIIQ* (Figure 6.12A), and *spoIIM* is in an operon with two hypothetical proteins and *nudF* (Figure 6.12B).

In *B. subtilis* SpoIIQ interacts with SpoIIIAH to form a complex that is likely to be a secretion system between the mother cell and the forespore. Several roles have been proposed for this secretion system: (1) the movement of peptidoglycan during engulfment when levels of the SpoIIDMP complex is low, (2) as a signaling channel to activate $\sigma^G$ and (3) a feeding channel to provide nutrients to the forespore (Broder & Pogliano, 2006, Camp & Losick, 2008, Meisner *et al.*, 2008, Camp & Losick, 2009, Doan *et al.*, 2009). The *spoIIIAA – AH* operon consists of eight genes, all of which are conserved in *Geobacillus*.

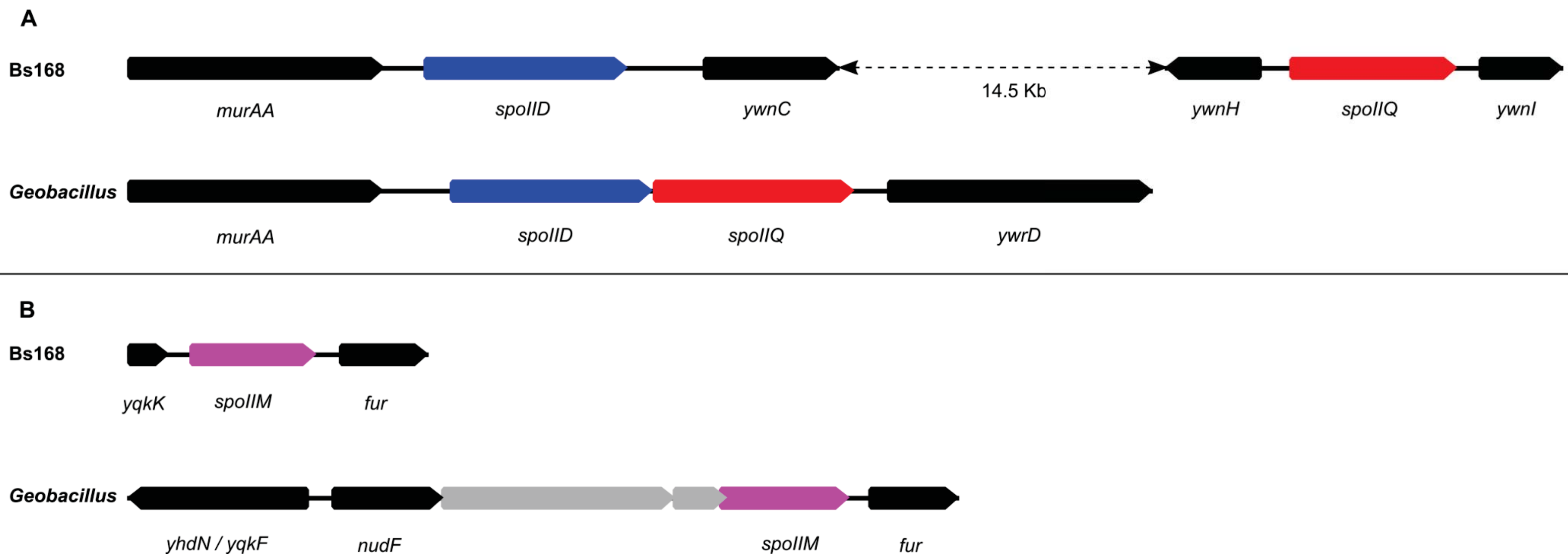



**Figure 6.12. Organisation of selected *B. subtilis* genes encoding for proteins involved with engulfment and their homologues in *Geobacillus*.** (**A**) *spoIID* and *spoIIQ*, (**B**) *spoIIM.* Strain *G. stearothermophilus* A1 was used for generation of this diagram, but the gene organisation of the putative *spoIID*, *spoIIQ* and *spoIIM* genes were syntenic across all of the *Geobacillus* genomes analyzed.

### *Cortex*

Stage IV of endospore formation is formation of the cortex, which involves the production of peptidoglycan. In addition, many proteins involved in spore germination are located within the cortex (e.g. YkvU). Many proteins required for cortex formation have been shown to be essential for *B. subtilis* endospore formation (Eichenberger *et al.*, 2003, Silvaggi *et al.*, 2004). Of these essential cortex proteins, StoA did not have a homologue in *Geobacillus* as defined using the BDBH algorithm (Figure 6.13). However, all of the *Geobacillus* genomes contained 1 – 2 genes annotated as *stoA* by Prokka. StoA belongs to a family of enzymes called thiol-disulphide oxidoreductases, which are members of COG0526. The function of this family is to reduce or oxidise disulphide bonds in a number of different proteins. In sporulating cells StoA appears to drive the breakage of disulphide bonds in SpoVD, which in turn initiates peptidoglycan synthesis, neccersary for the formation of the cortex (Liu and Hederstedt 2010). COG0526 contained 12 proteins in total from *B. subtilis* (Figure 6.13 and Table 6.10). In *Geobacillus* this ranged between 5 – 10 proteins, which included those encoded from ORFs annotated as *stoA*.

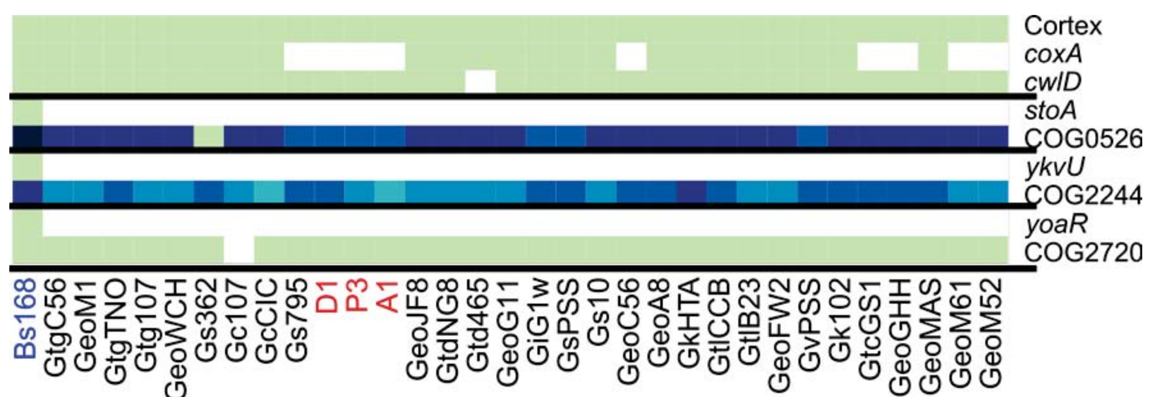



**Figure 6.13. Conservation of selected *B. subtilis* cortex genes in *Geobacillus* spp. and their equivalent COGs.** This heat-map shows whether homologues of 25 genes involved in cortex formation of *B. subtilis* are present or absent in the *Geobacillus* genus. Refer to Section 6.2.3 for determination of a homologue. The COG associated with each gene(s) is in the row below and is separated from the next gene and its associated COG by a black line. The COG analysis was only included for those genes that were variable or absent in the *Geobacillus* genus as defined by the BDBH algorithm. The row "Cortex" refers to the genes: *dacB, dacF, murB, murG, pbpG, pbpI, pdaA, sleB, spoIVA, spoVB, spoVD, spoVE, spoVIE, spoVIGA, spoVIGB, spoVR, yabQ, ypeB* and *yqfC*, which were present in all of the *Geobacillus* genomes analyzed.

**Table 6.10. Description of each COG, which contained proteins involved with cortex formation.**

| COG[a] | Description[b] | Gene names | |
|---|---|---|---|
| | | Spore cortex | Other[c] |
| **COG2244** | Membrane protein involved in the export of O-antigen and teichoic acid | *ykvU, spoVB* | *epsK, tuaB, yabM, ytgP* |
| **COG0526** | Thiol-disulfide isomerase or thioredoxin | *stoA* | *bdbA, resA, skfH, slp, ydbP, ydfQ, ykuV, yneN, yosR, ytpP, yusE* |
| **COG2720** | Vancomycin resistance protein YoaR (function unknown), contains peptidoglycan-binding and VanW domain | *yoaR* | No |

[a] Refer to CD: Chapter4/Results/COG/COG_LocusTag.xlsx for a list of locus tags within each COG

[b] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

[c] These genes listed are not known to be involved in cortex formation of *B. subtilis*, but they belong to the same COG as one of the cortex genes.

Most of the other genes involved in cortex formation, but not regarded as essential for sporulation were conserved in *Geobacillus*. Two exceptions were *yoaR*, which was absent and whose function is still largely unknown and *coxA* which was variable. CoxA is located in the cortex (Takamatsu *et al.*, 1998), but there is still limited understanding on its function. In *B. subtilis, coxA* is in an operon with *safA*, which is involved with inner coat formation. In *Geobacillus, coxA* was generally separated from *safA* by two other other genes (*nadE* and an ORF encoding a phosphotransferase) (Figure 6.14). Nine of the genomes did not contain *coxA* as defined by the BDBH algorithm; however, the organization of the genes in this region was syntentic with the other *Geobacillus* genomes and the gene annotated as a lipoprotein sat in the same orthologuous cluster as *coxA* when the *Geobacillus* genome C56-T4 was used as the reference genome. CwlD is located in the cortex, but is involved in germination. It was not found in one *Geobacillus* genome, *G. thermodenitrificans* DSM 465. In this case the surrounding genes were located on two different contigs and possibly the gene was “lost” because of fragmentation of the genome.

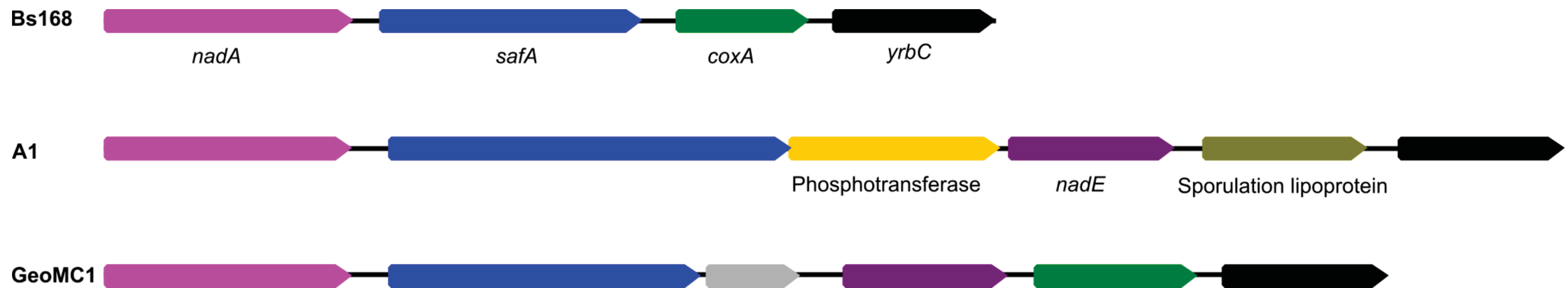



**Figure 6.14. Gene organisation of *coxA* (coloured in green) and *safA* (coloured in blue) in *B. subtilis* and their homologues in two strains of *Geobacillus*: *G. stearothermophilus* A1 and *Geobacillus sp.* Y4.1MC1 (Taxon group 10, *G. thermoglucosidasius*).** The surrounding genes are also included, each depicted by a different colour.

***Spore coat and outer surface***

The initial stages of coat formation begin during the engulfment phase, where coat proteins begin to assemble on the surface of the forespore (McKenney *et al.*, 2013). The spore coat can be divided up into four layers, with each layer having its own morphogenetic protein that drives the formation of that layer (McKenney *et al.*, 2013). Genes involved in the later stages of sporulation were less well conserved, particularly those involved in coat formation. The number of coat protein homologues varied within the *Geobacillus* genus, with homologues of approximately 60 – 80 % of the *B. subtilis* coat protein missing. This is not surprising given other studies have also found that the taxonomic distribution of many of the coat proteins is limited (de Hoon *et al.*, 2010, Galperin *et al.*, 2012). The exception is a group of proteins called the morphogenetic proteins, which drive coat formation. These were all conserved in the *Geobacillus* genus (Figure 6.15) and have previously been found to be conserved in most members of the *Bacillaceae* family (McKenney *et al.*, 2013).However, only one gene, *cotX*, was present that encoded a crust morphogenetic protein, which was only present in two genomes – *G. subterraneus* PSS2 and *G. icigianus* G1w1. In *B. subtilis*, the outer layer of the spore, which is termed the crust, covers the outer coat layer (Imamura *et al.*, 2010, McKenney *et al.*, 2010, Imamura *et al.*, 2011). To the author's knowledge there is no experimental evidence that spores of the *Geobacillus* genus form a crust. CotX also shows some similarity to the exosporium protein ExsK (Redmond *et al.*, 2004); therefore, this *cotX* homologue in *G. subterraneus* and *G. icigianus* could potentially be involved in formation of an exosporium instead of a crust.

Formation of the coat in *B.subtilis* involves approximately 70 proteins (McKenney *et al.*, 2013). Homologues of these remaining proteins were generally variable or not present within the *Geobacillus* genus (Figure 6.15 and CD[31]). There may be some additional coat proteins identified using the COG approach (Figure 6.15 and Table 6.11). However, many of these COG groups contain a number of proteins with similar functions but involved in different processes. For example CotSA belongs to COG0438, which contains glycosyltranferases that are involved in a variety of processes including exopolysaccharide production in biofilms.

---

[31] CD:Chapter6/SporulationGenes.xlsx

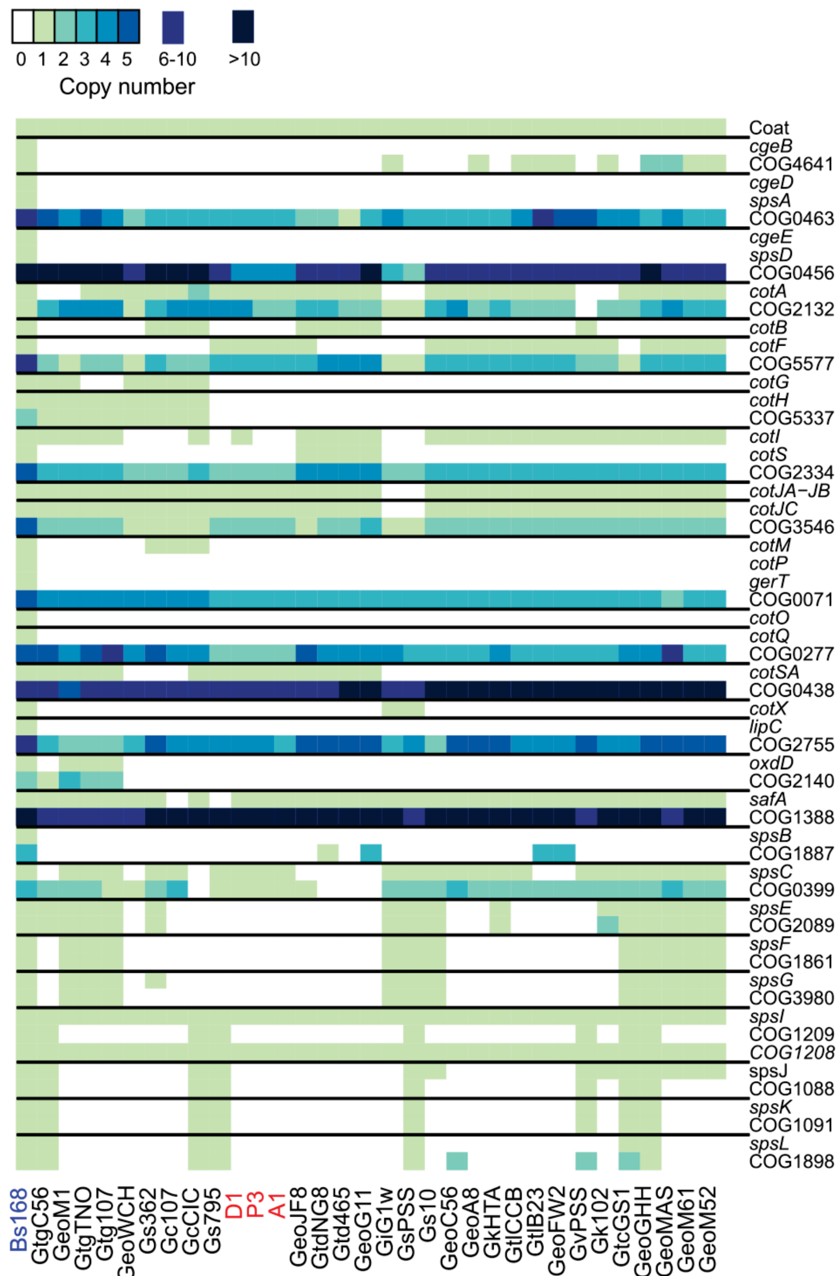



**Figure 6.15. Conservation of selected *B. subtilis* spore coat genes in *Geobacillus* spp. and their equivalent COGs.** This heat-map shows whether homologues of 71 genes that are believed to be involved in spore coat formation of *B. subtilis* are present or absent in the *Geobacillus* genus. Refer to Section 6.2.3 for determination of a homologue. The COG associated with each gene(s) is in the row below and is separated from the next gene and its associated COG by a black line. The COG analysis was only included for those genes that were variable or absent in the *Geobacillus* genus as defined by the BDBH algorithm. The row "Coat" refers to the genes: *cotD, cotE, cwlJ, gerQ, prkA, spoIVA, spoVID, spoVM, spsI, tgl, yaaH, yabG, yhaX, yhbB, yheC, yheD, ylbD, yppG, ysxE* and *yutH*, which were all present in the *Geobacillus* genomes analyzed. Those genes that did not have any homologues in the *Geobacillus* genus as defined by the BDBH algorithm as well as no members of its COG or did not belong to a COG (as highlighted below in bold) were not included in the heat-map. These included *cgeA* and its COG2067, *cotR* and its COG3621. The following proteins did not belong to a COG: CotB, **CotC**, CotG, CotJA, CotJB, **CotO**, **CotT**, **CotU**, **CotV**, **CotW**, CotX, **CotY**, **CotZ**.

**Table 6.11. Description of each COG which contained proteins involved with *B. subtilis* coat formation.**

| COG[a,b] | Description[c] | Gene names | | |
|---|---|---|---|---|
| | | **Coat** | **Unknown $\sigma^E$ and $\sigma^G$ controlled[c]** | **Other[d]** |
| **COG2067** | Long-chain fatty acid transport protein | *cgeA* | | No |
| **COG4641** | Spore maturation protein CgeB | *cgeB* | | No |
| **COG0463** | Glycosyltransferase involved in cell wall biosynthesis | *cgeD, spsA* | | *csbB, ggaA, ykcC, ykoT, yolJ, yycS* |
| **COG0456** | Ribosomal protein S18 acetylase RimI and related acetyltransferases | *cgeE, spsD* | *yodP, yycN* | *acuA, bltD, rimI, ybfA, ydgE, yfmK, yhdJ, yhfO, yjbC, ykwB, ykzC, yoaP, yobR, yocC, ypzK, yqkA, yqiY, yvbK, yxeL, yxbD, yyaR* |
| **COG2132** | Multicopper oxidase with three cupredoxin domains (includes cell division protein FtsP and spore coat CotA) | *cotA* | | No |
| **COG5577** | Spore coat protein CotF | *cotF* | *yhcQ, yraD, yraE, yraF, yraG, yusN, yvdQ* | |
| **COG5337** | Spore coat protein CotH | *cotH* | | *yisJ* |
| **COG2334** | Ser/Thr protein kinase RdoA involved in Cpx stress response, MazF antagonist | *cotI, cotS* | *ysxE, yutH* | *yerI* |
| **COG3546** | Mn-containing catalase (includes spore coat protein CotJC) | *cotJC* | *yjqC* | *ydbD, ydhU,* |
| **COG0071** | Molecular chaperone IbpA, HSP20 family | *cotM, cotP gerT* | *ypqA, yocM* | |
| **COG0277** | FAD/FMN-containing dehydrogenase | *cotQ* | | *glcD, ygaK, yitY, yvdP, yxiJ* |
| **COG3621** | Patatin-like phospholipase/acyl hydrolase | *cotR* | | No |
| **COG0438** | Glycosyltransferase involved in cell wall bisynthesis | *cotSA* | *ytcC* | *epsD, epsF, tagE, tuaH, ypjH, yqqM* |
| **COG2755** | Lysophospholipase L1 or related esterase | *lipC* | | *rhgT, yesY, ypmR, yqeF, yxiM* |
| **COG2140** | Oxalate decarboxylase/archaeal phosphoglucose isomerase, cupin superfamily | *oxdD* | | *oxdC* |
| **COG1388** | LysM repeat | *spoVID, safA* | *yaaH, ykvP* | *ipi, xlyB, ydhD, ykzQ, yneA, yocH, yoqH, ypbE, yqfZ* |
| **COG1887** | CDP-glycerol glycerophosphotransferase, TagB/SpsB family | *spsB* | | *tagB, tagF* |
| **COG0399** | dTDP-4-amino-4,6-dideoxygalactose transaminase | *spsC* | | *ntdA, epsN* |

| | | | |
|---|---|---|---|
| **COG2089** | Sialic acid synthase SpsE, contains C-terminal SAF domain | *spsE* | No |
| **COG1861** | Spore coat polysaccharide biosynthesis protein SpsF, cytidylyltransferase family | *spsF* | No |
| **COG3980** | Spore coat polysaccharide biosynthesis protein SpsG, predicted glycosyltransferase | *spsG* | No |
| **COG1209** | dTDP-glucose pyrophosphorylase | *spsH* | No |
| **COG1088** | dTDP-D-glucose 4,6-dehydratase | *spsI* | No |
| **COG1091** | dTDP-4-dehydrorhamnose reductase | *spsK* | No |

[a] The following proteins did not belong to a COG: CotB, CotC, CotG, CotJA, CotJB, CotO, CotT, CotU, CotV, CotW, CotX, CotY, CotZ

[b] Refer to CD: Chapter4/Results/COG/COG_LocusTag.xlsx for a list of locus tags within each COG

[c] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

[d] These genes listed are not known to be involved in the coat formation of *B. subtilis*. However, they are under control of one of the sporulation specific sigma factors.

[e] These genes listed are not known to be involved in the *B. subtilis* coat formation nor are they known to be under control of one of the sporulation specific sigma factors.

In addition to proteins, a small proportion of the coat contains exopolysaccharides (Driks, 1999). One operon that was believed to encode enzymes for the synthesis of some of these polysacharides was the *sps* operon; however, deletion of this operon resulted in no obvious phenotype (Driks, 1999). This operon is expressed late in sporulation and is under the control of $\sigma^K$ (Eichenberger *et al.*, 2004, Cangiano *et al.*, 2014). More recent work was not conclusive as to the location of the polysaccharides synthesized by this operon but given deletion of the *sps* operon resulted in a slight loss in ability to germinate as well as an increased ability to adsorb and a clumping phenotype, probably caused by a change in hydrophobicity, the authors concluded that the *sps* operon plays a role in spore surface formation (Cangiano *et al.*, 2014).

In *Geobacillus*, *spsI* was the only member of the *sps* operon that has a homologue in all of the *Geobacillus* genomes; although, the e-values of these homologues varied greatly ranging from $1e^{-93}$ to $1e^{-25}$. Four of the genes, *spsA*, *spsB*, *spsD* and *spsH* were not found in any of the genomes, with the others variable. Consequently there was no one *sps* operon in the *Geobacillus* genome. *Geobacillus* GHH01 contained the most *sps* homologues and these are encoded in two clusters, separated by approximately 216 Kb (Figure 6.16). *Geobacillus* Y412MC52 contained six homologues five of which (*spsCEFGJ*) were in the the same region of the genome. The dairy strains A1, P3 and D1 contained an orphan homologue of *spsC* as well as a homologue of *spsI* which clustered with three other genes (Figure 6.16).

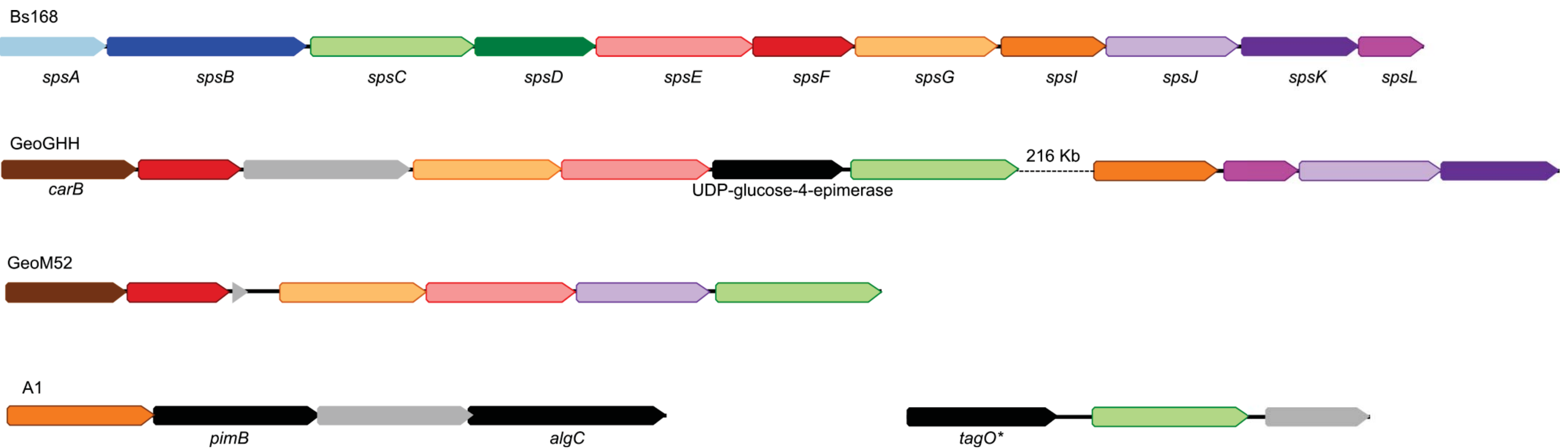



**Figure 6.16. Gene organisation of the *sps* operon in *B. subtilis* and putative *sps* operons in selected strains of *Geobacillus* spp.** Homologues are depicted with the same colour, with the exception of those genes coloured in black which are not homologues of the sps genes and do not have homologues in the other strains. Descriptions for the *Geobacillus* genes (named as per the Prokka annotations) are as follows: *carB*, Carbamoyl-phosphate synthase; *pimB*, GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol monomannoside mannosyltransferase; *algC*, phosphomannomutase/phosphoglucomutase and *tagO*, undecaprenyl-phosphate N-acetylglucosaminyl 1-phosphate transferase*strE,* dTDP-glucose 4,6-dehydratase, where *tagO* is marked with an * as it is a homologue of the *B. subtilis tagO.*

### *Exosporium*

In *B. subtilis* the outer layer is the spore crust. However, in some other Bacilli of which the *B. cereus* group has been most well characterized there is an additional layer called the exosporium. There is also evidence of one dairy strain of *G. stearothermophilus* that has an exosporium (Seale *et al.*, 2010). The exosporium of the *B. cereus* group is made up of two layers: the basal layer and the external hair-like nap (Sylvestre *et al.*, 2002, Todd *et al.*, 2003). The hair-like nap is predoiminantly comprised of the collagen-like glycoprotein BclA. At least 14 other proteins that are regarded as specific to the *B. cereus* group are associated with exosporium formation (Table 6.4). In addition, several coat proteins that have homologues in *B. subtilis* have been shown to be associated with the exosporium in the *B. cereus* group.

Of the *B. cereus* group specific proteins only two proteins, BclB and BxpA, had homologues in some members of the *Geobacillus* genus (Table 6.12). A homologue of *bclB* was most commonly found in the *Geobacillus* genomes. Like BclA, BclB is also a collagen-like glycoprotein and found on the surface of the spore. However, unlike BclA, which is evenly distributed around the exosporium, BclB tends to localize to one side of the exosporium (Thompson *et al.*, 2012). The BclB homologues in *Geobacillus* contained the characteristic collagen-like XXG repeats found in BclA and BclB of *B. anthracis* (Figure 6.17A). Sequences in the N-terminus of both BclA and BclB have been shown to be important for targeting BclA and BclB to the exosporium (Thompson & Stewart, 2008). The consensus sequence of the N-terminus region in BclA and BclB was not conserved in the N-terminus region of the BclB homologues in *Geobacillus* (Figure 6.17B).

BxpA has been less well characterized, compared with BclB and there is contradictory evidence as to whether this is actually a component of the exosporium or the cortex (Steichen *et al.*, 2003, Moody *et al.*, 2010). Homologues of this protein were only found in three *Geobacillus* genomes (Table 6.12). There was no evidence that any of the *G. stearothermophilus* strains had genes encoding for exosporium proteins.

**Table 6.12. Homologues of *B. cereus* group genes encoding for exosporium proteins.**

| Gene | Description | Presence in *Geobacillus* |
|---|---|---|
| ***bclB*** | Spore surface collagen-like glycoprotein | Gc[a] group, Gtg[b] group, Gs362, GsuPSS, GtlB23, GtlCCB, GkHTA, Gk102445, GeoMAS, GeoFW2, GeoGHH, GeoJF8 |
| ***bxpA*** | Exosporium and coat protein | Gs362, GeoWCH, GeoJF8 |

[a] Includes GcCIC and Gc10776

[b] Includes GtgC56, Gtg107, GtgTNO and GeoMC1

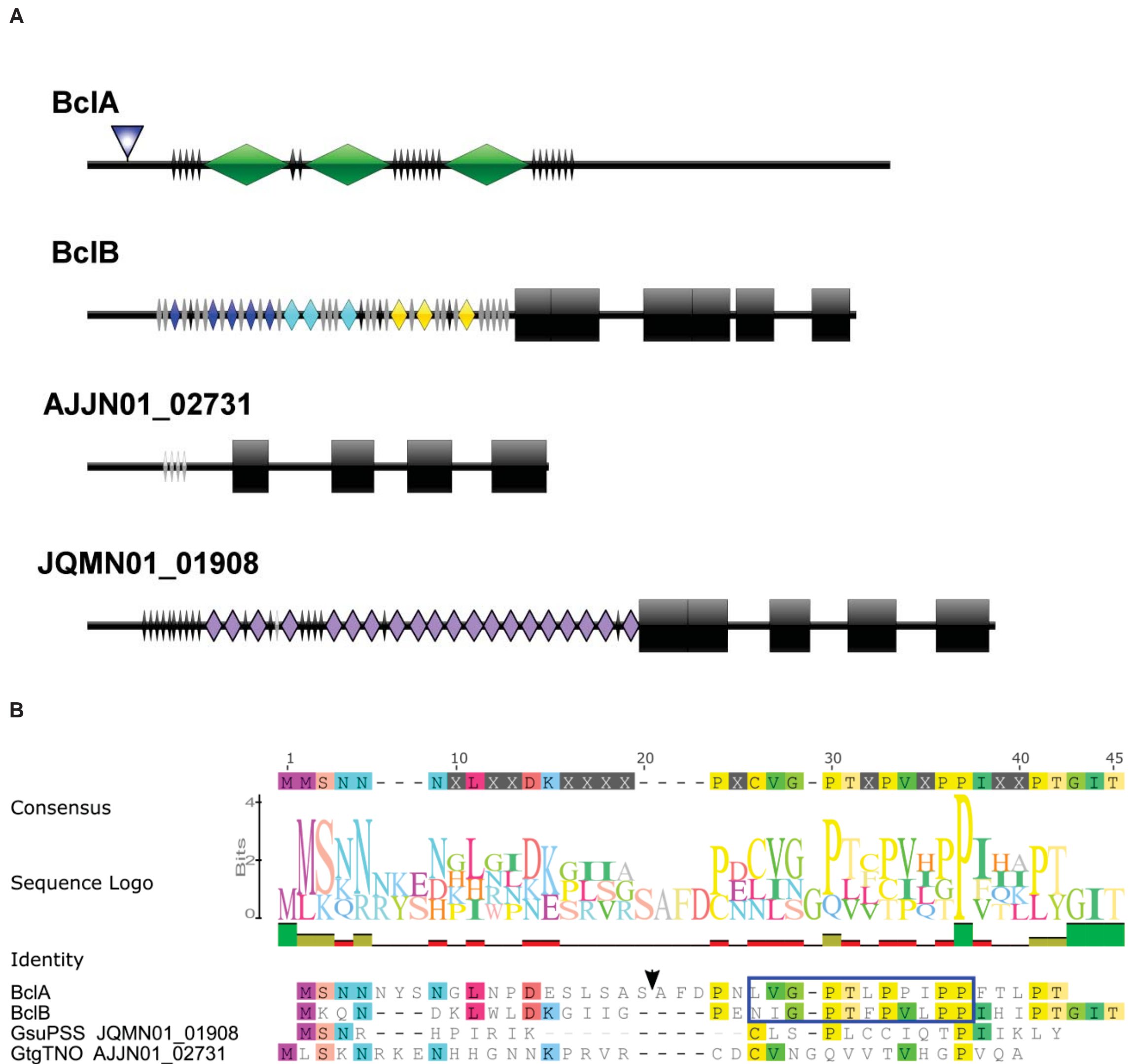



**Figure 6.17. Collagen-like glycoproteins. (A)** Schematic representation of the collagen-like proteins BclA and BclB (from *B. anthracis* Ames strain) and examples of the shortest and longest BclB homologues from two *Geobacillus* strains *G. thermoglucosidasius* TNO-09 (locus tag AJJN01_02731) and *G. subterraneus* PSS2 (locus tag JQMN01_01908) respectively. The collagen-like region consisting of triplet motifs of GXX are represented by diamonds where a black diamond represents GPT, a green diamond represents $[(GPT)_5GDTGTT]_2$, a dark blue diamond represents GATGIT, a light blue diamond represents (GPTGITGAT)2, a yellow diamond represents GLAGATGPT, a purple diamond represents GPTGPTGPA and other triplet motifs GXX are represented with a grey diamond. Additional transmembrane domains are represented by black squares and the blue triangle represents the cleavage site to generate the mature form of the BclA protein. BclB does not contain this cleavage site (Thompson & Stewart, 2008, Thompson *et al.*, 2012). **(B)** N-terminus sequence alignments of the four collagen-like proteins. Alignments were carried out using ClustalW as described in the Section 6.2.5.

### *Maturation*

The final stage of sporulation before lysis of the mother cell and release of the spore involves dehydration of the core, where water is replaced with dipicolinic acid (DPA). This is an important characteristic of spores as DPA stablises the spore as well as contributes to its resistance (Paidhungat *et al.*, 2000, Setlow *et al.*, 2006). DPA synthesis is driven by the *spoVF* operon and transportation from the mother-cell to the forespore has been shown to be dependent on at least the first five members of the *spoVA* operon (*spoVAA*, -*B*, -*C*, -*D*, -*EB*) but not on *spoVAEA* or *spoVAF* (Fort & Errington, 1985, Tovar-Rojo *et al.*, 2002, Li *et al.*, 2012, Perez-Valdespino *et al.*, 2014). In the forespore, DPA forms a complex with $Ca^{2+}$. The *spoVA* operon is also important for the release of DPA during germination (Tovar-Rojo *et al.*, 2002, Li *et al.*, 2012, Perez-Valdespino *et al.*, 2014, Setlow, 2014a).

Both the *spoVF* and *spoVA* operons were conserved in *Geobacillus*, although, *spoVAEA* was missing from three *Geobacillus* genomes. Other studies have also found *spoVAEA* missing from the genomes of some spore-formers belonging to the Bacillales and Clostridaiales orders (de Hoon *et al.*, 2010, Perez-Valdespino *et al.*, 2014), although Galperin *et al.* (2012) states *spoVAEA* has been missed from the annotation of some genomes and based on a COG analysis found it to be present in all spore-forming *Bacillaecae*.

A surprising difference between the *Geobacillus* genus and *B. subtilis* was the presence of an additional smaller *spoVA* cluster, which contained an additional *spoVAC, spoVAD* and *spoVAEB* with *spoAC* and *spoVAEB* showing closer homology (compared with the putative *spoVAC* and *spoVAEB* in the longer cluster) to the *B. subtilis spoVAC* and *spoVAEB* (Figure 6.18). de Hoon *et al.* (2010) also found that the *spoVAC* and *spoVAEB* in the smaller cluster of *G. kaustophilus* showed closer homology to the equivalent *B. subtilis* genes, but did not comment on the presence of an additional *spoVA* operon. The phylogenetic distribution of this additional *spoVA* cluster is unknown.

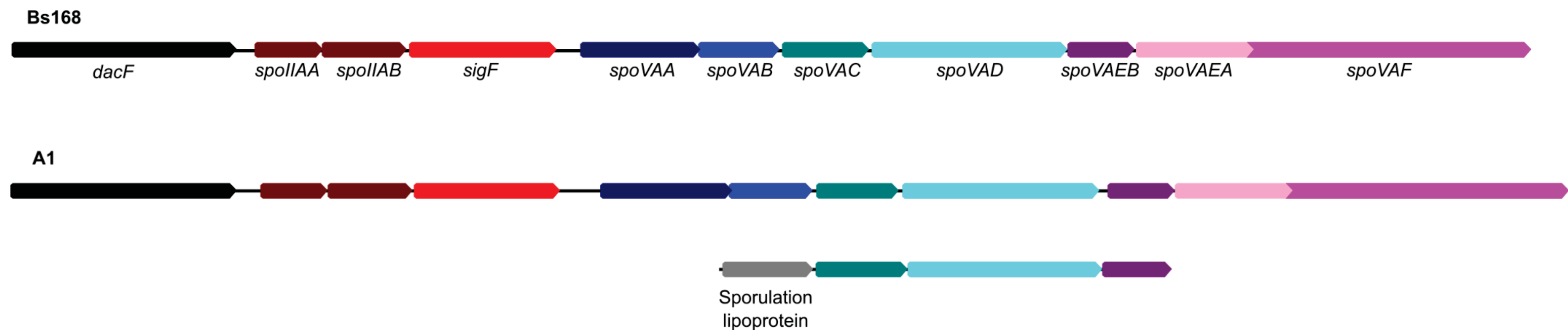



**Figure 6.18. Gene organisation of the *spoIIAA-sigF* and *spoVA* operons of *B. subtilis* and their homologues (as depicted by the different colours) in *Geobacillus*.** Strain *G. stearothermophilus* A1 was used for generation of this diagram, but the gene organisation was the same across all of the *Geobacillus* genomes analyzed, with the exception of *G. thermoglucosidasius* strains TNO, C56 and M1, which did not contain an ORF corresponding to *spoVAEA* as described above. All of the *Geobacillus* genomes contained two putative *spoVA* operons in different regions of the genome as shown in the figure. The second operon contained only four genes, with no homologue of the gene annotated as “sporulation lipoprotein” found in *B. subtilis*.

Could the additional putative *spoVA* operon contribute to the increased spore resistance of *Geobacillus* species? *Geobacillus* endospores are more heat resistant compared with those of *B. subtilis* (Mostert *et al.*, 1979, Burgess *et al.*, 2010), with *Geobacillus* spore counts being carried out after a heat-treatment of 100 °C as opposed to 70 – 80 °C for *B. subtilis* and other mesophilic *Bacillus* species. Sporulation at a higher temperature does result in lower core water content and therefore a higher resistance (Nicholson *et al.*, 2000) and $Ca^{2+}$-DPA uptake also contributes to the low core water content (Nicholson *et al.*, 2000, Setlow, 2006). SpoVAD, which is a soluable protein, as opposed to SpoVAC and SpoVAEB, which are integral membrane proteins, is able to bind to Ca-DPA (Li *et al.*, 2012). In the same study it was suggested that the other SpoVA proteins may form a complex with SpoVAD for DPA transport into the forespore. The role of the other SpoVA proteins and the mechanisms for uptake are yet to be determined. The additional putative *spoVA* cluster may also play a role in germination. The SpoVA proteins are also involved in the dispersion of Ca-DPA during germination (Vepachedu & Setlow, 2004, Vepachedu & Setlow, 2007a, Vepachedu & Setlow, 2007b). SpoVAC has properties similar to those found in other channel forming proteins and has been proposed to act as a channel to release the $Ca^{2+}$-DPA (Velasquez *et al.*, 2014). Both SpoVAD and SpoAEA have been shown to interact with germination receptors (Vepachedu & Setlow, 2007a).

In the smaller *Geobacillus spoVA* cluster, was an additional gene that was annotated as "Sporulation lipoprotein YhcN/YlaJ". The predicted protein encoded by this gene contained the Pfam Spore_YhcN_YlaJ domain (PF09580). Proteins containing this domain are predicted lipoproteins that have been detected as spore proteins but not vegetative proteins in *B. subtilis* and as indicated from the name this includes the proteins YhcN and YlaJ (Bagyan *et al.*, 1998, Kuwana *et al.*, 2002). The expression of these genes is believed to occur in the forespore and under the control of $\sigma^G$. The function of both these proteins in *B. subtilis* still remains unknown. A mutation in the *yhcN* gene has no effect on sporulation efficiency, resistance or germination but does show a slight defect in outgrowth (Bagyan *et al.*, 1998). It has been suggested that YhcN is related to the SASP proteins based on it containing a short sequence similar to the sequence in the α/β SASP types that is important for the break-down of SASPs during germination (Setlow, 1988, Bagyan *et al.*, 1998). YlaJ has been extracted from spores but to the author's knowledge no functional studies have been carried out (Kuwana *et al.*, 2002). YlaJ was not included in the initial analysis of this present study as it did not meet the criteria of being described in at least two studies. An additional check found that all of the *Geobacillus* genomes contained a *ylaJ* homologue but this was in a different location to the additional *spoVA* cluster. Despite all of the *Geobacillus* genomes containing multiple ORFs located in different regions of the genome that were annotated as the "Sporulation lipoprotein YhcN/YlaJ" there were no homologues of *yhcN* as defined by the BDBH algorithm. The presence of multiple ORFs in the *Geobacillus* genomes annotated as "Sporulation lipoprotein YhcN/YlaJ" suggests that they may play an important role is forespore development of *Geobacillus* spp.

### *DNA protection*

Spores contain two groups of SASPs: the α/β type and the γ type, both of which are expressed during sporulation. One of the main roles of the α/β type is to protect the DNA. The role of the γ type is less clear as although they are expressed in the forespore they do not localize to the core (Setlow, 1988). However, it may be involved with coat formation and both types play a role in germination (Setlow, 1988, Ruzal *et al*., 2013). Approximately 2 – 3 h into sporulation α/β type SASPs bind to the DNA resulting in protection of the DNA (Setlow, 1995). SASPs have been purified from a range of endospore forming bacteria. All of the members of the *Bacillaceae* family tested, including '*B. stearothermophilus*', contained the two major types of SASPs (Setlow, 1988). To date there are 17 genes known to encode for SASPs in *B. subtilis*, and with the exception of *sspG*, they are all expressed in the forespore. The three main types of SASPs, α, β and γ, are encoded by *sspA*, *sspB* and *sspE* respectively. The remaining types are all minor SASPs. In the *Geobacillus* genomes analyzed there were no homologues found of *sspA*, *sspC*, *sspG*, *sspJ*, *sspL* or *sspM*. However, based on the Prokka annotations an additional 2 or 3 genes encoding SASPs were located that belonged to the same family as *sspB* or *sspH* (Figure 6.19). One genome, *Geobacillus* WCH70, did not contain an *sspO* homologue, instead a putative transposase was present.

The number of SASPs identified in this present study did differ from the number of SASPs identified in *Geobacillus* species by previous studies (de Hoon *et al*., 2010, Galperin *et al*., 2012). For example, de Hoon *et al.* (2010) found *sspE* to be present in *G. kaustophilus* HTA426, but this present study did not find a homologue of *sspE* in the same strain. This is probably because of the less stringent criteria used by de Hoon *et al.* (2010) compared with this present study. However, in this present study, the Prokka annotations identified this additional SASP encoding gene. Galparin *et al.* (2012) found that the number of SASP genes varies between species and generally ranges from 11 – 22 in spore-formers from the *Bacillaceae* family. The authors suggest that a diverse range of SASPs is not required for efficient sporulation. The sequence similarity between the minor SASP genes is low and these proteins are very short so it is also possible that some novel SASPs in the *Geobacillus* genus may have been missed.

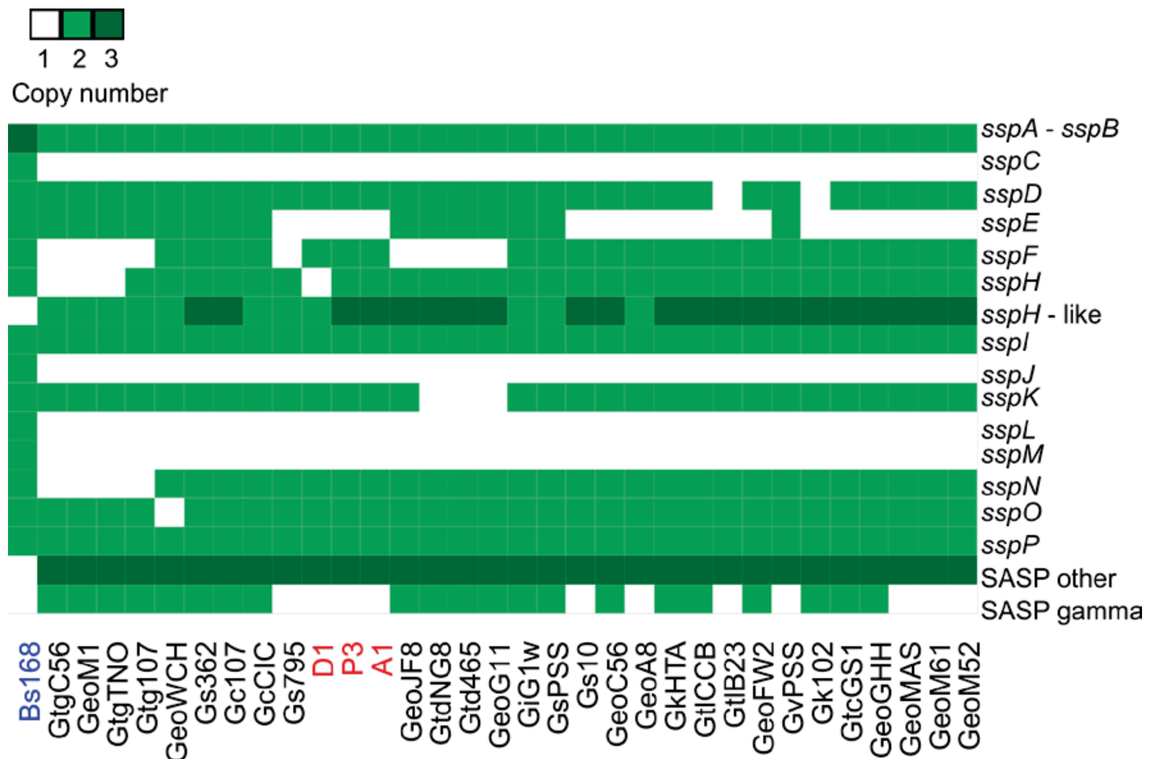



**Figure 6.19. Presence/absence of genes encoding for SASPs.** This heat-map shows the presence or absence of homologues of the *B. subtilis ssp* genes in the *Geobacillus* genus as well as other genes encoding for SASPs (SASP other and SASP gamma) as determined by the Prokka annotations. Refer to Section 6.2.3 for determination of a homologue.

***Mother cell lysis***

Once the endospore has formed the mother cell lyses and releases the spore. In *B. subtilis* two amidases CwlH and CwlC lyse the mother cell wall and the DNase NucB degrades the DNA released upon cell lysis (Nugroho *et al.*, 1999, Hosoya *et al.*, 2007). None of the *Geobacillus* genomes contained homologues of NucB or CwlH; however, other proteins in the same COG may be able to fulfill the same function (Figure 6.20 and Table 6.13).

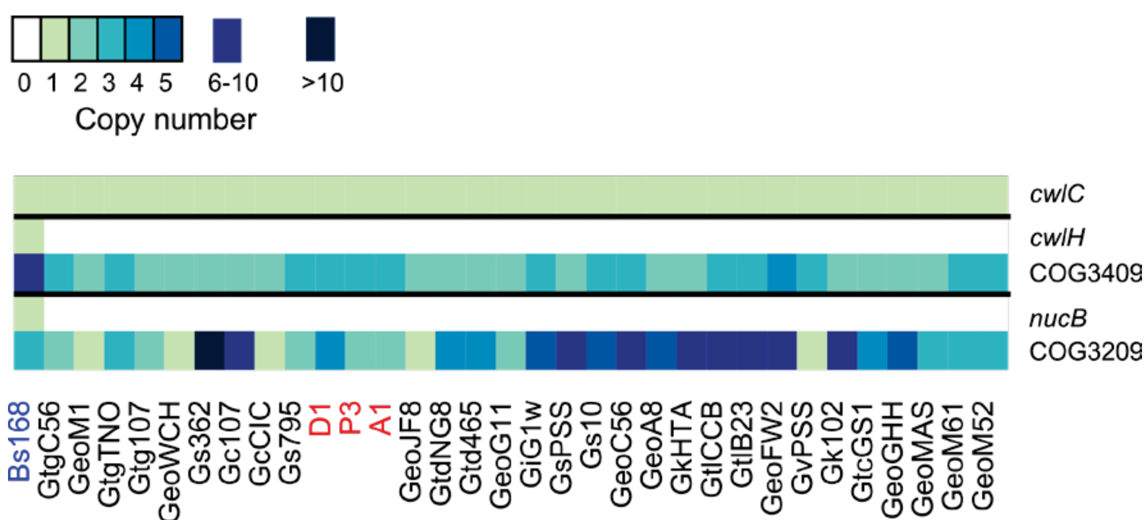



**Figure 6.20. Conservation of *B. subtilis* mother-cell lysis genes in *Geobacillus* spp. and their equivalent COGs.** This heat-map shows genes involved in mother-cell lysis of *B. subtilis* and the presence or absence of homologues in the *Geobacillus* genus. Refer to Section 6.2.3 for determination of a homologue. The COG associated with each gene(s) is in the row below and is separated from the next gene and its associated COG by a black line. The COG analysis was only included for *cwlH* and *nucB* which were absent in the *Geobacillus* genus as defined by the BDBH algorithm.

**Table 6.13. Description of COGs which contain proteins involved with mother-cell lysis.**

| COG[a] | Description[b] | Gene names | |
|---|---|---|---|
| | | Spore | Other |
| **COG3209** | Uncharacterized conserved protein RhaS, contains 28 RHS repeats | *nucB* | *nucA, wapA* |
| **COG3409** | Glycogen debranching enzyme (alpha-1,6-glucosidase) | *cwlH, sleB* | *csn, cwlA, fadG, xlyA, ybfG, ykzU, ykuH* |

[a] Refer to CD: Chapter4/Results/COG/COG_LocusTag.xlsx for a list of locus tags within each COG

[b] The descriptions originate from the COG database (Tatusov *et al.*, 2000, Galperin *et al.*, 2015)

[c] These genes listed are not known to be involved in mother-cell lysis but they belong to the same COG.

### 6.3.4 Summary of results

In summary, the key components of the phosphorelay and sporulation pathways of *B. subtilis* were well conserved in *Geobacillus*; whereas, components of the biofilm pathway were less well conserved as described in Table 6.14. There were some differences between the different *Geobacillus* taxa, but no notable differences were found between the three dairy strains that could account for the phenotypic differences seen in Chapter 3.

**Table 6.14. Summary of the key components of the phosphorelay, biofilm and sporulation pathways conserved in *Geobacillus* spp.**

| | Process | Genes | Presence/absence[a] in different *Geobacillus* taxa[b] compared with *B. subtilis* 168 (Bs) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Bs | 1. Gtl / Gka | 3. Gs | 4. Gic / Gsu | 6. Gtd | 7. Gc | 8. Geo 362 | 10. Gtg |
| Phosphorelay | Orphan HKs | *kinA, kinB, kinC, kinD, kinE* | 5 | 5-6 | 6 | 6 | 9 | 9 | 9 | 5 |
| | Control of kinases | *kapD, kbaA, kipI,sda* | P | P | P | P | P | P | P | P |
| | | *kapB, med* | P | A | A | A | A | A | A | A |
| | Phosphorelay | *spo0F, spo0B, spo0A* | P | P | P | P | P | P | P | P |
| | Dephosphorylation | Spo0A-P phosphatase: *spo0E, yisI or ynzD* | P | P | P | P | P | P | P | P |
| | | Response regulator of Spo0F-P*: rap* genes | P | SP | SP | SP | SP | SP | SP | SP |
| | | Inhibitor of Rap proteins*: phr* genes | P | A | A | A | A | A | A | A |
| Biofilm | Regulation | Two component system: *comA/comP* | P | V | A | P | A | A | A | P |
| | | Two component system: *degU/degS* | P | P | P | P | P | P | P | P |
| | | Transcriptional regulators: *slrR* or *sinR* | P | A | A | A | V | V | A | V |
| | | Antagonists: *slrA* or *sinI* | P | A | A | A | A | A | A | A |
| | | Global regulator *ymdB* | P | P | P | P | P | P | P | P |
| | Matrix – exopolysaccharide | *epsABCDEFGHIJKLMNO* | P | V 2-8[c] | V 2-3[d] | V 1-4 | V 4 | V 3-5 | SP 4 | V 3-8[c] |
| | Matrix – protein | *tapA – sipW – tasA* | P | A | A | A | A | A | SP | A |
| | Matrix - hydrophobicity | *bslA* | P | A | A | A | A | A | A | A |
| | Matrix – other | *pgsABC* | P | SP[d] | SP[d] | SP[d] | SP[d] | SP[d] | SP[d] | SP[d] |
| Sporulation | Regulation | *sigE, sigK, sigG, sigF, spoIIID, gerE, spoVT, rsfA, gerR* | P | P | P | P | P | P | P | P |
| | Regulation – additional transcriptional factors | *rsfA/gerR like* | A | P | P | P | P | P | P | P |
| | | *gerE* like | A | P | V[e] | P | A | A | A | A |
| | Assymetric division | *divIB, divIC, divIVA, ftsA, ftsH, ftsZ, racA, spoIIE, spoIIIE* | P | P | P | P | P | P | P | P |
| | Engulfment | *spoIID, spoIIIA* operon, *spoIIIJ, spoIIM, spoIIP* and *spoIIQ* | P | P | P | P | P | P | P | P |
| | Cortex | Refer to Figure 6.13 | P | MP | MP | MP | MP | MP | MP | MP |
| | Coat | Refer to Figure 6.15 | P | SP | SP | SP | SP | SP | SP | SP |
| | Exosporium | *bclA* | A | A | A | A | A | A | A | A |
| | | *bclB* | A | V | A | V | A | P | A | P |
| | DNA protection | *ssp* genes | P | MP | MP | MP | MP | MP | MP | MP |
| | Maturation | *spoVF* operon and *spoVA* operon | P | P | P | P | P | P | P | P |
| | | Additional putative spoVA operon | A | P | P | P | P | P | P | P |
| | Lysis – amidases | *cwlC* | P | P | P | P | P | P | P | P |
| | | *cwlH* | P | A | A | A | A | A | A | A |
| | Lysis – DNAse | *nucB* | P | A | A | A | A | A | A | A |

[a] The filled colours and letters refer to the following: green P, present; light green MP, most homologues present; cream SP, some homologues present; orange V, variable, where the number of homologues between strains within a taxon vary; red A, absent. Where there are numbers, this gives the number of homologues

[b] Refer to Table 6.2 for taxonomic groupings. Taxa have only been included that contain more than one strain, with the exception of Geobacillus sp. 3621, which differs from all the other strains in some cases.

[c] Large putative *eps* operons were found in strains GtlCCB, GeoC56 and GtgTNO.

[d] Two homologues were found in A1, P3 and Gs795 and three homologues in D1.

[e] Homologues of the *pgsABC* operon were only found of *pgsA* in all of the *Geobacillus* strains

[f] Additional *gerE* like genes were only found in strains A1and P3 for the *G. stearothermophilus* taxon

### 6.3.5 Reliability of results

This study used three approaches for identifying genes that might be involved in biofilm formation and sporulation of *Geobacillus* spp. A multi-faceted approach was used as each of the methods used has its own advantages and disadvantages. BDBH relies on defining a homologue by setting limits e.g. for coverage, percentage similarity and e-value. Initially, in this present study, when using the BDBH algorithm higher limits were set for coverage and percentage similarity (compared with those described in Section 6.2.3), but this resulted in some homologues being missed. Despite a low limit being set, it is possible that some more divergent homologues may have been missed. This can be overcome using a COG approach, which identifies proteins of similar function based on multiple PSI-BLAST searches as described in previous studies (Tatusov *et al.*, 2000, Makarova *et al.*, 2006, Galperin *et al.*, 2012). In some cases the function of the proteins in a particular COG group was very specific, which resulted in a small number of proteins in that group e.g. COG5337, which contained YisI and CotH. In other cases the COG group was very broad and contained a large number proteins all with a similar function but involved with different processes as seen in this present study and in the study by Galperin *et al.* (2012). For example, the spore surface polysaccharide protein SpsD belongs to COG0456, which contains over 20 proteins, with only two identified as being involved in sporulation. Another disadvantage of the COG approach is that some proteins (e.g. the SASPs) do not belong to a COG. In addition, the COG approach does not appear to be comparable between studies as the number of proteins assigned to a particular COG sometimes differed between this present study and the study by Galperin *et al.* (2012). For example, Galperin *et al.* (2012) assigned 18 proteins of *B. subtilis* to COG0456 whereas this present study assigned 26 proteins. These differences could also be due to difference COG datasets being used. In the study by Galperin *et al.* (2012) an older COG database was used as described by Tatusov *et al.* (2003), whereas this present study used a newer version as described by Galperin *et al.* (2015).

The BDBH and COG analyses both relied on the presence of an ORF predicted to encode a protein. In this present study, to ensure that all of the genomes were treated in the same way, all of the genomes were annotated using the program Prokka. However, it appears that Prokka could miss annotating short ORFs. This was seen with the *spoVM* gene, which encodes a very short protein of less than 30 aa, where it was only annotated in two of the *Geobacillus* genomes (*G. icigianus* G1w1 and *G. subterraneus* PSS2) by Prokka, hence homologues of the *B. subtilis spoVM* were not found in the other genomes using the BDBH algorithm. Short ORFs were

located in all of the *Geobacillus* genomes with a high similarity to *B. subtilis spoVM*. According to Galperin *et al.* (2012) this gene is often missed in genome annotations. The requirement of having an ORF predicted to encode a protein also means that split genes such as *sigK* are also missed. Split genes do not appear to be a common phenomenon but other sporulation genes have been found to be split including *spsN* in *B. subtilis* 168 (Abe *et al.*, 2014) and *spoVFB* in *Bacillus weihenstephanensis* (Abe *et al.*, 2013). Another disadvantage of using Prokka to re-annotate all of the genomes is that draft genomes can not be easily compared with those published by NCBI.

Comparative genomic analyses are sensitive to good annotations, making it difficult to make comparisons between studies. For example, it was difficult to make comparisons between this present study and the study by Galperin *et al.* (2012) given no locus tag or positon on the genome was provided for the homologues identified. For example, they claim that *G. kaustophilus* and *G. thermodenitrificans* contained four newly identified *cotD* genes. In this present study no *cotD* homologues were identified in any of the *Geobacillus* genomes. Given no locus tags or genome coordinates were provided no comparison could be made. In addition, although Galperin *et al.* (2012) listed nine *Geobacillus* genomes as being used in their study, only seven were included in their COG analysis and three of these were named similarly, so could not be distinguished.

Prokka annotations were also analyzed to determine putative sporulation genes as some predicted proteins may be missed by the first two approaches because they are paralogues (in the case of the BDBH algorithm), they do not belong to a COG or they do not have homologues in *B. subtilis*. The present study predominantly relied on similarity between *Geobacillus* and the model organism *B. subtilis*. The core components of the phosphorelay and stages I – IV of sporulation were well conserved and the later stages less well conserved. For example, *Geobacillus* spp. had a low number of coat protein homologues and is likely to contain additional unidentified coat proteins that do not contain homologues in *B. subtilis*. Analysis of the Prokka annotations was also carried out to identify genes that may not be found in *B. subtilis*. However, this approach relied on previously characterized proteins from other organisms that are present in the Prokka database. A transcriptomics approach would be needed to identify additional genes expressed under certain conditions e.g. when the bacterial cell sporulates. This present study can provide genomic insights into how *Geobacillus* might form biofilms and sporulate but it presents no information on whether a protein is actually functional.

## 6.4 Conclusions

This study was carried out to determine whether a genomics approach could be used to provide insights into the possible molecular mechanisms of biofilm formation and sporulation of *Geobacillus* spp., in particular, of the three dairy strains A1, P3 and D1. The model organism *B.*

*subtilis* has been well characterized for biofilm formation and sporulation; hence it was used as the main reference in this study.

In *B. subtilis*, biofilm formation and sporulation is initiated by the master regulator Spo0A via a phosphorelay. The core components of the phosphorelay were conserved; however, differences in the sensor HKs and some of the proteins required for controlling the phosphorelay suggests that the signals sensed by *Geobacillus* spp. and possibly some of the downstream processes regulated by the Spo0A master regulator could differ. Importantly, no KinD homologue was found in the *Geobacillus* spp., indicating the regulatory switch between these two processes is probably different.

Genes encoding regulatory proteins were well conserved, but structural components of both biofilms and to a lesser extent spores were not well conserved. Given the molecular mechanisms of biofilm formation vary between species and even between some strains of the same species the approach taken in this present study of gaining a genomic insight into biofilm formation of *Geobacillus* spp. was limited. The *Geobacillus* genus had homologues of some of the regulatory genes, but many of the genes required for matrix production were absent. In *B. subtilis* the matrix is composed of both exopolysaccharide and proteins (TasA). The number of putative *eps* genes varied between the *Geobacillus* strains, with the dairy strains containing the least number of homologues, and no *Geobacillus* strains contained a *tasA* homologue. This suggests the molecular mechanisms and composition of the matrix is different in *Geobacillus* spp.

The core components of the sporulation pathway were conserved in *Geobacillus* indicating that *Geobacillus* spp. undergo the same major steps as *B. subtilis* for sporulation to occur. Some of the structural components, particularly the coat proteins, were not well conserved. A key difference between *B. subtilis* and *Geobacillus* was the presence of an additional putative *spoVA* operon in all of the *Geobacillus* genomes analyzed. Based on the role of the *spoVA* operon in *B. subtilis* it is hypothesized that this additional putative operon may play a role in additional spore resistance of these thermophilic bacilli. Analysis of the genomes of the three dairy strains could provide no hypotheses for differences in both biofilm formation and sporulation as seen in Chapter 3 between these three strains. A transcriptomics approach would need to undertaken to determine which genes are involved, as well as the effect of milk and milk components on biofilm formation and sporulation.

Chapter 7

# 7 Conclusions and summary

This is the first study of its kind to examine both the phenotypic diversity of strains of thermophilic bacilli within a milk powder manufacturing plant as well as as gaining a genomics insight into how strains of *G. stearothermophilus* have the ability to survive and grow within a dairy environment.

## 7.1 Strain diversity within a dairy manufacturing plant

In Chapter 2, strains of thermophilic bacilli were isolated from ten different locations in the evaporator section of a milk powder manufacturing plant, with *G. stearothermophilus* being the dominant species that was isolated. Ten strains from each sample location were further characterised. It was hypothesized that these strains would have different phenotypic characteristics. This was found to be true where biochemical assays and fatty acid profiling were able to distinguish between certain *G. stearothermophilus* dairy strains that may represent various sources of contamination. Alternatively, there could be one source of diverse contamination early in the manufacturing process that might be detected with a larger pool of isolates that would allow us to group isolates based on the profiling methods used. It was also found that all of the dairy strains of *G. stearothermophilus* that were analyzed could grow solely on lactose, compared with the type strain of *G. stearothermohilus* ATCC 12980, which cannot grow on lactose. Lactose is the predominant sugar found in milk. Until this present study was carried out, the species *G. stearothermophilus* was described as a non-lactose producer.

The ability to produce spores and form biofilms, are two characteristics of the thermophilic bacilli which enable them to survive and grow within a milk powder manufacturing environment. In Chapter 3, it was found that the ten strains of *G. stearothermophilus* varied in their ability to form biofilms and produce spores. Three strains, A1, P3 (both isolated from evaporator 1) and D1 (isolated from evaporator 2), were selected for further analysis and it was found there were differences in biofilm morphology. The variability in the biofilm structure of *G. stearothermophilus* dairy isolates has not been previously reported. In addition, strains A1 and P3 were found to be greater acid producers compared with D1. The importance of how this biofilm variability could influence the control of *G. stearothermophilus* biofilms and the effect of acid production on biofilm and foulant formation in the dairy industry needs to be determined.

To determine whether any phenotypic differences could be linked to genotypic differences the genomes were sequenced of the three selected strains: A1, P3 and D1, as described in Chapter 4. Although, phenotypic differences could not be linked to genotypic differences between the three strains A1, P3 and D1, there were genotypic differences noted by the analysis of putative gene function using a COG approach as well as differences in the CRISPR arrays that showed

strain D1 was different to strains A1 and P3. In Chapter 5, core genome sequence comparisons provided further evidence that D1 was different from the other two strains A1 and P3.

## 7.2 Taxonomic classification

Given the phenotypic diversity seen in Chapter 2, phylogenomic comparisons were carried out to confirm that the three selected strains of *G. stearothermophilus* did belong to the *G. stearothermophilus* taxon. The three dairy strains A1, P3 and D1 grouped within the same taxon as *G. stearothermophilus* ATCC 7953 and the type strain ATCC 12980 as determined by rMLST, ANI and core genome comparisons in Chapter 5. Other strains named as *G. stearothermophilus* (*G. stearothermophilus* 10 and *G. stearothermophilus* 3621) were not found to be members of this taxon. In addition, this is the first study to show that a phylogenomics approach is able to resolve differences between species of the *Geobacillus* genus that cannot be resolved using the traditional approach of 16S rRNA gene sequence analysis or analysis of other house-keeping genes. Based on three different genomic approaches (rMLST, ANI and core genome comparisons), the *Geobacillus* genus could be divided into ten species for those *Geobacillus* strains that have genome sequences available.

The *G. stearothermophilus* taxon is also unique in that they have smaller genomes (2.6 – 3.0 Mb) compared with other *Geobacillus* spp. that have had their genomes sequenced. Gene loss may have occurred in the *G. stearothermophilus* taxon during evolution from the common ancestor of the *Geobacillus* genus. There is evidence of gene loss by the absence of clusters of genes (as opposed to single genes) that were involved in various functions, such as carbohydrate metabolism.

## 7.3 Genomics insight into growth and survival in a dairy manufacturing environment

The genome sequences of the three selected strains were further analyzed to determine if the genome could provide any further insight into the ability of certain *G. stearothermophilus* strains to grow and survive in a dairy environment; in particular, the ability of dairy strains of *G. stearothermophilus* to grow in milk (and utilization of lactose as a sole carbon source), form biofilms and produce highly heat-resistant spores.

In Chapter 4, the genome sequences of the three dairy strains were compared with other *Geobacillus* genomes to determine whether there was any gene gain or loss that could indicate dairy adaptation. This is the first study to provide evidence of dairy adaptation in a *Geobacillus* strain by the presence of a putative lactose operon in the dairy strains (A1, P3 and D1) that was not found in the other *Geobacillus* genomes analyzed. This conclusion was drawn despite the fragmented nature of the draft genome sequences. Other evidence of dairy adaptation, for example the presence of genomic islands and a more accurate number of pseudogenes, may have been lost due to the fragmented nature of the genomes.

The molecular mechanisms of biofilm formation and sporulation in *Geobacillus* spp is difficult to determine; however, comparative genomics with the model organism *B. subtilis*, which has been

well characterized for biofilm formation and sporulation, was used as the main reference to identify potential genes involved in biofilm formation and sporulation of *Geobacillus* spp. Although this approach does not provide any functional information, it gives an insight into potential mechanisms. For example, it was found that the core components of the phosphorelay, which are required for initiation of biofilm formation and sporulation in *B. subtilis*, are well conserved in the *Geobacillus* genus. However, differences in the sensor histidine kinases and some of the proteins required for controlling the phosphorelay suggests that the signals sensed by *Geobacillus* spp. and possibly some of the downstream processes regulated by the Spo0A master regulator could differ. Genes encoding regulatory proteins are well conserved, but structural components of both biofilms and to a lesser extent spores are not well conserved. Given the molecular mechanisms of biofilm formation vary between species and even between some strains of the same species, the approach taken in this present study of gaining a genomic insight into biofilm formation of *Geobacillus* spp. was limited. The *Geobacillus* genus has homologues of some of the regulatory genes, but many of the genes required for matrix production are absent.

The core components of the sporulation pathway are conserved in *Geobacillus* indicating that *Geobacillus* spp. undergo the same major steps as *B. subtilis* for sporulation to occur. Some of the structural components, particularly the coat proteins, are not well conserved. A key difference between *B. subtilis* and *Geobacillus* is the presence of an additional putative *spoVA* operon in all of the *Geobacillus* genomes analyzed. Based on the role of the *spoVA* operon in *B. subtilis,* it is hypothesized that this additional putative operon may play a role in additional spore resistance of these thermophilic bacilli or alternatively a novel mechanism for spore germination. Determination of the function of this additional *spoVA* operon could provide a knowledge platform for understanding how *Geobacillus* spores form highly heat resistant spores or germinate. Finally, analysis of the genomes of the three dairy strains could provide no hypotheses for phenotypic differences in both biofilm formation and sporulation between these three strains.

## 7.4 Summary

In summary, this study has shown that strains of *G. stearothermophilus* isolated from the same milk powder manufacturing plant can vary, phenotypically and genotypically. A comparative genomics approach was able to identify differences between the three selected dairy strains, but these could not be linked to phenotype. There was some evidence of dairy adaptation, but this needs to be clarified using a complete genome. It was difficult to gain any insight into the molecular mechanisms of biofilm formation of *Geobacillus* spp using a genomics approach; however, the sporulation pathway appeared to be well conserved with that of *B. subtilis*. Further work is required (as detailed in Section 7.6) to understand how these newly sequenced strains of *G. stearothermophilus* have adapted to a dairy environment, the molecular mechanisms of biofilm formation and to clarify differences in genes involved in sporulation between *B. subtilis* and the *Geobacillus* genus.

## 7.5 Practical contributions to the dairy industry

Highlights from this study that could have practical implications for the dairy industry are:

1. The diverse nature of strains of *G. stearothermophilus* within one milk powder manufacturing plant may have implications on cleaning regimes used in dairy manufacturing plants, where one type of cleaning regime may not be as effective on all strains of the same species. This highlights the need for using multiple strains when evaluating new cleaning regimes.
2. Acid production by strains of *G. stearothermophilus* varies, which may have consequences on the ability of a strain to form a biofilm and could contribute to the presence of milk foulant on equipment surfaces.
3. Adaptation of strains of *G. stearothermophilus* to a dairy environment. Further analysis on when and how quickly this adapation occurs could provide a knowledge platform for developing novel strategies for the control of thermophilic bacilli in dairy manufacturing plants.
4. Identification of an additional *spoVA* operon. Characterization of the proteins encoded for by this operon may provide information on the development of highly heat resistant spores or information on the germination mechanisms of the *Geobacillus* genus.

## 7.6 Future directions

**Applied:**

- Evaluation of different cleaning regimes on different strains of thermophilic bacilli that have different biofilm morphologies
- Determination of the effect of pH on biofilm formation of *G. stearothermohilus*
- Effect of CIP chemicals on spore resistance

**Fundamental:**

- Generation of a complete genome of one of the dairy strains of *G. stearothermophilus* using PACBio sequencing
- Determination of whether dairy adaptation occurs in the same way in other species of thermophilic bacilli (e.g. *A. flavithermus*) by genome sequencing and comparison of genomes with other strains
- Establishing a link between phenotype and genotype using the web-based program Phenolink, using a wider range of *G. stearothermophilus* strains isolated from a variety of envrionments and a wider range of biochemical assays
- Identification of genes involved in spore germination

- Use RNA sequencing to identify genes upregulated during growth in milk, biofilm formation and sporulation
- Identify additional genes that may be involved with sporulation, including the phosphorelay by carrying out a promoter search to identify putative members of the Spo0A, SigE, SigK, SigG and SigF regulons
- Expression of the additional *spoVA* operon, purification and characterisation of the SpoVA proteins to determine potential function *in-vitro*

# 8 Bibliography

Abe K, Yoshinari A, Aoyagi T, Hirota Y, Iwamoto K & Sato T (2013) Regulated DNA rearrangement during sporulation in *Bacillus weihenstephanensis* KBAB4. *Molecular Microbiology* **90**: 415-427.

Abe K, Kawano Y, Iwamoto K, Arai K, Maruyama Y, Eichenberger P & Sato T (2014) Developmentally-regulated excision of the SPb prophage reconstitutes a gene required for spore envelope maturation in *Bacillus subtilis*. *PloS Genetics* **10**: doi:10.1371/journal.pgen.1004636.

Abecasis AB, Serrano M, Alves R, Quintais L, Pereira-Leal JB & Henriques AO (2013) A genomic signature and the identification of new sporulation genes. *Journal of Bacteriology* **195**: 2101-2115.

Acinas SG, Marcelino LA, Klepac-Ceraj V & Polz MF (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of Bacteriology* **186**: 2629-2635.

Ackman RG & Sipos JC (1964) Application of specific response factors in the gas chromatographic analysis of methyl esters of fatty acid with flame ionization detectors. *Journal of the American Oil Chemists Society* **41**: 377-378.

Aguilar C, Vlamakis H, Guzman A, Losick R & Kolter R (2010) KinD is a checkpoint protein linking spore formation to extracellular-matrix production in *Bacillus subtilis* biofilms. *mBio* **1**: doi: 10.1128/mBio.00035-00010.

Ahmad I, Lamprokostopoulou A, Le Guyon S, Streck E, Barthel M, Peters V, Hardt W-D & Romling U (2011) Complex c-di-GMP signaling networks mediate transition between virulence properties and biofilm formation in *Salmonella enterica* serovar typhimurium. *PLoS One* **6**: doi:10.1371/journal.pone.0028351.

Ahmad S, Scopes RK, Rees GN & Patel BKC (2000) *Saccharococcus caldoxylosilyticus* sp nov., an obligately thermophilic, xylose-utilizing, endospore-forming bacterium. *International Journal of Systematic and Evolutionary Microbiology* **50**: 517-523.

Al-Hinai MA, Jones SW & Papoutsakis ET (2015) The *Clostridium* sporulation programs: Diversity and preservation of endospore differentiation. *Microbiology and Molecular Biology Reviews* **79**: 19-37.

Alcaraz LD, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L & Olmedo G (2010) Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics* **11**.

Aleksandrzak-Piekarczyk T, Kok J, Renault P & Bardowski J (2005) Alternative lactose catabolic pathway in *Lactococcus lactis* IL1403. *Applied and Environmental Microbiology* **71**: 6060-6069.

Allegrucci M, Hu FZ, Shen K, Hayes J, Ehrlich GD, Post JC & Sauer K (2006) Phenotypic characterization of *Streptococcus pneumoniae* biofilm development. *Journal of Bacteriology* **188**: 2325-2335.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**: 403-410.

Amitai G & Sorek R (2016) CRISPR-Cas adaptation: insights into the mechanism of action. *Nature Reviews Microbiology* **14**: 67-76.

Anantharaman V & Aravind L (2000) Cache- a signaling domain common to animal $Ca^{2+}$ channel subunits and a class of prokaryotic chemotaxis receptors. *Trends in Biochemical Sciences* **25**: 535-537.

Anantharaman V & Aravind L (2003) Application of comparative genomics in the identification and analysis of novel families of membrane-associated receptors in bacteria. *BMC Genomics* **4**.

Anantharaman V, Koonin EV & Aravind L (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *Journal of Molecular Biology* **307**: 1271-1292.

Ankolekar C & Labbe RG (2010) Physical characteristics of spores of food-associated isolates of the *Bacillus cereus* group. *Applied and Environmental Microbiology* **76**: 982-984.

Aravind L & Ponting CP (1999) The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. *FEMS Microbiology Letters* **176**: 111-116.

Archer NK, Mazaitis MJ, Costerton JW, Leid JG, Powers ME & Shirtliff ME (2011) *Staphylococcus aureus* biofilms Properties, regulation and roles in human disease. *Virulence* **2**: 445-459.

Aronson AI & Fitzjames P (1976) Structure and morphogenesis of bacterial spore coat. *Bacteriological Reviews* **40**: 360-402.

Ash C, Farrow J, Wallbanks S & Collins M (1991) Phylogenetic heterogeneity of the genus *Bacillus* revealed by comparative analysis of small subunit ribosomal RNA sequences. *Letters in Applied Microbiology* **13**: 202-206.

Atrih A & Foster SJ (2001) Analysis of the role of bacterial endospore cortex structure in resistance properties and demonstration of its conservation amongst species. *Journal of Applied Microbiology* **91**: 364-372.

Aziz RK, Bartels D, Best AA*, et al.* (2008) The RAST server: Rapid annotations using subsystems technology. *BMC Genomics* **9**: doi:10.1186/1471-2164-1189-1175.

Babu MM, Teichmann SA & Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of Molecular Biology* **358**: 614-633.

Bagyan I, Noback M, Bron S, Paidhungat M & Setlow P (1998) Characterization of *yhcN*, a new forespore-specific gene of *Bacillus subtilis*. *Gene* **212**: 179-188.

Banse AV, Hobbs EC & Losick R (2011) Phosphorylation of Spo0A by the histidine kinase KinD requires the lipoprotein Med in *Bacillus subtilis*. *Journal of Bacteriology* **193**: 3949-3955.

Barbe V, Cruveiller S, Kunst F*, et al.* (2009) From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* **155**: 1758-1775.

Barrow GI & Feltham RKA (2004) *Cowan and Steel's Manual for the Identification of Medical Bacteria*. Cambridge University Press, Cambridge, UK.

Basappa P (1974) Growth characteristics of thermophilic micro-organisms in milk. *Indian Journal of Dairy Science* **27**: 103-108.

Bayjanov JR, Molenaar D, Tzeneva V, Siezen RJ & van Hijum SAFT (2012) PhenoLink - a web-tool for linking phenotype to similar to omics data for bacteria: application to gene-trait matching for *Lactobacillus plantarum* strains. *BMC Genomics* **13**: doi: 10.1186/1471-2164-1113-1170.

Beaman TC & Gerhardt P (1986) Heat-resistance of bacterial spores correlated with protoplast dehydration, mineralization, and thermal adaptation. *Applied and Environmental Microbiology* **52**: 1242-1246.

Beaman TC, Pankratz HS & Gerhardt P (1988) Heat-shock affects permeability and resistance of *Bacillus stearothermophilus* spores. *Applied and Environmental Microbiology* **54**: 2515-2520.

Beaman TC, Pankratz HS & Gerhardt P (1989) Low heat resistance of *Bacillus sphaericus* spores correlated with high protoplast water content. *FEMS Microbiology Letters* **49**: 1-4.

Beauregard PB, Chai Y, Vlamakis H, Losick R & Kolter R (2013) *Bacillus subtilis* biofilm induction by plant polysaccharides. *Proceedings of the National Academy of Sciences of the United States of America* **110**: E1621-E1630.

Belduz AO, Dulger S, Demirbag Z & Erturk O (2000) Isolation and characterisation of *Bacillus flavothermus* strains from a hot spring in North-East Turkey. *Biologia* **55**: 249-258.

Ben-Yehuda S & Losick R (2002) Asymmetric cell division in *B. subtilis* involves a spiral-like intermediate of the cytokinetic protein FtsZ. *Cell* **109**: 257-266.

Ben-Yehuda S, Rudner DZ & Losick R (2003) RacA, a bacterial protein that anchors chromosomes to the cell poles. *Science* **299**: 532-536.

Bender GR & Marquis RE (1985) Spore heat-resistance and specific mineralization. *Applied and Environmental Microbiology* **50**: 1414-1421.

Berendsen EM, Krawczyk AO, Klaus V, de Jong A, Boekhorst J, Eijlander RT, Kuipers OP & Wells-Bennik MHJ (2015) *Bacillus thermoamylovorans* spores with very high level heat resistance germinate poorly in rich medium despite the presence of *ger* clusters but efficiently

upon exposure to calcium-dipicolinic acid. *Applied and Environmental Microbiology* **81**: 7791-7801.

Bhalla A, Kainth AS & Sani RK (2013) Draft genome sequence of lignocellulose-degrading thermophilic bacterium *Geobacillus* sp. strain WSUCF1. *Genome announcements* **1**: doi: 10.1128/genomeA.00595-00513.

Bick MJ, Lamour V, Rajashankar KR, Gordiyenko Y, Robinson CV & Darst SA (2009) How to switch off a histidine kinase: Crystal structure of *Geobacillus stearothermophilus* KinB with the inhibitor Sda. *Journal of Molecular Biology* **386**: 163-177.

Biggs PJ, Fearnhead P, Hotter G, Mohan V, Collins-Emerson J, Kwan E, Besser TE, Cookson A, Carter PE & French NP (2011) Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence type reveals multiple loci of different ancestral lineage. *PLoS One* **6**: doi:10.1371/journal.pone.0027121.

Biswas A, Fineran PC & Brown CM (2014) Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics* **30**: 1805-1813.

Biswas A, Gagnon JN, Brouns SJJ, Fineran PC & Brown CM (2013) CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biology* **10**: 817-827.

Blanchard K, Robic S & Matsumura I (2014) Transformable facultative thermophile *Geobacillus stearothermophilus* NUB321 as a host strain for metabolic engineering. *Applied Microbiology and Biotechnology* **98**: 6715-6723.

Bogaard PTC, Hols P, Kuipers OP, Kleerebezem M & Vos WMd (2004) Sugar utilisation and conservation of the *gal-lac* gene cluster in *Streptococcus thermophilus*. *Systematic and Applied Microbiology* **27**: 10-17.

Bolotin A, Quinquis B, Renault P*, et al.* (2004) Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nature Biotechnology* **22**: 1554-1558.

Bongiorni C, Ishikawa S, Stephenson S, Ogasawara N & Perego M (2005) Synergistic regulation of competence development in *Bacillus subtilis* by two Rap-Phr systems. *Journal of Bacteriology* **187**: 4353-4361.

Boonmark C, Takahasi Y & Morikawa M (2013) Draft genome sequence of *Geobacillus thermoleovorans* strain B23. *Genome announcements* **1**: doi:10.1128/genomeA.00944-00913.

Bos R, van der Mei HC & Busscher HJ (1999) Physico-chemistry of initial microbial adhesive interactions - its mechanisms and methods for study. *FEMS Microbiology Reviews* **23**: 179-230.

Branda SS, Gonzalez-Pastor JE, Ben-Yehuda S, Losick R & Kolter R (2001) Fruiting body formation by *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 11621-11626.

Branda SS, Chu F, Kearns DB, Losick R & Kolter R (2006) A major protein component of the *Bacillus subtilis* biofilm matrix. *Molecular Microbiology* **59**: 1229-1238.

Branda SS, Gonzalez-Pastor JE, Dervyn E, Ehrlich SD, Losick R & Kolter R (2004) Genes involved in formation of structured multicellular communities by *Bacillus subtilis*. *Journal of Bacteriology* **186**: 3970-3979.

Bremer PJ, Seale RB, Flint SH & Palmer J (2009) Biofilms in dairy processing. *Biofilms in the food and beverage industries,*(Fratamico PM, Annous BA & Gunter NW, eds.), pp. 396-431. Woodhead Publishing Limited, Oxford.

Bremer PJ, Flint SH, Brooks J & Palmer J (2015) Introduction to Biofilms: Definition and basic concepts. *Biofilms in the Dairy Industry,*(Teh KH, Flint SH, Brooks J & Knight GC, eds.), pp. 1-12. John Wiley & Sons, Ltd., Chichester, West Sussex, UK.

Broder DH & Pogliano K (2006) Forespore engulfment mediated by a ratchet-like mechanism. *Cell* **126**: 917-928.

Brumm P, Land ML, Hauser LJ, Jeffries CD, Chang YJ & Mead DA (2015a) Complete genome sequences of *Geobacillus* sp Y412MC52, a xylan-degrading strain isolated from obsidian hot spring in Yellowstone National Park. *Standards in Genomic Sciences* **10**.

Brumm PJ, De Maayer P, Mead DA & Cowan DA (2015b) Genomic analysis of six new *Geobacillus* strains reveals highly conserved carbohydrate degradation architectures and strategies. *Frontiers in Microbiology* **6**.

Brunsing RL, La Clair C, Tang S, Chiang C, Hancock LE, Perego M & Hoch JA (2005) Characterization of sporulation histidine kinases of *Bacillus anthracis*. *Journal of Bacteriology* **187**: 6972-6981.

Bryanskaya AV, Rozonov AS, Logacheva MD, Kotenko AV & Peltek SE (2014) Draft genome sequence of *Geobacillus icigianus* strain G1w1$^{T}$ isolated from hot springs in the valley of geysers, Kamchatka (Russian Federation). *Genome announcements* **2**: doi: 10.1128/genomeA.01098-01014.

Bryanskaya AV, Rozanov AS, Slynko NM, Shekhovtsov SV & Peltek SE (2015) *Geobacillus icigianus* sp nov., a thermophilic bacterium isolated from a hot spring. *International Journal of Systematic and Evolutionary Microbiology* **65**: 864-869.

Bryant D & Moulton V (2004) Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* **21**: 255-265.

Burgess SA, Lindsay D & Flint SH (2010) Thermophilic bacilli and their importance in dairy processing. *International Journal of Food Microbiology* **144**: 215-225.

Burgess SA, Flint SH & Lindsay D (2014a) Characterization of thermophilic bacilli from a milk powder processing plant. *Journal of Applied Microbiology* **11**: 350-359.

Burgess SA, Lindsay D & Flint SH (2014b) Biofilms of thermophilic bacilli isolated from dairy processing plants and efficacy of sanitizers. *Microbial biofilms - Methods and protocols,* Vol. 1147 (Gianfrancok D, ed.) pp. 367-377. Springer, New York.

Burgess SA, Brooks JD, Rakonjac J, Walker KM & Flint SH (2009) The formation of spores in biofilms of *Anoxybacillus flavithermus*. *Journal of Applied Microbiology* **107**: 1012-1018.

Burkholder WF, Kurtser I & Grossman AD (2001) Replication initiation proteins regulate a developmental checkpoint in *Bacillus subtilis*. *Cell* **104**: 269-279.

Burmistrz M, Dudek B, Staniec D, Martinez JIR, Bochtler M, Potempa J & Pyrc K (2015) Functional analysis of *Porphyromonas gingivalis* W83 CRISPR-Cas systems. *Journal of Bacteriology* **197**: 2631-2641.

Bylund G (1995) *Dairy processing handbook*. Tetra Pak Processing Systems AB.

Cairns LS, Hobley L & Stanley-Wall NR (2014) Biofilm formation by *Bacillus subtilis*: new insights into regulatory strategies and assembly mechanisms. *Molecular Microbiology* **93**: 587-598.

Camp AH & Losick R (2008) A novel pathway of intercellular signalling in *Bacillus subtilis* involves a protein with similarity to a component of type III secretion channels. *Molecular Microbiology* **69**: 402-417.

Camp AH & Losick R (2009) A feeding tube model for activation of a cell-specific transcription factor during sporulation in *Bacillus subtilis*. *Genes & Development* **23**: 1014-1024.

Cangiano G, Mazzone A, Baccigalupi L, Isticato R, Eichenberger P, De Felice M & Ricca E (2010) Direct and indirect control of late sporulation genes by GerR of *Bacillus subtilis*. *Journal of Bacteriology* **192**: 3406-3413.

Cangiano G, Sirec T, Panarella C, Isticato R, Baccigalupi L, De Felice M & Ricca E (2014) The *sps* gene products affect the germination, hydrophobicity, and protein adsorption of *Bacillus subtilis* spores. *Applied and Environmental Microbiology* **80**: 7293-7302.

Carniol K, Ben-Yehuda S, King N & Losick R (2005) Genetic dissection of the sporulation protein SpoIIE and its role in asymmetric division in *Bacillus subtilis*. *Journal of Bacteriology* **187**: 3511-3520.

Caspers MP, Boekhorst J, Abee T, Siezen RJ & Kort R (2013) Complete genome sequence of *Anoxybacillus flavithermus* TNO-09.00, a thermophilic sporeformer associated with a dairy-processing environment. *Genome announcements* **1**: doi:10.1128/genomeA.00010-00013.

Cavanagh D, Casey A, Altermann E, Cotter PD, Fitzgerald GF & McAuliffe O (2015) Evaluation of *Lactococcus lactis* isolates from nondairy sources with potential dairy applications reveals extensive phenotype-genotype disparity and implications for a revised species. *Applied and Environmental Microbiology* **81**: 3961-3972.

Chai Y, Kolter R & Losick R (2009) A widely conserved gene cluster required for lactate utilization in *Bacillus subtilis* and its involvement in biofilm formation. *Journal of Bacteriology* **191**: 2423-2430.

Chandler DP, Brown J, Bruckner-Lea CJ, Olson L, Posakony GJ, Stults JR, Valentine NB & Bond LJ (2001) Continuous spore disruption using radially focused, high-frequency ultrasound. *Analytical Chemistry* **73**: 3784-3789.

Chastanet A & Losick R (2007) Engulfment during sporulation in *Bacillus subtilis* is governed by a multi-protein complex containing tandemly acting autolysins. *Molecular Microbiology* **64**: 139-152.

Chauhan K, Seale RB, Deeth HC & Turner MS (2014) Hypervariable pili and flagella genes provide suitable new targets for DNA high-resolution melt-based genotyping of dairy *Geobacillus* spp. *Journal of Food Protection* **77**: 1715-1722.

Chauhan K, Dhakal R, Seale RB, Deeth HC, Pillidge CJ, Powell IB, Craven H & Turner MS (2013) Rapid identification of dairy mesophilic and thermophilic sporeforming bacteria using DNA high resolution melt analysis of variable 16S rDNA regions. *International Journal of Food Microbiology* **165**: 175-183.

Checinska A, Burbank M & Paszczynski AJ (2012) Protection of *Bacillus pumilus* spores by catalases. *Applied and Environmental Microbiology* **78**: 6413-6422.

Chen L, Daniel RM & Coolbear T (2003) Detection and impact of protease and lipase activities in milk and milk powders. *International Dairy Journal* **13**: 255-275.

Chen L, Coolbear T & Daniel RM (2004) Characteristics of proteinases and lipases produced by seven *Bacillus* sp. isolated from milk powder production lines. *International Dairy Journal* **14**: 495-504.

Chen SQ, Tang QY, Zhang XD, Zhao GH, Hu XS, Liao XJ, Chen F, Wu JH & Xiang H (2006) Isolation and characterization of thermo-acidophilic endospore-forming bacteria from the concentrated apple juice-processing environment. *Food Microbiology* **23**: 439-445.

Chen Y, Cao S, Chai Y, Clardy J, Kolter R, Guo J-h & Losick R (2012) A *Bacillus subtilis* sensor kinase involved in triggering biofilm formation on the roots of tomato plants. *Molecular Microbiology* **85**: 418-430.

Chen Y, Yan F, Chai Y, Liu H, Kolter R, Losick R & Guo J-h (2013) Biocontrol of tomato wilt disease by *Bacillus subtilis* isolates from natural environments depends on conserved genes mediating biofilm formation. *Environmental Microbiology* **15**: 848-864.

Chiu K-C, Lin C-J & Shaw G-C (2014) Transcriptional regulation of the L-lactate permease gene *lutP* by the LutR repressor of *Bacillus subtilis* RO-NN-1. *Microbiology* **160**: 2178-2189.

Chopra AK & Mathur DK (1984) Isolation, screening and characterization of thermophilic *Bacillus* species isolated from dairy products. *Journal of Applied Bacteriology* **57**: 263-271.

Chun J & Rainey FA (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *International Journal of Systematic and Evolutionary Microbiology* **64**: 316-324.

Claus D & Berkeley RCW (1986) Genus *Bacillus* Cohn 1872. *Bergey's Manual of Systematic Bacteriology,* Vol. 2 (Sneath PHA, Mair NS, Sharpe ME & Holt JG, eds.), pp. 1105-1139.

Coenye T & Vandamme P (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters* **228**: 45-49.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR & Tiedje JM (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* **42**: D633-D642.

Contreras-Moreira B & Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* **79**: 7696-7701.

Cooney PH & Freese E (1976) Commitment to sporulation in *Bacillus megaterium* and uptake of specific compounds. *Journal of General Microbiology* **95**: 381-390.

Coorevits A, Dinsdale AE, Halket G, Lebbe L, De Vos P, Van Landschoot A & Logan NA (2012) Taxonomic revision of the genus *Geobacillus*: emendation of *Geobacillus*, *G. stearothermophilus*, *G. jurassicus*, *G. toebii*, *G. thermodenitrificans* and *G. thermoglucosidans* (nom. corrig., formerly '*thermoglucosidasius*'); transfer of *Bacillus thermantarcticus* to the genus as *G. thermantarcticus* comb. nov.; proposal of *Caldibacillus debilis* gen. nov., comb. nov.; transfer of *G. tepidamans* to *Anoxybacillus* as *A. tepidamans* comb. nov.; and proposal of *Anoxybacillus caldiproteolyticus* sp nov. *International Journal of Systematic and Evolutionary Microbiology* **62**: 1470-1485.

Coorevits A, De Jonghe V, Vandroemme J, Reekemans R, Heyrman J, Messens W, De Vos P & Heyndrickx M (2008) Comparative analysis of the diversity of aerobic spore-forming bacteria in raw milk from organic and conventional dairy farms. *Systematic and Applied Microbiology* **31**: 126-140.

Cortezzo DE & Setlow P (2005) Analysis of factors that influence the sensitivity of spores of *Bacillus subtilis* to DNA damaging chemicals. *Journal of Applied Microbiology* **98**: 606-617.

Cortezzo DE, Koziol-Dube K, Setlow B & Setlow P (2004) Treatment with oxidizing agents damages the inner membrane of spores of *Bacillus subtilis* and sensitizes spores to subsequent stress. *Journal of Applied Microbiology* **97**: 838-852.

Cosentino S, Mulargia AF, Pisano B, Tuveri P & Palmas F (1997) Incidence and biochemical characteristics of *Bacillus* flora in Sardinian dairy products. *International Journal of Food Microbiology* **38**: 235-238.

Costerton JW, Lewandowski Z, Caldwell DE, Korber DR & Lappinscott HM (1995) Microbial biofilms. *Annual Review of Microbiology* **49**: 711-745.

Costerton JW, Lewandowski Z, DeBeer D, Caldwell D, Korber D & James G (1994) Biofilms, the customized microniche. *Journal of Bacteriology* **176**: 2137-2142.

Couvigny B, Therial C, Gautier C, Renault P, Briandet R & Guedon E (2015) *Streptococcus thermophilus* biofilm formation: A remnant trait of ancestral commensal life? *PLoS One* **10**.

Crawshaw AD, Serrano M, Stanley WA, Henriques AO & Salgado PS (2014) A mother cell-to-forespore channel: current understanding and future challenges. *FEMS Microbiology Letters* **358**: 129-136.

Crielly EM, Logan NA & Anderton A (1994) Studies on the *Bacillus* flora of milk and milk products. *Journal of Applied Bacteriology* **77**: 256-263.

Cucksey S (2002) Rapid enumeration of thermophiles. Evaluation of the BacTrac 4100 micro-organism growth analyser in Pahiatua's milk powder factory. Master of Dairy Science and Technology thesis Thesis, Massey University, Palmerston North.

Daniel RA & Errington J (1993) Cloning, DNA-sequence, functional-analysis and transcriptional regulation of the genes encoding dipicolinic acid synthetase required for sporulation in *Bacillus subtilis*. *Journal of Molecular Biology* **232**: 468-483.

Darling ACE, Mau B, Blattner FR & Perna NT (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**: 1394-1403.

Dartois V, Djavakhishvili T & Hoch JA (1997) KapB is a lipoprotein required for KinB signal transduction and activation of the phosphorelay to sporulation in *Bacillus subtilis*. *Molecular Microbiology* **26**: 1097-1108.

De Clerck E, Rodriguez-Diaz M, Vanhoutte T, Heyrman J, Logan NA & De Vos P (2004a) *Anoxybacillus contaminans* sp nov and *Bacillus gelatini* sp nov., isolated from contaminated gelatin batches. *International Journal of Systematic and Evolutionary Microbiology* **54**: 941-946.

De Clerck E, Vanhoutte T, Hebb T, Geerinck J, Devos J & De Vos P (2004b) Isolation, characterization, and identification of bacterial contaminants in semifinal gelatin extracts. *Applied and Environmental Microbiology* **70**: 3664-3672.

de Hoon MJL, Eichenberger P & Vitkup D (2010) Hierarchical evolution of the bacterial sporulation network. *Current Biology* **20**: R735-R745.

De Jonghe V, Coorevits A, De Block J, Van Coillie E, Grijspeerdt K, Herman L, De Vos P & Heyndrickx M (2010) Toxinogenic and spoilage potential of aerobic spore-formers isolated from raw milk. *International Journal of Food Microbiology* **136**: 318-325.

De Maayer P, Williamson CE, Vennard CT, Danson MJ & Cowan DA (2014) Draft genome sequences of *Geobacillus* sp. strains CAMR5420 and CAMR12739. *Genome announcements* **2**: doi:10.1128/genomeA.00567-00514.

Delihas N (2011) Impact of small repeat sequences on bacterial genome evolution. *Genome Biology and Evolution* **3**: 959-973.

den Besten HMW, van der Mark E-J, Hensen L, Abee T & Zwietering MH (2010) Quantification of the effect of culturing temperature on salt-induced heat resistance of *Bacillus* species. *Applied and Environmental Microbiology* **76**: 4286-4292.

Denny CB (1981) Thermophilic organisms involved in food spoilage - introduction. *Journal of Food Protection* **44**: 144-145.

Devi SN, Vishnoi M, Kiehler B, Haggett L & Fujita M (2015) *In vivo* functional characterization of the transmembrane histidine kinase KinC in *Bacillus subtilis*. *Microbiology* **161**: 1092-1104.

Diaz AR, Stephenson S, Green JM, Levdikov VM, Wilkinson AJ & Perego M (2008) Functional role for a conserved aspartate in the Spo0E signature motif involved in the dephosphorylation of the *Bacillus subtilis* sporulation regulator Spo0A. *Journal of Biological Chemistry* **283**: 2962-2972.

Diaz AR, Core LJ, Jiang M, Morelli M, Chiang CH, Szurmant H & Perego M (2012) *Bacillus subtilis* RapA phosphatase domain interaction with its substrate, phosphorylated Spo0F, and its inhibitor, the PhrA peptide. *Journal of Bacteriology* **194**: 1378-1388.

Diethmaier C, Newman JA, Kovacs AT, Kaever V, Herzberg C, Rodrigues C, Boonstra M, Kuipers OP, Lewis RJ & Stuelke J (2014) The YmdB phosphodiesterase is a global regulator of late adaptive responses in *Bacillus subtilis*. *Journal of Bacteriology* **196**: 265-275.

Dinsdale AE, Halket G, Coorevits A, Van Landschoot A, Busse H-J, De Vos P & Logan NA (2011) Emended descriptions of *Geobacillus thermoleovorans* and *Geobacillus thermocatenulatus*. *International Journal of Systematic and Evolutionary Microbiology* **61**: 1802-1810.

Doan T, Morlot C, Meisner J, Serrano M, Henriques AO, Moran CP, Jr. & Rudner DZ (2009) Novel Secretion Apparatus Maintains Spore Integrity and Developmental Gene Expression in *Bacillus subtilis*. *PLoS Genetics* **5**: doi:10.1371/journal.pgen.1000566.

Doi K, Mori K, Martono H, Nagayoshi Y, Fujino Y, Tashiro K, Kuhara S & Ohshima T (2013) Draft genome sequence of *Geobacillus kaustophilus* GBlys, a lysogenic strain with Bacteriophage OH2. *Genome announcements* **1**: doi: 10.1128/genomeA.00634-00613.

Donk PJ (1920) A highly resistant thermophilic organism. *Journal of Bacteriology* **5**: 373-374.

Driks A (1999) *Bacillus subtilis* spore coat. *Microbiology and Molecular Biology Reviews* **63**: 1-20.

Driks A (2002) Maximum shields: the assembly and function of the bacterial spore coat. *Trends in Microbiology* **10**: 251-254.

Dy RL, Pitman AR & Fineran PC (2013) Chromosomal targeting by CRISPR-Cas systems can contribute to genome plasticity in bacteria. *Mobile Genetic Elements* **3** e26831.

Earl AM, Losick R & Kolter R (2007) *Bacillus subtilis* genome diversity. *Journal of Bacteriology* **189**: 1163-1170.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792-1797.

Eduardo Gonzalez-Pastor J (2011) Cannibalism: a social behavior in sporulating *Bacillus subtilis*. *FEMS Microbiology Reviews* **35**: 415-424.

Edwards U, Rogall T, Blocker H, Emde M & Bottger EC (1989) Isolation and direct complete nucleotide determination of entire genes - characterization of a gene coding for 16S-Ribosomal RNA. *Nucleic Acids Research* **17**: 7843-7853.

Eichenberger P, Jensen ST, Conlon EM*, et al.* (2003) The $\sigma^E$ regulon and the identification of additional sporulation genes in *Bacillus subtilis*. *Journal of Molecular Biology* **327**: 945-972.

Eichenberger P, Fujita M, Jensen ST*, et al.* (2004) The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. *PLoS Biology* **2**: 1664-1683.

Eid J, Fehr A, Gray J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

Eijlander RT, de Jong A, Krawczyk AO, Holsappel S & Kuipers OP (2014) SporeWeb: an interactive journey through the complete sporulation cycle of *Bacillus subtilis*. *Nucleic Acids Research* **42**: D685-D691.

Errington J (2003) Regulation of endospore formation in *Bacillus subtilis*. *Nature Reviews Microbiology* **1**: 117-126.

Eswaramoorthy P & Fujita M (2010) Systematic domain deletion analysis of the major sporulation kinase in *Bacillus subtilis*. *Journal of Bacteriology* **192**: 1744-1748.

Eswaramoorthy P, Guo T & Fujita M (2009) *In vivo* domain-based functional analysis of the major sporulation sensor kinase, KinA, in *Bacillus subtilis*. *Journal of Bacteriology* **191**: 5358-5368.

Eswaramoorthy P, Duan D, Dinh J, Dravis A, Devi SN & Fujita M (2010) The threshold level of the sensor histidine kinase KinA governs entry into sporulation in *Bacillus subtilis*. *Journal of Bacteriology* **192**: 3870-3882.

Etoa FX & Michiels L (1988) Heat induced resistance of *Bacillus stearothermophilus* spores. *Letters in Applied Microbiology* **6**: 43-45.

Fabret C, Feher VA & Hoch JA (1999) Two-component signal transduction in *Bacillus subtilis*: How one organism sees its world. *Journal of Bacteriology* **181**: 1975-1983.

Faille C, Tauveron G, Gentil L, vre LC & Slomianny C (2007) Occurrence of *Bacillus cereus* spores with a damaged exosporium: consequences on the spore adhesion on surfaces of food processing lines. *Journal of Food Protection* **70**: 2346-2353.

Faille C, Jullien C, Fontaine F, Bellon-Fontaine MN, Slomianny C & Benezech T (2002) Adhesion of Bacillus spores and Escherichia coli cells to inert surfaces: role of surface hydrophobicity. *Canadian Journal of Microbiology* **48**: 728-738.

Faille C, Benezech T, Midelet-Bourdin G, Lequette Y, Clarisse M, Ronse G, Ronse A & Slomianny C (2014) Sporulation of *Bacillus* spp. within biofilms: A potential source of contamination in food processing environments. *Food Microbiology* **40**: 64-74.

Fazzini MM, Schuch R & Fischetti VA (2010) A novel spore protein, ExsM, regulates formation of the exosporium in *Bacillus cereus* and *Bacillus anthracis* and affects spore size and shape. *Journal of Bacteriology* **192**: 4012-4021.

Feng L, Wang W, Cheng J*, et al.* (2007) Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 5602-5607.

Fernandez-No IC, Boehme K, Diaz-Bao M, Cepeda A, Barros-Velazquez J & Calo-Mata P (2013) Characterisation and profiling of *Bacillus subtilis*, *Bacillus cereus* and *Bacillus licheniformis* by MALDI-TOF mass fingerprinting. *Food Microbiology* **33**: 235-242.

Feucht A, Evans L & Errington J (2003) Identification of sporulation genes by genome-wide analysis of the $\sigma^E$ regulon of *Bacillus subtilis*. *Microbiology* **149**: 3023-3034.

Feucht A, Magnin T, Yudkin MD & Errington J (1996) Bifunctional protein required for asymmetric cell division and cell-specific transcription in *Bacillus subtilis*. *Genes & Development* **10**: 794-803.

Finley N & Fields ML (1962) Heat activation and heat-induced dormancy of *Bacillus stearothermophilus* spores. *Applied Microbiology* **10**: 231-&.

Finn RD, Bateman A, Clements J*, et al.* (2014) Pfam: the protein families database. *Nucleic Acids Research* **42**: D222-D230.

Flint S, Walker K, Waters B & Crawford R (2007) Description and validation of a rapid (1h) flow cytometry test for enumerating thermophilic bacteria in milk powders. *Journal of Applied Microbiology* **102**: 909-915.

Flint S, Drocourt JL, Walker K, Stevenson B, Dwyer M, Clarke I & McGill D (2006) A rapid, two-hour method for the enumeration of total viable bacteria in samples from commercial milk powder and whey protein concentrate powder manufacturing plants. *International Dairy Journal* **16**: 379-384.

Flint SH, Bremer PJ & Brooks JD (1997) Biofilms in dairy manufacturing plant - Description, current concerns and methods of control. *Biofouling* **11**: 81-97.

Flint SH, Ward LJH & Walker KMR (2001a) Functional grouping of thermophilic *Bacillus* strains using amplification profiles of the 16S-23S internal spacer region. *Systematic and Applied Microbiology* **24**: 539-548.

Flint SH, Palmer J, Bloemen K, Brooks J & Crawford R (2001b) The growth of *Bacillus stearothermophilus* on stainless steel. *Journal of Applied Microbiology* **90**: 151-157.

Fort P & Errington J (1985) Nucleotide-sequence and complementation analysis of a polycistronic sporulation operon, *spoVA*, in *Bacillus subtilis*. *Journal of General Microbiology* **131**: 1091-1105.

Fortina MG, Mora D, Schumann P, Parini C, Manachini PL & Stackebrandt E (2001) Reclassification of *Saccharococcus caldoxylosilyticus* as *Geobacillus caldoxylosilyticus* (Ahmad et al. 2000) comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **51**: 2063-2071.

Fox GE, Pechman KR & Woese CR (1977) Comparative cataloging of 16S ribosomal ribonucleic-acid - molecular approach to procaryotic systematics. *International Journal of Systematic Bacteriology* **27**: 44-57.

Frank JF & Yousef AE (2004) Thermophilic bacteria. *Standard methods for the examination of dairy products,*(Wehr MH & Frank JF, eds.), pp. 230-231. American Public Health Association, Washington.

Fujita M & Losick R (2005) Evidence that entry into sporulation in *Bacillus subtilis* is governed by a gradual increase in the level and activity of the master regulator Spo0A. *Genes & Development* **19**: 2236-2244.

Fujita M, Gonzalez-Pastor JE & Losick R (2005) High- and low-threshold genes in the Spo0A regulon of *Bacillus subtilis*. *Journal of Bacteriology* **187**: 1357-1368.

Fujita Y, Fujita T, Miwa Y, Nihashi J & Aratani Y (1986) Organization and transcription of the gluconate operon, *gnt*, of *Bacillus subtilis*. *Journal of Biological Chemistry* **261**: 3744-3753.

Galperin MY (2004) Bacterial signal transduction network in a genomic perspective. *Environmental Microbiology* **6**: 552-567.

Galperin MY (2013) Genome diversity of spore-forming firmicutes. *Microbiology spectrum* **1**: doi:10.1128/microbiolspectrum.TBS-0015-2012.

Galperin MY, Makarova KS, Wolf YI & Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research* **43**: D261-D269.

Galperin MY, Mekhedov SL, Puigbo P, Smirnov S, Wolf YI & Rigden DJ (2012) Genomic determinants of sporulation in Bacilli and Clostridia: towards the minimal set of sporulation-specific genes. *Environmental Microbiology* **14**: 2870-2890.

Gao T, Greenwich J, Li Y, Wang Q & Chai Y (2015a) The bacterial tyrosine kinase activator TkmA contributes to biofilm formation largely independently of the cognate kinase PtkA in *Bacillus subtilis*. *Journal of Bacteriology* **197**: 3421-3432.

Gao T, Foulston L, Chai Y, Wang Q & Losick R (2015b) Alternative modes of biofilm formation by plant-associated *Bacillus cereus*. *MicrobiologyOpen* **4**: 452-464.

Gao X, Mukherjee S, Matthews PM, Hammad LA, Kearns DB & Dann CE, III (2013) Functional characterization of core components of the *Bacillus subtilis* Cyclic-Di-GMP signaling pathway. *Journal of Bacteriology* **195**: 4782-4792.

Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen HB & Nahm MH (2015) Pneumococcal capsules and their types: Past, present, and future. *Clinical Microbiology Reviews* **28**: 871-899.

Gerwig J, Kiley TB, Gunka K, Stanley-Wall N & Stuelkel J (2014) The protein tyrosine kinases EpsB and PtkA differentially affect biofilm formation in *Bacillus subtilis*. *Microbiology-Sgm* **160**: 682-691.

Ghosh S & Setlow P (2009) Isolation and Characterization of Superdormant Spores of *Bacillus* Species. *Journal of Bacteriology* **191**: 1787-1797.

Ghosh S, Ramirez-Peralta A, Gaidamakova E, Zhang P, Li YQ, Daly MJ & Setlow P (2011) Effects of Mn levels on resistance of *Bacillus megaterium* spores to heat, radiation and hydrogen peroxide. *Journal of Applied Microbiology* **111**: 663-670.

Gillaspy AE, Worrel V, Orvis J, Roe BA, Dyer DW & Iandol JJ (2006) The *Staphylococcus aureus* NCTC 8325 genome. *Gram positive pathogens,*(Fischetti VA, Novick J, Ferrette J, Portnoy D & Rood JI, eds.), pp. 381-412. ASM Press.

Gilmour A & Rowe MT (1990) Micro-organisms associated with milk. *Dairy Microbiology,*(Robinson RK, ed.) pp. 37 - 76. Elsevier Applied Science, London.

Goodhead I & Darby AC (2015) Taking the pseudo out of pseudogenes. *Current Opinion in Microbiology* **23**: 102-109.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P & Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* **57**: 81-91.

Goto K, Fujita R, Kato Y, Asahara M & Yokota A (2004) Reclassification of *Brevibacillus brevis* strains NCIMB 13288 and DSM 6472 (=NRRL NRS-887) as *Aneurinibacillus danicus* sp nov and *Brevibacillus limnophilus* sp nov. *International Journal of Systematic and Evolutionary Microbiology* **54**: 419-427.

Granger AC, Gaidamakova EK, Matrosova VY, Daly MJ & Setlow P (2011) Effects of Mn and Fe levels on *Bacillus subtilis* spore resistance and effects of $Mn^{2+}$, other divalent cations,

orthophosphate, and dipicolinic acid on protein resistance to ionizing radiation. *Applied and Environmental Microbiology* **77**: 32-40.

Grau RR, de Ona P, Kunert M*, et al.* (2015) A duo of potassium-responsive histidine kinases govern the multicellular destiny of *Bacillus subtilis*. *mBio* **6**: doi: 10.1128/mBio.00581-00515.

Greene RA & Slepecky RA (1972) Minimal requirements for commitment to sporulation in *Bacillus megaterium*. *Journal of Bacteriology* **111**: 557-565.

Griffiths MW (1995) Foodborne illness caused by *Bacillus* spp. other than *B. cereus* and their importance to the dairy industry. *Bulletin of the International Dairy Federation* **302**: 3-6.

Grissa I, Vergnaud G & Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research* **35**: W52-W57.

Guizelini BP, Vandenberghe LPS, Sella S & Soccol CR (2012) Study of the influence of sporulation conditions on heat resistance of *Geobacillus stearothermophilus* used in the development of biological indicators for steam sterilization. *Archives of Microbiology* **194**: 991-999.

Gundogan N & Arik MT (2004) Comparison of the protease activity of psychrotrophic and thermophilic *Bacillus* spp. isolated from raw milk samples. *Indian Veterinary Journal* **81**: 1013-1015.

Guttenplan SB, Blair KM & Kearns DB (2010) The EpsE flagellar clutch is bifunctional and synergizes with EPS biosynthesis to promote *Bacillus subtilis* biofilm formation. *PLoS Genetics* **6**: doi:10.1371/journal.pgen.1001243.

Ha D-G & O'Toole GA (2015) c-di-GMP and its effects on biofilm formation and dispersion: A *Pseudomonas aeruginosa* review. *Microbiology Spectrum* **3**: doi:10.1128/microbiolspec.MB-0003-2014.

Hackett RH & Setlow P (1988) Properties of spores of *Bacillus subtilis* strains which lack the major small, acid-soluble protein. *Journal of Bacteriology* **170**: 1403-1404.

Hall BG (2001) Predicting evolutionary potential. I. Predicting the evolution of a lactose-PTS system in *Escherichia coli*. *Molecular Biology and Evolution* **18**: 1389-1400.

Hammer P, Suhren G & Heeschen W (1996a) The presence of heat resistant mesophilic sporeformers in UHT milk - the current state of research. *DMZ, Lebensmittelindustrie und Milchwirtschaft* **117**: 738-744.

Hammer P, Lembke F, Suhren G & Heeschen W (1996b) Characterization of a heat resistant mesophilic *Bacillus* species affecting the quality of UHT-milk. *Heat Treatments & Alternative Methods* 9-16.

Hammer P, Herman L, Heyndricks M*, et al.* (2000) *Bacillus sporothermodurans* - a bacillus forming highly heat-resistant spores. *Bulletin of the International Dairy Federation* 3-27.

Han J, Seale RB, Silcock P, McQuillan AJ & Bremer PJ (2011) The physico-chemical characterization of casein-modified surfaces and their influence on the adhesion of spores from a Geobacillus species. *Biofouling* **27**: 459-466.

Hayes PR (1985) Quality assurance and production control. *Food Microbiology and Hygiene,*(Hayes PR, ed.) pp. 255. Elsevier Applied Science Publishers, London.

Hayrapetyan H, Abee T & Groot MN (2016) Sporulation dynamics and spore heat resistance in wet and dry biofilms of Bacillus cereus. *Food Control* **60**: 493-499.

Hayrapetyan H, Muller L, Tempelaars M, Abee T & Nierop Groot M (2015) Comparative analysis of biofilm formation by *Bacillus cereus* reference strains and undomesticated food isolates and the effect of free iron. *International Journal of Food Microbiology* **200**: 72-79.

Head NE & Yu HW (2004) Cross-sectional analysis of clinical and environmental isolates of *Pseudomonas aeruginosa*: Biofilm formation, virulence, and genome diversity. *Infection and Immunity* **72**: 133-144.

Heinen W, Lauwers AM & Mulders JWM (1982) *Bacillus flavothermus*, a newly isolated facultative thermophile. *Antonie van leeuwenhoek Journal of Microbiology* **48**: 265-272.

Henry JT & Crosson S (2011) Ligand-binding PAS domains in a genomic, cellular, and structural context. *Annual Review of Microbiology, Vol 65,* Vol. 65 (Gottesman S & Harwood CS, eds.), pp. 261-286.

Herbert S, Ziebandt A-K, Ohlsen K, Schaefer T, Hecker M, Albrecht D, Novick R & Goetz F (2010) Repair of global regulators in *Staphylococcus aureus* 8325 and comparative analysis with other clinical isolates. *Infection and Immunity* **78**: 2877-2889.

Heyndrickx M & Scheldeman P (2002) Bacilli associated with spoilage in dairy products and other food. *Applications and sytematics of Bacillus and relatives,*(Berkeley RCW, Heyndrickx M, Logan NA & De Vos P, eds.), pp. Blackwell Publishing Company.

Heyndrickx M, Lebbe L, Kersters K, De Vos P, Forsyth C & Logan NA (1998) *Virgibacillus*: a new genus to accommodate *Bacillus pantothenticus* (Proom and Knight 1950). Emended description of *Virgibacillus pantothenticus*. *International Journal of Systematic Bacteriology* **48**: 99-106.

Heyndrickx M, Vandemeulebroecke K, Hoste B, Janssen P, Kersters K, DeVos P, Logan NA, Ali N & Berkeley RCW (1996) Reclassification of *Paenibacillus* (formerly *Bacillus*) *pulvifaciens* (Nakamura 1984) Ash et al 1994, a later subjective synonym of *Paenibacillus* (formerly *Bacillus*) *larvae* (White 1906) Ash et al 1994, as a subspecies of *P. larvae*, with emended descriptions of *P. larvae* as *P. larvae* subsp *larvae* and *P. larvae* subsp *pulvifaciens*. *International Journal of Systematic Bacteriology* **46**: 270-279.

Heyndrickx M, Lebbe L, Vancanneyt M, Kersters K, DeVos P, Logan NA, Forsyth G, Nazli S, Ali N & Berkeley RCW (1997) A polyphasic reassessment of the genus *Aneurinibacillus*, reclassification of *Bacillus thermoaerophilus* (Meier-Stauffer et al 1996) as *Aneurinibacillus*

*thermoaerophilus* comb nov, and emended descriptions of *A. aneurinilyticus* corrig, *A. migulanus*, and *A. thermoaerophilus*. *International Journal of Systematic Bacteriology* **47**: 808-817.

Heyrman J, Logan NA, Busse HJ, Balcaen A, Lebbe L, Rodriguez-Diaz M, Swings J & De Vos P (2003) *Virgibacillus carmonensis* sp nov., *Virgibacillus necropolis* sp nov and *Virgibacillus picturae* sp nov., three novel species isolated from deteriorated mural paintings, transfer of the species of the genus *Salibacillus* to *Virgibacillus*, as *Virgibacillus marismortui* comb. nov and *Virgibacillus salexigens* comb. nov., and emended description of the genus *Virgibacillus*. *International Journal of Systematic and Evolutionary Microbiology* **53**: 501-511.

Higgins D & Dworkin J (2012) Recent progress in *Bacillus subtilis* sporulation. *FEMS Microbiology Reviews* **36**: 131-148.

Hill BM (2004) Thermophilic spores: UHT and retort processing of recombined milks.pp. 184-193. U.S. Dairy Export Council, Cancun, Mexico.

Hill BM & Smythe BW (1994) Progress in understanding the behaviour of thermophilic bacteria during milk powder manufacture. *Milk powders for the future II,*pp. 19-26. The Dunmore Press Limited, Palmerston North.

Hill BM & Smythe BW (2012) Endospores of thermophilic bacteria in ingredient milk powders and their significance to the manufacture of sterilized milk products: An industrial perspective. *Food Reviews International* **28**: 299-312.

Hinton AR, Trinh KT, Brooks JD & Manderson GJ (2002) Thermophile survival in milk fouling and on stainless steel during cleaning. *Food and Bioproducts Processing* **80**: 299-304.

Hintze PE & Nicholson WL (2010) Single-spore elemental analyses indicate that dipicolinic acid-deficient *Bacillus subtilis* spores fail to accumulate calcium. *Archives of Microbiology* **192**: 493-497.

Hobley L, Ostrowski A, Rao FV, Bromley KM, Porter M, Prescott AR, MacPhee CE, Van Aalten DMF & Stanley-Wall NR (2013) BslA is a self-assembling bacterial hydrophobin that coats the *Bacillus subtilis* biofilm. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 13600-13605.

Holt C (1985) The milk salts: their secretion, concentrations and physical chemistry *Developments in dairy chemistry - 3,*(Fox PF, ed.) pp. 143-181. Elsvier Applied Science, London, England.

Horvath P & Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**: 167-170.

Horvath P, Coute-Monvoisin A-C, Romero DA, Boyaval P, Fremaux C & Barrangou R (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *International Journal of Food Microbiology* **131**: 62-70.

Horvath P, Romero DA, Coute-Monvoisin A-C, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C & Barrangou R (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriology* **190**: 1401-1412.

Hosoya S, Lu ZL, Ozaki Y, Takeuchi M & Sato T (2007) Cytological analysis of the mother cell death process during sporulation in *Bacillus subtilis*. *Journal of Bacteriology* **189**: 2561-2565.

Huang S-S, Chen D, Pelczar PL, Vepachedu VR, Setlow P & Li Y-Q (2007) Levels of Ca($2^+$)-dipicolinic acid in individual *Bacillus* spores determined using microfluidic raman tweezers. *Journal of Bacteriology* **189**: 4681-4687.

Huang ZJ, Edery I & Rosbash M (1993) PAS is a dimerization domain common to *Drosophila* period and several transcription factors. *Nature* **364**: 259-262.

Huson DH & Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**: 254-267.

Iciek J, Papiewska A & Molska M (2006) Inactivation of *Bacillus stearothermophilus* spores during, thermal processing. *Journal of Food Engineering* **77**: 406-410.

Imamura D, Kuwana R, Takamatsu H & Watabe K (2010) Localization of proteins to different layers and regions of *Bacillus subtilis* spore coats. *Journal of Bacteriology* **192**: 518-524.

Imamura D, Kuwana R, Takamatsu H & Watabe K (2011) Proteins involved in formation of the outermost layer of *Bacillus subtilis* spores. *Journal of Bacteriology* **193**: 4075-4080.

Ivanova N, Sorokin A, Anderson I*, et al.* (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* **423**: 87-91.

Jacq C, Miller JR & Brownlee GG (1977) Pseudogene structure in 5S-DNA of *Xenopus laevis*. *Cell* **12**: 109-120.

Jeurnink TJM, Walstra P & deKruif CG (1996) Mechanisms of fouling in dairy processing. *Netherlands Milk and Dairy Journal* **50**: 407-426.

Jiang D, Li P, Jiang Z, Dai X, Zhang R, Wang Y, Guo Q & Wang Y (2015) Chemolithoautotrophic arsenite oxidation by a thermophilic *Anoxybacillus flavithermus* strain TCC9-4 from a hot spring in Tengchong of Yunnan, China. *Frontiers in Microbiology* **6**: doi:10.3389/fmicb.2015.00360.

Jiang M, Grau R & Perego M (2000a) Differential processing of propeptide inhibitors of Rap phosphatases in *Bacillus subtilis*. *Journal of Bacteriology* **182**: 303-310.

Jiang M, Shao WL, Perego M & Hoch JA (2000b) Multiple histidine kinases regulate entry into stationary phase and sporulation in *Bacillus subtilis*. *Molecular Microbiology* **38**: 535-542.

Johnson MJ, Todd SJ, Ball DA, Shepherd AM, Sylvestre P & Moir A (2006) ExsY and CotY are required for the correct assembly of the exosporium and spore coat of *Bacillus cereus*. *Journal of Bacteriology* **188**: 7905-7913.

Johnston N (2004) Debaffling biofilms - studies follow transformations and detail a major signal. *Scientist* **18**: 34-35.

Johnstone K, Ellar DJ & Appleton TC (1980) Location of metal-ions in *Bacillus megaterium* spores by high-resolution electron-probe x-ray-microanalysis. *FEMS Microbiology Letters* **7**: 97-101.

Jolley KA, Bliss CM, Bennett JS*, et al.* (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**: 1005-1015.

Jones CA, Padula NL & Setlow P (2005) Effect of mechanical abrasion on the viability, disruption and germination of spores of *Bacillus subtilis*. *Journal of Applied Microbiology* **99**: 1484-1494.

Jones RG, Ober CK, Hodge P, Kratochvil P, Moad G & Vert M (2013) Terminology for aggregation and self-assembly in polymer science (IUPAC Recommendations 2013). *Pure and Applied Chemistry* **85**: 463-490.

Kalogridouvassiliadou D (1992) Biochemical activities of *Bacillus* species isolated from flat sour evaporated milk. *Journal of Dairy Science* **75**: 2681-2686.

Karunakaran E & Biggs CA (2011) Mechanisms of *Bacillus cereus* biofilm formation: an investigation of the physicochemical characteristics of cell surfaces and extracellular proteins. *Applied Microbiology and Biotechnology* **89**: 1161-1175.

Kaul A & Singh RS (1982) Production of stable *Bacillus stearothermophilus* spores. *Journal of Food Protection* **45**: 795-796.

Kemp JT, Driks A & Losick R (2002) FtsA mutants of *Bacillus subtilis* impaired in sporulation. *Journal of Bacteriology* **184**: 3856-3863.

Khalil A, Sivakumar N & Qarawi S (2015) Genome sequence of *Anoxybacillus flavithermus* Strain AK1, a thermophile isolated from a hot spring in Saudi Arabia. *Genome announcements* **3**: doi:10.1128/genomeA.00604-00615.

Kiley TB & Stanley-Wall NR (2010) Post-translational control of *Bacillus subtilis* biofilm formation mediated by tyrosine phosphorylation. *Molecular Microbiology* **78**: 947-963.

Kim J & Foegeding PM (1990) Effects of heat treatment, $CaCl_2$ treatment and ethanol treatment on activation of *Bacillus* spores. *Journal of Applied Bacteriology* **69**: 414-420.

Kim M, Oh H-S, Park S-C & Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* **64**: 346-351.

King N, Dreesen O, Stragier P, Pogliano K & Losick R (1999) Septation, dephosphorylation, and the activation of $\sigma^F$ during sporulation in *Bacillus subtilis*. *Genes & Development* **13**: 1156-1167.

Kirisits MJ, Prost L, Starkey M & Parsek MR (2005) Characterization of colony morphology variants isolated from *Pseudomonas aeruginosa* biofilms. *Applied and Environmental Microbiology* **71**: 4809-4821.

Knight GC & Weeks MG (2008) Conditions for inactivation of thermophilic spores in NaOH solutions for reuse applications. *Australian Journal of Dairy Technology* **63**: 82-86.

Knight GC, Nicol RS & McMeekin TA (2004) Temperature step changes: a novel approach to control biofilms of *Streptococcus thermophilus* in a pilot plant-scale cheese-milk pasteurisation plant. *International Journal of Food Microbiology* **93**: 305-318.

Kobayashi K & Iwano M (2012) BslA(YuaB) forms a hydrophobic layer on the surface of *Bacillus subtilis* biofilms. *Molecular Microbiology* **85**: 51-66.

Kobayashi K, Shoji K, Shimizu T, Nakano K, Sato T & Kobayashi Y (1995) Analysis of a suppressor mutation *ssb* (*kinC*) of *sur0B20* (*spo0A*) mutation in *Bacillus subtilis* reveals that *kinC* encodes a histidine protein kinase. *Journal of Bacteriology* **177**: 176-182.

Kobayashi K, Ehrlich SD, Albertini A*, et al.* (2003) Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 4678-4683.

Kolodkin-Gal I, Elsholz AKW, Muth C, Girguis PR, Kolter R & Losick R (2013) Respiration control of multicellularity in *Bacillus subtilis* by a complex of the cytochrome chain with a membrane-embedded histidine kinase. *Genes & Development* **27**: 887-899.

Koren S & Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23**: 110-120.

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH & Phillippy AM (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* **14**: doi: 10.1186/gb-2013-1114-1189-r1101.

Kuisiene N, Raugalas J & Chitavichius D (2004) *Geobacillus lituanicus* sp nov. *International Journal of Systematic and Evolutionary Microbiology* **54**: 1991-1995.

Kuisiene N, Raugalas J & Chitavichius D (2009) Phylogenetic, inter, and intraspecific sequence analysis of *spo0A* gene of the genus *Geobacillus*. *Current Microbiology* **58**: 547-553.

Kunkel B, Losick R & Stragier P (1990) The *Bacillus subtilis* gene for the developmental transcription factor Sigma-K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes & Development* **4**: 525-535.

Kuwana R, Kasahara Y, Fujibayashi M, Takamatsu H, Ogasawara N & Watabe K (2002) Proteomics characterization of novel spore proteins of *Bacillus subtilis*. *Microbiology* **148**: 3971-3982.

Lagesen K, Hallin P, Rodland EA, Staerfeldt H-H, Rognes T & Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**: 3100-3108.

Langeveld AN, van der Klijn W & van der Waals CB (1990) Bacteriegroei in indampers voor melk en wei. *Voedingamiddelentechnologie* **23**: 13-17.

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-360.

Lazarevic V, Soldo B, Medico N, Pooley H, Bron S & Karamata D (2005) *Bacillus subtilis* alpha-phosphoglucomutase is required for normal cell morphology and biofilm formation. *Applied and Environmental Microbiology* **71**: 39-45.

Ledeaux JR & Grossman AD (1995) Isolation and characterization of *kinC*, a gene that encodes a sensor kinase homologous to the sporulation sensor kinases *kinA* and *kinB* in *Bacillus subtilis*. *Journal of Bacteriology* **177**: 166-175.

Ledeaux JR, Yu NY & Grossman AD (1995) Different roles for KinA, KinB, and KinC in the initiation of sporulation in *Bacillus subtilis*. *Journal of Bacteriology* **177**: 861-863.

Lennox J (2011) Biofilm growth and development. *Biofilms: The Hypertextbook,*(Cunningham AB, Lennox JE & Rockford JR, eds.), http://www.hypertextbookshop.com/biofilmbook/working_version/contents/chapters/chapter002/section002/blue/page001.html.

Leuschner RGK & Lillford PJ (1999) Effects of temperature and heat activation on germination of individual spores of *Bacillus subtilis*. *Letters in Applied Microbiology* **29**: 228-232.

Lewis S (2003) Optimal separation temperature. Master of Dairy Science and Technology Thesis Thesis, Massey University, Palmerston North.

Li L, Stoeckert CJ & Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**: 2178-2189.

Li Y, Davis A, Korza G, Zhang P, Li Y-q, Setlow B, Setlow P & Hao B (2012) Role of a SpoVA protein in dipicolinic acid uptake into developing spores of *Bacillus subtilis*. *Journal of Bacteriology* **194**: 1875-1884.

Lin S-H & Liao Y-C (2013) CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes. *PLoS One* **8**: doi: 10.1371/journal.pone.0060843.

Lindsay D & Von Holy A (1999) Different responses of planktonic and attached *Bacillus subtilis* and *Pseudomonas fluorescens* to sanitizer treatment. *Journal of Food Protection* **62**: 368-379.

Lindsay D, Brozel VS & von Holy A (2005) Spore formation in *Bacillus subtilis* biofilms. *Journal of Food Protection* **68**: 860-865.

Lindsay D, Brozel VS, Mostert JF & von Holy A (2000) Physiology of dairy-associated *Bacillus* spp. over a wide pH range. *International Journal of Food Microbiology* **54**: 49-62.

Liszewski Zilla M, Chan YGY, Lunderberg JM, Schneewind O & Missiakas D (2015) LytR-CpsA-Psr enzymes as determinants of *Bacillus anthracis* secondary cell wall polysaccharide assembly. *Journal of Bacteriology* **197**: 343-353.

Liu W, Xie Y, Ma J*, et al.* (2015) IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**: 3359-3361.

Logan NA, De Vos P & Dinsdale AE (2009) Genus *Geobacillus*. *Bergey's Manual of Systematic Bacteriology,* Vol. 3 (De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH & Whitman WB, eds.), pp. 144-160. Springer, New York, USA.

Loman NJ & Pallen MJ (2015) Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* **13**: 787-794.

Lopez D & Kolter R (2010) Extracellular signals that define distinct and coexisting cell fates in *Bacillus subtilis*. *FEMS Microbiology Reviews* **34**: 134-149.

Lopez D, Vlamakis H & Kolter R (2009a) Generation of multiple cell types in *Bacillus subtilis*. *FEMS Microbiology Reviews* **33**: 152-163.

Lopez D, Vlamakis H, Losick R & Kolter R (2009b) Cannibalism enhances biofilm development in *Bacillus subtilis*. *Molecular Microbiology* **74**: 609-618.

Lopez D, Fischbach MA, Chu F, Losick R & Kolter R (2009c) Structurally diverse natural products that cause potassium leakage trigger multicellularity in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 280-285.

Luecking G, Stoeckel M, Atamer Z, Hinrichs J & Ehling-Schulz M (2013) Characterization of aerobic spore-forming bacteria associated with industrial dairy processing environments and product spoilage. *International Journal of Food Microbiology* **166**: 270-279.

Makarova K, Slesarev A, Wolf Y*, et al.* (2006) Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 15611-15616.

Makarova KS, Wolf YI, Alkhnbashi OS*, et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology* **13**: 722-736.

Mallidis CG & Scholefield J (1987) Relation of the heat resistance of bacterial spores to chemical composition and structure I. Relation to core components. *Journal of Applied Bacteriology* **62**: 65-69.

Marchant R, Banat IM, Rahman TJ & Berzano M (2002) The frequency and characteristics of highly thermophilic bacteria in cool soil environments. *Environmental Microbiology* **4**: 595-602.

Marquis RE & Bender GR (1985) Mineralization and heat-resistance of bacterial spores. *Journal of Bacteriology* **161**: 789-791.

Matsutani M, Shirakihara Y, Imada K, Yakushi T & Matsushita K (2013) Draft genome sequence of a thermophilic member of the Bacillaceae, *Anoxybacillus flavithermus* strain Kn10, isolated from the Kan-nawa Hot Spring in Japan. *Genome announcements* **1**: doi:10.1128/genomeA.00311-00313.

McGuiggan JTM, McCleery DR, Hannan A & Gilmour A (2002) Aerobic spore-forming bacteria in bulk raw milk: factors influencing the numbers of psychrotrophic, mesophilic and thermophilic *Bacillus* spores. *International Journal of Dairy Technology* **55**: 100-107.

McKenney PT & Eichenberger P (2012) Dynamics of spore coat morphogenesis in *Bacillus subtilis*. *Molecular Microbiology* **83**: 245-260.

McKenney PT, Driks A & Eichenberger P (2013) The *Bacillus subtilis* endospore: assembly and functions of the multilayered coat. *Nature Reviews Microbiology* **11**: 33-44.

McKenney PT, Driks A, Eskandarian HA, Grabowski P, Guberman J, Wang KH, Gitai Z & Eichenberger P (2010) A distance-weighted interaction map reveals a previously uncharacterized layer of the *Bacillus subtilis* spore coat. *Current Biology* **20**: 934-938.

McLoon AL, Kolodkin-Gal I, Rubinstein SM, Kolter R & Losick R (2011a) Spatial regulation of histidine kinases governing biofilm formation in *Bacillus subtilis*. *Journal of Bacteriology* **193**: 679-685.

McLoon AL, Guttenplan SB, Kearns DB, Kolter R & Losick R (2011b) Tracing the domestication of a biofilm-forming bacterium. *Journal of Bacteriology* **193**: 2027-2034.

Meisner J, Wang X, Serrano M, Henriques AO & Moran CP, Jr. (2008) A channel connecting the mother cell and forespore during bacterial endospore formation. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 15100-15105.

Meyer P, Gutierrez J, Pogliano K & Dworkin J (2010) Cell wall synthesis is necessary for membrane dynamics during sporulation of *Bacillus subtilis*. *Molecular Microbiology* **76**: 956-970.

Michna RH, Commichau FM, Toedter D, Zschiedrich CP & Stuelke J (2014) SubtiWiki - a database for the model organism *Bacillus subtilis* that links pathway, interaction and expression information. *Nucleic Acids Research* **42**: D692-D698.

Mika F & Hengge R (2014) Small RNAs in the control of RpoS, CsgD, and biofilm architecture of *Escherichia coli*. *Rna Biology* **11**: 494-507.

Miller RA, Kent DJ, Boor KJ, Martin NH & Wiedmann M (2015a) Different management practices are associated with mesophilic and thermophilic spore levels in bulk tank raw milk. *Journal of Dairy Science* **98**: 4338-4351.

Miller RA, Kent DJ, Watterson MJ, Boor KJ, Martin NH & Wiedmann M (2015b) Spore populations among bulk tank raw milk and dairy powders are significantly different. *Journal of Dairy Science* **98**: 8492-8504.

Minana-Galbis D, Pinzon DL, Gaspar Loren J, Manresa A & Oliart-Ros RM (2010) Reclassification of *Geobacillus pallidus* (Scholz et al. 1988) Banat et al. 2004 as *Aeribacillus pallidus* gen. nov., comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **60**: 1600-1604.

Mirouze N, Parashar V, Baker MD, Dubnau DA & Neiditch MB (2011) An atypical Phr peptide regulates the developmental switch protein RapH. *Journal of Bacteriology* **193**: 6197-6206.

Miyamoto M, Motooka D, Gotoh K*, et al.* (2014) Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics* **15**: doi: 10.1186/1471-2164-1115-1699.

Moody KL, Driks A, Rother GL, Cote CK, Brueggemann EE, Hines HB, Friedlander AM & Bozue J (2010) Processing, assembly and localization of a *Bacillus anthracis* spore protein. *Microbiology* **156**: 174-183.

Morlot C, Uehara T, Marquis KA, Bernhardt TG & Rudner DZ (2010) A highly coordinated cell wall degradation machine governs spore morphogenesis in *Bacillus subtilis*. *Genes & Development* **24**: 411-422.

Mostert JF, Luck H & Husmann RA (1979) Isolation identification and practical properties of *Bacillus* sp. from ultra high temperature and sterilized milk. *South African Journal of Dairy Technology* **11**: 125-132.

Muir DD, Griffiths MW, Phillips JD, Sweetsur AWM & West IG (1986) Effect of the bacterial quality of raw milk on the bacterial quality and some other properties of low-heat and high-heat dried milk. *Journal of the Society of Dairy Technology* **39**: 115-118.

Murphy PM, Lynch D & Kelly PM (1999) Growth of thermophilic spore forming bacilli in milk during the manufacture of low heat powders. *International Journal of Dairy Technology* **52**: 45-50.

Murray EJ, Kiley TB & Stanley-Wall NR (2009) A pivotal role for the response regulator DegU in controlling multicellular behaviour. *Microbiology* **155**: 1-8.

Murugan B & Villi RA (2009) Lipolytic activity of *Bacillus* species isolated from milk and dairy products. *Indian Veterinary Journal* **86**: 80-81.

Nazina TN, Tourova TP & Poltaraus AB (2001) Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *Geobacillus stearothermophilus*, *Geobacillus thermocatenulatus*, *Geobacillus thermoleovorans*, *Geobacillus kaustophilus*, *Geobacillus thermoglucosidasius* and *Geobacillus thermodenitrificans*. *International Journal of Systematic and Evolutionary Microbiology* **51**: 433 - 446.

Nazina TN, Lebedeva EV, Poltaraus AB, Tourova TP, Grigoryan AA, Sokolova DS, Lysenko AM & Osipov GA (2004) *Geobacillus gargensis* sp nov., a novel thermophile from a hot spring, and the reclassification of *Bacillus vulcani* as *Geobacillus vulcani* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **54**: 2019-2024.

Nazina TN, Sokolova DS, Grigoryan AA, Shestakova NM, Mikhailova EM, Poltaraus AB, Tourova TP, Lysenko AM, Osipov GA & Belyaev SS (2005) *Geobacillus jurassicus* sp. nov., a new thermophilic bacterium isolated from a high-temperature petroleum reservoir, and the validation of the *Geobacillus* species. *Systematic and Applied Microbiology* **28**: 43-53.

Nicholson WL, Munakata N, Horneck G, Melosh HJ & Setlow P (2000) Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial environments. *Microbiology and Molecular Biology Reviews* **64**: 548-572.

Nold SC, Kopczynski ED & Ward DM (1996) Cultivation of aerobic chemoorganotrophic proteobacteria and Gram-positive bacteria from a hot spring microbial mat. *Applied and Environmental Microbiology* **62**: 3917-3921.

Nugroho FA, Yamamoto H, Kobayashi Y & Sekiguchi J (1999) Characterization of a new Sigma-K-dependent peptidoglycan hydrolase gene that plays a role in *Bacillus subtilis* mother cell lysis. *Journal of Bacteriology* **181**: 6230-6237.

O'Gara JP (2007) *ica* and beyond: biofilm mechanisms and regulation in *Staphylococcus epidermidis* and *Staphylococcus aureus*. *FEMS Microbiology Letters* **270**: 179-188.

O'Neill E, Pozzi C, Houston P, Smyth D, Humphreys H, Robinson DA & O'Gara JP (2007) Association between methicillin susceptibility and biofilm regulation in *Staphylococcus aureus* isolates from device-related infections. *Journal of Clinical Microbiology* **45**: 1379-1388.

Oh MH & Cox JM (2009) Toxigenic bacilli associated with food poisoning. *Food Science and Biotechnology* **18**: 594-603.

Oke V & Losick R (1993) Multilevel regulation of the sporulation transcription factor $\sigma^K$ in *Bacillus subtilis*. *Journal of Bacteriology* **175**: 7341-7347.

Onyenwoke RU, Brill JA, Farahi K & Wiegel J (2004) Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). *Archives of Microbiology* **182**: 182-192.

Oomes SJCM, Jonker MJ, Wittink FRA, Hehenkamp JO, Breit TM & Brul S (2009) The effect of calcium on the transcriptome of sporulating *B. subtilis* cells. *International Journal of Food Microbiology* **133**: 234-242.

Paidhungat M, Ragkousi K & Setlow P (2001) Genetic requirements for induction of germination of spores of *Bacillus subtilis* by $Ca^{2+}$-dipicolinate. *Journal of Bacteriology* **183**: 4886-4893.

Paidhungat M, Setlow B, Driks A & Setlow P (2000) Characterization of spores of *Bacillus subtilis* which lack dipicolinic acid. *Journal of Bacteriology* **182**: 5505-5512.

Palmer J (2008) Surface characteristics of an adhesive thermophilic spore-forming *Bacillus* isolated from milk powder. PhD thesis Thesis, Massey University, Palmerston North, New Zealand.

Palmer JS, Flint SH, Schmid J & Brooks JD (2010) The role of surface charge and hydrophobicity in the attachment of *Anoxybacillus flavithermus* isolated from milk powder. *Journal of Industrial Microbiology & Biotechnology* **37**: 1111-1119.

Palop A, Manas P & Condon S (1999a) Sporulation temperature and heat resistance of Bacillus spores: A review. *Journal of Food Safety* **19**: 57-72.

Palop A, Sala FJ & Condon S (1999b) Heat resistance of native and demineralized spores of *Bacillus subtilis* sporulated at different temperatures. *Applied and Environmental Microbiology* **65**: 1316-1319.

Parashar V, Mirouze N, Dubnau DA & Neiditch MB (2011) Structural basis of response regulator dephosphorylation by Rap phosphatases. *PLoS Biology* **9**: doi: 10.1371/journal.pbio.1000589.

Paredes-Sabja D, Setlow P & Sarker MR (2011) Germination of spores of Bacillales and Clostridiales species: mechanisms and proteins involved. *Trends in Microbiology* **19**: 85-94.

Paredes CJ, Alsaker KV & Papoutsakis ET (2005) A comparative genomic view of clostridial sporulation and physiology. *Nature Reviews Microbiology* **3**: 969-978.

Parkar SG, Flint SH & Brooks JD (2003) Physiology of biofilms of thermophilic bacilli - potential consequences for cleaning. *Journal of Industrial Microbiology & Biotechnology* **30**: 553-560.

Parkar SG, Flint SH, Palmer JS & Brooks JD (2001) Factors influencing attachment of thermophilic bacilli to stainless steel. *Journal of Applied Microbiology* **90**: 901-908.

Pasvolsky R, Zakin V, Ostrova I & Shemesh M (2014) Butyric acid released during milk lipolysis triggers biofilm formation of *Bacillus* species. *International Journal of Food Microbiology* **181**: 19-27.

Pavic S, Brett M, Petric N, Lastre D, Smoljanovic M, Atkinson M, Kovacic A, Cetinic E & Ropac D (2005) An outbreak of food poisoning in a kindergarten caused by milk powder containing toxigenic *Bacillus subtilis* and *Bacillus licheniformis*. *Archiv Fur Lebensmittelhygiene* **56**: 20-22.

Pei A, Li H, Oberdorf WE*, et al.* (2012) Diversity of 5S rRNA genes within individual prokaryotic genomes. *FEMS Microbiology Letters* **335**: 11-18.

Pei AY, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM & Pei Z (2009) Diversity of 23S rRNA genes within individual prokaryotic genomes. *PLoS One* **4**: doi: e543710.541371/journal.pone.0005437.

Pei AY, Oberdorf WE, Nossa CW*, et al.* (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. *Applied and Environmental Microbiology* **76**: 3886-3897.

Perego M (2001) A new family of aspartyl phosphate phosphatases targeting the sporulation transcription factor Spo0A of *Bacillus subtilis*. *Molecular Microbiology* **42**: 133-143.

Perego M (2013) Forty years in the making: understanding the molecular mechanism of peptide regulation in bacterial development. *PLoS Biology* **11**.

Perego M, Glaser P & Hoch JA (1996) Aspartyl-phosphate phosphatases deactivate the response regulator components of the sporulation signal transduction system in *Bacillus subtilis*. *Molecular Microbiology* **19**: 1151-1157.

Perez-Valdespino A, Ghosh S, Cammett EP, Kong L, Li YQ & Setlow P (2013) Isolation and characterization of *Bacillus subtilis* spores that are superdormant for germination with dodecylamine or $Ca^{2+}$-dipicolinic acid. *Journal of Applied Microbiology* **114**: 1109-1119.

Perez-Valdespino A, Li Y, Setlow B*, et al.* (2014) Function of the SpoVAEa and SpoVAF proteins of *Bacillus subtilis* spores. *Journal of Bacteriology* **196**: 2077-2088.

Pettersson B, Lembke F, Hammer P, Stackebrandt E & Priest FG (1996) *Bacillus sporothermodurans*, a new species producing highly heat-resistant endospores. *International Journal of Systematic Bacteriology* **46**: 759-764.

Pettit LJ, Browne HP, Yu L*, et al.* (2014) Functional genomics reveals that *Clostridium difficile* Spo0A coordinates sporulation, virulence and metabolism. *BMC Genomics* **15**.

Pikuta E, Cleland D & Tang J (2003) Aerobic growth of *Anoxybacillus pushchinoensis* K1(T) : Emended descriptions of *A. pushchinoensis* and the genus *Anoxybacillus*. *International Journal of Systematic and Evolutionary Microbiology* **53**: 1561-1562.

Pikuta E, Lysenko AM, Chuvilskaya N, Mendrock U, Hippe H, Suzina N, Nikitin D, Osipov GA & Laurinavichius K (2000) *Anoxybacillus pushcinensis* gen. nov., sp nov., a novel anaerobic, alkaliphilic, moderately thermophilic bacterium from manure, and description of *Anoxybacillus flavothermus* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **50**: 2109-2117.

Poli A, Laezza G, Gul-Guven R, Orlando P & Nicolaus B (2011) *Geobacillus galactosidasius* sp nov., a new thermophilic galactosidase-producing bacterium isolated from compost. *Systematic and Applied Microbiology* **34**: 419-423.

Pore SD, Arora P & Dhakephalkar PK (2014) Draft genome sequence of *Geobacillus* sp. strain FW23, isolated from a formation water sample. *Genome announcements* **2**: doi:10.1128/genomeA.00352-00314.

Postollec F, Mathot A-G, Bernard M, Divanac'h M-L, Pavan S & Sohier D (2012) Tracking spore-forming bacteria in food: From natural biodiversity to selection by processes. *International Journal of Food Microbiology* **158**: 1-8.

Price CE, Zeyniyev A, Kuipers OP & Kok J (2012) From meadows to milk to mucosa - adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. *FEMS Microbiology Reviews* **36**: 949-971.

Priest FG, Goodfellow M & Todd C (1988) A numerical classification of the genus *Bacillus*. *Journal of General Microbiology* **134**: 1847-1882.

Qinglong W, Cheung CKW & Nagendra PS (2015) Towards galactose accumulation in dairy foods fermented by conventional starter cultures: challenges and strategies. *Trends in Food Science & Technology* **41**: 24-36.

Rainey FA, Fritze D & Stackebrandt E (1994) The phylogenetic diversity of thermophilic members of the genus *Bacillus* as revealed by 16S rDNA analysis. *FEMS Microbiology Letters* **115**: 205-211.

Read TD, Peterson SN, Tourasse N*, et al.* (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**: 81-86.

Redmond C, Baillie LWJ, Hibbs S, Moir AJG & Moir A (2004) Identification of proteins in the exosporium of *Bacillus anthracis*. *Microbiology* **150**: 355-363.

Refstrup E (1998) Heat treatment of milk prior to evaporation and drying. pp. 228-237. International Dairy Federation, Denmark.

Refstrup E (2000) Evaporation and drying technology developments. *International Journal of Dairy Technology* **53**: 163-167.

Reginensi SM, Gonzalez MJ, Olivera JA, Sosa M, Juliano P & Bermudez J (2011) RAPD-based screening for spore-forming bacterial populations in Uruguayan commercial powdered milk. *International Journal of Food Microbiology* **148**: 36-41.

Ren PG & Zhou PJ (2005) *Tenuibacillus multivorans* gen. nov., sp nov., a moderately halophilic bacterium isolated from saline soil in Xin-Jiang, China. *International Journal of Systematic and Evolutionary Microbiology* **55**: 95-99.

Richter M & Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 19126-19131.

Ricker N, Qian H & Fulthorpe RR (2012) The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics* **100**: 167-175.

Rode TM, Langsrud S, Holck A & Moretro T (2007) Different patterns of biofilm formation in *Staphylococcus aureus* under food-related stress conditions. *International Journal of Food Microbiology* **116**: 372-383.

Romero D, Aguilar C, Losick R & Kolter R (2010) Amyloid fibers provide structural integrity to *Bacillus subtilis* biofilms. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 2230-2234.

Romero D, Vlamakis H, Losick R & Kolter R (2011) An accessory protein required for anchoring and assembly of amyloid fibres in *B. subtilis* biofilms. *Molecular Microbiology* **80**: 1155-1168.

Romling U, Galperin MY & Gomelsky M (2013) Cyclic di-GMP: the first 25 Years of a universal bacterial second messenger. *Microbiology and Molecular Biology Reviews* **77**: 1-52.

Ronimus RS, Parker LE & Morgan HW (1997) The utilization of RAPD-PCR for identifying thermophilic and mesophilic *Bacillus* species. *FEMS Microbiology Letters* **147**: 75-79.

Ronimus RS, Rueckert A & Morgan HW (2006) Survival of thermophilic spore-forming bacteria in a 90(+) year old milk powder from Ernest Shackelton's Cape Royds Hut in Antarctica. *Journal of Dairy Research* **73**: 235-243.

Ronimus RS, Parker LE, Turner N, Poudel S, Ruckert A & Morgan HW (2003) A RAPD-based comparison of thermophilic bacilli from milk powders. *International Journal of Food Microbiology* **85**: 45-61.

Ronner U, Husmark U & Henriksson A (1990) Adhesion of *Bacillus* spores in relation to hydrophobicity. *Journal of Applied Bacteriology* **69**: 550-556.

Rossler D, Ludwig W, Schleifer KH, Lin CZ, McGill TJ, Wisotzkey JD, Jurtshuk P & Fox GE (1991) Phylogenetic diversity in the genus *Bacillus* as seen by 16S ribosomal-RNA sequencing studies. *Systematic and Applied Microbiology* **14**: 266-269.

Rowland SL, Burkholder WF, Cunningham KA, Maciejewski MW, Grossman AD & King GF (2004) Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in *Bacillus subtilis*. *Molecular Cell* **13**: 689-701.

Rozanov AS, Bryanskaya AV, Kotenko AV, Malup TK & Peltek SE (2014) Draft genome sequence of *Anoxybacillus flavithermus* strain 25, isolated from the Garga Hot Spring in the Barguzin Valley, Baikal Region, Russian Federation. *Genome announcements* **2**: doi:10.1128/genomeA.01258-01214.

Rueckert A, Ronimus RS & Morgan HW (2004) A RAPD-based survey of thermophilic bacilli in milk powders from different countries. *International Journal of Food Microbiology* **96**: 263-272.

Rueckert A, Ronimus RS & Morgan HW (2005a) Rapid differentiation and enumeration of the total, viable vegetative cell and spore content of thermophilic bacilli in milk powders with reference to *Anoxybacillus flavithermus*. *Journal of Applied Microbiology* **99**: 1246-1255.

Rueckert A, Ronimus RS & Morgan HW (2005b) Development of a rapid detection and enumeration method for thermophilic bacilli in milk powders. *Journal of Microbiological Methods* **60**: 155-167.

Rueckert A, Ronimus RS & Morgan HW (2006) Development of a real-time PCR assay targeting the sporulation gene, *spo0A*, for the enumeration of thermophilic bacilli in milk powder. *Food Microbiology* **23**: 220-230.

Ruzal SM, Bustos PL & Sanchez-Rivas C (2013) "A La Recherche" of functions for the spore protein SASP-E from *Bacillus subtilis*. *Journal of Microbiology and Biotechnology* **23**: 15-21.

Ryder C, Byrd M & Wozniak DJ (2007) Role of polysaccharides in *Pseudomonas aeruginosa* biofilm development. *Current Opinion in Microbiology* **10**: 644-648.

Sadiq FA, Li Y, Liu T, Flint S, Zhang G & He G (2016) A RAPD based study revealing a previously unreported wide range of mesophilic and thermophilic spore formers associated with milk powders in China. *International Journal of Food Microbiology* **217**: 200-208.

Sanchez-Salas J-L, Setlow B, Zhang P, Li Y-q & Setlow P (2011) Maturation of released spores is necessary for acquisition of full spore heat resistance during *Bacillus subtilis* sporulation. *Applied and Environmental Microbiology* **77**: 6746-6754.

Sang W & Sim SJ (2006) Increased heat resistance of *Geobacillus stearothermophilus* spores heat-shocked during sporulation. *Journal of Microbiology and Biotechnology* **16**: 633-636.

Scaramozzino F, White A, Perego M & Hoch JA (2009) A unique GTP-dependent sporulation sensor histidine kinase in *Bacillus anthracis*. *Journal of Bacteriology* **191**: 687-692.

Scheldeman P, Herman L, Foster S & Heyndrickx M (2006) *Bacillus sporothermodurans* and other highly heat-resistant spore formers in milk. *Journal of Applied Microbiology* **101**: 542-555.

Schwarzenbach RR & Hill BM (1999) Do bacterial spores in milk powders survive UHT processing. pp. 53-58. New Zealand Dairy Research Institute, Palmerston North, 30 June - 2 July 1999.

Scott SA (2005) The growth of thermophilic bacteria in a milk powder plant and the formation of spores in biofilms of the dairy thermophile *Anoxybacillus flavithermus*. MSc Thesis, Massey University, Palmerston North.

Scott SA, Brooks JD, Rakonjac J, Walker KMR & Flint SH (2007) The formation of thermophilic spores during the manufacture of whole milk powder. *International Journal of Dairy Technology* **60**: 109-117.

Seale B, Burgess SA, Flint SH, Brooks J, Bremer PJ & Parkar SG (2015) Thermophilic spore-forming bacilli in the dairy industry. *Biofilms in the Dairy Industry,*(Teh KH, Flint SH, Brooks J & Knight GC, eds.), pp. 112-137. John Wiley & Sons, Ltd., Chichester, West Sussex, UK.

Seale RB (2009) The surface characteristics of spores from thermophilic bacilli isolated from a milk powder production line and their influence on adhesion to surfaces. PhD Thesis, The University of Otago, Dunedin, New Zealand.

Seale RB, Flint SH, McQuillan AJ & Bremer PJ (2008) Recovery of spores from thermophilic dairy bacilli and effects of their surface characteristics on attachment to different surfaces. *Applied and Environmental Microbiology* **74**: 731-737.

Seale RB, Bremer PJ, Flint SH & McQuillan AJ (2010) Characterization of spore surfaces from a *Geobacillus* sp. isolate by pH dependence of surface charge and infrared spectra. *Journal of Applied Microbiology* **109**: 1339-1348.

Seale RB, Flint SH, McQuillan AJ & Bremer PJ (2011) Effect of NaOH (caustic wash) on the viability, surface characteristics and adhesion of spores of a *Geobacillus* sp. isolated from a milk powder production line. *Letters in Applied Microbiology* **52**: 104-108.

Seale RB, Dhakal R, Chauhan K, Craven HM, Deeth HC, Pillidge CJ, Powell IB & Turner MS (2012) Genotyping of present-day and historical *Geobacillus* species isolates from milk powders by High-Resolution Melt Analysis of Multiple-Variable-Number Tandem-Repeat Loci. *Applied and Environmental Microbiology* **78**: 7090-7097.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068-2069.

Setlow B, Atluri S, Kitchel R, Koziol-Dube K & Setlow P (2006) Role of dipicolinic acid in resistance and stability of spores of *Bacillus subtilis* with or without DNA-protective α/β type small acid-soluble proteins. *Journal of Bacteriology* **188**: 3740-3747.

Setlow P (1988) Small, acid-soluble spore proteins of *Bacillus* species - structure, synthesis, genetics, function, and degradation. *Annual Review of Microbiology* **42**: 319-338.

Setlow P (1995) Mechanisms for the prevention of damage to DNA in spores of *Bacillus* species. *Annual Review of Microbiology* **49**: 29-54.

Setlow P (2006) Spores of *Bacillus subtilis:* their resistance to and killing by radiation, heat and chemicals. *Journal of Applied Microbiology* **101**: 514-525.

Setlow P (2007) I will survive: DNA protection in bacterial spores. *Trends in Microbiology* **15**: 172-180.

Setlow P (2014a) Germination of spores of *Bacillus* species: What we know and do not know. *Journal of Bacteriology* **196**: 1297-1305.

Setlow P (2014b) Spore resistance properties. *Microbiology Spectrum* **2**: doi:10.1128/microbiolspec.TBS-0003-2012.

Setlow P, Liu J & Faeder JR (2012) Heterogeneity in bacterial spore populations. *Bacterial Spores,*(Abel-Santos E, ed.) pp. 199-214. Caister Academic Press, Norfolk, UK.

Severson KM, Mallozzi M, Bozue J, Welkos SL, Cote CK, Knight KL & Driks A (2009) Roles of the *Bacillus anthracis* spore protein ExsK in exosporium maturation and germination. *Journal of Bacteriology* **191**: 7587-7596.

Shariat N & Dudley EG (2014) CRISPRs: Molecular signatures used for pathogen subtyping. *Applied and Environmental Microbiology* **80**: 430-439.

Shariat N, Sandt CH, DiMarzio MJ, Barrangou R & Dudley EG (2013) CRISPR-MVLST subtyping of *Salmonella enterica* subsp *enterica* serovars Typhimurium and Heidelberg and application in identifying outbreak isolates. *BMC Microbiology* **13**.

Shariat N, Timme RE, Pettengill JB, Barrangou R & Dudley EG (2015) Characterization and evolution of *Salmonella* CRISPR-Cas systems. *Microbiology* **161**: 374-386.

Shemesh M & Chai Y (2013) A combination of glycerol and manganese promotes biofilm formation in *Bacillus subtilis* via histidine kinase KinD signaling. *Journal of Bacteriology* **195**: 2747-2754.

Shemesh M, Kolter R & Losick R (2010) The biocide chlorine dioxide stimulates biofilm formation in *Bacillus subtilis* by activation of the histidine kinase KinC. *Journal of Bacteriology* **192**: 6352-6356.

Shida O, Takagi H, Kadowaki K & Komagata K (1996) Proposal for two new genera, *Brevibacillus* gen nov and *Aneurinibacillus* gen nov. *International Journal of Systematic Bacteriology* **46**: 939-946.

Shintani M, Ohtsubo Y, Fukuda K*, et al.* (2014) Complete genome sequence of the thermophilic polychlorinated biphenyl degrader *Geobacillus* sp. strain JF8 (NBRC 109937). *Genome announcements* **2**: doi:10.1128/genomeA.01213-01213.

Siddiqui M, Rashid N, Ayyampalayam S & Whitman WB (2014) Draft genome sequence of *Geobacillus thermopakistaniensis* strain MAS1. *Genome announcements* **2**: doi:10.1128/genomeA.00559-00514.

Siezen RJ, Renckens B, van Swam I, Peters S, van Kranenburg R, Kleerebezem M & de Vos WM (2005) Complete sequences of four plasmids of *Lactococcus lactis* subsp cremoris SK11 reveal extensive adaptation to the dairy environment. *Applied and Environmental Microbiology* **71**: 8371-8382.

Siezen RJ, Bayjanov JR, Felis GE, van der Sijde MR, Starrenburg M, Molenaar D, Wels M, van Hijum SAFT & Vlieg JETvH (2011) Genome-scale diversity and niche adaptation analysis of *Lactococcus lactis* by comparative genome hybridization using multi-strain arrays. *Microbial Biotechnology* **4**: 383-402.

Silvaggi JM, Popham DL, Driks A, Eichenberger P & Losick R (2004) Unmasking novel sporulation genes in *Bacillus subtilis*. *Journal of Bacteriology* **186**: 8089-8095.

Smokvina T, Wels M, Polka J, Chervaux C, Brisse S, Boekhorst J, Vlieg JETvH & Siezen RJ (2013) *Lactobacillus paracasei* comparative genomics: Towards species pan-genome definition and exploitation of diversity. *PLoS One* **8**: doi:10.1371/journal.pone.0068731.

Sneath PHA (1986) Endospore-forming Gram-positive rods and cocci. *Bergey's Manual of Systematic Bacteriology,* Vol. 2 (Sneath PHA, Mair NS, Sharpe ME & Holt JG, eds.), pp. 1104-1105.

Somerton B, Flint S, Palmer J, Brooks J & Lindsay D (2013) Preconditioning with cations increases the attachment of *Anoxybacillus flavithermus* and *Geobacillus* species to stainless steel. *Applied and Environmental Microbiology* **79**: 4186-4190.

Somerton B, Lindsay D, Palmer J, Brooks J & Flint S (2015) Changes in sodium, calcium, and magnesium ion concentrations that inhibit *Geobacillus* biofilms have no effect on *Anoxybacillus flavithermus* biofilms. *Applied and Environmental Microbiology* **81**: 5115-5122.

Sorek R, Kunin V & Hugenholtz P (2008) CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology* **6**: 181-186.

Splittstoesser DF, Lee CY & Churey JJ (1998) Control of *Alicyclobacillus* in the juice industry. *Dairy, Food and Environmental Sanitation* **18**: 585-587.

Stadhouders J, Hup G & Hassing F (1982) The conceptions index and indicator organisms discussed on the basis of the bacteriology of spray-dried milk powder. *Netherlands Milk and Dairy Journal* **36**: 231-260.

Stanley NR & Lazazzera BA (2005) Defining the genetic differences between wild and domestic strains of *Bacillus subtilis* that affect poly-gamma-DL-glutamic acid production and biofilm formation. *Molecular Microbiology* **57**: 1143-1158.

Steichen C, Chen P, Kearney JF & Turnbough CL (2003) Identification of the immunodominant protein and other proteins of the *Bacillus anthracis* exosporium. *Journal of Bacteriology* **185**: 1903-1910.

Steil L, Serrano M, Henriques AO & Volker U (2005) Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of *Bacillus subtilis*. *Microbiology* **151**: 399-420.

Stephenson K & Hoch JA (2001) PAS-A domain of phosphorelay sensor kinase A: A catalytic ATP-binding domain involved in the initiation of development in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 15251-15256.

Stephenson K & Hoch JA (2002) Evolution of signalling in the sporulation phosphorelay. *Molecular Microbiology* **46**: 297-304.

Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genetics* **26**: 335-340.

Stewart GC (2015) The exosporium layer of bacterial spores: A connection to the environment and the infected host. *Microbiology and Molecular Biology Reviews* **79**: 437-457.

Stoodley P, Lewandowski Z, Boyle JD & Lappin-Scott HM (1999) The formation of migratory ripples in a mixed species bacterial biofilm growing in turbulent flow. *Environmental Microbiology* **1**: 447-455.

Stoodley P, Sauer K, Davies DG & Costerton JW (2002) Biofilms as complex differentiated communities. *Annual Review of Microbiology* **56**: 187-209.

Stragier P & Losick R (1996) Molecular genetics of sporulation in *Bacillus subtilis*. *Annual Review of Genetics* **30**: 297-341.

Stragier P, Kunkel B, Kroos L & Losick R (1989) Chromosomal rearrangement generating a composite gene for a developmental transcription factor. *Science* **243**: 507-512.

Studholme DJ (2015) Some (bacilli) like it hot: genomics of *Geobacillus* species. *Microbial Biotechnology* **8**: 40-48.

Suihko ML, Sinkko H, Partanen L, Mattila-Sandholm T, Salkinoja-Salonen M & Raaska L (2004) Description of heterotrophic bacteria occurring in paper mills and paper products. *Journal of Applied Microbiology* **97**: 1228-1235.

Sun DL, Jiang X, Wu QLL & Zhou NY (2013) Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Applied and Environmental Microbiology* **79**: 5962-5969.

Sung MH, Kim H, Bae JW*, et al.* (2002) *Geobacillus toebii* sp nov., a novel thermophilic bacterium isolated from hay compost. *International Journal of Systematic and Evolutionary Microbiology* **52**: 2251-2255.

Sunna A, Tokajian S, Burghardt J, Rainey F, Antranikian G & Hashwa F (1997) Identification of *Bacillus kaustophilus*, *Bacillus thermocatenulatus* and *Bacillus* strain HSR as members of *Bacillus thermoleovorans*. *Systematic and Applied Microbiology* **20**: 232-237.

Sylvestre P, Couture-Tosi E & Mock M (2002) A collagen-like surface glycoprotein is a structural component of the *Bacillus anthracis* exosporium. *Molecular Microbiology* **45**: 169-178.

Sylvestre P, Couture-Tosi E & Mock M (2005) Contribution of ExsFA and ExsFB proteins to the localization of BclA on the spore surface and to the stability of the *Bacillus anthracis* exosporium. *Journal of Bacteriology* **187**: 5122-5128.

Tai SK, Lin HPP, Kuo J & Liu JK (2004) Isolation and characterization of a cellulolytic *Geobacillus thermoleovorans* T4 strain from sugar refinery wastewater. *Extremophiles* **8**: 345-349.

Takamatsu H, Hiraoka T, Kodama T, Koide H, Kozuka S, Tochikubo K & Watabe K (1998) Cloning of a novel gene *yrbB*, encoding a protein located in the spore integument of *Bacillus subtilis*. *FEMS Microbiology Letters* **166**: 361-367.

Takami H, Nishi S, Lu J, Shinamura S & Takaki Y (2004a) Genomic characterization of thermophilic *Geobacillus* species isolated from the deepest sea mud of the Mariana trench. *Extremophiles* **8**: 351-356.

Takami H, Takaki Y, Chee GJ, Nishi S, Shimamura S, Suzuki H, Matsui S & Uchiyama I (2004b) Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. *Nucleic Acids Research* **32**: 6292-6303.

Takemaru K, Mizuno M, Sato T, Takeuchi M & Kobayashi Y (1995) Complete nucleotide sequence of a skin element excised by DNA rearrangement during sporulation in *Bacillus subtilis*. *Microbiology* **141**: 323-327.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731-2739.

Tatusov RL, Galperin MY, Natale DA & Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**: 33-36.

Tatusov RL, Fedorova ND, Jackson JD*, et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: doi: 10.1186/1471-2105-1184-1141.

Teeling H, Meyerdierks A, Bauer M, Amann R & Glockner FO (2004a) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology* **6**: 938-947.

Teeling H, Waldmann J, Lombardot T, Bauer M & Glockner FO (2004b) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**: doi: 10.1186/1471-2105-5-163.

Thompson BM & Stewart GC (2008) Targeting of the BclA and BclB proteins to the *Bacillus anthracis* spore surface. *Molecular Microbiology* **70**: 421-434.

Thompson BM, Hoelscher BC, Driks A & Stewart GC (2011) Localization and assembly of the novel exosporium protein BetA of *Bacillus anthracis*. *Journal of Bacteriology* **193**: 5098-5104.

Thompson BM, Hoelscher BC, Driks A & Stewart GC (2012) Assembly of the BclB glycoprotein into the exosporium and evidence for its role in the formation of the exosporium cap' structure in *Bacillus anthracis*. *Molecular Microbiology* **86**: 1073-1084.

Thompson LS, Beech PL, Real G, Henriques AO & Harry EJ (2006) Requirement for the cell division protein DivIB in polar cell division and engulfment during sporulation in *Bacillus subtilis*. *Journal of Bacteriology* **188**: 7677-7685.

Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W & Kaempfer P (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology* **60**: 249-266.

Todd SJ, Moir AJG, Johnson MJ & Moir A (2003) Genes of *Bacillus cereus* and *Bacillus anthracis* encoding proteins of the exosporium. *Journal of Bacteriology* **185**: 3373-3378.

Tojo S, Hirooka K & Fujita Y (2013) Expression of *kinA* and *kinB* of *Bacillus subtilis*, necessary for sporulation initiation, is under positive stringent transcription control. *Journal of Bacteriology* **195**: 1656-1665.

Tourova TP, Korshunova AV, Mikhailova EM, Sokolova DS, Poltaraus AB & Nazina TN (2010) Application of *gyrB* and *parE* sequence similarity analyses for differentiation of species within the genus *Geobacillus*. *Microbiology* **79**: 356-369.

Tovar-Rojo F, Chander M, Setlow B & Setlow P (2002) The products of the *spoVA* operon are involved in dipicolinic acid uptake into developing spores of *Bacillus subtilis*. *Journal of Bacteriology* **184**: 584-587.

Trach KA & Hoch JA (1993) Multisensory activation of the phosphorelay initiating sporulation in *Bacillus subtilis:* identification and sequence of the protein kinase of the alternate pathway. *Molecular Microbiology* **8**: 69-79.

Tyson GW & Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environmental Microbiology* **10**: 200-207.

van Baar BLM (2000) Characterisation of bacteria by matrix-assisted laster desorption/ionisation and electrospray mass spectrometry. *FEMS Microbiology Reviews* **24**: 193-219.

Van der Heiden E, Delmarcelle M, Lebrun S, Freichels R, Brans A, Vastenavond CM, Galleni M & Joris B (2013) A pathway closely related to the D-Tagatose pathway of Gram-Negative Enterobacteria identified in the Gram-Positive bacterium *Bacillus licheniformis*. *Applied and Environmental Microbiology* **79**: 3511-3515.

Velasquez J, Schuurman-Wolters G, Birkner JP, Abee T & Poolman B (2014) *Bacillus subtilis* spore protein SpoVAC functions as a mechanosensitive channel. *Molecular Microbiology* **92**: 813-823.

Vepachedu VR & Setlow P (2004) Analysis of the germination of spores of *Bacillus subtilis* with temperature sensitive *spo* mutations in the *spoVA* operon. *FEMS Microbiology Letters* **239**: 71-77.

Vepachedu VR & Setlow P (2007a) Analysis of interactions between nutrient germinant receptors and SpoVA proteins of *Bacillus subtilis* spores. *FEMS Microbiology Letters* **274**: 42-47.

Vepachedu VR & Setlow P (2007b) Role of SpoVA proteins in release of dipicolinic acid during germination of *Bacillus subtilis* spores triggered by dodecylamine or lysozyme. *Journal of Bacteriology* **189**: 1565-1572.

Vercoe RB, Chang JT, Dy RL, Taylor C, Gristwood T, Clulow JS, Richter C, Przybilski R, Pitman AR, Fineran PC. 2013. Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genetics* **9** doi: 10.1371/journal.pgen.1003454.

Verhamme DT, Murray EJ & Stanley-Wall NR (2009) DegU and Spo0A jointly control transcription of two loci required for complex colony development by *Bacillus subtilis*. *Journal of Bacteriology* **191**: 100-108.

Visser H & Jeurnink TJM (1997a) General aspects of fouling and cleaning. *Bulletin of the International Dairy Federation* 5-6.

Visser J & Jeurnink TJM (1997b) Fouling of heat exchangers in the dairy industry. *Experimental Thermal and Fluid Science* **14**: 407-424.

Vlamakis H & Kolter R (2011) Biofilms. *Bacterial Stress Responses,*(G. S & Hengge R, eds.), pp. 365-373. ASM Press, Washington, DC.

Vlamakis H, Aguilar C, Losick R & Kolter R (2008) Control of cell fate by the formation of an architecturally complex bacterial community. *Genes & Development* **22**: 945-953.

Vlamakis H, Chai Y, Beauregard P, Losick R & Kolter R (2013) Sticking together: building a biofilm the *Bacillus subtilis* way. *Nature Reviews Microbiology* **11**: 157-168.

Voort Mvd & Abee T (2013) Sporulation environment of emetic toxin-producing *Bacillus cereus* strains determines spore size, heat resistance and germination capacity. *Journal of Applied Microbiology* **114**: 1201-1210.

Vries YPd (2006) *Bacillus cereus* spore formation, structure, and germination. PhD Thesis, Wageningen University.

Waino M, Tindall BJ, Schumann P & Ingvorsen K (1999) *Gracilibacillus* gen. nov., with description of *Gracilibacillus halotolerans* gen. nov., sp. nov.; transfer of *Bacillus dipsosauri* to *Gracilibacillus dipsosauri* comb. nov., and *Bacillus salexigens* to the genus *Salibacillus* gen. nov., as *Salibacillus salexigens* comb. nov. *International Journal of Systematic Bacteriology* **49**: 821-831.

Wang L, Grau R, Perego M & Hoch JA (1997) A novel histidine kinase inhibitor regulating development in *Bacillus subtilis*. *Genes & Development* **11**: 2569-2579.

Wang L, Perpich J, Driks A & Kroos L (2007) One perturbation of the mother cell gene regulatory network suppresses the effects of another during sporulation of *Bacillus subtilis*. *Journal of Bacteriology* **189**: 8467-8473.

Wang L, Fabret C, Kanamaru K, Stephenson K, Dartois V, Perego M & Hoch JA (2001) Dissection of the functional and structural domains of phosphorelay histidine kinase a of *Bacillus subtilis*. *Journal of Bacteriology* **183**: 2795-2802.

Wang ST, Setlow B, Conlon EM, Lyon JL, Imamura D, Sato T, Setlow P, Losick R & Eichenberger P (2006) The forespore line of gene expression in *Bacillus subtilis. Journal of Molecular Biology* **358**: 16-37.

Wang Y-T, Missiakas D & Schneewind O (2014) GneZ, a UDP-GlcNAc 2-epimerase, is required for s-layer assembly and vegetative growth of *Bacillus anthracis*. *Journal of Bacteriology* **196**: 2969-2978.

Warnecke F (2001) The ecology of thermophilic bacilli of milk powder processing plants. M.Phil. Thesis, University of Waikato, Hamilton.

Warriner K & Waites WM (1999) Enhanced sporulation in *Bacillus subtilis* grown on medium containing glucose : ribose. *Letters in Applied Microbiology* **29**: 97-102.

Watkinson WJ (2008) Chemistry of detergents and disinfectants. *Cleaning-in-Place: Dairy, Food and Beverage Operations,*(Tamime A, ed.) pp. 56-80. Blackwell Publishing, Oxford.

Weng FY, Chiou CS, Lin PHP & Yang SS (2009) Application of *recA* and *rpoB* sequence analysis on phylogeny and molecular identification of *Geobacillus* species. *Journal of Applied Microbiology* **107**: 452-464.

Wheeler D & Bhagwat M (2007) BLAST QuickStart: example-driven web-based BLAST tutorial. *Methods in molecular biology (*Clifton NJ, ed.*)* **395**: 149-176.

White D, Sharp RJ & Priest FG (1994) A polyphasic taxonomic study of thermophilic bacilli from a wide geographical area. *Antonie van Leeuwenhoek* **64**: 357-386.

Whitman WB (2015) Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Systematic and Applied Microbiology* **38**: 217-222.

Wiedenbeck J & Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews* **35**: 957-976.

Wiegand S, Rabausch U, Chow J, Daniel R, Streit WR & Liesegang H (2013) Complete genome sequence of *Geobacillus* sp. strain GHH01, a thermophilic lipase-secreting bacterium. *Genome announcements* **1**: doi:10.1128/genomeA.00092-00013.

Wiencek KM, Klapes NA & Foegeding PM (1990) Hydrophobicity of *Bacillus* and *Clostridium* spores. *Applied and Environmental Microbiology* **56**: 2600-2605.

Wiencek KM, Klapes NA & Foegeding PM (1991) Adhesion of *Bacillus* spores to inanimate materials: effects of substratum and spore hydrophobicity. *Biofouling* **3**: 139-149.

Williams SG, Greenwood JA & Jones CW (1992) Molecular analysis of the *lac* operon encoding the binding-protein-dependent lactose transport system and beta-galactosidase in *Agrobacterium radiobacter*. *Molecular Microbiology* **6**: 1755-1768.

Wisotzkey JD, Jurtshuk P, Fox GE, Deinhard G & Poralla K (1992) Comparative sequence analyses on the 16S ribosomal-RNA (rDNA) of *Bacillus acidocaldarius*, *Bacillus acidoterrestris*, and *Bacillus cycloheptanicus* and proposal for creation of a new genus, *Alicyclobacillus* gen nov. *International Journal of Systematic Bacteriology* **42**: 263-269.

Wu LJ & Errington J (2000) Identification and characterization of a new prespore-specific regulatory gene, *rsfA*, of *Bacillus subtilis*. *Journal of Bacteriology* **182**: 418-424.

Yao N, Ren Y & Wang W (2013) Genome sequence of a thermophilic bacillus, *Geobacillus thermodenitrificans* DSM465. *Genome announcements* **1**: doi:10.1128/genomeA.01046-01013.

Yi X & Setlow P (2010) Studies of the commitment step in the germination of spores of *Bacillus* species. *Journal of Bacteriology* **192**: 3424-3433.

Yoon JH, Kang KH & Park YH (2002) *Lentibacillus salicampi* gen. nov., sp nov., a moderately halophilic bacterium isolated from a salt field in Korea. *International Journal of Systematic and Evolutionary Microbiology* **52**: 2043-2048.

Yoon JH, Oh TK & Park YH (2004) Transfer of *Bacillus halodenitrificans* Denariaz et al. 1989 to the genus *Virgibacillus* as *Virgibacillus halodenitrificans* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **54**: 2163-2167.

Young JPW, Downer HL & Eardly BD (1991) Phylogeny of the phototrophic rhizobium strain BTAi1 by polymerase chain reaction-based sequencing of a 16S rRNA gene segment. *Journal of Bacteriology* **173**: 2271-2277.

Yuan D, Liu G, Ren DC, Zhang D, Zhao L, Kan C, Yang Y, Ma W, Li Y & Zhang L (2012) A survey on occurrence of thermophilic bacilli in commercial milk powders in China. *Food Control* **25**: 752-757.

Zarilla KA & Perry JJ (1987) *Bacillus thermoleovorans*, sp. nov. a species of obligately thermophilic hydrocarbon utilizing endospore-forming bacteria. *Systematic and Applied Microbiology* **9**: 258-264.

Zeigler DR (2005) Application of a *recN* sequence similarity analysis to the identification of species within the bacterial genus *Geobacillus*. *International Journal of Systematic and Evolutionary Microbiology* **55**: 1171-1179.

Zeigler DR (2011) The genome sequence of *Bacillus subtilis* subsp. *spizizenii* W23: insights into speciation within the *B. subtilis* complex and into the history of *B. subtilis* genetics. *Microbiology* **157**: 2033-2041.

Zeigler DR (2014) The *Geobacillus* paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? *Microbiology* **160**: 1-11.

Zerbino DR & Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**: 821-829.

Zhang J-q, Griffiths KK, Cowan A, Setlow P & Yu J (2013) Expression level of *Bacillus subtilis* germinant receptors determines the average rate but not the heterogeneity of spore germination. *Journal of Bacteriology* **195**: 1735-1740.

Zhao Y, Caspers MP, Abee T, Siezen RJ & Kort R (2012) Complete genome sequence of *Geobacillus thermoglucosidans* TNO-09.020, a thermophilic sporeformer associated with a dairy-processing environment. *Journal of Bacteriology* **194**: 4118-4118.

Zhao Y, Caspers MPM, Metselaar KI, de Boer P, Roeselers G, Moezelaar R, Groot MN, Montijn RC, Abee T & Kort R (2013) Abiotic and microbiotic factors controlling biofilm formation by thermophilic sporeformers. *Applied and Environmental Microbiology* **79**: 5652-5660.

Zheng B, Zhang F, Chai L*, et al.* (2014) Permanent draft genome sequence of *Geobacillus thermocatenulatus* strain GS-1. *Marine Genomics* **18**: 129-131.

Zhou TT, Dong ZT, Setlow P & Li YQ (2013) Kinetics of germination of individual spores of *Geobacillus stearothermophilus* as measured by raman spectroscopy and differential interference contrast microscopy. *PLoS One* **8**: doi:10.1371/journal.pone.0074987.

Zilla ML, Lunderberg JM, Schneewind O & Missiakas D (2015) Bacillus a*nthracis lcp* genes support vegetative growth, envelope assembly, and spore formation. *Journal of Bacteriology* **197**: 3731-3741.

# Appendix 1 Preparing a BLAST database

The commands used to prepare the BLAST database are described below. The options for preparing a BLAST database are described in Table A1.1. The reference sequences used are described in Section 4.2.6. A description for each step is included where the sentence starts with a ##.

**Table A1.1. Options for preparing a BLAST database.**

| Option name | Parameter |
|---|---|
| - in | Name of input file |
| **-** out | Name of output file |
| **-** dbtype | Type of database e.g. nucleotide (nucl) or protein (prot) |
| **-** db | Name of database |
| **-query** | Name of query file |
| **-outfmt** | Format of the output file. In this case “6” was selected, which is a tabular output format that can then be opened as a spread-sheet. |

### Example A1.1 BLASTN Database

```
## A database was prepared for each reference sequence.

makeblastdb -in Gst_23S.fa -dbtype nucl -out Gst_23S

makeblastdb -in C56T3_16S.fa -dbtype nucl -out C56T3_23S


## The CISA assemblies for each strain were concatenated into one
## file.

cat ./Geobacillus_[A1|P3|D1]_trimmed_*/contigs.fa >> GeoP3_trimmed.fa


## A BLASTN search was carried out using the reference sequence

## databases as the subject and the concatanated assemblies

## as the query.

blastn –db Gst_23S –query Geo[A1|P3|D1]__trimmed.fa -out\
GeoNameOfStrain_trimmed.txt –outfmt 6
```

| Query ID | Subject ID | %ID | Alignment length | Mismatch | Gaps | Subject Start | Subject End | Query Start | Query End | E value | Bit value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NODE_8_length_2066_ | gi\|40110\|emb\|X01387.1\| | 99.78 | 2236 | 5 | 0 | 1 | 2236 | 693 | 2928 | 0 | 4102 |
| NODE_503_length_110 | gi\|40110\|emb\|X01387.1\| | 95.63 | 1349 | 30 | 26 | 1 | 1339 | 1067 | 2396 | 0 | 2137 |
| NODE_467_length_109 | gi\|40110\|emb\|X01387.1\| | 95.6 | 1341 | 30 | 26 | 1 | 1331 | 1071 | 2392 | 0 | 2122 |
| NODE_435_length_109 | gi\|40110\|emb\|X01387.1\| | 95.57 | 1333 | 30 | 26 | 1 | 1323 | 1075 | 2388 | 0 | 2108 |
| NODE_435_length_109 | gi\|40110\|emb\|X01387.1\| | 95.57 | 1333 | 30 | 26 | 1 | 1323 | 1075 | 2388 | 0 | 2108 |
| NODE_432_length_109 | gi\|40110\|emb\|X01387.1\| | 95.55 | 1325 | 30 | 26 | 1 | 1315 | 1079 | 2384 | 0 | 2093 |
| NODE_432_length_109 | gi\|40110\|emb\|X01387.1\| | 95.55 | 1325 | 30 | 26 | 1 | 1315 | 1079 | 2384 | 0 | 2093 |
| NODE_436_length_108 | gi\|40110\|emb\|X01387.1\| | 95.52 | 1317 | 30 | 26 | 1 | 1307 | 1083 | 2380 | 0 | 2078 |
| NODE_436_length_108 | gi\|40110\|emb\|X01387.1\| | 95.52 | 1317 | 30 | 26 | 1 | 1307 | 1083 | 2380 | 0 | 2078 |
| NODE_436_length_108 | gi\|40110\|emb\|X01387.1\| | 95.49 | 1309 | 30 | 26 | 1 | 1299 | 1087 | 2376 | 0 | 2063 |
| NODE_436_length_108 | gi\|40110\|emb\|X01387.1\| | 95.49 | 1309 | 30 | 26 | 1 | 1299 | 1087 | 2376 | 0 | 2063 |
| NODE_214_length_121 | gi\|40110\|emb\|X01387.1\| | 99.77 | 1304 | 1 | 2 | 3 | 1306 | 1 | 1302 | 0 | 2390 |

**Figure A1.1. Example of the the top 12 hits from a BLASTN output, sorted by bit score.**

### Example A1.2 TBLASTX database

In this example a TBLASTX was carried out where the query nucleotide sequence was translated and compared with the database translated nucleotide sequence. This was because in this example, gene sequences (*spoIVB* and *spoIIIC*) were used from *B. subtilis*, which does not show enough similarity at the nucleotide level with *Geobacillus* strains to use a BLASTN.

```
makeblastdb -in spoIVB_spoIIIC.fasta -dbtype nucl -out spoIVB_spoIIIC

tblastx –db spoIVB_spoIIIC –query\
~/prokka41_Sara/Geobacillus/basicAnnotation/PROKKAon_A1/A1.fsa -out\
A1_sigK.txt –outfmt 6
```

# Appendix 2 Mapping using Bowtie 2

The commands used to map the next generation sequence reads to a consensus sequence is described below. A description for each step is included where the sentence starts with a ##. The options for mapping the reads to the index using Bowtie 2 are described in Table A2.1

**Table A2.1. Options for using Bowtie2.**

| Option name | Parameter |
|---|---|
| -build | Name of input file to be used for generating an index |
| - x | Name of index |
| - 1 | Paired-end read file 1 |
| - 2 | Paired-end read file 2 |
| - S | SAM mode with name of output file |

```
## An index was generated using the reference sequence

~/software/bowtie2-2.1.0/bowtie2-build [A1|P3|D1]__consensus16S.fasta\
P3_consensus


## The sequence reads were mapped to the index

~/software/bowtie2-2.1.0/bowtie2 –x [A1|P3|D1]__consensus16S -1\
processed_[A1|P3|D1]__S3_L001_R1_001.fastq -2\
processed_[A1|P3|D1]__S3_L001_R2_001.fastq -S [A1|P3|D1]__16S.sam


## A new SAM file was generated containing only those reads that ##
mapped to the index

grep "rRNA_[A1|P3|D1]_" [A1|P3|D1]__16S.sam >\
[A1|P3|D1]__16S_Specific.sam
```

# Appendix 3 Identifying orthologous clusters

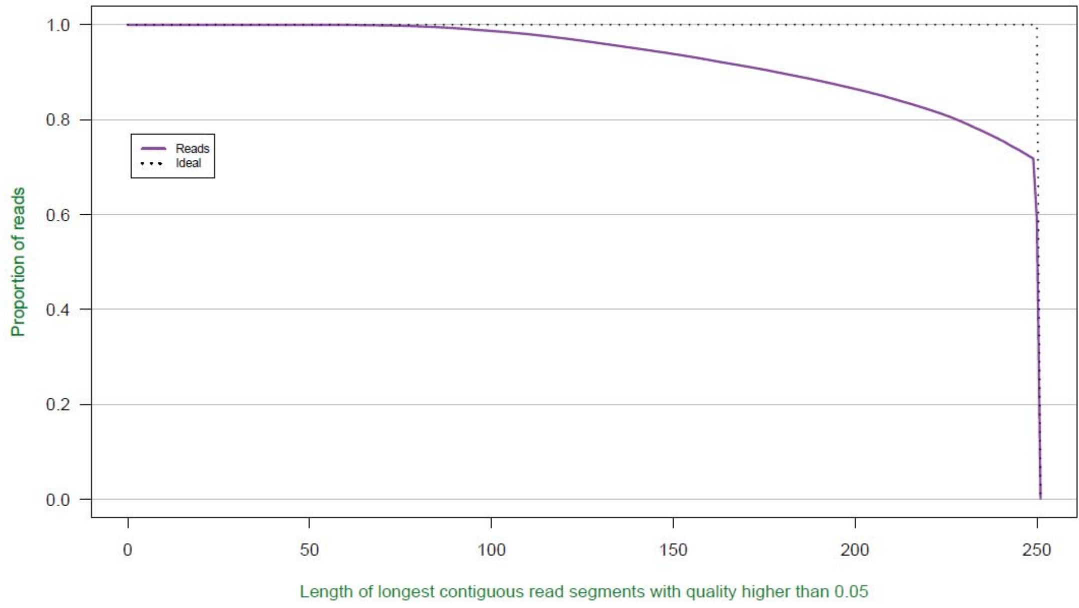
Two MySQL tables were created containing information on the orthologous clusters identified by OrthoMCL. The first table called `prokka_Geobacillus_Perfect` only contained clusters with either no genes or one gene from each genome. The second table, called `prokka_Geobacillus_nonOrtho`, contained clusters with paralogues. These two tables were used to identify gene clusters that contained only the three dairy strains of *G. stearothermophilus*, but not in the other *Geobacillus* strains and *vice versa*. Gene clusters that only contained the newly sequenced *G. stearothermophilus* strains were identified by searching for sets that were not empty (`!=''`, where `!` means "not" and `''` means "empty") for the three *G. stearothermophilus* strains and empty sets (`''`) for the remaining *Geobacillus* strains. An example of the MySQL code is given below. The opposite was carried out to identify gene clusters that did not contain the three *G. stearothermophilus* strains.

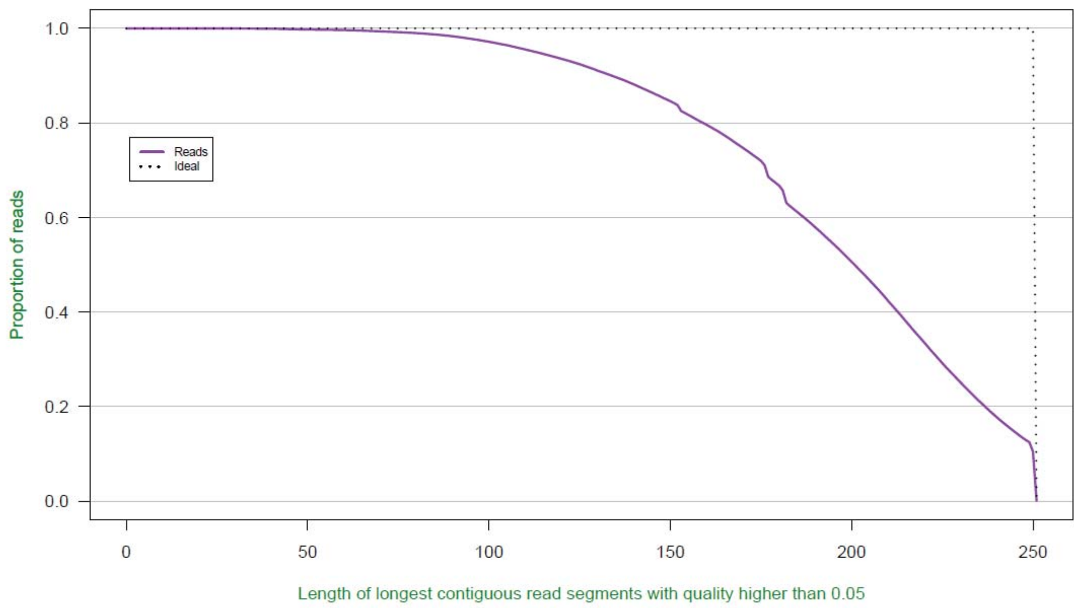
```
## Identification of clusters from the prokka_Geobacillus_Perfect
table that contained sets for A1, P3 and D1, but not the other
Geobacillus spp. genomes.

select * from prokka_Geobacillus_Perfect where A1 !='' and P3 !='' and
D1 !='' and ABVH01 ='' and AJJN01 ='' and AMRO01 ='' and AOT201 =''
and ATCO01 ='' and AUXP01 ='' and AYKT01 ='' and AYSF01 ='' and BASG01
='' and BATY01 ='' and BAWO01 ='' and BAWP01 ='' and BBJV01 ='' and
Bst10 ='' and JALS01 ='' and JFHZ01 ='' and JGCJ01 ='' and JHUR01 =''
and JHUS01 ='' and JP0101 ='' and JQMN01 ='' and NC_006510 ='' and
NC_009328 ='' and NC_012793 ='' and NC_013411 ='' and NC_014206 =''
and NC_014650 ='' and NC_014915 ='' and NC_015660 ='' and NC_016593
='' and NC_020210 ='' and NC_022080 ='' and JPYA01 ='';
```

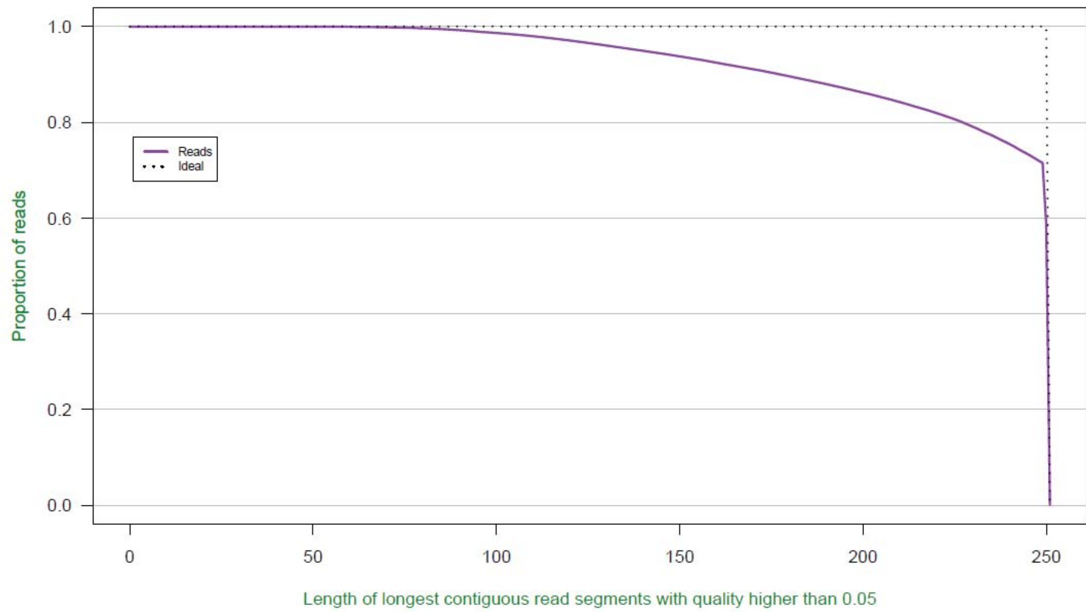
# Appendix 4 Cumulative frequency of trimmed read lengths

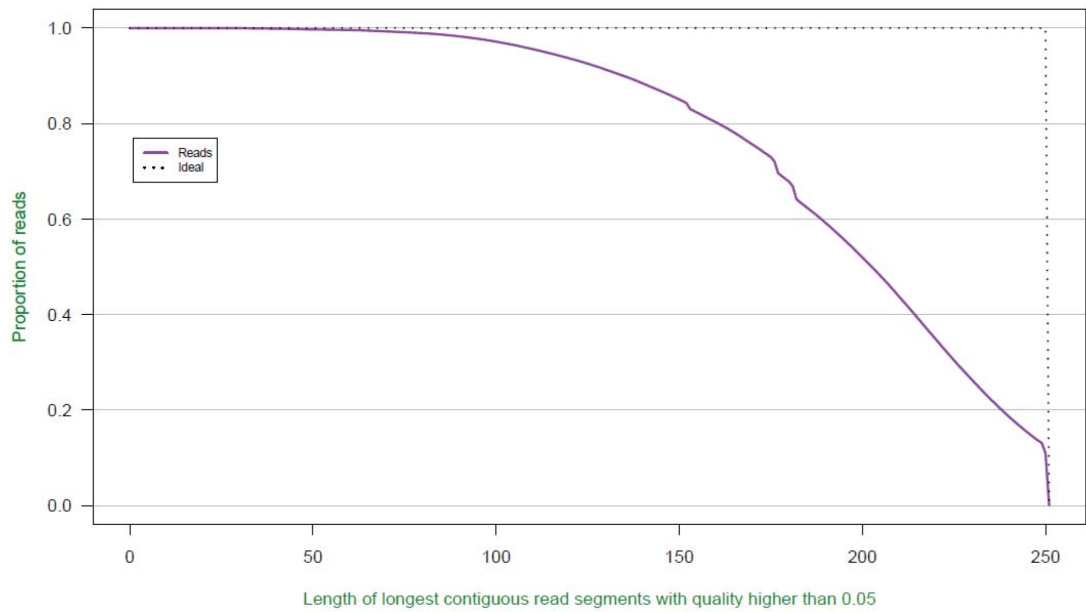
NZGL carried out a number of processes to check the quality of the sequence reads generated. As part of one of these processes the cumulative frequency of trimmed read lengths was plotted as a line graph compared with the perfect results plotted as a dotted line. Each graph below shows what proportion of the sequence reads are a certain length. The graphs were named by NZGL as follows:

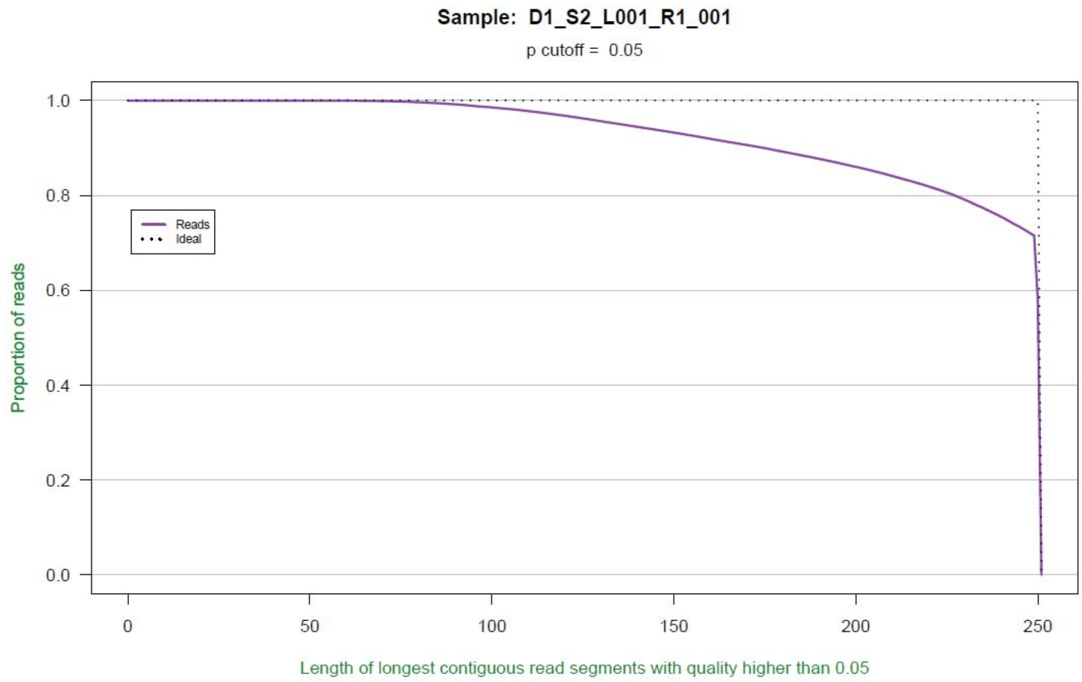
(sample name)_(sample number)_L001_(read number)_001

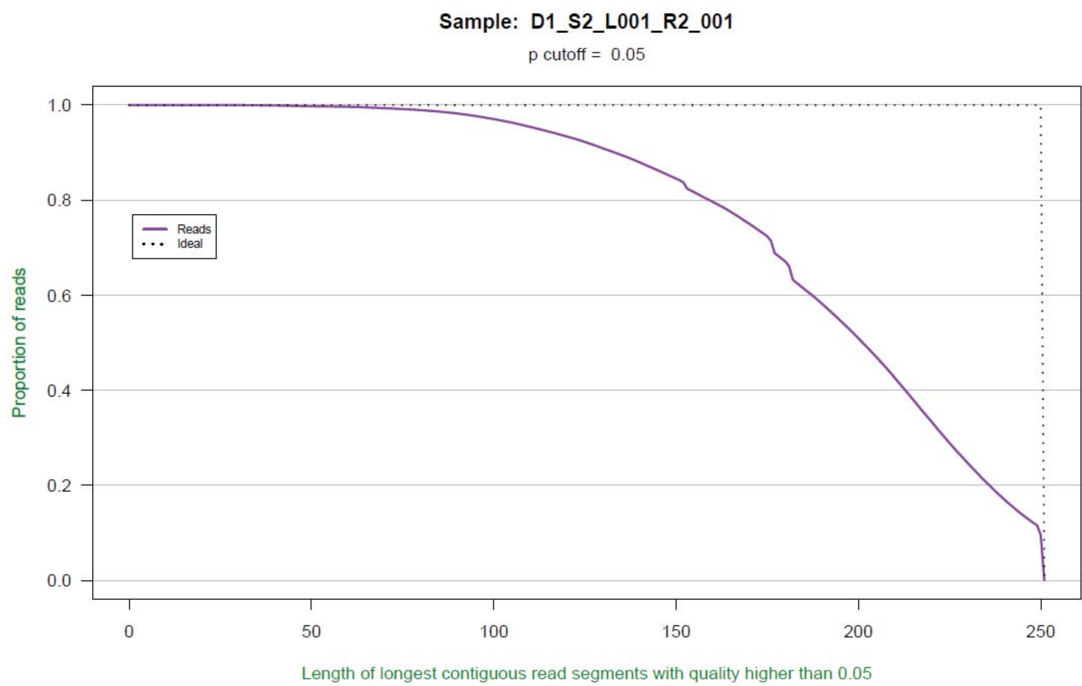

Sample: A1_S1_L001_R1_001
p cutoff = 0.05
1.0
0.8
0.6
0.4
0.2
0.0
Proportion of reads
Reads
Ideal
0
50
100
150
200
250
Length of longest contiguous read segments with quality higher than 0.05


Sample: A1_S1_L001_R2_001
p cutoff = 0.05
1.0
0.8
0.6
0.4
0.2
0.0
Proportion of reads
Reads
Ideal
0
50
100
150
200
250
Length of longest contiguous read segments with quality higher than 0.05


Sample: P3_S3_L001_R1_001
p cutoff = 0.05
1.0
0.8
0.6
0.4
0.2
0.0
Reads
Ideal
Proportion of reads
0
50
100
150
200
250
Length of longest contiguous read segments with quality higher than 0.05


Sample: P3_S3_L001_R2_001
p cutoff = 0.05
1.0
0.8
0.6
0.4
0.2
0.0
Reads
Ideal
Proportion of reads
0
50
100
150
200
250
Length of longest contiguous read segments with quality higher than 0.05


Sample: D1_S2_L001_R1_001
p cutoff = 0.05
Reads
Ideal
Proportion of reads
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Length of longest contiguous read segments with quality higher than 0.05


Sample: D1_S2_L001_R2_001
p cutoff = 0.05
Reads
Ideal
Proportion of reads
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
Length of longest contiguous read segments with quality higher than 0.05

# Appendix 5 General genomic features of published *Geobacillus* strains

| Abbreviated name | Locus ID | Genome size (bp) | Number of contigs | GC content | Repeat regions | Number of CDS | Number of tRNA |
|---|---|---|---|---|---|---|---|
| GcCIC | AMRO01 | 3826547 | 82 | 44.17 | 0 | 3772 | 73 |
| Gc10776 | BAWO01 | 3798752 | 130 | 43.92 | 0 | 3793 | 82 |
| GkHTA | NC_006510 | 3544776 | 1 | 52.09 | 3 | 3498 | 88 |
| GkGbl | BASG01 | 3541481 | 216 | 52.05 | 6 | 3546 | 88 |
| Gk102445 | BBJV01 | 3446385 | 118 | 51.99 | 2 | 3471 | 83 |
| Gs10 | Bst10 | 3594224 | 120 | 52.56 | 4 | 3539 | 87 |
| Gs7953 | JALS01 | 2787229 | 129 | 52.39 | 1 | 2940 | 85 |
| Gs3621 | AOT201 | 3621385 | 10 | 44.38 | 1 | 3656 | 76 |
| GthdNG8 | NC_009328 | 3550319 | 1 | 49.01 | 2 | 3493 | 88 |
| Gthd465 | AYKT01 | 3400546 | 76 | 49.05 | 1 | 3352 | 63 |
| GthgC56 | NC_015660 | 3893306 | 1 | 43.95 | 5 | 3872 | 90 |
| GthgTNO | AJJN01 | 3737238 | 31 | 43.86 | 6 | 3703 | 88 |
| Gthg10776 | BAWP01 | 3871162 | 68 | 43.69 | 7 | 3811 | 81 |
| GthlCCB | NC_016593 | 3596620 | 1 | 52.28 | 2 | 3537 | 89 |
| GthlB23 | BATY01 | 3353053 | 209 | 52.29 | 6 | 3338 | 84 |
| GsuPSS | JQMN01 | 3746741 | 2 | 51.58 | 9 | 3601 | 90 |
| GthcGS1 | JFHZ01 | 3519600 | 155 | 52.11 | 9 | 3551 | 67 |
| GvPSS | JP0101 | 3389115 | 1 | 52.4 | 3 | 3367 | 86 |
| GeoWCH | NC_012793 | 3464618 | 1 | 42.84 | 5 | 3402 | 92 |
| GeoMC61 | NC_013411 | 3622844 | 1 | 52.42 | 6 | 3556 | 87 |
| GeoMC52 | NC_014915 | 3628883 | 1 | 52.43 | 6 | 3557 | 87 |
| GeoC56 | NC_014206 | 3650813 | 1 | 52.49 | 3 | 3593 | 88 |
| GeoMC1 | NC_014650 | 3840330 | 1 | 44.02 | 5 | 3768 | 90 |
| GeoGHH | NC_020210 | 3583134 | 1 | 52.28 | 5 | 3493 | 88 |
| GeoJF8 | NC_022080 | 3446282 | 1 | 52.87 | 2 | 3404 | 89 |
| GeoG11 | ABVH01 | 3545187 | 31 | 48.8 | 2 | 3601 | 72 |
| GeoFW2 | JGCJ01 | 3486830 | 245 | 52.24 | 5 | 3496 | 91 |
| GeoCAMR1 | JHUR01 | 3414745 | 74 | 52.21 | 2 | 5147 | 48 |
| GeoCAMR5 | JHUS01 | 3499823 | 96 | 51.89 | 4 | 3460 | 89 |
| GicG1w | JPYA01 | 3457810 | 207 | 52.03 | 7 | 3405 | 84 |
| GeoWSU | ATCO01 | 3402383 | 346 | 52.21 | 5 | 3691 | 75 |
| GeoMAS | AYSF01 | 3497407 | 121 | 52.2 | 3 | 3592 | 95 |
| GeoA8 | AUXP01 | 3348314 | 173 | 52.41 | 5 | 3376 | 87 |

# Appendix 6 Core genome sequence comparisons

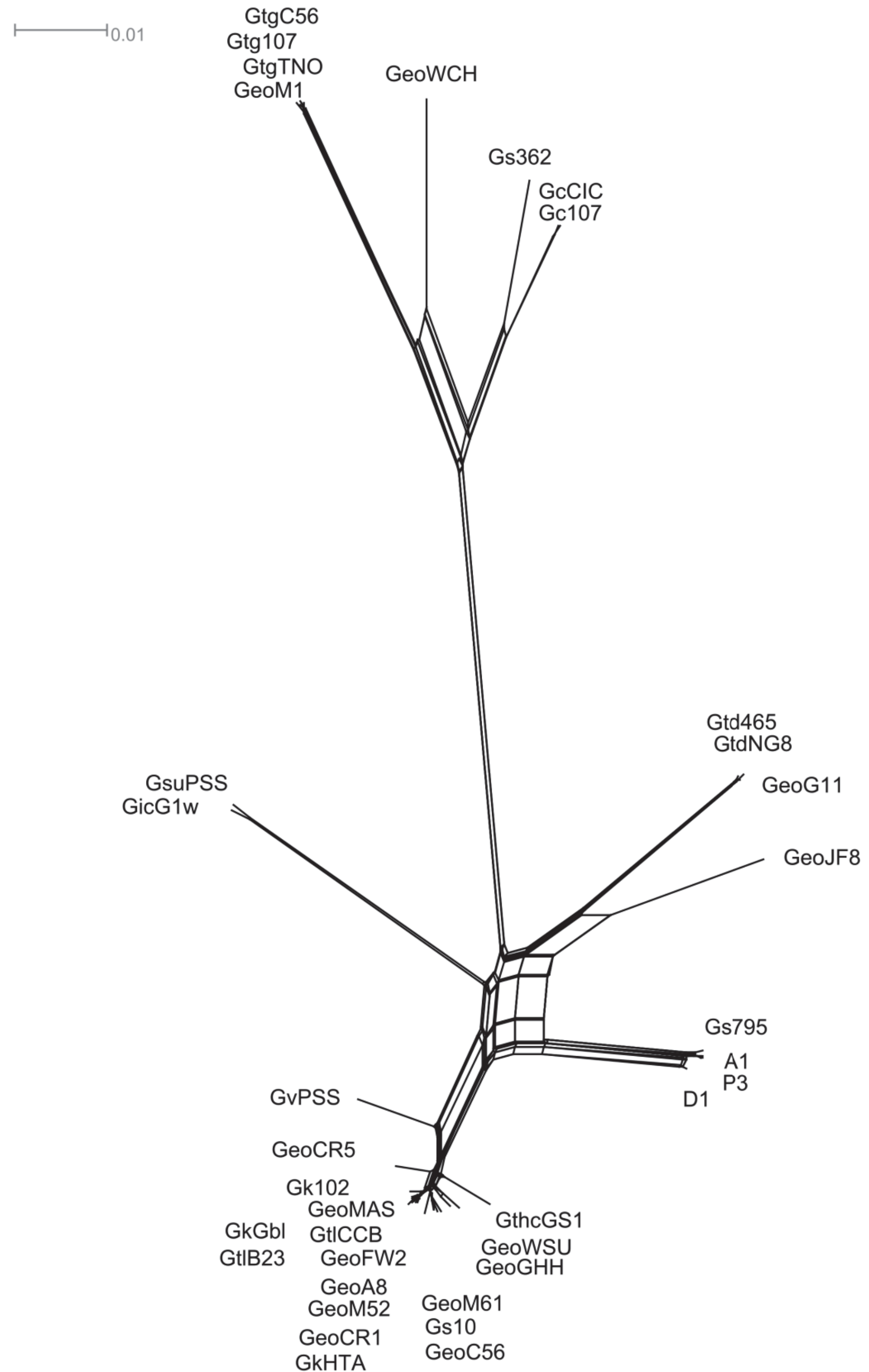



**Figure A6.1. Core genome sequence comparisons.** The phylogenetic networks were generated using the NeighbourNet algorithm in Splits Tree (v. 4.13.1). The orthologous groups were defined using the program OrthoMCL (v. 2.0.9) and the analyses were based on those genes that have orthologous gene members with the same length.

# Appendix 7 Heat-maps to visualize the presence/absence of homologues in *Geobacillus* spp.

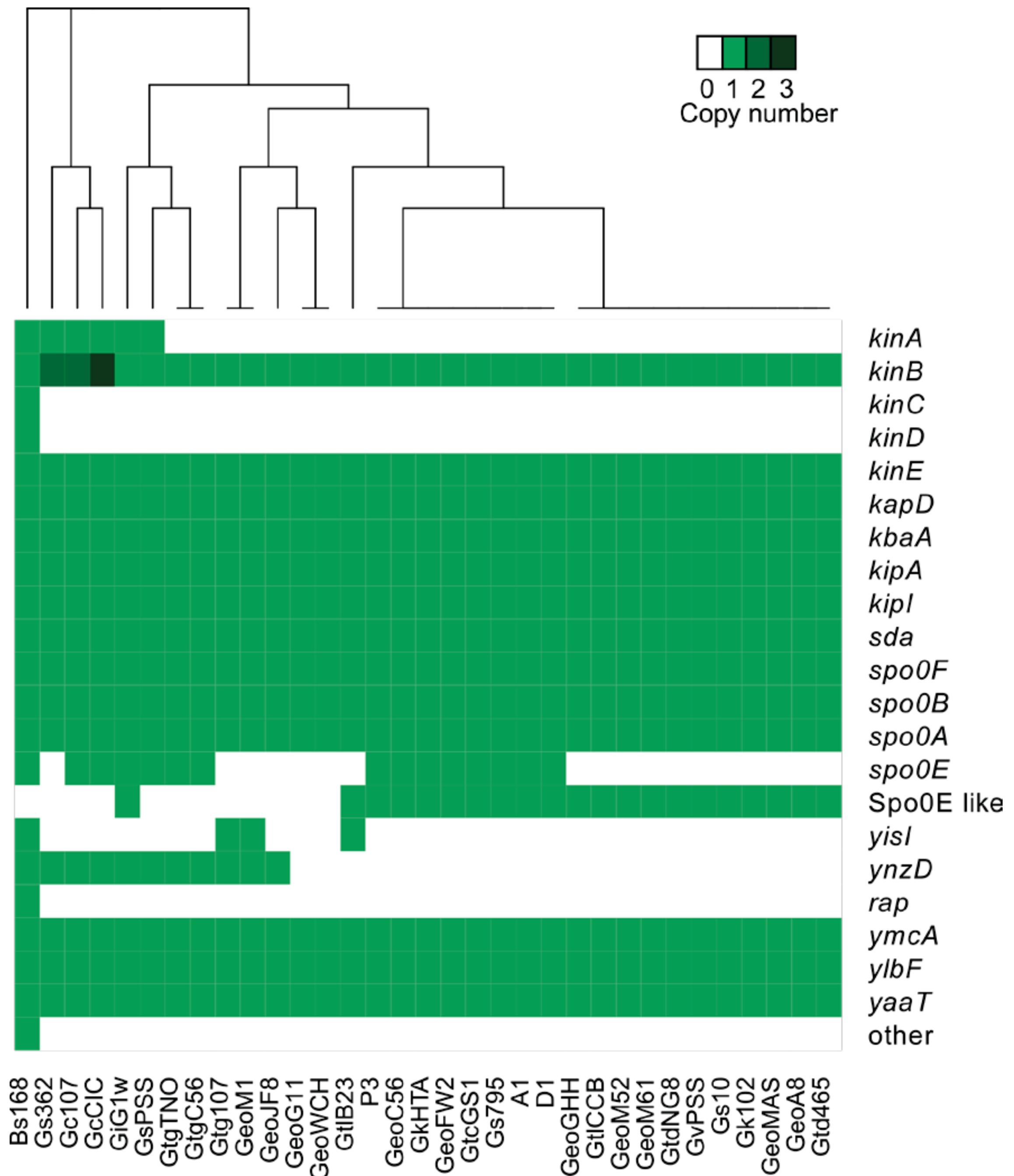



**Figure A7.1. Conservation of selected *Bacillus subtilis* phosphorelay genes in the *Geobacillus* genus.** The presence of a homologue was determined by using the BDBH algorithm as described in Section 6.2.3, with the exception of the row labelled Spo0E like which was based on additional genes annotated as *spo0E* by Prokka. "*rap*" includes the genes *rapA*, *rapB*, *rapE* and *rapH*. "other" refers to other genes that were absent across the *Geobacillus* genus, which included: *phrA, phrC, phrE, med* and *kapB*.

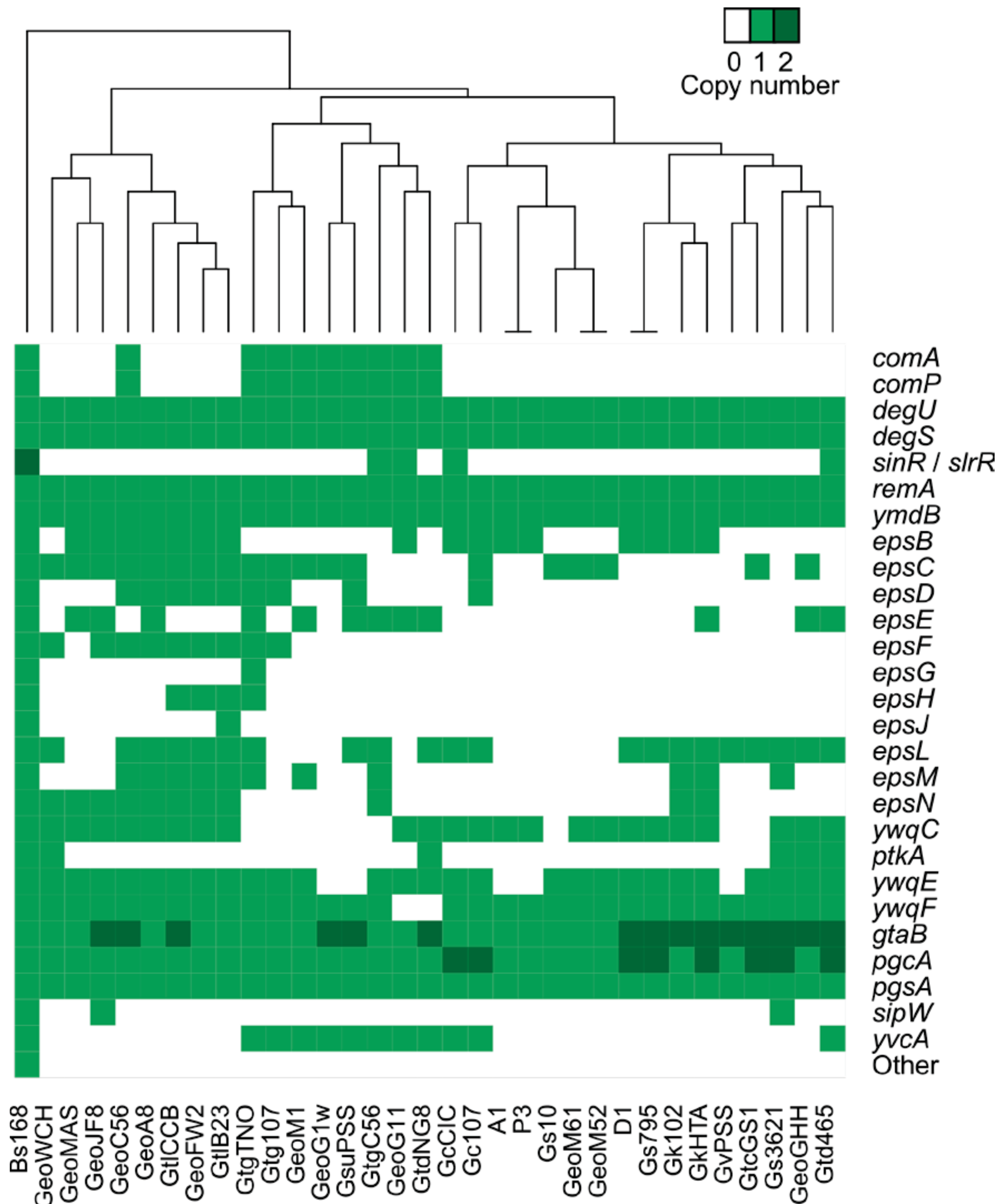



**Figure A7.2. Conservation of selected *Bacillus subtilis* biofilm genes in the *Geobacillus* genus.** The presence of a homologue was determined by using the BDBH algorithm as described in Section 6.2.3. “Other” refers to those genes that were absent across the *Geobacillus* genus, which included: *epsA, epsI, epsK, epsO, pgsB, pgsC, tapA, tasA, bslA, sinI, degQ, remB,* and *slrA.*

# Appendix 8 Gene organisation of the *skin* elements

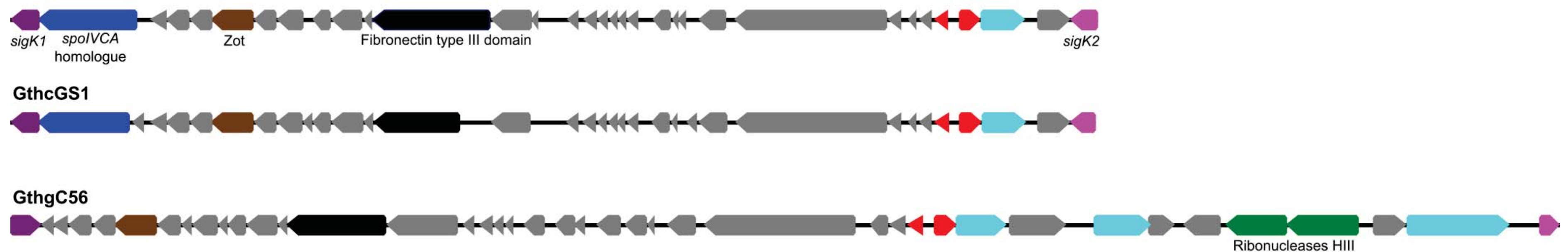



**Figure A8.1. Gene organization of the *skin* element in three *Geobacillus* strains, *Geobacillus* MAS1, *G. thermocatenulatus* GS-1 and *G. thermoglucosidasius* C56T3.** The *skin* element sits between the N-terminal (*sigK1*, coloured in purple) and C-terminal (*sigK2*) regions of the *sigK* gene. Homologues of *B. subtilis spoIVCB* were found in the *skin* elements of two of the strains (coloured in dark blue), other genes encoding for recombinases/resolvases are coloured in light blue. Most of the genes found within the *skin* elements were annotated as hypothetical proteins by Prokka (coloured in grey). In addition, there were genes, annotated by Prokka, encoding a Zonular occludens toxin (Zot, coloured in brown), Fibronectin type III domain protein (coloured in black), transcriptional regulators (coloured in red) and Ribonucleases HIII (coloured in green).

# Appendix 9 BLASTX analysis of the *rsfA* and *gerR* genes

**Table A9.1. The top three BLASTX hits using *B. subtilis* 168 *rsfA* as the query and *Geobacillus* sp. C56-T3 as the subject**

| Coverage | E value | Identity similarity | Protein accession | Prokka annotation |
|---|---|---|---|---|
| **98 %** | 4e-73 | 55 % | WP_013146646 | rsfA 1 |
| **98 %** | 2e-44 | 39 % | WP_013144650 | rsfA 2 |
| **89 %** | 4e-32 | 36 % | WP_ 013145858 | rsfA 3 |

**Table A9.2. The top three BLASTX hits using *B. subtilis* 168 *gerR* as the query and *Geobacillus* sp. C56-T3 as the subject**

| Coverage | E value | Identity similarity | Protein accession | Prokka annotation |
|---|---|---|---|---|
| **93 %** | 2e-46 | 43 % | WP_013146646 | rsfA 1 |
| **92 %** | 3e-44 | 47 % | WP_013145858 | rsfA 3 |
| **92 %** | 6e-38 | 41 % | WP_013144650 | rsfA 2 |