

Synthetic biology approaches to the metabolic engineering of *Geobacillus thermoglucosidans* for isobutanol production

A thesis submitted for the degree of Doctor of Philosophy

Elena Martinez-Klimova

December 4th, 2014

Imperial College London
Department of Bioengineering
Centre for Synthetic Biology and Innovation

Supervisors: Dr. Tom Ellis and Prof. David Leak

DECLARATION

I hereby certify that everything included in this thesis is my own work, and in the instances where work has been used from other sources, it is acknowledged accordingly.

Elena Martinez-Klimova

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

ABSTRACT

Renewable green alternatives to fossil fuels need to be sought in order to address the challenges of environmental and energy crises. Up until now, ethanol has been the major biofuel. *Geobacillus thermoglucosidans* is a thermophilic bacterium that is capable of producing bioethanol in an industrial setting at high temperatures and is capable of metabolizing pentoses and hexoses commonly found in lignocellulosic biomass. Due to these attractive properties, the aim of this work has been to construct a toolbox of genetic components to develop *G. thermoglucosidans* as the leading thermophile chassis for synthetic biology and metabolic engineering. The toolbox is composed of shuttle vectors that have higher transformation efficiencies than previous existing vectors and are modular, where the presence of restriction sites separating each of the components allows users to exchange parts easily and efficiently. Also included in the toolbox are the fluorescent reporters sfGFP, mCherry and BsFbFP that will permit the characterization of promoters.

As a proof-of-principle application to demonstrate the effectivity of the toolbox for the production of valuable compounds, this work explores the production of isobutanol by the thermophile bacteria *Geobacillus thermoglucosidans*. Isobutanol is a higher chain alcohol that is a significantly better fuel molecule than ethanol, both for energy content and infrastructure compatibility. The *Geobacillus* host was able to produce isobutanol in amounts of around 50 mg/L via the conversion of isobutyryl-CoA to isobutyraldehyde by an (ALDH) and from isobutyraldehyde to isobutanol by an alcohol dehydrogenase (ADH). It was observed that supplementing the growth medium with an intermediate of the valine biosynthesis pathway, 2-ketoisovalerate, resulted in the production of isobutanol and overexpressing ALDH increased the isobutanol titres.

ACKNOWLEDGMENTS

First and foremost, my gratitude goes to my supervisors, Prof. David Leak and Dr. Tom Ellis. A heartfelt thank you for all the time and guidance, and for all the effort invested in this project. Thank you for keeping track of the progress so closely and for all those e-mails at ungodly hours.

I am also gratefully indebted to all the incredible scientists I've had the privilege of working with these past four years. Rochelle Aw, thanks for being my friend and postdoc. Ben Reeve, thanks for the thermophile chats and for giving me a thumbs-up when I start cursing *Geobacillus*; take care of the Hot Box. Thank you Charlotte Ward for all the help you've sent my way since 2009. Likewise, thanks to Jeremy Bartosiak-Jentys and Nicola Crowhurst for always being available to offer a hand. Additional thanks go to Kate Royle, Beata Lisowska, Ali Hussein and Agripina Banda for making Lab. 326 such an enjoyable place to work in. Thank you Kealan Exley for teaching me Western Blotting and Protein Purification. Thanks to George Pothoulakis and Felix Jonas for helping me with Microscopy, FACS and FlowJo. Not least important are Francesca Ceroni, Arturo Casini, Ali Awan, Dejana Jovicevic, Ben Blount, Ollie Wright, Tim Weenink and Chris Hirst: thanks for all the various ways in which all of you have helped me these past two years and for making CSynBI a place I have nothing but good memories from. To everyone who tried to convince me I could get it done: thank you.

Special thanks to James Mansfield for teaching me the ropes of HPLC analysis. Thanks to Shyam K. Masakapalli and Matthew Styles from the University of Bath for helping me with last-minute HPLC and GC-MS analyses.

The financial sponsorship by CONACyT Mexico is gratefully acknowledged.

Y finalmente, gracias a mi familia por siempre hacerlo todo posible.

TABLE OF ABBREVIATIONS

| | |
|------------------|---|
| ABE | Acetone Butanol Ethanol |
| ADH | Alcohol dehydrogenase |
| AdhE | alcohol dehydrogenase from <i>E. coli</i> |
| adh2 | gene encoding an alcohol dehydrogenase in <i>S. cerevisiae</i> |
| Ald | gene encoding an acylating aldehyde dehydrogenase ALDH |
| ALDH | acylating aldehyde dehydrogenase |
| alsS | gene encoding an acetohydroxy acid synthase (AHAS I) from <i>B. subtilis</i> |
| ampR | gene that encodes for a β -lactamase, which confers resistance to ampicillin |
| BsFbFP | gene encoding a <i>Bacillus subtilis</i> Flavin based Fluorescent Protein (BsFbFP) |
| catE | gene encoding for a chloramphenicol acetyltransferase |
| CBP | Consolidated BioProcessing |
| CDS | Coding Sequence |
| ColE1 | also named <i>pUC18_ori</i> . Origin of replication for <i>E. coli</i> |
| FbFP | Flavin-based Fluorescent Proteins |
| FMN | Flavin MonoNucleotide |
| fnr | global transcription factor for anaerobic growth involved in fumarate and nitrate reduction |
| frdAB | gene encoding a fumarate reductase |
| GC-MS | Gas Chromatography-Mass Spectrometry |
| HPLC | High-Performance Liquid Chromatography |
| ilvA | gene encoding a threonine deaminase |
| ilvC | gene encoding an acetohydroxy acid isomeroreductase |
| ilvD | gene encoding a Dihydroxyacid dehydratase |
| ilvE | gene encoding a transaminase B |
| ilvH | gene encoding an acetohydroxy acid synthase (AHAS I) |
| IPTG | Isopropyl β -D-1-thiogalactopyranoside |
| kanR | gene encoding for a kanamycin nucleotidyltransferase |
| Kdc | 2-Keto acid decarboxylase |
| kivd | <i>L. lactis</i> gene encoding a 2-keto acid decarboxylase (Kdc) |
| lacZ | gene encoding for β -galactosidase |
| ldh | gene encoding a lactate dehydrogenase (Ldh) |
| ldhA | gene encoding a D-lactate dehydrogenase |
| leuA | gene encoding a 2-isopropylmalate synthase |
| Lrp | Leucine responsive protein |
| LOV | Light Oxygen Voltage |
| MCS | Multiple Cloning Site |
| panB | 3-Methyl-2-oxobutanoate hydroxymethyl transferase |
| pfl | gene encoding a pyruvate formate lyase (PFL) |
| pflA | gene encoding a pyruvate formate lyase (PFL) |
| pheB | gene encoding a 2,3-catechol dioxygenase |
| PpFbFP | gene encoding for <i>Pseudomonas putida</i> Flavin based Fluorescent Protein |
| pta | gene encoding a phosphate acetyltransferase (PTA) |
| pUC18_ori | also named ColE1. Origin of replication for <i>E. coli</i> |
| pUP | uracil phosphoribosyltransferase gene |
| pyrE | gene encoding an orotate phosphoribosyltransferase |
| RBS | Ribosome Binding Site |
| RplS | gene encoding for the 50S ribosomal subunit protein L19 |
| repB | Origin of replication derived from plasmid pUB190 |
| SEVA | Standard European Vector Architecture |
| sfGFP | gene encoding superfolder GFP (sfGFP) |
| SSF | Simultaneous Saccharification and Fermentation |
| Q-PCR | Quantitative PCR |
| YtvA | non-fluorescent version of BsFbFP |
| yqhD | gene encoding for an alcohol dehydrogenase in <i>E. coli</i> |

TABLE OF CONTENTS

| | |
|--|----|
| Abstract | 3 |
| Table of abbreviations | 5 |
| Index of figures | 10 |
| Index of tables | 15 |
| 1. Chapter 1: motivation and introduction | 16 |
| 1.1. Motivation | 16 |
| 1.2. Research objectives | 18 |
| 1.3. Introduction | 18 |
| 1.3.1. Organic chemistry vs. synthetic biology | 19 |
| 1.3.2. Definition of synthetic biology | 21 |
| 1.3.3. Metabolic engineering | 21 |
| 1.3.4. How synthetic biology and metabolic engineering come together | 22 |
| 1.3.5. The biological parts that constitute devices | 24 |
| 1.3.6. Promoters | 25 |
| 1.3.7. Ribosome binding site (RBS) | 26 |
| 1.3.8. Transcriptional terminators | 26 |
| 1.3.9. Chassis | 28 |
| 1.3.10. Characteristics of a chassis for synthetic biology | 28 |
| 1.3.11. Selecting a host for biofuel production | 29 |
| 1.3.12. Heat tolerance in industry | 30 |
| 1.3.13. Valuable thermophiles | 31 |
| 1.3.14. Genus <i>Geobacillus</i> | 33 |
| 1.3.15. <i>Geobacillus thermoglucosidans</i> | 35 |
| 1.3.16. <i>Geobacillus thermoglucosidans</i> as a chassis | 36 |
| 1.3.17. A synthetic biology toolbox: characterization of bioparts | 38 |
| 1.3.18. Standardization and modularization of bioparts | 39 |
| 1.3.19. Reporter proteins | 40 |
| 1.3.20. Plasmids/shuttle vectors | 41 |
| 1.3.21. Guideline: shuttle vectors for <i>Clostridium</i> spp. | 42 |
| 1.4. Hypothesis | 42 |
| 1.5. Aims of the project | 43 |
| 2. Chapter 2: Materials and Methods | 44 |
| 2.1. Bacterial strains and media | 44 |
| 2.2. Bacterial growth conditions | 45 |
| 2.3. Frozen glycerol stocks | 46 |
| 2.4. Sterilization | 46 |
| 2.5. Antibiotic selection | 46 |
| 2.6. Plasmid purification | 46 |
| 2.7. Polymerase Chain Reaction (PCR) | 47 |
| 2.8. Gibson one-step isothermal DNA assembly | 48 |
| 2.9. Restriction digests and ligations | 48 |
| 2.10. Preparation of chemically competent <i>E. coli</i> | 48 |
| 2.11. Transformation of chemically competent <i>E. coli</i> | 49 |
| 2.12. Preparation of electrocompetent <i>G. thermoglucosidans</i> | 49 |
| 2.13. Transformation of <i>G. thermoglucosidans</i> | 49 |
| 2.14. Growing <i>G. thermoglucosidans</i> anaerobically and microaerobically | 50 |
| 2.15. Optical density analysis | 50 |
| 2.16. Lysing | 51 |
| 2.17. SDS-PAGE | 51 |
| 2.18. Western Blotting | 52 |
| 2.19. Total RNA extraction | 53 |
| 2.20. Reverse Transcription PCR (RT-PCR) | 53 |
| 2.21. High-Performance Liquid Chromatography analysis (HPLC) | 53 |

| | |
|--|-----|
| 2.22. Gas-chromatography and Mass Spectrometry (GC-MS) | 54 |
| 2.23. Microscopy | 54 |
| 2.24. Flow cytometry | 55 |
| 3. Chapter 3: Promoters and reporter proteins for <i>Geobacillus thermoglucosidans</i> | 56 |
| 3.1. Introduction | 56 |
| 3.1.1. Promoters for <i>Geobacillus</i> spp. | 57 |
| 3.1.2. Reporters | 58 |
| 3.1.2.1. The pheB assay for <i>Geobacillus</i> spp. | 58 |
| 3.1.2.2. Fluorescent proteins | 62 |
| 3.1.2.2.1. Green fluorescent protein (GFP) | 62 |
| 3.1.2.2.2. Red fluorescent protein (mCherry) | 65 |
| 3.1.2.2.3. Anaerobic fluorescent proteins: LOV-domain reporters | 65 |
| 3.1.2.2.3.1. BsFbFP and PpFbFP | 67 |
| 3.1.2.2.3.2. hotLOV | 70 |
| 3.1.2.2.3.3. hbLOV | 71 |
| 3.1.3. Salis RBS Calculator | 74 |
| 3.2. Results | 76 |
| 3.2.1. Superfolder Green Fluorescent Protein (sfGFP) | 76 |
| 3.2.2. The Ldh promoter in <i>E. coli</i> | 79 |
| 3.2.3. sfGFP in the absence of oxygen | 80 |
| 3.2.4. BsFbFP in <i>E. coli</i> and <i>G. thermoglucosidans</i> | 81 |
| 3.2.5. PpFbFP | 87 |
| 3.2.6. hotLOV | 90 |
| 3.2.7. hbLOV | 90 |
| 3.2.8. Flow cytometry | 91 |
| 3.3. Discussion | 97 |
| 3.3.1. Promoters for <i>G. thermoglucosidans</i> | 97 |
| 3.3.2. Reporters for <i>G. thermoglucosidans</i> | 98 |
| 3.3.2.1. Superfolder Green Fluorescent Protein (sfGFP) | 98 |
| 3.3.2.2. Red Fluorescent Protein (mCherry) | 99 |
| 3.3.2.3. Anaerobic Fluorescent Proteins: LOV-domain reporters | 99 |
| 3.4. Conclusions | 103 |
| 3.5. Future work | 103 |
| 4. Chapter 4: Shuttle vectors for <i>E. coli</i> / <i>G. thermoglucosidans</i> | 105 |
| 4.1. Introduction | 105 |
| 4.1.1. Existing shuttle vectors: pUCG18 and pUCG3.8 | 106 |
| 4.1.2. Are two antibiotic resistance genes really required? | 109 |
| 4.1.3. Chloramphenicol resistance in thermophiles | 110 |
| 4.1.4. pMK3 and pMK4 | 110 |
| 4.2. Results | 112 |
| 4.2.1. Analysis of the DNA sequence of existing plasmids | 112 |
| 4.2.2. The problem with <i>kanR</i> | 113 |
| 4.2.3. Design of the shuttle vectors | 116 |
| 4.2.3.1. pSEVA | 117 |
| 4.2.3.2. First modular plasmid design | 118 |
| 4.2.3.3. Assembly of the plasmids or DNA synthesis? | 120 |
| 4.2.4. Construction of the first modular plasmid | 121 |
| 4.2.5. Construction of a second modular plasmid | 126 |
| 4.2.6. Construction of p11AC1 | 129 |
| 4.2.7. Chloramphenicol resistance in <i>G. thermoglucosidans</i> | 130 |
| 4.2.8. Construction of p12AC1 | 131 |
| 4.2.9. Construction of p12AK1 | 132 |
| 4.2.10. Construction of p13AK1 | 132 |
| 4.2.11. Three-part minimal modular plasmids | 133 |
| 4.2.12. Characterization of parts: KanR | 133 |
| 4.2.13. Plasmid compatibility | 135 |

| | |
|---|-----|
| 4.2.14. Characterization of parts: <i>E. coli</i> origin of replication | 136 |
| 4.2.15. Characterization of parts: <i>Geobacillus</i> origins of replication | 137 |
| 4.3. Discussion | 139 |
| 4.3.1. Chloramphenicol resistance in <i>G. thermoglucosidans</i> | 139 |
| 4.3.2. Kanamycin resistance in <i>G. thermoglucosidans</i> | 141 |
| 4.3.3. Plasmid (in)compatibility | 142 |
| 4.3.4. Multiple Cloning Site (MCS) design | 143 |
| 4.4. Conclusion | 143 |
| 4.5. Future work | 144 |
| 5. Chapter 5: genomic knock-in/outs in <i>G. thermoglucosidans</i> | 146 |
| 5.1. Introduction | 146 |
| 5.1.1. The known knock-out procedure | 147 |
| 5.1.2. pyrE: counter-selection for second cross-overs | 150 |
| 5.2. Results | 152 |
| 5.2.1. Knock-in of sfGFP in the <i>G. thermoglucosidans</i> genome | 152 |
| 5.2.1.1. Construction of the knock-out/in plasmid | 155 |
| 5.2.2. pyrE knock-out in <i>G. thermoglucosidans</i> | 160 |
| 5.3. Discussion | 163 |
| 5.3.1. Why is it important to know the parts of the vector in use? | 163 |
| 5.3.2. Knock-out/in in <i>G. thermoglucosidans</i> | 164 |
| 5.3.3. pyrE knock-out in <i>G. thermoglucosidans</i> DL33 | 165 |
| 5.4. Conclusion | 165 |
| 5.5. Future work | 165 |
| 6. Chapter 6: Isobutanol production by <i>Geobacillus thermoglucosidans</i> | 166 |
| 6.1. Introduction | 166 |
| 6.1.1. Butanols as fuels | 167 |
| 6.1.2. Toxicity of butanols | 168 |
| 6.1.3. Natural producers of butanols | 170 |
| 6.1.4. Heterologous producers of butanols | 170 |
| 6.1.5. Sources of 2-keto acid decarboxylases (Kdc) | 171 |
| 6.1.6. Isobutanol production in <i>B. subtilis</i> | 172 |
| 6.1.7. Isobutanol production in <i>C. glutamicum</i> | 172 |
| 6.1.8. Biosynthesis of 2-ketoisovalerate in <i>E. coli</i> and <i>C. glutamicum</i> | 174 |
| 6.1.9. Keto acid decarboxylases (Kdc) for <i>G. thermoglucosidans</i> | 176 |
| 6.1.9.1. Keto acid decarboxylase Kdc_CAB34226 | 176 |
| 6.1.10. Testing keto acid decarboxylases for activity | 177 |
| 6.1.11. Isobutanol production in <i>G. thermoglucosidans</i> | 178 |
| 6.1.12. Alternative pathway for isobutanol production | 179 |
| 6.1.13. Sources of alcohol dehydrogenases (Adh) | 181 |
| 6.2. Results | 182 |
| 6.2.1. Tolerance of <i>G. thermoglucosidans</i> to isobutanol | 182 |
| 6.2.2. Expression of <i>Lactococcus lactis</i> Kdc | 183 |
| 6.2.3. Finding a thermostable acylating aldehyde dehydrogenase (ALDH) | 191 |
| 6.2.4. Expression of acylating aldehyde dehydrogenase (ALDH) | 195 |
| 6.2.5. HPLC analysis to detect isobutanol production | 199 |
| 6.2.6. Overproduction of 2-ketoisovalerate | 211 |
| 6.3. Discussion | 212 |
| 6.3.1. ALDH and cloning | 214 |
| 6.3.2. Isobutanol yields | 215 |
| 6.3.3. Isobutanol as a biofuel | 216 |
| 6.4. Future work | 217 |
| 7. General discussion | 220 |
| 7.1. Synthetic biology | 220 |
| 7.2. Abstraction in synthetic biology: bioparts | 221 |
| 7.3. Synthetic biology and industry | 221 |
| 7.4. Toolbox | 222 |

| | |
|--|-----|
| 7.5. Heterologous expression by <i>G. thermoglucosidans</i> | 224 |
| 7.6. Improvements to the toolbox | 224 |
| 7.7. The importance of predictive software for synthetic biology | 224 |
| 7.8. New selection markers | 225 |
| 7.9. Lessons to be learnt from this project | 226 |
| 7.10. Bioparts: <i>E. coli</i> ↔ <i>G. thermoglucosidans</i> | 227 |
| 7.11. What the future may hold for <i>G. thermoglucosidans</i> synthetic biology | 229 |
| 7.12. Isobutanol production by <i>G. thermoglucosidans</i> | 230 |
| 8. Chapter 8: Conclusions | 231 |
| 9. Chapter 9: References | 232 |
| 10. Chapter 10: Appendix | 243 |

INDEX OF FIGURES

| | |
|---|-----------|
| Figure 1. Organic synthesis and cellular metabolism | 20 |
| Figure 2. Arrangement of parts to produce a protein generator device | 24 |
| Figure 3. Hairpin formed by a rho-independent transcriptional terminator | 27 |
| Figure 4. A phylogenetic tree of the domain Bacteria | 33 |
| Figure 5. Neighbour-joining tree of <i>Geobacillus</i> spp. | 34 |
| Figure 6. The role of synthetic biology | 38 |
| Figure 7. 1 kb DNA molecular weight marker | 47 |
| Figure 8. Prestained protein weight marker | 52 |
| Figure 9. Plasmid map of "pGR002" | 60 |
| Figure 10. Photograph of <i>G. thermoglucosidans</i> and <i>E. coli</i> cell pellets expressing pheB with or without the addition of catechol | 61 |
| Figure 11. Schematic view of the structure of the <i>B. subtilis</i> YtvA | 68 |
| Figure 12. Phylogenetic tree of bacterial species that contain genes that encode for putative flavin-binding photosensors | 69 |
| Figure 13. Comparison of the genomic sequence of <i>Thermosynechococcus elongatus</i> corresponding to the LOV-domain protein with the codon optimized sequence of hotLOV. | 71 |
| Figure 14. Three-dimensional structure of hbLOV. | 72 |
| Figure 15. Hypothetical optimal melting temperatures for mesophilic and thermophilic proteins. | 73 |
| Figure 16. Colony-PCR products looking for the presence of sfGFP in <i>E. coli</i> transformants analyzed on a 0.8% agarose gel. | 77 |
| Figure 17. Cassette used to express sfGFP under the control of the <i>ldh</i> promoter, in <i>G. thermoglucosidans</i> DL44 | 78 |
| Figure 18. Expression of sfGFP by <i>G. thermoglucosidans</i> . | 79 |
| Figure 19. Schematic representation of the primers used to amplify <i>ytvA</i> . | 81 |
| Figure 20. Primer pairs and PCR products used to amplify <i>ytvA</i> from the genome of <i>B. subtilis</i> 168 and convert it to BsFbFP. | 82 |
| Figure 21. Analysis of PCR-amplification of the p43 and <i>ldh</i> promoters in an 0.8% agarose gel. | 82 |
| Figure 22. Cassette used to express BsFbFP under the control of the p43 promoter, in both <i>E. coli</i> JM109 and <i>G. thermoglucosidans</i> DL44. | 83 |
| Figure 23. Fluorescence of BsFbFP in <i>E. coli</i> . | 83 |
| Figure 24. Cassette used to express BsFbFP under the control of the <i>Ldh</i> promoter, in <i>G. thermoglucosidans</i> DL44. | 85 |
| Figure 25. Colony-PCR of <i>ldh</i> +BsBfFP analyzed on a 0.8% agarose gel. | 85 |
| Figure 26. 0.8% agarose gel showing the RT_PCR of BsFbFP. | 86 |
| Figure 27. Comparison of the amino acid sequences among BsFbFP and PpFbFP using protein BLAST | 87 |
| Figure 28. Cassette used to express PpFbFP under the control of the RplS promoter, in both <i>E. coli</i> DH10B and <i>G. thermoglucosidans</i> DL44. | 88 |
| Figure 29. Western blot detecting the PpFbFP protein from cultures of <i>G. thermoglucosidans</i> DL44 containing the p11AK+RplS+PpFbFP construct. | 89 |
| Figure 30. Fluorescence observed for each of the reporter proteins tested in this study when exposing pellets to blue light in a blue-light transilluminator. | 90 |
| Figure 31. Histograms A-O obtained from flow cytometry analysis of the various promoter /reporter protein combinations. | 91 |
| Figure 32. Shuttle vector pUCG18. | 108 |
| Figure 33. Plasmid map of pUCG3.8. | 108 |
| Figure 34. Plasmid map of pMK3. | 111 |
| Figure 35. Plasmid map of pMK4. | 111 |
| Figure 36. Secondary structures of transcriptional terminators. | 114 |
| Figure 37. Plasmid map of pUCG18. | 115 |

| | |
|---|-----|
| Figure 38. The basic design of a pSEVA plasmid. | 117 |
| Figure 39. Plasmid map of the first modular plasmid. | 119 |
| Figure 40. Design of the synthetic MCS to be included in the modular plasmids. | 119 |
| Figure 41. Secondary structures of transcriptional terminators. | 120 |
| Figure 42. Primers <i>AscI</i> - <i>SwaI</i> -pUC_ori and RC-kanR- <i>AscI</i> were used to amplify pUCG18. | 122 |
| Figure 43. Plasmid resulting from the <i>AscI</i> self-ligation of the PCR product from Figure 64. | 122 |
| Figure 44. M13F and M13R primers were used to PCR-amplify the sequence corresponding to the synthetic MCS | 122 |
| Figure 45. Plasmid resulting from the ligation of the plasmid shown in Figure 65 digested with <i>AscI</i> and <i>SwaI</i> | 123 |
| Figure 46. PCR-amplification of the plasmid from Figure 67 | 123 |
| Figure 47. PCR-amplification of the plasmid from Figure 67 | 123 |
| Figure 48. Plasmid resulting from the ligation of the PCR products from Figures 68 and 69 | 124 |
| Figure 49. Primers <i>NotI</i> -repBST1 and RC-AmpR- <i>NotI</i> were used to amplify the entire vector shown in Figure 70 | 124 |
| Figure 50. Schematic view of the new minimal modular plasmid p11AK1. | 124 |
| Figure 51. 1% agarose gel of test-digest with <i>NotI</i> / <i>PmeI</i> to demonstrate the functionality of two of the boundary restriction sites in p11AK1. | 125 |
| Figure 52. 1% agarose gel of test-digest with <i>PmeI</i> / <i>AscI</i> to demonstrate the functionality of these two boundary restriction sites in p11AK1. | 125 |
| Figure 53. Restriction map of the empty Multiple Cloning Site (MCS) and location of the “rho1” and “rho2” transcriptional terminators. | 126 |
| Figure 54. Schematic representation of the larger vector set constructed to be part of the <i>E. coli</i> / <i>G. thermoglucosidans</i> toolbox. | 128 |
| Figure 55. Schematic representation of the smaller vector set constructed to be part of the <i>E. coli</i> / <i>G. thermoglucosidans</i> toolbox. | 129 |
| Figure 56. Transformation efficiencies of <i>E. coli</i> obtained with the modular vectors. | 134 |
| Figure 57. Transformation efficiencies of <i>G. thermoglucosidans</i> obtained with the modular vectors. | 135 |
| Figure 58. Microscopy imaging detection of sfGFP and mCherry fluorescence to test for plasmid (in)compatibility. | 136 |
| Figure 59. Average DNA concentration obtained after extracting each plasmid from overnight <i>E. coli</i> . | 137 |
| Figure 60. Average DNA concentration obtained after extracting each plasmid from overnight <i>G. thermoglucosidans</i> . | 138 |
| Figure 61. Growth comparison among strain p12AC1 (<i>repB</i> <i>pmk4_catE</i>) and strain p11*AC1 (<i>repBST1</i> <i>pmk4_catE</i>). | 139 |
| Figure 62. Illustration of the sequence of events leading to the first cross-over event | 148 |
| Figure 63. Illustration of the sequence of events leading to the second cross-over event | 149 |
| Figure 64. Creation of a <i>pyrE</i> knock-out strain. | 151 |
| Figure 65. Knock-in of <i>lacZ</i> into the chromosome of <i>Clostridium acetobutylicum</i> | 152 |
| Figure 66. Sequence of events leading to the first cross-over event. | 153 |
| Figure 67. Sequence of events leading to the second cross-over event. | 154 |
| Figure 68. PCR-amplification of the first homology region | 155 |
| Figure 69. PCR-amplification of the second homology region | 155 |
| Figure 70. PCR-amplification of the gene encoding for <i>sfGFP</i> | 156 |
| Figure 71. PCR-amplification of the Gibson isothermal assembly reaction | 156 |
| Figure 72. Plasmid resulting from the <i>PstI</i> / <i>SacI</i> ligation | 157 |
| Figure 73. RBS calculations to determine the suitability of the <i>ldh</i> native RBS with sfGFP. | 158 |
| Figure 74. Primers SEQ-Ldh-KO and SEQ-RC-Ldh-KO anneal to regions adjacent | 159 |

| | |
|---|-----|
| to the Ldh locus in the genome of <i>G. thermoglucosidans</i> DL44. | |
| Figure 75. Colony-PCR screening results of the Ldh locus | 160 |
| Figure 76. Genome map of <i>Geobacillus</i> sp. Y4.1MC1 highlighting (red box) the location of the <i>pyrE</i> gene in the genome. | 161 |
| Figure 77. Primers Ascl-400 and RC-400 were designed to amplify the first 400 bp homology region. | 162 |
| Figure 78. PCR-amplification of the 400 bp region using primers Ascl-400 and RC-400. | 162 |
| Figure 79. PCR-amplification of the 1200 bp region using primers AvrII-1200 and RC-1200-Bsu36I. | 163 |
| Figure 80. Atsumi and Liao's 2008 strategy to produce isobutanol in <i>E. coli</i> . | 171 |
| Figure 81. Map of the biosynthesis of the branched-chain amino acid valine in <i>E. coli</i> . | 175 |
| Figure 82. Pathway for isobutanol production as proposed by Donaldson <i>et al.</i> [2010]. | 180 |
| Figure 83. Growth of DL44 in the presence of 0, 1, 1.5 and 2% isobutanol monitored over the course of 49 hours. | 182 |
| Figure 84. Accl/PstI test-digest to determine the presence of the pLysE and the kdc plasmids analyzed on a 1% agarose gel. | 184 |
| Figure 85. Accl/PstI test-digest to determine the loss of the pLysE plasmid analyzed on a 1% agarose gel. | 185 |
| Figure 86. SDS-PAGE gel electrophoresis of intracellular protein from cultures of BL21 containing the T7+RBS+kdc construct. | 186 |
| Figure 87. Western blot analysis of the yields of each of the two wash steps and four elution steps carried out to purify the 6XHis-tagged proteins. | 187 |
| Figure 88. Absorbance readings of the coupled decarboxylase assay. | 188 |
| Figure 89. Design of the cassette used to express Kdc by <i>G. thermoglucosidans</i> DL44. | 189 |
| Figure 90. Cell lysate of <i>G. thermoglucosidasius</i> DL44 analyzed on a Western blot. | 190 |
| Figure 91. Degradation pathway of valine in <i>G. thermoglucosidans</i> . | 191 |
| Figure 92. Concise BLAST results of the search against the NCBI database of the <i>T. thermophilus</i> ALDH. | 192 |
| Figure 93. Conserved domains identified by NCBI Blast for the acylating aldehyde dehydrogenase of <i>T. thermophilus</i> . | 192 |
| Figure 94. Alignment between the ALDH amino acid sequences of <i>T. thermophilus</i> ALDH (Subject) and the <i>G. thermoglucosidans</i> ALDH (Query). | 193 |
| Figure 95. Sequence comparison among the codon-optimized <i>T. thermophilus</i> Ald gene (Query) and the non-codon optimized version of the same gene (Subject). | 194 |
| Figure 96. Design of the cassette used to express Tt_Aldh by <i>G. thermoglucosidans</i> DL44. | 195 |
| Figure 97. PCR-amplification of gene <i>Ald</i> from <i>G. thermoglucosidans</i> DL33. | 196 |
| Figure 98. Primers used to amplify Geoth_ALDH from <i>G. thermoglucosidans</i> DL33 | 197 |
| Figure 99. Design of the cassette used to express Geoth_Aldh by <i>G. thermoglucosidans</i> DL44. | 197 |
| Figure 100. Cell lysate of <i>G. thermoglucosidasius</i> DL44 analyzed on a Western blot. | 198 |
| Figure 101. HPLC chromatogram of the sterile ASM medium used to grow <i>G. thermoglucosidans</i> . | 199 |
| Figure 102. HPLC chromatogram of the isobutanol 0.1 M standard. | 200 |
| Figure 103. HPLC chromatogram of the supernatant of DL44 p11AK+kdc grown in ASM supplemented with 0.05 M ketoisovalerate. | 201 |
| Figure 104. HPLC chromatogram of the supernatant of DL44 p11AK+Tt_ALDH grown in ASM supplemented with 0.05 M ketoisovalerate. | 202 |
| Figure 105. HPLC chromatogram of the supernatant of DL44 p11AK+Geoth_ALDH grown in ASM supplemented with 0.05M ketoisovalerate. | 203 |

| | |
|---|-----|
| Figure 106. HPLC chromatogram of the supernatant of DL44 grown in ASM supplemented with 0.05M ketoisovalerate. | 204 |
| Figure 107. HPLC chromatogram of the supernatant of DL44 grown in ASM not supplemented with ketoisovalerate. | 205 |
| Figure 108. HPLC chromatograms obtained with RI detection from the analysis of decreasing concentrations of isobutanol. | 206 |
| Figure 109. GC-MS chromatogram of the supernatant of DL44 grown in ASM supplemented with 0.05M ketoisovalerate. | 208 |
| Figure 110. GC-MS chromatogram of an 0.1M isobutanol standard in HPLC grade toluene. | 208 |
| Figure 111. Average production of moles of isobutanol per moles of glucose achieved by strains Geoth_ALDH and Tt_ALDH in comparison to DL44. | 209 |
| Figure 112. Design of artificial operon to overproduce 2-ketoisovalerate | 211 |
| Figure 113. Synthetic biology's input to biotechnology. | 212 |
| Figure 114. Synthetic biology's input to biotechnology | 220 |
| Figure A115. DNA sequence of <i>Ldh</i> promoter and sfGFP as found in "pUCG18+pLdh+sfGFP" | |
| Figure A116. Highlighted in yellow are the results of the reverse engineering analysis to determine the strength of the pheB RBS | |
| Figure A117. Reverse engineering analysis of the native RBS found downstream from the <i>G. stearothermophilus</i> Ldh CDS. | |
| Figure A118. Results of the reverse engineering analysis to determine the strength of the native RBS of the <i>G. stearothermophilus</i> Ldh promoter. | |
| Figure A119. DNA and amino acid sequence of YtvA as found in the genome of <i>B. subtilis</i> | 168 |
| Figure A120. DNA and amino acid sequences of BsFbFP. | |
| Figure A121. Detail of sequence corresponding to the p43 promoter and BsFbFP | |
| Figure A122. Detail of sequence corresponding to the <i>Ldh</i> promoter and BsFbFP | |
| Figure A123. Detailed sequence of synthetic MCS containing the RplS promoter | |
| Figure A124. Comparison between the size and agglomeration of <i>E. coli</i> and <i>G. thermoglucosidans</i> when compared by flow cytometry. | |
| Figure A125. Detailed sequence of <i>repBST1</i> as found in pUCG18. | |
| Figure A126. Detailed sequence of <i>kanR</i> as found in pUCG18. | |
| Figure A127. Detailed sequence of <i>ampR</i> as found in pUCG18. | |
| Figure A128. Detailed sequence of <i>repB</i> as found in pUB190. | |
| Figure A129. Complete sequence of plasmid p11AK0. | |
| Figure A130. Sequence detail of the mutations produced in <i>repBST1</i> (p11AC1) | |
| Figure A131. Sequence detail of the temperature sensitive plasmid used to knock-in <i>sfGFP</i> into the <i>ldh</i> locus. | |
| Figure A132. <i>Lactococcus lactis</i> subsp. <i>lactis kivd</i> gene for alpha-ketoisovalerate decarboxylase, | |
| Figure A133. Codon-optimized for <i>G. stearothermophilus</i> version of the <i>Lactococcus lactis</i> subsp. <i>lactis kivd</i> gene | |
| Figure A134. Sequence detail of the T7 promoter+RBS+ <i>kdc</i> +his-tag construct used to express Kdc | |
| Figure A135. Sequencing of the <i>kdc</i> gene as in BL21 cells after successive subculturing that resulted in the cells loss of pLysE. | |
| Figure A136. Sequence detail of the ALDH gene from <i>Thermus thermophilus</i> as obtained from NCBI. | |
| Figure A137. Sequence detail of the ALDH gene from <i>Geobacillus</i> sp. Y4.1MC1 as obtained from NCBI. | |
| Figure A138. Sequence detail of the ALDH gene from <i>Thermus thermophilus</i> as obtained from NCBI after codon optimization by GeneART, | |
| Figure A139. Sequencing results of the <i>T. thermophilus</i> ALDH gene after it was incorporated to plasmid p11AK+Tt_ALDH. | |

Figure A140. Sequencing results of the *G. thermoglucosidans* ALDH gene after it was incorporated to plasmid p11AK+Tt_ALDH.

Figure A141. HPLC chromatogram of glucose standard 0.05 M.

Figure A142. HPLC chromatogram of 2-ketoisovalerate standard 0.05 M.

Figure A143. HPLC chromatogram of sterile ASM without additional glucose or 2-ketoisovalerate.

Figure A144. HPLC chromatogram of sterile ASM that has been supplemented with 0.05M ketoisovalerate and 0.1M glucose.

Figure A155. Close up of the isobutyraldehyde peak (26.032 min) and the isobutanol peak (34.644 min) observed from cultures of DL44 grown in the presence of 0.05M 2-ketoisovalerate.

Figure A156. Colony-PCR screening of *E. coli* DH10B transformant colonies.

Figure A157. Colony-PCR screening of *G. thermoglucosidans* DL44 transformant colonies.

Figure A158. Continuation of Figure A170.

INDEX OF TABLES

Table 1. Selected examples of where synthetic biology has come together with metabolic engineering to achieve the production of useful compounds via the heterologous expression of proteins from diverse kingdoms in *E. coli* or *S. cerevisiae*.

Table 2. Useful enzymes derived from thermophiles that have applications in industry.

Table 3. Potential thermophiles being proposed as platforms for the industrial production of valuable compounds.

Table 4. Strains of *E. coli* and *G. thermoglucosidans* used in this study.

Table 5. Media used to grow *E. coli* and *G. thermoglucosidans*

Table 6. Mutations of three of the variants of GFP of in comparison to the wild-type GFP

Table 7. Maximal excitation and emission wavelengths for the fluorescent reporter proteins tested for *E. coli* and *G. thermoglucosidans*

Table 8. Naming convention for modular toolbox plasmids based on parts composition

Table 9. Naming given to modular toolbox plasmids according to naming convention

Table 10. Average number of colonies obtained after transforming with 15 ng of DNA

Table 11. Physical and chemical properties of isobutanol

Table 12. Tolerance to alcohols by natural and non-natural producers of ethanol and butanol.

Table 13. Kdc and Adh enzymes screened by Atsumi and Liao [2008]

Table 14. Relevant characteristics of engineered isobutanol producing strains

Table 15. Isobutanol production titres achieved by Lin *et al.* [2014] in *G. thermoglucosidans* using the Kdc_LLKF_1386.

Table 16. Results from the calibration curve obtained after analyzing decreasing concentrations by HPLC

Table 17. Peak areas extracted from the HPLC analysis of the supernatant ASM of *G. thermoglucosidans* (nRIU*S)

Table 18. Moles of isobutanol obtained per 1 mole of glucose

Table A19. List of primers used in this study.

CHAPTER 1:

MOTIVATION AND INTRODUCTION

1.1. MOTIVATION

Science and industry rely on chemistry to synthesize complex valuable compounds, such as chemicals, fuels and pharmaceuticals, using simple organic compounds as starting materials in reactions. Such synthesis frequently requires very particular physical or chemical conditions, such as high temperatures or pressure to allow reactions to take place or alternatively, they rely on expensive synthetic catalysts. These chemical processes are costly and often, many of the compounds that are being chemically synthesized are already synthesized by nature, but in lower yields or in complicated mixtures with other molecules [Wallace and Balskus, 2014]. Our exploration of biology through science has allowed us to screen nature for useful enzymes, but our increasing understanding of the natural world has revealed that it is possible to explore the full potential of living cells and adapt and engineer these to human requirements [Church *et al.* 2014].

The field of metabolic engineering aims to improve the yields of relevant biochemical compounds by modifying the existing metabolic pathways of organisms that already naturally produce a desired molecule [Ladkau 2014]. This is complemented by synthetic biology, a novel approach that takes metabolic engineering a step beyond. Synthetic biology is an emerging field that enables the application of rational engineering principles to the design of biology. The aim of synthetic biology is to make genetic engineering approaches predictable by breaking genetic arrangements into their minimal basic components and study how these components work together and can be made to perform new tasks. Synthetic biology aims to create parts that can be put together into devices and pathways that result in finely tuned biological functions [Keasling *et al.* 2012]. Such devices can yield a more thorough understanding of the biology of organisms and can also create new industrially-relevant phenotypes, such as cells that perform the production of valuable compounds like pharmaceuticals or fuels from cheap feedstocks.

The idea of applying engineering principles to biology requires reducing biology to its most basic construction elements. Understanding can only be achieved once the ability has been gained to modify each of the basic building block elements in genetic constructions whose outcomes we can observe and measure [Kelwick *et al.* 2014]. The more rational the comparisons, the better the understanding that can be obtained from matching experimental work with mathematical modeling, leading ultimately to the design of optimized parts and devices for each particular required function.

Synthetic biology efforts so far have focused largely on traditional hosts such as *Escherichia coli* or *Saccharomyces cerevisiae* because these are well-known cells and tools are available for both that permit genetic modification for desired functions. However, certain applications (e.g. growth in extreme conditions) would pose a major challenge to the existing machinery of these cells and so to overcome this limitation, specialized hosts need to be sought that already have the natural capacity to perform in unique environments but can also be engineered [Fischer *et al.* 2008].

To engineer a non-standard organism requires first 'domesticating' it, by learning to culture and manipulate it in the lab. Working with a poorly characterized host for whom efficient genetic tools do not exist is challenging but necessary work towards achieving industrially-viable biotechnologies that can work in specific environments. One of the main requirements that enable domestication of an organism is the availability of appropriate genetic tools that work with its cells. Developing a modular genetic toolkit by following the principles of synthetic biology allows a large number of interchangeable DNA parts to be quickly developed for genetic engineering and accelerate domestication [Nikel *et al.* 2014].

Geobacillus thermoglucosidans is a thermophilic bacterium that has been used for the industrial scale production of ethanol from lignocellulosic biomass: a renewable, cheap, untapped feedstock. The high temperatures of growth permit the easy separation of the desired volatile by-product ethanol. *G. thermoglucosidans* is an interesting organism because it has naturally evolved the capacity to feed on biomass, and it has the natural capacity to produce ethanol. Metabolic engineering efforts have so far focused on improving the yields of ethanol production [Taylor, 2009].

We propose that *G. thermoglucosidans* is a uniquely specialized organism whose ability to grow on lignocellulose and withstand harsh industrial conditions should be exploited further, but to be able to do this, appropriate genetic tools need to be constructed and tested.

1.2. RESEARCH OBJECTIVES

The research objectives of this project are therefore to contribute to developing *Geobacillus thermoglucosidans*, an industrially-relevant thermophilic bacteria, as a chassis organism for synthetic biology by the creation of appropriate and effective genetic tools. The suitability of the tools created is tested here, by the introduction of genes for isobutanol production; a proof-of-concept application.

1.3. INTRODUCTION

- The present work is a combination of metabolic engineering and synthetic biology. It is the application of the principles of synthetic biology to achieve a metabolic engineering application.
- Synthetic biology uses engineering to design and compose biological parts and devices for useful applications, and metabolic engineering is the modification of metabolic pathways for the production or consumption of organic molecules.
- Genetic tools are a requirement for genetic modification of any organism and are lacking for thermophilic microbes, hindering efficient metabolic engineering in these organisms.
- The useful application being considered in this project is the genetic modification of thermophiles for the production of valuable organic compounds such as biofuels.
- The organism being modified is *Geobacillus thermoglucosidans* because it is a thermophile capable of producing ethanol at an industrial scale using lignocellulosic biomass as a feedstock.

1.3.1. ORGANIC CHEMISTRY VS. SYNTHETIC BIOLOGY

Over the past two hundred years, synthetic organic chemists have developed a huge array of reagents, catalysts, reaction conditions, methodologies and solvents to permit converting one compound into another. The product of one reaction serves as the starting point for the next reaction, in a series of transformations that lead to the formation of a final product of interest. In contrast, cells have evolved pathways that also lead to the biosynthesis of compounds. Cells act as a reaction vessel and contain enzymes that modify metabolites, under regulated conditions, in a series of reactions to also achieve the production of compounds of interest [Wallace and Balskus, 2014].

Instead of optimizing reaction conditions, for biology, it is important to find or modify enzymes to confer them the reactivity and selectivity needed to produce the target molecule. Advances in enzyme engineering and synthetic biology will expand the breadth of organisms' biosynthetic capabilities [Keasling *et al.* 2012].

Advances in the fields of metabolic engineering and synthetic biology raise the question as to which approach (chemistry or biology) is more suitable for the production of target molecules. It is questionable whether the synthetic capabilities of organisms will ever match or outcompete those of synthetic organic chemistry for certain target molecules [Wallace and Balskus, 2014], but there are certainly some advantages that biological biosynthesis holds over chemical synthesis:

- economic viability of production by using cheap energy and carbon sources
- reduction of environmental impact
- improvement of selectivity by enzyme engineering
- improvement of catalytic efficiency by enzyme engineering
- mild reaction conditions

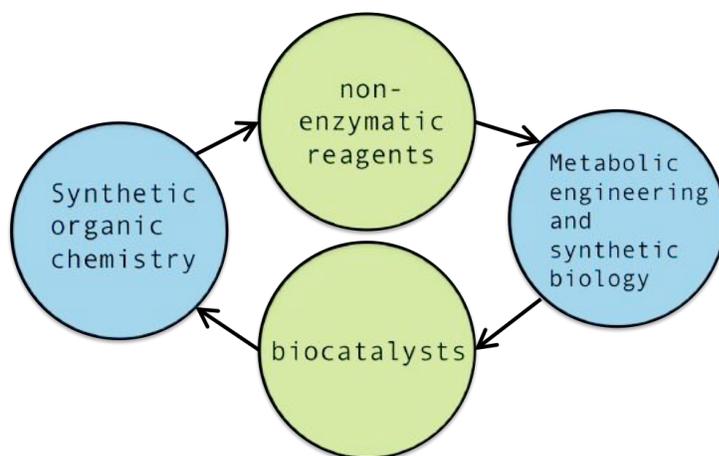


Figure 1. Wallace and Balskus [2014] propose that it is time to merge organic synthesis with cellular metabolism. Synthetic biology and metabolic engineering expand the breadth of reactions and pathways and organic chemistry produces available synthetic compounds which could be used as precursors for enzymatic biocatalysis *in vivo*.

The use of biocatalysts has become increasingly important in synthetic chemistry (see **Figure 1**) and likewise, non-enzymatic reagents are being incorporated into synthetic biology and metabolic engineering efforts [Wallace and Balskus, 2014]. Synthetic organic chemistry for commodity chemical production has small profit margins. In the case of biofuels, especially, the oil price instability squeeze those margins even further, which resulted in the abandonment of low margin products by refineries. [Keasling, 2012]

Synthetic biology is bringing new tools and approaches to biotechnology [Church et al. 2014] and biotechnology is becoming important for the economy. In 2007, biotechnology was contributing to 1% of US GDP [Carlson, 2007] and in 2014, products of genetically modified cells were contributing to 2% of US GDP [Thomas Smith *et al.* 2014]. The economic contribution of biotechnology depends on the capabilities of new biological technologies. As the sophistication of biological engineering increases, it will provide new goods and services at lower costs and higher efficiencies [Carlson, 2007]. The fields of synthetic biology and metabolic engineering, described below, are responsible for producing genetically modified organisms.

1.3.2. DEFINITION OF SYNTHETIC BIOLOGY

The introduction of engineering to biology may enable a discipline similar to the one that resulted in modern aviation or information technologies: combining assembly tools with modeling and an existing design framework permits the assembly of airplanes and cars, and ideally the same engineering principles can also be applied to biology [Carlson, 2007].

Synthetic biology is the engineering-driven building of increasingly complex biological entities for novel useful applications [Heinemann 2006]. In order to be able to construct biological systems that can be understood, designed and tuned to meet specific performance criteria, parts are assembled into larger devices and systems and introduced into cells to change their behavior to solve specific problems.

The building of novel biological entities involves the rational design and engineering of biological parts that can be assembled into devices, circuits and systems. Devices are formed by the combination of parts, circuits are a combination of devices and systems are the combination of circuits or are metabolic pathways composed of many interacting enzyme-encoding genes [Kelwick *et al.* 2014].

1.3.3. METABOLIC ENGINEERING

According to Ladkau [2014] metabolic engineering aims at pathway optimization for:

- Metabolite overproduction.
- Maximizing precursor supply.
- Expansion of the target product portfolio.
- Introduction of heterologous pathways.
- Optimization of pathway flux.
- Optimizing the production of native target products.
- Generate new target products.
- Use cheap growth substrates.

The field of metabolic engineering aims to reinforce what nature does best by correcting inadequacies and removing dead-ends of a metabolic pathway by "hijacking" cellular systems and converting them into cell factories [Trantas *et al.* 2015]. In order to do so, metabolites must be rerouted, regulation must be altered and new enzymatic pathways must be introduced and balanced [Abatemarco *et al.* 2013].

Metabolic engineering so far has been responsible for the production of biofuels, pharmaceuticals and in bioremediation [Lee *et al.* 2011]. Advances in the efficient microbial production of chemicals have been shaped by the development of tools that allow understanding and manipulating cellular metabolism; tools which permit transferring genes, modulating gene expression and engineering proteins [Clomburg and Gonzalez, 2010]. Overexpression of a pathway enzyme is a common strategy in metabolic engineering as it is an obvious way to achieve higher fluxes. The expression of heterologous genes can extend existing pathways to produce novel products. When a host is producing its own secondary metabolites, and genes encoding enzymes that produce novel products are introduced, there will however, be a shift in metabolic flux [Holtz and Keasling, 2010].

The project introduced here will focus on the production of useful, efficient appropriate genetic tools, by following synthetic biology principles. These tools will enable the production of useful compounds such as biofuels in *Geobacillus*, as they can be used to manipulate cellular metabolism to direct it towards the production of useful compounds.

1.3.4. HOW SYNTHETIC BIOLOGY AND METABOLIC ENGINEERING COME TOGETHER

The study of nature permits us to define what the characteristic properties are for living systems. The fields of metabolic engineering and synthetic biology come together to determine how such properties can be improved and how to further exploit them. The combination of synthetic biology and metabolic engineering has made possible the heterologous production of many low and high-value chemicals. These chemicals exist in nature but the ability to make them can be transferred between various organisms, so long as genomic information and tools to introduce such functions into heterologous hosts are available.

Advances in synthetic biology can enhance the field of metabolic engineering by designing methods for engineering genetic control. This emerging branch of biology provides key tools that aid in the design, assembly and implementation of synthetic pathways. An approach that combines the strategies and concepts from both synthetic biology and metabolic engineering can be implemented to develop efficient microbial platforms for the production of biofuels [Clomburg and Gonzalez 2010].

There are many examples of research papers in which synthetic biology and metabolic engineering have come together for the production of valuable chemical compounds. The examples summarized in **Table 1** were selected because they involve the expression of heterologous proteins to achieve the production of a target compound.

| Table 1. Selected examples of where synthetic biology has come together with metabolic engineering to achieve the production of useful compounds via the heterologous expression of proteins from diverse kingdoms in <i>E. coli</i> or <i>S. cerevisiae</i>. | | | |
|--|--|---|--|
| Chassis | Pathway | Target product | Utility of target product |
| <i>E. coli</i> [Keasling, 2012] | The mevalonate pathway of <i>S. cerevisiae</i> was coupled with the artemisinic acid pathway from <i>Artemisia annua</i> (Plantae: Asteraceae) | Artemisinic acid | Precursor to the antimalarial drug artemisinin |
| <i>E. coli</i> and <i>S. cerevisiae</i> [Peralta <i>et al.</i> 2011] | The mevalonate pathway of <i>S. cerevisiae</i> was coupled with a bisabolene synthase from the gymnosperm <i>Abies grandis</i> (Plantae: Embryophyta) | Bisabolene | Precursor to bisabolane, a terpene-based potential biofuel |
| <i>S. cerevisiae</i> [Li <i>et al.</i> 2013] | Xylanase and β -xylosidase from <i>Aspergillus terreus</i> (Fungi:Ascomycota) was paired with xylose reductase from <i>Candida tropicalis</i> were paired with promoters and terminators from <i>S. cerevisiae</i> . | Xylitol | Sugar substitute in food industry |
| <i>E. coli</i> [Ajikumar <i>et al.</i> 2010] | Taxadiene pathway obtained from <i>Taxus canadensis</i> , <i>Taxus brevifolia</i> , and <i>Taxus cuspidata</i> (Plantae:Pinophyta) were paired with a native isoprenoid pathway from <i>E. coli</i> | Taxadiene | Precursor to taxol (anticancer compound) |
| <i>E. coli</i> Zhu <i>et al.</i> 2014 | Farnesene synthase obtained from <i>Malus x domestica</i> (Plantae:Rosaceae) was coupled to the mevalonate pathway of <i>S. cerevisiae</i> | Farnesene | Potential precursor of jet fuel |
| <i>E. coli</i> Atsumi and Liao [2008] | Overexpression of the amino acid biosynthesis pathways coupled with the expression of a keto acid decarboxylase from <i>Lactococcus lactis</i> and knocking-out of competing | 1-Butanol, isobutanol, 1-propanol, 2-methyl-1-butanol, 2- | Potential biofuels |

| | | | |
|--|--|---------------------------------------|--------------------|
| | pathways (ethanol, acetate, lactate, succinate and formate). | phenylethanol, 3-methyl-1-butanol | |
| <i>E. coli</i> [Howard <i>et al.</i> 2013] | Free fatty acids were converted by an acyl-ACP reductase and an aldehyde decarboxylase from <i>Nostoc punctiforme</i> (cyanobacteria). | n-alkanes, iso-alkanes, and n-alkenes | Potential biofuels |

One of the major goals of synthetic biology is to design biological systems for useful applications [Chen *et al.* 2015]. The examples shown in **Table 1** demonstrate that synthetic biology offers the potential to transfer genetic material between diverse organisms to construct engineered cells capable of the production of useful compounds such as fuels, pharmaceuticals, specialty chemicals, polymer precursors, or complex full-length polymers.

1.3.5. THE BIOLOGICAL PARTS THAT CONSTITUTE DEVICES

The work carried out in this thesis focused exclusively on the rational engineering of devices. Therefore, explained below are the principles behind the functional bioparts that constitute these.

It has been explained previously how bioparts are discrete biological sequences of known behavior that have a standard descriptive function [Kelwick *et al.* 2014]. The most common biological parts used to control gene expression in synthetic biology are promoters, which initiate transcription; ribosome binding sites (RBSs), which initiate translation and terminators, which attenuate transcription. These are combined with protein coding sequences to fine-tune the expression of genes of interest [Kosuri *et al.* 2013].

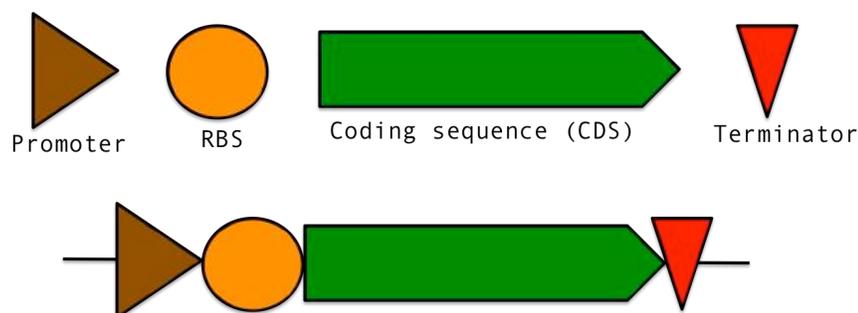


Figure 2. Arrangement of bioparts to produce a typical gene – in this case a "protein generator" device. The promoter regulates transcription and the RBS regulates translation of the gene. The terminator dissociates the transcriptional elongation complex. Protein coding sequences (CDS) encode the protein to be made.

The simplest device in synthetic biology is a "protein generator", depicted in **Figure 2**, consisting of a promoter, an RBS, a protein coding sequence and a terminator. The purpose of this device is to produce the protein encoded by the CDS at a defined expression level [Baldwin *et al.* 2012].

1.3.6. PROMOTERS

Promoters are specific sequences of DNA located upstream from a protein coding sequence (gene). They are necessary to regulate transcription because they are composed of recognition sequences to which RNA polymerases bind to initiate transcription of mRNA [Madigan *et al.* 2009]. The major components of prokaryotic promoter sequences are the -35 and -10 boxes as well as transcription factor binding sites [Singh, 2014].

There are two main kinds of promoters:

1. Constitutive promoters are useful because they are used for continual protein production.
2. Inducible promoters contain operators that allow controlling the timing of transcription by adding a compound or changing the cell environment, which will recruit transcription factors to the operator [Baldwin *et al.* 2012]. Inducible promoters are advantageous over constitutive promoters because they permit the cells to grow, while carrying a synthetic construct that's fully repressed and can then be switched on once the cells have reached stationary phase [Keasling, 2012]. Other kinds of promoters are those that are induced upon starvation of a certain compound or promoters that are induced upon early stationary phase [Keasling, 2008]. *In yeast for example, galactose inducible promoters are used as well as a native copper-inducible promoter. Inducible promoters are desirable when a gene poses a burden on the cell when expressed [Da Silva and Srikrishnan, 2012].*

Much of the utilization of promoters in synthetic biology involves the use of either wild-type or slightly modified inducible or constitutive promoters of different strengths. A library of promoters can be constructed by modifying the DNA sequence of the promoter to achieve different levels of binding by transcription factors and so different levels of transcription from the RNA polymerase [Ellis *et al.* 2009].

1.3.7. RIBOSOME BINDING SITE (RBS)

The Ribosome Binding Site (RBS) is a sequence found in the mRNA of bacteria and it acts to recruit ribosomes to initiate translation. In prokaryotes, the core sequence in the middle of an RBS is also known as the Shine-Dalgarno sequence. The sequence of the RBS affects the strength of the ribosome binding and the rate of translation initiation. This therefore directly affects the amount of protein produced from each mRNA as translation initiation is usually the rate limiting step of the whole process of translation of RNA into a functioning folded protein [Madigan *et al.* 2009].

1.3.8. TRANSCRIPTIONAL TERMINATORS

Transcriptional termination is essential to gene expression [Peters *et al.* 2011]. In prokaryotes, transcription starts when the RNA polymerase binds to the promoter and continues until RNAP reaches a transcriptional terminator [Hoon *et al.* 2005]. Terminators are important for mRNA half-life and stopping transcription. Terminators are regions in the 3'-end of a gene that encode a functional element that influences the 3'-end processing of mRNA, mRNA stability and translational efficiency, all of which have an effect on protein production because terminators are determinant of mRNA abundance [Redden *et al.* 2015].

Two types of bacterial terminators are known [Peters *et al.* 2011]:

1. Rho-independent transcriptional terminators dissociate the elongation complex is dissociated by interactions of DNA and RNA with RNAP without the assistance of auxiliary factors.
2. Rho-dependent transcriptional terminators dissociate transcription complexes with the assistance of an RNA helicase called Rho.

Rho-independent transcriptional regulators are found at the end of operons and upstream from genes where they regulate transcription via attenuation [Peters *et al.* 2011]. Rho-independent transcriptional terminators are the norm in *E. coli* [Kingsford *et al.* 2007], where they have a highly variable thymine content from 2-90% [Reynold *et al.* 1992].

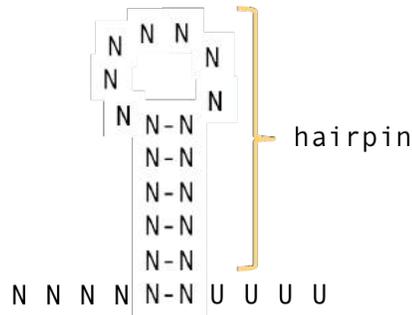


Figure 3. Secondary hairpin structure formed by a rho-independent transcriptional terminator. Rho-independent transcriptional terminators normally include, apart from the complementary palindromic base pairs that form the stem-loop structure of the hairpin, regions, a uracil-rich region that causes the elongation complex to pause in order to permit the formation of the hairpin [Peters, 2011]. *Figure modified from Lesnik, 2011.*

As shown in **Figure 3**, rho-independent terminators consist of two motifs: a short RNA hairpin followed by a U-rich sequence. Both motifs are required for efficient termination. The hairpin must form within the RNA exit channel of the RNA polymerase and the U-rich sequence is necessary because it causes transcription to pause, which in turn permits the hairpin to form and disrupt the elongation complex binding between the U-rich region of RNA and the template DNA [Peters *et al.* 2011, Chen *et al.* 2013].

RNA hairpin formation is crucial for dissociation of the elongation complex. The A rich and T rich regions are important for pausing the complex to give time for the hairpin to form. The free energy for hairpin formation can be calculated using the nearest-neighbour thermodynamic model [Lesnik *et al.* 2001], which takes into account the energy in the interaction among neighbouring base pairs [Owczarzy *et al.* 2011]

Following these principles and using modelling of RNA folding, Chen *et al.* [2013] used characterization of terminators for *E. coli*, to predict design rules for synthetic terminator design. The demand for terminators for *E. coli* is increasing with the growing construction of synthetic devices and circuits: there are very few reliable terminators for *E. coli* available and they are constantly being reused within designs. Weak terminators are sometimes inadvertently used, which has undesirable consequences because the polymerase continues transcription after the terminator.

Similar to the modelling software used for RNA folding by Chen *et al.* [2013], OligoAnalyzer is an online tool for nucleic acid folding that operates on nearest-neighbour

parameters¹. The tool was designed predict the formation of secondary structures of short sequences of DNA or RNA. It is normally used to analyze synthetic oligonucleotides for unwanted secondary structures like hairpins, homodimers or heterodimers. In this work it is used to analyze hairpin formation in short sequences of RNA located at the 3'-end of genes. If hairpin formation is detected, then these sequences are possible transcriptional terminators.

1.3.9. CHASSIS

“Chassis” is an engineering concept. It is the abstract term for the “biological host” in synthetic biology. A chassis is a physical internal framework that accommodates the execution of a synthetic system and usually is a cell. "Framework" refers to the metabolic environment, energy sources, transcription and translation machinery [Kelwick *et al.* 2014].

There are two kinds of chassis in synthetic biology: cells and cell-free systems. Cell-free systems are not feasible for industrial production of compounds because they are more expensive than *in vivo* production, with the major cost being the requirement for specialized high-energy metabolites for system operation. In contrast cells rely on less expensive energy-rich molecules like glucose or tryptone and as such as cheaper to ‘run’ [Thomas Smith *et al.* 2014].

1.3.10. CHARACTERISTICS OF A CHASSIS FOR SYNTHETIC BIOLOGY

Escherichia coli, *Saccharomyces cerevisiae* and *Bacillus subtilis* all share certain characteristics that have made them model lab organisms. Such characteristics are also necessary to develop an industrially relevant strain [Fischer *et al.* 2008]:

- The first requirement is that they are amenable to genetic manipulation. The strain must accept foreign DNA in a controllable fashion and high transformation efficiencies are desirable and particularly necessary for the successful screening of combinatorial libraries.

¹ <http://eu.idtdna.com/Calc/Analyzer/Home/Definitions>

- The second trait that makes laboratory strains attractive is the availability of well-characterized "modules" that allow manipulation of the genotype, *i.e.* promoters of various strengths, termination sequences, repressor-inducer systems, plasmids, chromosome integration cassettes, etc.
- A third feature is the availability of "-omics", *i.e.* platforms and models for genome-wide characterization of cellular responses to different manipulations and environments.
- Finally, the selection of a bacterial species to be a chassis for synthetic biology should be based on the naturally pre-endowed physiological and metabolic properties of the bacteria. Selecting one bacterial species as a chassis can be viewed as the selection of an animal species of interest for domestication [Nikel *et al.* 2014].

1.3.11. SELECTING A HOST FOR BIOFUEL PRODUCTION

The optimal host for biofuel production would degrade lignocellulosic components, ferment the resulting hexoses and pentoses at high rates and with high yields and tolerate the end-products at high titres. In addition to this an ideal host would also be able to work at high temperatures and extreme pH, to avoid costs associated with cooling needed during fermentation after hydrolysis.

Research towards an optimal host has opted to tackle this problem by combining as many of the desirable characteristics as possible [Fischer *et al.* 2008] but most organisms don't suit all the needs. *S. cerevisiae* for example, cannot naturally ferment pentose sugars like xylose [Sonderegger *et al.* 2004], which is a major component of lignocellulosic biomass. And common thermophilic ethanologens like *Clostridium thermosaccharolyticum* cannot tolerate ethanol concentrations higher than 4 % (v/v) [Fong *et al.* 2006].

The important aspect associated with large-scale fermentations to consider is the cost of maintaining sterile conditions. The use of antibiotics is undesirable, not only for economic reasons but also to avoid development of resistance. Running fermentations at a low pH is a possible solution [Fischer *et al.* 2008], but this implies using acidophilic microbes.

Microorganisms that naturally produce the desired metabolite are rarely highly tolerant to acid, heat or similar restrictive conditions. Environmental tolerance is generally a

phenotype dictated by many genes and those traits are hard to introduce to a desired host. For instance, the phenotype an organism has that confers heat-resistance is a property of the system in its entirety, encoded into the peptide sequence of all folding proteins. Environmental tolerance can be improved with respect to the wild-type strain, but the limit depends on the phenotype being sought [Fischer *et al.* 2008].

1.3.12. HEAT TOLERANCE IN INDUSTRY

There is considerable demand for a new generation of stable enzymes that are able to withstand the severe conditions in industrial processes [Elleuche *et al.* 2014]. Extremozymes are enzymes derived from extremophilic organisms. Extremozymes are able to withstand harsh conditions in industrial processes.

The majority of enzymes currently used in industry are obtained from fungi or mesophiles and only a few enzymes, like DNA-polymerases from *Thermus aquaticus*, are commercialized [Elleuche *et al.* 2014].

Mesophiles like *E. coli* or *Bacillus* spp. are commonly used to express extremozymes [Elleuche *et al.* 2014], but this is not ideal because:

1. Lower temperatures can also cause incorrect folding of some thermophilic proteins [Suzuki *et al.* 2013],
2. Different codon usage may be a burden for *E. coli* and reduce productivity [Turner, 2007].

So far, only *Thermus* spp. and *Sulpholobus solfataricus* [Albers *et al.* 2006] are used for the expression of proteins.

The application of extremophiles as cellular factories has attracted attention in recent years, especially for the production of bioethanol following a Consolidated Bioprocessing approach due to several advantages [Lin and Xu 2013, Taylor 2009]:

- Reduction of energy input for cooling (and thus reduction of costs).
- Faster chemical reactions can be achieved.

- Minimization of microbial contamination since heat acts as a crude form of sterilization as the normal environmental microbes that contaminate large scale fermentations are mesophiles that cannot grow at elevated temperatures.
- Facilitation of the downstream recovery of volatile metabolic products.
- Certain lignocellulose degrading enzymes exhibit higher catalytic activity at temperatures of 50 °C [Georgieva *et al.* 2008].

Among the most popular extremozymes currently being studied for the production of useful compounds are (Table 2) [Elleuche *et al.* 2014]:

| Enzyme | Product | Utility of product |
|-------------------------------|--|--|
| Amylases | Starch is degraded into glucose | Additive in food |
| Cellulases and hemicellulases | Cellulose and hemicellulose are degraded into fermentable sugars | For the bioethanol production from lignocellulose |
| Chitinases | Decomposition of chitin | Chitinases are used as biofungicide and bioinsecticide |

1.3.13. VALUABLE THERMOPHILES

Thermophilic bacteria that naturally produce valuable compounds are listed in Table 3.

| Organism | Useful compound produced | Reference |
|--|--|---------------------------------------|
| <i>Carboxydotherrmus hydrogenoformans</i> (Firmicutes: Clostridia) | Capable of converting carbon monoxide and water to hydrogen and carbon dioxide, and it has been postulated as a potential industrial scale producer of cheap hydrogen gas and acetate. | [Henstra and Stams, 2011] |
| <i>Caldicellulosiruptor saccharolyticus</i> (Firmicutes: Clostridia) | Production of hydrogen from hemicellulose-containing waste | [Willquist, Zedan and van Niel, 2010] |
| <i>Thermoanaerobacter sp. X514</i> (Firmicutes: Clostridia) | Production of ethanol | [Lin <i>et al.</i> 2013] |
| <i>Chelatococcus daeguensis</i> (Proteobacteria: α -Proteobacteria) | Production of polyhydroxyalkanoates, which are suitable alternatives to petrochemical-based plastics | [Xu <i>et al.</i> 2014] |

| | | |
|--------------------------------------|--|----------------------|
| <i>Clostridium</i> spp. | Production of ethanol, butanol and isopropanol | [Lin and Xu, 2013] |
| <i>Geobacillus thermoglucosidans</i> | Production of bioethanol | [Cripps et al. 2009] |

There are a few thermophiles proposed to be used in the industrial production of biofuels for their capacity to digest the pentoses and hexoses from lignocellulosic biomass and among them is *Clostridium thermocellum*. However, *Clostridium thermocellum* is not as tolerant to ethanol as *Geobacillus thermoglucosidans* (0.7 vs. 10% v/v) [Akinosho et al. 2014, Tang et al. 2010]. So far, the only thermophile that is used in an industrial environment in *Geobacillus thermoglucosidans* [Cripps et al. 2009], so the research in this project will be focused on this organism.

The advantages of *Geobacillus* spp. for industrial processes are:

- They can be cultivated on a variety of carbon sources [Taylor, 2009].
- They form endospores that allow them to survive unfavorable conditions and facilitate their storage and distribution [Blanchard et al. 2014].
- Ease of cultivation and readily transformable.
- Their high growth rate allows high cell densities and consequently high production of products [Suzuki et al. 2013].
- And most importantly, ethanol is easily separated from the reaction mixture due to its volatility. Biofuel biosynthesis by microbes is a complicated process because biofuels are difficult to purify from the mix of other metabolites of the cell, such as hydrogen, acetic, lactic and propionic acids, acetone, isopropanol and ethanol (in butanol forming *Clostridium* spp. for example). Such separation increases the cost of downstream product purification [Jin et al. 2014 and references therein].

The aforementioned reasons highlight the huge biotechnological potential [Kananavičiūtė and Čitavičius 2015] of *Geobacillus* spp. and for this reason it is necessary to know what are the characters that define the genus *Geobacillus*.

1.3.14. GENUS *Geobacillus*

Several organisms have evolved the capacity to use components of lignocellulose as a feedstock; among them are the *Geobacillus* spp. which are common inhabitants of decaying plant material. The genus *Geobacillus* belongs to the superkingdom Bacteria, phylum Firmicutes, class Bacilli, order Bacillales and family Bacillaceae².

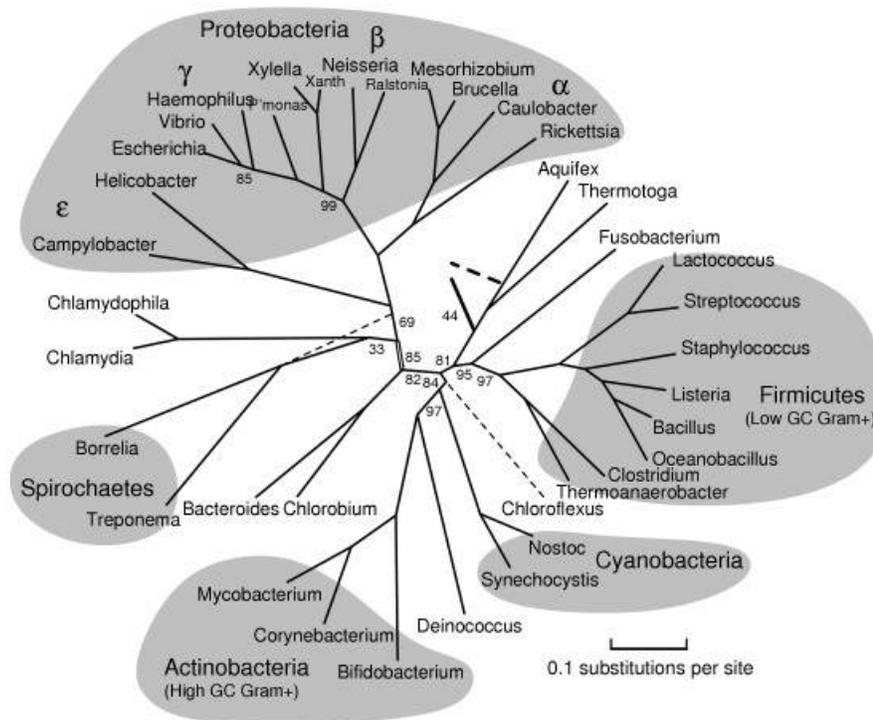


Figure 4. A phylogenetic tree of the domain Bacteria. The tree was obtained by a maximum likelihood approach after analyzing the sequence of 30 representative proteins. *Aeropyrum* and *Methanopyrus* are the outgroup used to root the tree. Numbers indicate bootstrap support values of 100 repetitions.

Figure taken from Bern and Goldberg, 2005.

Apart from *Geobacillus*, the phylum Firmicutes includes other endospore-forming soil bacteria such as *Bacillus* and *Clostridium*, as seen in **Figure 4**. Cells in this phylum sporulate into endospores in stationary phase, after the depletion of essential nutrients. Endospores are resistant to heat, chemicals and radiation and they permit the survival of the species until the environmental conditions permit their growth. Such structures are easily dispersed by wind or water and can remain viable for at least several decades [Madigan, 2009].

² [<http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1426>].

The genus *Geobacillus* was created by Nazina *et al.* in 2001 to accommodate the thermophiles of family Bacillaceae, which were previously part of the genus *Bacillus*. The most recent taxonomic revision for the genus is that of Coorevits *et al.* [2011]. It is the most extensive taxonomic revision of the genus to date and was done using a very thorough comparison of 16S rRNA sequences, as well as polar lipids and fatty acids, phenotypic characterization and DNA-DNA hybridization experiments.

Taxonomic studies have indicated that the genus *Geobacillus* is a monophyletic group and included in this group are: *G. stearothermophilus*, *G. thermodenitrificans*, *G. toebii*, *G. thermoglucosidans*, *G. jurassicus*, *G. subteraneus*, *G. thermoleovorans*, *G. thermocatenulatus* and *G. caldxylosilyticus* [Coorevits *et al.* 2011]. A recent phylogenetic tree of the genus is included in **Figure 5**.

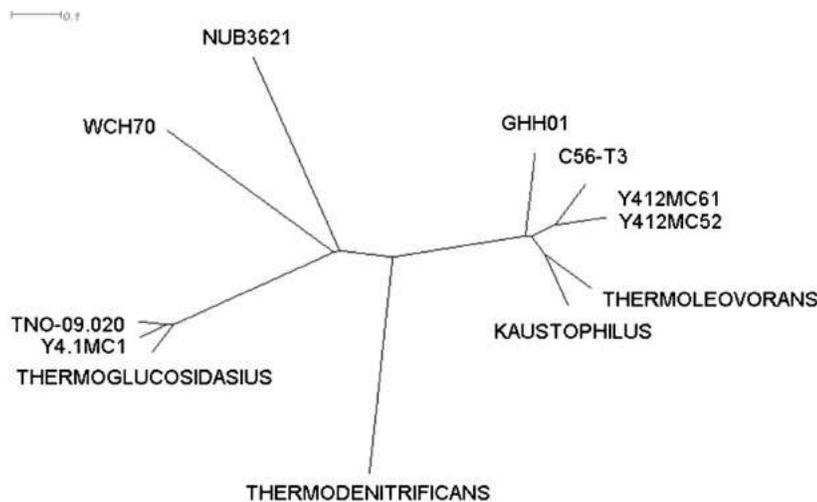


Figure 5. Neighbour-joining phylogenetic tree of *Geobacillus* spp. based on maximal unique matches of their genome sequences from NCBI. *G.thermoglucosidans* appears to be most closely related to *Geobacillus* spp. Y4.1MC1 and TNO-09.020.
Figure taken from Blanchard et al. [2014]

Geobacillus spp. are commonly isolated from compost heaps and hot springs [Coorevits *et al.* 2011] but have also been found deep within the Mariana's trench [Takami *et al.* 2004]. They are chemo-organotrophs [Taylor *et al.* 2009] and they also require trace metal elements for growth, including Fe, V, Mo, W, Mn, Co, Ni, Cu and Zn [Bartosiak-Jentys, 2009]. They grow aerobically or facultatively anaerobically, occasionally using nitrate as an electron acceptor [Coorevits *et al.* 2011]. *Geobacillus* spp. can also be used for the

production of thermostable enzymes, exopolysaccharides or bacteriocins. *Geobacillus* spp. also have the capacity to metabolize hydrocarbons in high temperature oil fields [Nazina *et al.* 2005].

Apart from their use as whole-cell biocatalysts, other applications for *Geobacillus* spp. are as sources of enzymes relevant for biotechnology, for example proteases, amylases, lipases, DNA polymerases or reverse transcriptases [Kananavičiūtė and Čitavičius 2015 and references cited therein].

The most commonly used *Geobacillus* strains in research are: *G. stearotherophilus*, *G. kaustophilus* and *G. thermoglucosidans*, but the present work focuses exclusively on *G. thermoglucosidans*, so it will be described below.

1.3.15. *GEOBACILLUS THERMOGLUCOSIDANS*

Geobacillus thermoglucosidasius was renamed as *G. thermoglucosidans* in 2011 by Coorevits *et al.* based on 16S rRNA sequencing data and due to its ability to hydrolyze gelatin. Before 2001 this organism was known as *Bacillus thermoglucosidasius* [Nazina *et al.* in 2001]. In 2009, Tang *et al.* reported that *G. thermoglucosidans* ferments a range of pentoses and hexoses and that it is tolerant to ethanol concentrations of up to 10 % (v/v), highlighting it as a potentially valuable organism for consolidated bioprocessing for biofuels production. In addition to this desirable phenotype, *G. thermoglucosidans* has also proved to be a "tractable" organism [Taylor *et al.* 2008, Cripps *et al.* 2009] where growth can be done in a standard lab and protocols exist for genetic modification of the cell. In a first step towards establishing *G. thermoglucosidasius* as a cell for biofuel production, the UK company TMO Renewables Ltd developed strains of *G. thermoglucosidasius* NCIMB11955 with modifications to its genome that enable it to perform industrial-scale ethanol production [Cripps *et al.* 2009, Taylor *et al.* 2008]. As the optimal temperature of growth of this organism is 60 °C [Tang *et al.* 2009 and Coorevits *et al.* 2009] it is well suited for industrial fermentation. At this temperature, volatile fermentation by-products such as ethanol are partially evaporated, and their separation from the mixture is significantly easier [Cripps *et al.* 2009].

Currently, research is being carried out by the Leak Group (Bath University) to develop *G. thermoglucosidasius* as an optimal Consolidated Bioprocessing platform for the production of biofuels [Bartosiak-Jentys *et al.* 2013]. A CBP approach would reduce the costs of industrially producing biofuels if based on lignocellulosic feedstocks because the energy input is low. In addition, contamination of the industrial bioreactors by mesophiles, which can potentially compromise the yields and health of the cultures, is no longer a threat if the entire process is carried out at 60 °C [Taylor *et al.* 2009, Cripps *et al.* 2009].

A further advantageous property of *G. thermoglucosidans* as a CBP organism, is that at higher temperatures there is also higher motility between cellulose and hemicellulose, which makes the sugars more easily accessible to the enzymes used for hydrolysis and saccharification [Gilbert and Hazlewood, 1993].

G. thermoglucosidans was selected in this project to be developed as a chassis for synthetic biology because for production processes where the cost of goods is critical (especially for the production of fuels and commodity chemicals) it is justifiable to use a chassis that it is capable of synthesizing necessary macromolecules from an inexpensive carbon source and simple minimal additional nutrients. *G. thermoglucosidans* has already found a niche in industrial chemical/fuel production, where it is a robust production platform [Cripps *et al.* 2009], and it is increasingly a better-characterized system to work with. Using *G. thermoglucosidans* for microbial fuel production from biomass aids the challenge of condensing the energy content of biomass into a high-energy fuel [Fischer *et al.* 2008].

1.3.16. *GEOBACILLUS THERMOGLUCOSIDANS* AS A CHASSIS

The Taylor *et al.* and Cripps *et al.* reports from 2008 and 2009, respectively, are the work that started the development of *G. thermoglucosidans* as a chassis for synthetic biology: Taylor *et al.* created the first shuttle vector that enabled efficient transformation of *G. thermoglucosidans*, while Cripps *et al.* [2009] created the first knock-out vector for this microorganism. However, much more work is necessary to permit the development of *G. thermoglucosidans* into a chassis for advanced molecular engineering, namely, an efficient toolbox of genetic components is necessary.

Furthermore, the use of *G. thermoglucosidans* in industry (TMO Renewables, UK) has so far been limited to ethanol production. An investment into developing this organism as a metabolic engineering chassis would go hand-in-hand with the exploration of all the potential it has to offer for the production of valuable chemicals, in particular biofuels beyond ethanol. With that in mind, this project had the following aims:

The implementation of an organism to make it convert it to a cell factory for any desired chemical requires that [Keasling, 2012]:

- the synthetic constructs can be stably maintained in the organism
- the expression of whatever genes are in the construct do not stress the cells to a point that does not permit their survival.
- the growth should ideally be achievable in a minimal (non-complex) growth medium to minimize production costs and
- that the intermediates or final products are non-toxic at economic production levels so that commercial production is feasible.

Enzymes heterologous to the production host are used to construct engineered metabolic pathways and remove undesired regulatory checkpoints in natural pathways by replacement with synthetic elements. The right bioparts could minimize metabolite accumulation and reduce potential toxicity in heterologous pathways in chassis not adapted to cope with high concentrations of these compounds [Dueber *et al.* 2009].

Bioparts and tools are necessary to control gene expression because it may not always be ideal to maximize production of a target protein, but rather, the enzymes need to be produced in amounts sufficient to transform at sufficient rates so as not to rob the cell of metabolites that might compromise growth (and thus yield) or which might be too toxic [Dueber *et al.* 2009].

1.3.17. A SYNTHETIC BIOLOGY TOOLBOX: CHARACTERIZATION OF BIOPARTS

One of the main issues that synthetic biology has encountered is the challenging nature of rationally designing devices, circuits and systems because they behave in a context-dependent manner, *i.e.* their properties depend on the combination of other elements used in the device, circuit or system and can also be affected by external factors like growth conditions [Galanie *et al.* 2013]. For instance, Gorochowski *et al.* [2013] found striking changes regarding transcription and translation rates in *E. coli* in response to temperature and media composition. One example of a content-dependent effect is that the RBS interacts with the promoter and the gene of interest and may lead to the formation of secondary structures, which in turn may strongly affect expression levels [Chen 2015]. If different combinations of parts have different activities in response to various environmental and physiological conditions, then it is important to characterize the performance of each biopart in a manner that can be directly comparable [Heinemann 2006, Kosuri 2013]. For this purpose, the engineering principles of standardization and modularization must be applied for the construction of synthetic devices (see Figure 6).

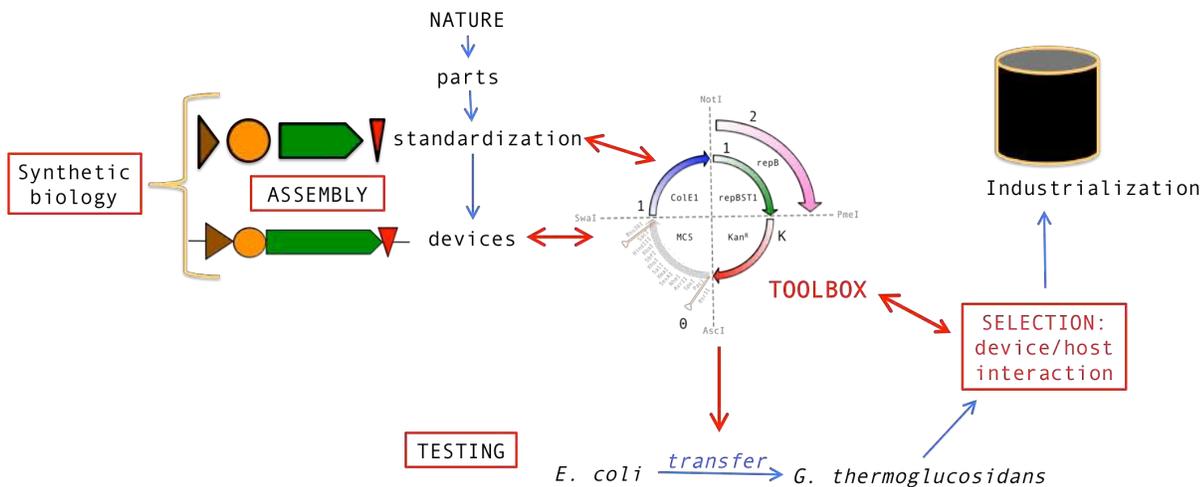


Figure 6. The role of synthetic biology. Synthetic biology participates in many processes ranging from the identification of parts from nature, to the standardization and assembly of devices. A genetic toolbox is a central part of synthetic biology and it is required for the assembly, testing and selection stages that are vital to construct a process that can be industrialized.

1.3.18. STANDARDIZATION AND MODULARIZATION OF BIOPARTS

The core engineering principles adopted in synthetic biology, according to Kelwick *et al.* [2014], Lee *et al.* [2011], Endy *et al.* [2005], Heinemann [2006] are:

1. Standardization: libraries of standard interchangeable parts are one hallmark of modern technology [Canton *et al.* 2008].
2. Modularization can help establish platforms for the sharing and reuse of bioparts, it is required to create a "plug in and play" compatibility.
3. Abstraction: Engineering principles enable the separation of complexity. It has been explained previously how bioparts are discrete biological sequences of known behavior that have a standard descriptive function.

To enable synthetic biology to improve the reliability of parts to allow more predictable behaviors when constructing larger systems, significant efforts so far have revolved around standardization [Kosuri *et al.* 2013]. The importance of "plug-and-play" compatibility in synthetic biology is clear in order to obtain directly comparable characterization results [Nickel *et al.* 2014]. For instance, when determining the activity of a library of promoters, in order to obtain results that are directly comparable, special attention must be paid to ensuring that absolutely all factors within the design as well as the conditions remain the same except for the promoter library. Likewise, to determine whether an RBS is the cause for poor expression for example, all factors must remain the same except for the RBS, or if there are optimal media/temperature/growth conditions must be sought for the production of a particular protein or metabolite, the DNA parts must remain exactly the same and the only thing that can change are the conditions. The use of well-characterized regulatory elements that behave in predictable ways when combined will enable to achieve predictable levels of expression of heterologous proteins when engineering biological systems [Kosuri *et al.* 2013]. Prediction and standardization of parts is necessary because it is economically not feasible to synthesize new constructs for every experiment, and also with the advent of standardization and modularity, it means there is no need and parts can be replaced simply and efficiently [Lee *et al.* 2011, Martínez-García *et al.* 2014].

If the goal of synthetic biology is to gain a better understanding of how to coordinate and regulate gene expression in organisms [Redden *et al.* 2015], then powerful standardized tools are needed for all steps along the design-build-test-learn cycle [Breitling *et al.* 2015].

1.3.19. REPORTER PROTEINS

A reporter is ideally a non-native protein that is synthesized by the cell in response to a predetermined external stimulus, and it can be measured for activity or amount non-invasively [Roelof and Belkin, 2010]. An ideal reporter should be easy to detect and can be assayed in a quantitative way, so that increases in expression by 5 or 10% can be discerned above noise and background. When a reporter CDS is placed downstream from a promoter and RBS, the output is used to characterize the expression level determined by the promoter and RBS. By then changing the promoter and RBS sequences individually and measuring the reporter levels subsequently, the contribution each part makes to the total gene expression output can be determined [Salis, 2011].

When these parts are put together, they create a device that has a specific, user-defined function. Such specific function depends both on the function of the individual parts and of the coordination between them. The construction of synthetic pathways, *i.e.* a series of sequential reactions catalyzed by enzymes [Shin *et al.* 2013], requires a good understanding of the behaviour of each of the parts composing a construct. The most important component that will drive the expression of any gene is actually the combined effect of the promoter plus the RBS sequence. There are enough idiosyncratic interactions and context-dependent effects that it is necessary to construct several variants of a circuit or device to obtain one with the desired functions [Kosuri, 2013].

Kosuri *et al.* [2013] characterized a vast number of combinations of promoter and RBS sequences and they characterized transcription and translation rates of each. They argue that constructs need to be synthesized and tested instead of relying absolutely on predictable models in order to characterize context dependent interactions. The large library they created however, used GFP and mCherry as outputs. So changing the CDS would still result in context dependent idiosyncratic results that can only be known once it is tested, and the construction of such does need to rely on predictable models such as the Salis RBS

Calculator because it is impractical to build entire libraries of promoters and RBSs every time a CDS is changed.

Placing a synthetic device in a multi-copy plasmid increases the level of gene expression, but it also implies that the cellular machinery has to be shared between the plasmid and the rest of the cellular functions required for survival and multiplication. Gene expression is also influenced by the balance achieved between transcription and translation. Strong overexpression causes competition for ribosomes and may lead to cell death [Dong *et al.* 1995].

A key factor for any genetic manipulations are the vectors used to insert DNA into a host organism. Ideally, a vector should have a marker that ensures that all the cells in the culture produce the vector, it also has to be minimal and of consistent copy number, it has to have the ability to carry large fragments of DNA. Low copy plasmids offer a benefit over high copy plasmids when the cells struggle to express the contents of the vector when the copies are too many or when there is imbalance in transcription/translation, for example, when too many mRNAs are created, but the ribosomes are not transcribing them all, thus wasting the cell's energy. Also, plasmids that permit the integration of foreign DNA into the chromosome allow the replacement of native DNA regions such as regulatory regions and thus alter gene expression [Keasling, 2008].

1.3.20. PLASMIDS/SHUTTLE VECTORS

One crucial aim for synthetic biology is the expression of heterologous genes. There are two options for this purpose:

1. Plasmid-bound: requires plasmids
2. Genome-bound: requires integration vectors

Vectors are essential tool for any genetic manipulation [Keasling, 2012]. They are required to carry and harbor the transforming DNA in the host. According to Redden *et al.* [2015] plasmids have three attributes that can affect gene expression:

1. The origin of replication
2. The selection marker
3. The Multiple Cloning Site (MCS)

1.3.21. GUIDELINE: SHUTTLE VECTORS FOR *CLOSTRIDIUM* SPP.

Clostridium is a relevant genus for the production of useful chemicals from renewable energy sources: *C. acetobutylicum* and *C. beijerinckii* are saccharolytic species capable of producing biobutanol. *C. thermocellum*, *C. cellulolyticum* and *C. phytofermentans* are cellulolytic ethanol producing strains, *C. noyVII* and *C. sporogenes* are being pursued as possible delivery vehicles for tumour therapy as clostridial spores selectively germinate in tumours [Minton 2003].

To increase the already high profile of the genus, Heap *et al.* constructed in 2009 a set of modular shuttle vectors for *Clostridium* spp. and *E. coli* in which each of the components was delimited by a rarely occurring enzyme to facilitate combinatorial designs. The *Clostridium/E. coli* shuttle vectors were composed of the standard elements that comprise plasmids, namely origins of replication for both: *Clostridium* spp. and *E. coli* selectable markers as well as a conjugal transfer function. Considering that the *Clostridium/E. coli* shuttle vector system addresses needs similar to what we need to address for the construction of a shuttle vector system for *G. thermoglucosidans/E. coli*, it has been decided for this project to use the approach of Heap *et al.* as an example to guide the design of the much needed *Geobacillus* set of modular vectors.

More specific information regarding the bioparts and tools such as shuttle vectors that currently exist for *G. thermoglucosidans* have been included in the introduction sections of Chapters 3, 4 and 5 of the present work.

1.4. HYPOTHESIS:

To develop *Geobacillus thermoglucosidans* as a thermophile chassis for synthetic biology, it must have a functioning set of genetic tools that allow users to apply synthetic biology principles, specifically:

1. If shuttle vectors constructed based on modular design of abstract functional parts yield improved transformation efficiencies than the reference plasmids, then synthetic biology principles were correctly applied and can continue being applied in further work that involves the genetic modification of this organism.

2. If bioparts such as promoters, RBS sequences, reporters and transcriptional terminators can be efficiently included in the toolbox and show intended activity, then the toolbox can be used for the future characterization of bioparts and their use in systems for metabolic engineering.
3. If the toolbox can be used for the heterologous expression of a protein that results in the production of a useful compound, then the toolbox can be used in future efforts to develop *G. thermoglucosidans* as a platform for the production of valuable chemicals and/or proteins.

1.5. AIMS OF THE PROJECT

The first aim was to contribute to developing *G. thermoglucosidans* as a chassis for synthetic biology by creating a toolbox of genetic components that will enable efficient engineering of the cell. In particular, reliable plasmids are required, as are promoters for overexpression and reporter genes for quantifying gene expression.

The second aim was to make use of the toolbox to produce a strain of *G. thermoglucosidans* that is capable of producing an advanced biofuel – and in particular isobutanol, which is potentially a drop-in fuel for current petrol engines.

CHAPTER 2: MATERIALS AND METHODS

All reagent-grade chemicals were acquired from Sigma Aldrich (Dorset, UK) or Fisher Scientific (Leicestershire, UK). Yeast extract, tryptone and soy peptone were purchased from Merck (Darmstadt, Germany).

2.1. BACTERIAL STRAINS AND MEDIA

All bacterial strains used in this work are listed in **Table 4**. Cultures were grown in a MaxQ 6000 (Thermo Scientific) incubator shaker. Overnight cultures were inoculated from a frozen stock. *E. coli* was grown at either 30 or 37 °C, and *G. thermoglucosidasius* was routinely grown at 55 °C.

| Strain | Description | Source/reference |
|---------------------------------|---|---------------------------------|
| <i>E. coli</i> JM109 | <i>endA1, recA1, gyrA96, thi, hsdR17</i> (rk ⁻ , mk ⁺), <i>relA1, supE44, Δ(lac-proAB)</i> , [F' <i>traD36, proAB, lacIqZΔM15</i>] | Promega (Southampton, UK) |
| <i>E. coli</i> DH5α | F ⁻ , <i>endA1, glnV44, thi-1, recA1, relA1, gyrA96, deoR, Φ80dlacZΔM15, Δ(lacZYA-argF)U169, hsdR17</i> (rk ⁻ mk ⁺), λ ⁻ | Life Technologies (Paisley, UK) |
| <i>E. coli</i> DH10B | <i>Δ(ara,leu)7697, araD139, tonA, ΔlacX74, DH10Bdup, galK16, galE15,d eoR, 14, mcrA, galU, Φ80dlacZΔM15, recA1, relA1, endA1, Tn10.10, nupG, rpsL, rph, spoT1, Δ(mrr-hsd-RMS-mcrBC)</i> [Durfee <i>et al.</i> 2008] | Life Technologies (Paisley, UK) |
| <i>E. coli</i> BioBLUE | <i>recA1, endA1, gyrA96, thi-1, hsdR17</i> (rk ⁻ mk ⁺), <i>supE44, relA1, lac</i> [F' <i>proAB lacIqZΔM15 Tn10(Tetr)</i>] | Bioline (London, UK) |
| <i>G.thermoglucosidans</i> DL33 | Wild-type isolate | TMO-Renewables (Surrey, UK) |
| <i>G.thermoglucosidans</i> DL44 | <i>ldh</i> ⁻ knockout of DL33 | Taylor, 2008 |

2.2. BACTERIAL GROWTH CONDITIONS

All growth media used in this study for both *E. coli* and *G. thermoglucosidans* are listed in **Table 5**.

| Table 5. Media used to grow <i>E. coli</i> and <i>G. thermoglucosidans</i> | | |
|---|--|--|
| Medium | Components and specifications | Organisms cultured |
| LB | Tryptone 10 g/L, Yeast extract 5 g/L, NaCl 5 g/L. Adjusted to pH7 with HCl or NaOH | <i>E. coli</i> <i>B. subtilis</i> 168 |
| 2TY | Tryptone 16 g/L, Yeast extract 10 g/L, NaCl 5 g/L. Adjusted to pH7 with HCl or NaOH | <i>G. thermoglucosidans</i> |
| TGP | Tryptone 17 g/L, Soy peptone 3 g/L, NaCl 5 g/L, K ₂ HPO ₄ 2.5 g/L. Post-autoclave addition of glycerol 4 mL/L and C ₃ H ₃ NaO ₃ 4 g/L. Adjusted to pH7 with HCl or NaOH | <i>G. thermoglucosidans</i> |
| BCM | Tryptone 17 g/L, Soy peptone 3 g/L, Yeast extract 5 g/L, NaCl 5 g/L, K ₂ HPO ₄ 2.5 g/L. Post-autoclave addition of Glucose 10 g/L, HEPES 2 mM. Adjusted to pH7 with HCl or NaOH | <i>G. thermoglucosidans</i> |
| ASM | NaH ₂ PO ₄ 20 mM, K ₂ SO ₄ 10 mM, Citric acid 8 mM, MgSO ₄ 5 mM, CaCl ₂ 0.08 mM, Na ₂ MoO ₄ 1.65 μM, (NH ₄) ₂ SO ₄ 25 mM, Tryptone 0.1%, Sulphate Trace Elements Solution (TE) 5 mL/L; where TE: H ₂ SO ₄ ·7H ₂ O 5 mL/L, ZnSO ₄ ·7H ₂ O 1.44 g/L, FeSO ₄ ·7H ₂ O 5.56 g/L, MnSO ₄ ·H ₂ O 1.69 g/L, CuSO ₄ ·5H ₂ O 0.25 g/L, CoSO ₄ ·7H ₂ O 0.56 g/L, NiSO ₄ ·6H ₂ O 0.89 g/L, H ₃ BO ₃ 0.08 g/L. Post autoclave additions of: Tryptone 1%, Glucose 0.1M Adjusted to pH7 with NaOH | <i>G. thermoglucosidans</i> |

2.3. FROZEN GLYCEROL STOCKS

Glycerol stocks of the strains used in this study were prepared by mixing 750 μL of 50% glycerol with 750 μL of culture in a cryogenic vial (Thermo Scientific). Stocks were kept at $-80\text{ }^{\circ}\text{C}$.

2.4. STERILIZATION

All media, instruments and containers were sterilized before use by autoclaving $121\text{ }^{\circ}\text{C}/103\text{ MPa}$ 15 minutes. Heat-sensitive ingredients were dissolved and filtered through a Minisart 0.2 μm syringe filter (Sartorius, Goettigen, Germany). Gamma-irradiated disposable spreaders and loops were used.

2.5. ANTIBIOTIC SELECTION

For selection based on antibiotic resistance in *E. coli*, ampicillin was used at a concentration of 100 $\mu\text{g}/\text{mL}$ and kanamycin used at a final concentration of 50 or 100 $\mu\text{g}/\text{mL}$. For selection in *G. thermoglucosidans*, a final concentration of 25 $\mu\text{g}/\text{mL}$ kanamycin was used. Chloramphenicol was used at a final concentration of 12 $\mu\text{g}/\text{mL}$ in *G. thermoglucosidasius* and 6 $\mu\text{g}/\text{mL}$ in *E. coli*. Ampicillin and chloramphenicol were diluted in 100% ethanol, whereas kanamycin was diluted in water.

2.6. PLASMID PURIFICATION

Plasmids were purified from *E. coli* and *G. thermoglucosidasius* using the QIAprep Spin Miniprep Kit (QIAGEN, Crawley, UK) following the manufacturer's protocol. In the case of *G. thermoglucosidasius*, lysozyme was added to P1 buffer at a final concentration of 10 mg/mL and the cells were mixed with P1/lysozyme and incubated at $37\text{ }^{\circ}\text{C}$ for 48 hours. DNA was quantified using a NanoDrop 1000 (Thermo Scientific). Sequencing was carried out by Source Bioscience.

2.7. POLYMERASE CHAIN REACTION (PCR)

Oligonucleotides were purchased from Invitrogen (Life Technologies, Paisley, UK) and Integrated DNA Technologies (Leuven, Belgium). PCR reactions were carried out in a G Storm thermal cycler (Gene Technologies Ltd, Essex, UK) or in a MasterCycler Gradient (Eppendorf, Hamburg, Germany). Phusion Polymerase (Thermo Scientific, MA, USA) and Q5 HF polymerase (New England Biolabs, Hertfordshire, UK) were used according to manufacturer's instructions. The annealing temperatures were calculated using the NEB T_m determination website [<http://tmcalculator.neb.com/#/>] and an elongation time of 30 sec per 1000 bp was used. For analytical colony PCRs, ReadyMix REDtaq PCR mix (Sigma-Aldrich) was used.

PCR amplified DNA was electrophoresed through a 0.8 or 1 % agarose gel in TAE Buffer (4.85 g Tris, 1.14 g acetic acid and 2 mL 0.5M sodium EDTA per litre) and, when required, gel-extracted and purified using the QiaQuick gel extraction kit (QIAGEN, Crawley, UK). The sizes of the fragments were compared against the Fermentas 1 kb DNA ladder (Figure 7).

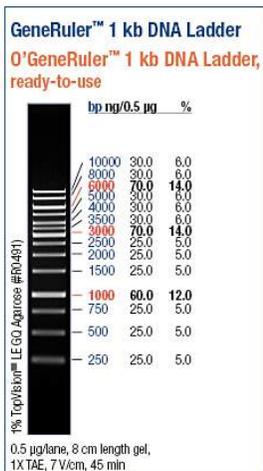


Figure 7. 1 kb DNA ladder used as the molecular weight marker (Thermo Scientific)

2.8. GIBSON ONE-STEP ISOTHERMAL DNA ASSEMBLY

A 5X isothermal reaction buffer was prepared according to the protocol of Gibson *et al.* [2009], containing 3 mL of 1M Tris-HCl pH7.5, 150 μ L of 2M MgCl₂, 60 μ L of 100 mM of each of the four dNTPs, 300 μ L 1M DTT, 1.5 g PEG-8000 and 300 μ L 100 mM NAD and molecular grade water to a final volume of 6 mL. The assembly mastermix contained 320 μ L of the 5X isothermal reaction buffer plus 0.64 μ L of 10 U/ μ L T5 exonuclease, 20 μ L of 2 U/ μ L Phusion DNA polymerase, 160 μ L of 40 U/ μ L Taq ligase and water to a final volume of 1.2 mL, which was then divided into 15 μ L aliquots. For every Gibson reaction, one aliquots was mixed with 0.15 pmoles of DNA of each part. 1 μ L of the Gibson reaction was used to transform BioBLUE chemically competent cells (Bioline, London, UK).

2.9. RESTRICTION DIGESTS AND LIGATIONS

Restriction enzymes were purchased from New England Biolabs. Digests were carried out following the manufacturer's protocol at 37 °C overnight. DNA was electrophoresed through a 1 % agarose gel and, when required, gel-extracted and purified. The ligation reactions were carried out by mixing 5 μ L of T4 ligation buffer (Fermentas) with 1 μ L of 1000 CEU/ μ L T4 DNA ligase and incubating at 10 °C overnight following a denaturation step at 65°C for 20 minutes and desalting using 0.25 μ m nitrocellulose filters. 5 μ L of the desalted ligation were routinely used to transform 100 μ L of chemically competent DH10B.

2.10. PREPARATION OF CHEMICALLY COMPETENT *E. COLI*

Chemically competent *E. coli* JM109 or DH10B cells were prepared after a frozen stock was used to inoculate 5 mL of LB media. The culture was grown overnight at 37 °C with shaking at 250 rpm. 500 μ L of this culture was used to inoculate 50 mL of LB media in a 250 mL baffled conical flask, which was subsequently incubated at 37 °C in a rotary shaker for 2-3 h until it reached an OD₆₀₀ of 0.5. The culture was cooled on ice for 10 min and the cells were harvested by centrifugation (4°C, 4000 rpm, 5 min). The cells were resuspended in 5 mL of chilled CCMB80 buffer (80mM Calcium Chloride, 20mM Manganese Chloride, 10mM

Magnesium Chloride, 10mM Potassium Acetate, 10% Glycerol) and divided into 200 μ L aliquots that were stored at -80 °C.

2.11. TRANSFORMATION OF CHEMICALLY COMPETENT *E. COLI*

100-300 ng of purified plasmid DNA was mixed with a 100 μ L aliquot of chemically competent *E. coli* cells. The vial was allowed to rest on ice for 30 min, after which the cells were incubated at 42 °C for 30 seconds and then placed back on ice for 2 minutes. Following this, the cells were diluted with 320 μ L of SOC media (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl₂, 10 mM MgSO₄, and 20 mM glucose) and allowed to recover for 1 hour at 37 °C, shaking, after which they were plated on an LB agar plate containing the appropriate antibiotic.

2.12. PREPARATION OF ELECTROCOMPETENT *G. THERMOGLUCOSIDANS*

Electrocompetent *G. thermoglucosidans* cells were prepared according to the protocol of Taylor *et al.* [2008]: A sample from a frozen stock of *G. thermoglucosidans* 11955, DL33 or DL44 was used to inoculate 5 mL of pre-warmed 2TY or TGP and grown overnight at 55 °C in a rotary shaker at 250 rpm. 500 μ L of this culture were used to inoculate 50 mL of pre-warmed 2TY or TGP in a 250 mL baffled conical flask, which was subsequently incubated at 55 °C with shaking for 5-6 hours until it reached an OD₆₀₀ of 1.4. The culture was allowed to cool on ice for 10 minutes and then the cells were harvested by centrifugation (4 °C, 4000 rpm, 10 min). The cells were washed three times with 10 mL of electroporation buffer (sorbitol 0.5 M, mannitol 0.5 M and glycerol 10% v/v). The final cell pellet was resuspended in 800 μ L of electroporation buffer and divided into 300 μ L aliquots that were immediately stored at -80 °C.

2.13. TRANSFORMATION OF *G. THERMOGLUCOSIDANS*

Transformation of *Geobacillus* was done following the protocol in Taylor *et al.* [2008]: 100-300 ng of plasmid DNA were mixed with 140 μ L of electrocompetent cells and transferred to an electroporation cuvette. A Bio-Rad gene pulser electroporator was used to

pulse cells at 2500V, 10 μ F capacitance and 600 Ω resistance. Time constants ranged from 4-6 ms. Immediately after the pulse, 320 μ L of pre-warmed TGP were added to the cells and they were allowed to recover for 1 h at 55 °C with shaking at 250 rpm. Transformants were selected on TGP or 2TY plates containing kanamycin at 55 °C.

2.14. GROWING *G. THERMOGLUCOSIDANS* ANAEROBICALLY AND MICROAEROBICALLY

Anaerobic growth was achieved by inoculating a single colony grown overnight on a TGP or 2TY/kanamycin plate into 1.5 mL of ASM media under sterile conditions. After vortexing, 1 mL of this culture was injected through the suba-seal to 15 mL of ASM in a sterile Hungate tube. The culture was then sparged by injecting filter sterilized oxygen free nitrogen through a needle that pierced the suba-seal for 5-10 minutes while allowing oxygen to be displaced and removed from the vial by a second unconnected needle that was also placed piercing the suba-seal. A test for oxygen concentration after sparging was not carried out.

Microaerobic growth was achieved by inoculating a single colony grown overnight on a TGP or 2TY/kanamycin plate to a 50 mL Falcon tube that was filled with 45 mL of BCM media. The cap was tightly screwed and the Falcon tube was placed in a rotary shaker horizontally at 55 °C overnight. This allowed initial aerobic growth until the oxygen was consumed but a test for oxygen concentration after sparging was not carried out.

2.15. OPTICAL DENSITY ANALYSIS

Optical density was routinely measured on a Jasco Genova Jenway spectrophotometer. Optical density at 600 nm (OD_{600}) was used as an indicator of the biomass in the culture. Cultures were diluted 1:10 and sterile media was used as a blank to obtain an accurate reading.

2.16. LYSING

Cells were routinely lysed by using a Sonics Vibracell (Sonics and Materials Inc., Newtown, CT, USA) sonicator. Six fifteen-second bursts of sonication were interspersed between six fifteen-seconds without sonication (amplitude 50 %). The soluble fraction of the cell extract was obtained by centrifugation at 14 000 rpm for 3 minutes.

2.17. SDS-PAGE

A 5 mL resolving 12% polyacrylamide gel consisted of 1.6 mL water, 2 mL acrylamide mix (30 % acrylamide and 0.8% bis-acrylamide), 1.3 mL Tris-HCl 1.5 M pH 6.8, 100 μ L 10% SDS, 30 μ L 10% Ammonium persulphate, 4 μ L TEMED. The stacking gel consisted of 680 μ L water, 170 μ L acrylamide mix, 130 μ L Tris-HCl 1 M pH 8.8, 10 μ L 10% SDS, 10 μ L 10% ammonium persulphate, 1 μ L TEMED.

40 μ L of cell lysate were mixed with 5X SDS Laemmli buffer (60 mM TrisHCl pH 6.8, 2% SDS, 10% glycerol, 5% β -mercaptoethanol, 0.01% bromophenol blue). The mix was incubated at 95 °C for 5 minutes and loaded onto the gel together with the Prestained Molecular Weight Marker (Thermo Scientific), shown in **Figure 8**.

Gels were run in a Mini-PROTEAN Tetra Cell (BioRad, Hampstead, UK), at 100 V for one hour in 1X SDS-Running buffer; where 50X SDS Running buffer: 30 g/L Tris, 144 g/L glycine and 10 g/L SDS.

SDS-PAGE gels were incubated with Coomassie protein staining dye (0.2% Coomassie blue R-250, 40% ethanol, 10% acetic acid) for one hour. The gel was then transferred to destaining solution (40 % ethanol, 10% acetic acid) overnight.

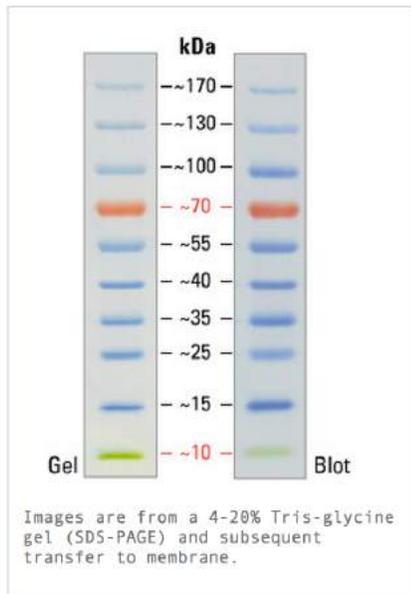


Figure 8. Prestained molecular protein marker included in SDS-PAGE gels and Western blots (Thermo Scientific)³.

2.18. WESTERN BLOTTING

An SDS-PAGE gel was blotted to a membrane optimized for protein transfer (Amersham Hybond-LFP, GE Healthcare, UK) using a Transfer Blot SD semi-dry Transfer Cell (BioRad, CA, USA). The membrane was activated by suspension in methanol and afterwards, soaked in Transfer Buffer (48 mM Tris, 39 mM glycine, 0.04% SDS and 20% methanol). Proteins were transferred onto the membrane for 30 minutes at 25 V. The membrane was subsequently soaked in Blocking Solution and the Novex Western Breeze Immunodetection kit (Life Technologies) was used according to manufacturer's instructions. The membrane was incubated with primary 6X His Anti-Mouse antibody (Invitrogen) overnight, followed by two washes in Antibody Wash solution. The membrane was then incubated with a secondary anti-mouse antibody for an hour. All incubations were carried out with shaking at 60 rpm. The membrane was subsequently washed three times with water and incubated with Novex Chromogenic dye and the bands were allowed to reveal for 20 minutes, after which the membrane was washed with water and imaged.

³ Figure taken from [<http://www.thermoscientificbio.com/protein-electrophoresis/pageruler-prestained-protein-ladder/>]

2.19. TOTAL RNA EXTRACTION

The RNeasy Protect Bacteria Mini Kit (QIAGEN) was used in conjunction with the RNase free DNase kit (QIAGEN) to extract total RNA from *G. thermoglucosidans*. RNase-free tips, microcentrifuge tubes and water were used throughout. For each isolation, 1 mL of a culture in exponential phase (OD_{600} 0.5-0.8) was used. 10 mg/mL lysozyme were added to TE buffer and Proteinase K was added to inactivate nucleases according to the manufacturer's instructions. RNase-free DnaseI (Fermentas, York, UK) was used to remove DNA from the extractions.

2.20. REVERSE TRANSCRIPTION PCR (RT-PCR)

The purified and DNase I-treated total RNA was converted to cDNA using the High Capacity cDNA Reverse Transcription kit (Applied Biosystems, Warrington, UK) following the manufacturer's instructions. cDNA was subsequently used as a template for PCR.

2.21. HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY ANALYSIS (HPLC)

The extracellular presence in the supernatant of the growth medium ASM of isobutanol, isobutyraldehyde, 2-ketoisovalerate as well as glucose, acetate, lactate, ethanol, formate, pyruvate, citrate and succinate in the supernatant of cultures of *G. thermoglucosidans* grown in ASM were analyzed using high performance liquid chromatography (HPLC). Overnight cultures were harvested by centrifugation and the supernatant was filtered through a Minisart 0.2 μ m syringe filter (Sartorius) to remove physical particles and cell debris. 1 mL of the filtered supernatant was transferred to a chromatography glass vial and sealed with septum and cap (Thermo Fisher Scientific, Chromacol, Hertfordshire, UK).

Samples were analyzed on a Agilent 1200 series HPLC system which consisted of the following components: a PU2080 plus pump, AS-2051 plus intelligent sampler, a UV075 plus UV/Vis spectrophotometer measuring absorbance at a wavelength of 210 nm, a RI-2031 plus refractive index (RI) detector. The column is Phenomenex Rezex RHM, Monosaccharide

H+ 300 x 7.8. The temperature of the column was set to 65 °C. A 10 µl volume of the sample was injected into the HPLC system and carried by a mobile phase of 10 mM sulphuric acid at a flow rate of 0.6 ml/min. Compounds were separated over an elution time of 35 minutes at 60 °C. Detection was done by Refractive Index (RI) only.

2.22. GAS-CHROMATOGRAPHY AND MASS SPECTROMETRY (GC-MS)

Gas Chromatography and Mass Spectrometry (GC-MS) was used to detect the presence of isobutanol in *G. thermoglucosidans* cultures grown overnight in ASM. Overnight cultures were harvested by centrifugation and the supernatant was filtered through a Minisart 0.2 µm syringe filter (Sartorius) to remove physical particles and cell debris. 1 mL of the filtered supernatant was mixed with 1 mL of toluene and shaken for 30 minutes, after which, the mixture was centrifuged and the upper organic layer was transferred to a chromatography glass vial sealed with septum and cap (Thermo Fisher Scientific). The sample was analyzed on an Agilent (7890B GC with a 5977 MSD) GC-MS system. The boiling point column used was a DB-FFAP (30m x 250µm x 0.25µm). The method was as follows: Isothermal at 40°C for 1 min, ramp to 200°C with the rate of 10°C/min, isothermal at 200°C for 5 minutes.

2.23. MICROSCOPY

Cells were diluted in water and a 5 µL sample was added to a microscopy slide. The microscope used was the Nikon Eclipse Ti inverted microscope with the 60x CPI60 objective. The excitation, emission and exposure times for each acquisition were for the GFP channel 480nm excitation and 535nm emission while for the Cy3 (Red) channel 532nm and 590nm respectively. **The images were viewed using the software** NIS-Elements Microscope Imaging Software (Nikon).

2.24. FLOW CYTOMETRY

1 μ L of a culture that reached stationary phase of *E. coli* and *G. thermoglucosidans* were diluted in 1 mL of water and analyzed on a modified Becton-Dickinson FACScan flow cytometer. Samples were run on high flow rate until 30000 events were observed or 30 seconds had elapsed. The gate setting for *G. thermoglucosidans* applied were the *E. coli* parameters. Flow cytometry settings used were all taken on the log scale; FSC sensor E01, SSC voltage 350, FL1 voltage 700, threshold 52 (SSC). Data was analyzed using FlowJo with a tight forward scatter/side scatter gate being applied to select for a homogeneous population size.

3. CHAPTER 3: PROMOTERS AND REPORTER PROTEINS FOR *GEOBACILLUS* *THERMOGLUCOSIDANS*

The key contributions and results of this chapter are:

1. It has been demonstrated that sfGFP is an ideal fluorescent reporter for *G. thermoglucosidans* when the growth of cells occurs under aerobic conditions because it is thermostable at high temperatures.
2. It has been demonstrated that mCherry could be used as an alternative to sfGFP as a fluorescent reporter under aerobic conditions but mCherry is not as thermostable as sfGFP.
3. It has been demonstrated that BsFbFP could be used as a fluorescent reporter under anaerobic conditions as long as the fluorescence is measured with flow cytometry using the appropriate excitation wavelength.

3.1. INTRODUCTION

To be able to use *G. thermoglucosidans* as the thermophilic platform of production, it is necessary to create and characterize the bioparts required for transcriptional and translational regulation. This chapter in particular focuses on new reporter proteins that can be paired up with existing *G. thermoglucosidans* promoters.

The aim of the toolbox is to create an expression vector that permits to carry out initial screenings of promoter activity and therefore protein expression in *E. coli* and subsequently, to characterize the strength of promoter activity in *G. thermoglucosidans*. In the present chapter we describe the most relevant promoters that have been previously used for transcriptional control in *G. thermoglucosidans* as well as the non-fluorescent reporter proteins used to characterize them. In addition, we introduce potential fluorescent proteins that would be ideal reporters for *Geobacillus thermoglucosidans*.

3.1.1. PROMOTERS

Promoters are important for this work because:

- They are crucial bioparts required for the construction of synthetic biology devices.
- Promoters that work in *G. thermoglucosidans* are necessary to express proteins in *G. thermoglucosidans*.
- Ideal promoters for *G. thermoglucosidans* should also be active in both *E. coli* and *G. thermoglucosidans* to facilitate the preliminary screening of colonies in *E. coli*.

Compared to the model organisms used widely in synthetic biology and metabolic engineering (e.g. *E. coli* and *S. cerevisiae*), there are very few characterized promoters for *Geobacillus* spp. and they are: p43, *ldh*, pUP and RpIS, all introduced below.

The p43 promoter was obtained from the genome of *B. subtilis* 168 by Wang and Doi [1984] where it is known to result in strong gene expression in *B. subtilis* and *E. coli* [Wang and Doi 1984, Zhang *et al.* 2007]. N. Crowhurst [2010] confirmed that the p43 promoter is active in *G. thermodenitrificans* under aerobic and microanaerobic conditions.

The *ldh* promoter was developed as a promoter for expression in *Geobacillus* spp. by Bartosiak-Jentys [2012]. It was obtained from the genomic *ldh* gene sequence of *G. stearothermophilus* NCA1503 and in its natural setting it controls the expression of lactate dehydrogenase (Ldh), responsible for the conversion of pyruvate to lactate [Taylor, 2007].

pUP is a constitutive promoter whose sequence is taken from the genomic uracil phosphoribosyltransferase gene of *G. thermoglucosidans* TM89. The RpIS promoter is also a constitutive promoter from the *Geobacillus* genome and in its natural setting it transcribes the mRNA for the 50S ribosomal protein L19. Both of these promoters were isolated by Benjamin Reeve at Imperial College London [publication pending]. Unlike the *ldh* promoter, both of these promoters are active both in *E. coli* and *G. thermoglucosidans*.

It was decided that the expression of novel reporter proteins should be directed by the strongest promoters available for *G. thermoglucosidans*, namely: the *ldh*, p43 and RpIS promoters because the work was aimed at testing whether the novel reporter proteins are (1)

expressed by *G. thermoglucosidans* in good yields to allow detection and (2) whether the proteins are thermostable when expressed at 55 °C given that denaturation by heat is a common problem of mesophilic proteins.

3.1.2. REPORTERS

This work focuses on the search for novel fluorescent reporter proteins for *G. thermoglucosidans* because:

- Fluorescent reporters are easier to measure than colorimetric reporters using flow cytometry.
- Flow cytometry is a more effective method for measuring reporter expression because it is more accurate than spectrophotometer readings.

Reporter proteins are used to characterize promoter activity, particularly to measure promoters and RBS interactions in a variety of culture conditions. Reporter proteins are also often used as tags for protein localization studies. An ideal reporter protein should be easy to detect and assay in a way that allows reliable quantification that can be repeatable. The most common reporter proteins for use in mesophiles like *E. coli* are the green fluorescent protein (GFP), red fluorescent protein (RFP), β -galactosidase (LacZ) and variants of luciferases (e.g. LuxAB). While reporter systems for Gram-positive thermophiles like *G. thermoglucosidans*, were until very recently [Bartosiak-Jentys, 2012] [Blanchard, 2014], unavailable. This lack of reporter genes meant that promoters for *Geobacillus* spp. could not be characterized and progress in developing *Geobacillus* as a synthetic biology chassis was severely hampered.

3.1.2.1. THE PHEB ASSAY FOR *GEOBACILLUS* SPP.

So far the only reporter proteins that have been used in *Geobacillus* spp. are pheB and lacZ, but both of which require a colorimetric assay to characterize promoters, which is a tedious assay. Bartosiak-Jentys developed a reporter system [2009, 2012] based on the *xyIE* homologue *pheB* isolated from *G. stearothermophilus*. The pheB assay is a colorimetric assay: when 100 mM catechol is added to cells expressing *xyIE*, they change colour and become bright yellow because the enzyme catechol 2,3 dioxygenase cleaves catechol to 2-

hydroxymuconic semialdehyde, which is a yellow compound. The colour change can be measured by a spectrophotometer at 375 nm within minutes of the assay being initiated. Colonies or cell extract containing pheB can be identified by their yellow colouration after being sprayed with 100 mM catechol.

In order to express pheB in *G. thermoglucosidans*, Bartosiak-Jentys [2009] developed a plasmid termed "pGR002", which was composed of the pUCG18 backbone and within the plasmid backbone, as seen in **Figure 9**, placed between the *PstI* and *XbaI* restriction enzyme recognition sites, was included the NCA1503 *ldh* promoter between the *PstI* and *XbaI* sites of the multiple cloning site (MCS) and the reporter pheB, which encodes the catechol 2,3 dioxygenase from *G. stearothermophilus* (NCBI accession number DQ 146476.2) between *XbaI* and *SacI*. The RBS sequence that was placed downstream from the *ldh* promoter and used to express pheB was the one natively found upstream from the start codon of the *pheB* gene in *G. stearothermophilus*. Hereinafter, it will be referred to as "pheB-RBS". The sequence of pGR002, including that of pheB-RBS can be found in the **Appendix Figure A114**.

The advantage of pheB as a reporter is that the colour change is independent of the conditions used. The pheB protein as well as the catechol and 2-hydroxymuconic semialdehyde are thermostable so the assay can be carried out at either 55 °C or at room temperature and it does not require oxygen for maturation, therefore it can be used to characterize promoters anaerobically [Bartosiak-Jentys 2012].

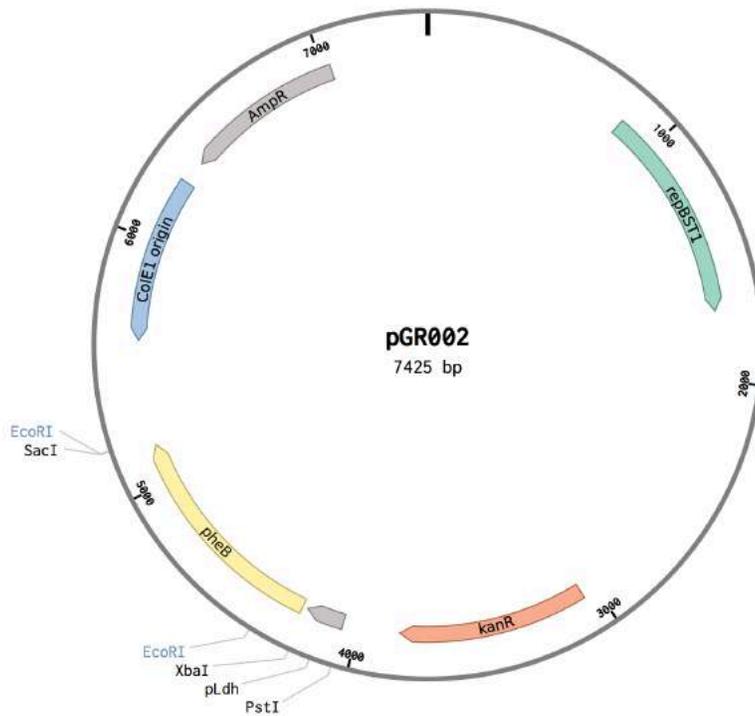


Figure 9. Plasmid map of pGR002. The *G. stearothermophilus* NCA1503 *ldh* promoter is located between the *PstI* and *XbaI* restriction enzyme recognition sites, whereas the reporter *pheB* is located between the *XbaI* and *SacI* sites.

However, *pheB* is not an ideal reporter because the assay is time-consuming. The absorbance readings depend strongly on the OD_{600} of the cultures, and even when cell lysate is used, calibration is required (i.e. protein quantification using a Bradford assay). The timing of the reading is also important, because the yellow colour turns to brown with time and temperature. In addition, calibration curves need to be routinely compiled in order to obtain readings as accurate as possible.

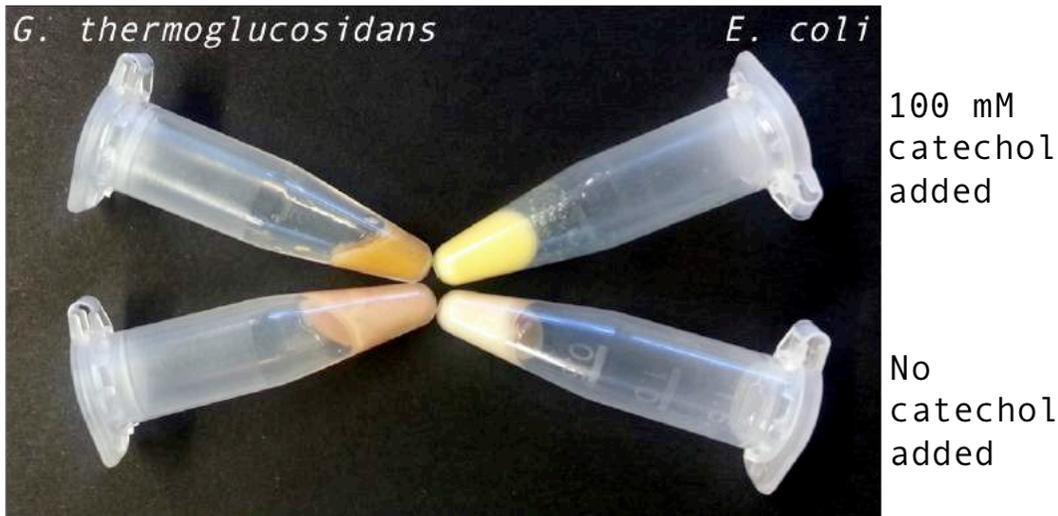


Figure 10. Photograph demonstrating the colour change exhibited by *G. thermoglucosidans* and *E. coli* cells expressing pheB with and without the addition of catechol. The two microcentrifuge tubes at the left contain cell pellets of *G. thermoglucosidans*. The cells at the top left exhibit orange coloration, product of the pheB-catalyzed degradation of catechol into 2-hydroxymuconic semialdehyde (which is a yellow compound), that distinguishes them from the bottom left cells, that have not had catechol added to them but are also expressing pheB. The two microcentrifuge tubes at the right contain cell pellets of *E. coli* cells that have also had 100 mM catechol added to them (top right) and are yellow in comparison to the cells that haven't had catechol added to them (top right). The cells were mixed with 100 mM catechol by pipetting and incubated for 10 minutes at room temperature.

Figure 10 vividly illustrates why pheB is not an ideal reporter. Promoter activity is measured by the degree of "yellowness" exhibited by the cells after the addition of catechol and it's not comparable between *G. thermoglucosidans* and *E. coli*. In the thermophile, the colouring becomes more orange/brown when catechol is added compared to a salmon/pink color when no catechol has been added. For every absorbance reading, the negative needs to be subtracted, but it is uncertain the exact degree to which the salmon/pink colour overlaps with the reading for the yellow colour. A spectral scan would be necessary to determine the difference in readings obtained with the different colorations.

Lin *et al.* [2014] used the *G. stearothermophilus* β -galactosidase as a reporter to characterize promoter activities in *G. thermoglucosidans*. Blanchard *et al.* [2014], also used LacZ, but in *G. stearothermophilus*. The LacZ assay is colorimetric as well and presents the same problems as the pheB assay, therefore more direct reporters needs to be sought.

More effective reporter proteins have been used for mesophiles that permit to analyze whole cells in vivo and which produce more accurate results. Such is the case of fluorescent

proteins like GFP. Quantifying fluorescent reporters using flow cytometry bypasses the problems encountered for the pheB assay because it permits to determine accurately the fluorescence of every single cell. Cells can be processed *in vivo* and whole and the final OD₆₀₀ has no effect on the fluorescence reading and calibrations are not necessary per run.

3.1.2.2. FLUORESCENT PROTEINS

Three kinds of fluorescent proteins will be investigated as potential fluorescent reporters in *G. thermoglucosidans*:

- Green Fluorescent Protein (GFP)
- Red Fluorescent Protein (mCherry)
- LOV-domain proteins

3.1.2.2.1. GREEN FLUORESCENT PROTEIN (GFP)

sfGFP is being investigated as a reporter in *G. thermoglucosidans* because:

- It is a widely used reporter in synthetic biology.
- Colonies are easy to screen visually using a table-top blue-light transilluminator.
- It has more robust folding kinetics than any of the other variants of GFP.

To the author's knowledge, no successful fluorescent reporters have been reported for *G. thermoglucosidans* and they are necessary because flow cytometry is the most used method in synthetic biology for the characterization of promoters and in the assembly of circuits (e.g. REF), but it is unknown how thermostable or thermophilic these proteins are when expressed by *G. thermoglucosidans*. GFP has however, been expressed by thermophiles before, for instance, Cava *et al.* reported in 2008 that a superfolder variant of **GFP** (hereafter referred to as **sfGFP**) works at high temperatures in the thermophile *Thermus thermophilus*. Blanchard *et al.* reported in 2014 on the successful expression of sfGFP in *G. stearothermophilus*. The use of **sfGFP** in *G. thermoglucosidans* has not been reported up to now, but its improved folding kinetics leads us to believe that it would be a good candidate to test for fluorescence at 55 °C in *G. thermoglucosidans*.

As detailed in **Table 6**, the superfolder variant of GFP (sfGFP) differs from a previously used variant of GFP named **frGFP** (folding-reporter GFP) by six mutations (twelve mutations that make it different from the wild-type GFP), which increase fluorescence and make folding kinetics more robust [Pédelacq *et al.* 2006]. GFP has been used as a fusion tag in protein folding experiments. Fused proteins can reduce the folding of GFP variants that have poor folding kinetics. The advantage of using sfGFP to study protein folding is that it is unaffected by fusion partner misfolding [Pédelacq *et al.* 2006], thus making it ideal to study protein expression.

Table 6. Mutations of three of the variants of GFP of in comparison to the wild-type GFP

| Variant of GFP | Mutations |
|------------------------------|--|
| enhanced GFP (eGFP) | F64L, S65T |
| folding reporter GFP (frGFP) | F64L, S65T, F99S, M153T, V163A |
| superfolder GFP (sfGFP) | F64L, S65T, F99S, M153T, V163A, S30R, Y39N, N105T, Y145F, I171V, A206V, Q80R |

A correctly folded GFP consists of an internal chromophore surrounded by a tight β -barrel. The chromophore of sfGFP is formed from residues T65-Y66-G67 and it is buried in the middle of the central helix, which in turn contains the backbone cyclized chromophore. Around the chromophore, there must be a formation of a β -barrel structure for the correct formation of the chromophore [Andrews *et al.* 2007]. GFP does not require cellular chaperones for folding, but to exhibit fluorescence, the tight β -barrel structure must be maintained, and sfGFP forms its β -barrel faster than any other variant of GFP, which is what makes it more robust [Aronson, 2011].

Directed evolution approaches were required to obtain variants of GFP with better kinetic and folding properties that would allow to express them heterologously in other hosts because wild-type GFP misfolds when expressed by *E. coli* due to the highly oxidizing periplasmic conditions in *E. coli*, which generate the formation of disulphide bridges involving the cysteine residues (C49 and C71) located in the interior of the β -barrel and flanking the chromophore, since both residues are exposed during folding and can bind to other folding GFP molecules or cysteine-containing proteins. Interestingly, two of the mutations of sfGFP (S30R, Y39N) occur before the cysteine residues, which must help confer the rapid folding and stability [Aronson *et al.* 2011].

The formation of the chromophore in GFP is a posttranslational modification. Barondeau *et al.* in 2003, reported that under aerobic conditions, the chromophore of GFP cyclizes and incorporates an oxygen atom. This does not happen under anaerobic conditions. Without molecular oxygen, the oxidation reaction cannot take place and the conjugated bonds required to form the chromophore do not occur, therefore, the final three-dimensional protein structure to exhibit fluorescence.

E. coli expressing sfGFP were twofold more fluorescent than cells expressing frGFP and 50 fold more fluorescent when fused to a poorly folded protein (ferritin), and is more stable [Pédélecq *et al.* 2006], makes believe that it would be ideal to test as a reporter protein in *G. thermoglucosidans*. If the protein exhibits fluorescence at high temperatures, ideally at 55 °C in *G. thermoglucosidans*, it could be used as a reporter to characterize promoter activity both in *E. coli* and *G. thermoglucosidans*.

The excitation wavelength of sfGFP is 485 nm and its emission 530 nm. Apart from sfGFP, other reporter proteins have been developed with different excitation/emission wavelengths. This would permit to express simultaneously various reporter proteins in the same cell and be able to analyze each separately. Different colors can be used to image different parts of the cell simultaneously and allow real-time tracking of parallel cellular processes. When a cell is expressing various fluorescent proteins simultaneously, in order to avoid obtaining overlapping signals, having proteins that have different excitation/emission wavelengths allows the controlled observation of each reporter at a time so it is necessary to look for reporters other than GFP that have excitation and emission wavelengths that do not overlap with sfGFP for when it is necessary to express both reporters simultaneously, in a circuit for example, or when two plasmids need to be transformed into *G. thermoglucosidans*, each carrying a different cargo that has a specific function. It would be an easy way to confirm that the cells have retained both plasmids, or to test for the efficiency of the circuit if it can be confirmed that both proteins are being expressed simultaneously, but their signals need to not overlap. Various fluorescent proteins of different colours have been created for this purpose, they have emission and excitation spectra, depending on their colour, that overlap little or not at all with that of GFP. mCherry is one of the most popular reporters used simultaneously with sfGFP to avoid readings of overlapping signals.

3.1.2.2.2. RED FLUORESCENT PROTEIN (mCherry)

mCherry is being investigated as a reporter in *G. thermoglucosidans* because:

- Its excitation/emission wavelengths do not overlap with sfGFP.
- mCherry is already known to be used as a reporter in *E. coli*.
- It would be ideal if mCherry, in addition to sfGFP, could also be used as a reporter for *G. thermoglucosidans*, for cases when two reporters need to be expressed within the same cell in a circuit or to test for plasmid compatibility.

mHoneydew, mBanana, mOrange, tdTomato, mTangerine, mStrawberry and mCherry, are some of the improved monomeric red, orange and yellow fluorescent proteins derived from the mRFP1 monomer, from the marine anemone *Discosoma striata* DsRed fluorescent tetramer and they were developed by Shaner *et al.* in 2004. They were created by subjecting mRFP1 to directed evolution to obtain variants that are different in colour as well as excitation and emission wavelengths and improved in brightness.

mCherry has an excitation maximum of 587 nm and an emission maximum of 610 nm. Among all the variants tested (mHoneydew, mBanana, mOrange, tdTomato, mTangerine, mStrawberry and mCherry), mCherry offered the longest wavelengths, the highest photostability, the fastest maturation and excellent pH resistance [Shaner *et al.* 2004, Tramier, 2006].

3.1.2.2.3. ANAEROBIC FLUORESCENT PROTEINS: LOV-DOMAIN REPORTERS

LOV-domain proteins need to be investigated as potential reporters to characterize promoter activity under anaerobic conditions, a task which neither GFP nor mCherry are capable of due to the need of molecular oxygen for the formation of their chromophores and hence fluorescence:

- LOV-domain proteins are fluorescent and do not require oxygen to exhibit fluorescence.
- Four variants of LOV-domain protein are to be tested: BsFbFP, PpFbFP, hotLOV and hbLOV. Each was selected for the specific reasons explained in detail below.

- Due to the absence of reports, to the best of the author's knowledge, regarding the successful expression of any LOV-domain protein in a thermophile, several variants of LOV-domain proteins were selected for testing of expression and correct folding in *E. coli* and *G. thermoglucosidans*.
- For this project, it is relevant to know that the LOV-domain proteins are involved in a stress response in *Bacillus subtilis* because when expressed heterologously in a different host, such as *G. thermoglucosidans*, the LOV-domain protein might trigger a physiological response as well.

The strict requirement of molecular oxygen for the correct folding of their chromophore means that the proteins of the GFP family can't be used to characterize the anaerobic activity of promoters. Because anaerobic growth is important for the downstream use of *G. thermoglucosidans*, we began a search for another fluorescent protein that does not require oxygen for maturation. This search highlighted the flavin mononucleotide (FMN)-based fluorescent protein family as a promising avenue.

Plants have FMN-based photoreceptors that are involved in phototropism, chloroplast movement, leaf expansion, stomatal opening and even flowering. These plant photoreceptors contain two Light, Oxygen, Voltage (LOV) domains that act as signaling modules and exhibit sensitivity to blue light. The chromophore is a flavin mononucleotide (FMN): the FMN molecule absorbs blue light photons, which cause it to excite and form a covalent the molecule bond with a specific cysteine residue in the LOV domain, and once this bond has formed, the protein is in signaling state. All the subsequent processes that lead to the physiological response after the molecule has achieved signaling state are under investigation [Losi, 2006].

Photosynthetic and also non-photosynthetic bacteria, possess homologues to the plant FMN-based phototropins. Among the non-photosynthetic prokaryotes, these proteins are found in *Pseudomonas* spp., *Brucella* spp., *Listeria* spp. and interestingly for this project, also in *B. subtilis*. In fact, YtvA from *B. subtilis*, was the first LOV protein discovered when plant LOV domains were being compared to bacterial sequences [Huala, 1997]. YtvA undergoes exactly the same blue-light induced biochemical reaction as plant phototropins, although in *B. subtilis*, the signaling state of the protein triggers the general stress response [Gaidenko, 2006]. It is not well understood why this is induced by blue light [Krauss, 2006]. It

is known, however, that blue light is absorbed by porphyrins. All bacteria synthesize porphyrins as heme or chlorophyll precursors. So it is possible that blue-light stress sensors evolved in response to changing environmental conditions [Ghetti *et al.* 1992], which might have evolved from the expression of photosynthesis genes in cyanobacteria [Stephan and Gabriele, 2004].

3.1.2.2.3.1. *BACILLUS SUBTILIS* FLAVIN BASED FLUORESCENT PROTEIN (BSFBFP) AND *PSEUDOMONAS PUTIDA* FLAVIN BASED FLUORESCENT PROTEIN (PPFBFP)

The first two LOV-domain proteins tested as potential anaerobic fluorescent reporters for *G. thermoglucosidans* will be BsFbFP and PpFbFP because:

- They have been shown to work as anaerobic fluorescent reporters in mesophiles.
- What they have in common is that they both possess a LOV-domain.
- What's different between them is their amino acid sequence.

Drepper, *et al.* reported in 2007 the development of BsFbFP and PpFbFP. BsFbFP is a mutated version of the *B. subtilis* 168 YtvA protein (see **Figure 11**). These two proteins are FMN-based proteins and contain the LOV domain, but they exhibit cyan-green fluorescence at an excitation wavelength of 450 nm, under both anaerobic and aerobic conditions, thanks to a single mutation in their amino acid sequence (C62A). In the case of YtvA, it was reported that by mutating the cysteine residue in position 62, which is the one to which FMN binds, to a non-polar alanine that won't bind FMN, the protein becomes fluorescent, thus conferring it potential to be exploited as a fluorescent reporter. PpFbFP is another variant of the YtvA protein, but it is found in *Pseudomonas putida*. PpFbFP is the version mutated by Drepper *et al.* in 2007 of the LOV-photosensor protein SB2, discovered by Krauss *et al.* in 2005. Both BsFbFP and PpFbFP showed a maximal light absorption at 450 nm and a maximal emission at 495 nm upon excitation. **Table 7** shows a comparison among the excitation and emission wavelength for the three reporter proteins tested.

| Table 7. Maximal excitation and emission wavelengths for the fluorescent reporter proteins tested for <i>E. coli</i> and <i>G. thermoglucosidans</i> | | |
|--|------------------------------------|----------------------------------|
| Reporter protein | Maximal excitation wavelength (nm) | Maximal emission wavelength (nm) |
| sfGFP | 485 | 530 |
| mCherry | 587 | 610 |
| LOV-domain proteins | 450 | 495 |

No reports exist so far for the BsFbFP protein being used as a reporter in thermophiles, but given that *B. subtilis* is a close relative of *Geobacillus* spp. and their codon usage **should be** similar, it was decided it would be interesting to find out whether this protein can be implemented as a reporter in thermophiles.

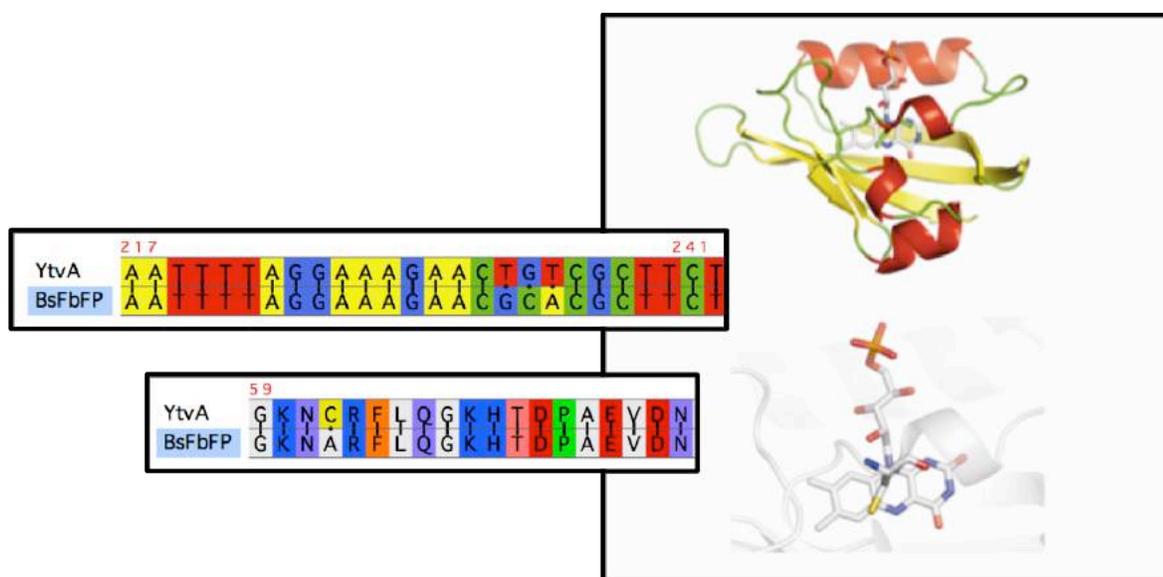


Figure 11. Schematic view of the structure of the *B. subtilis* YtvA. The top red alpha helix is the LOV domain. The structure at the bottom is a closer look at the FMN-chromophore that interacts with the LOV domain. The cysteine residue in position 62 is at the top of the FMN-chromophore. It is observed in the top structure how this cysteine residue interacts with the LOV domain. The sequence at the top left shows the codon mutation required at positions 233-235. At the bottom left is the amino acid change from cysteine to alanine in position 62.

Figure taken and modified from Krauss, 2007.

LOV proteins can be found in the genomes of 35 bacterial species Losi *et al.* [2006]. A phylogenetic analysis based on on the alignment of the LOV sequences can be seen in **Figure 12**.

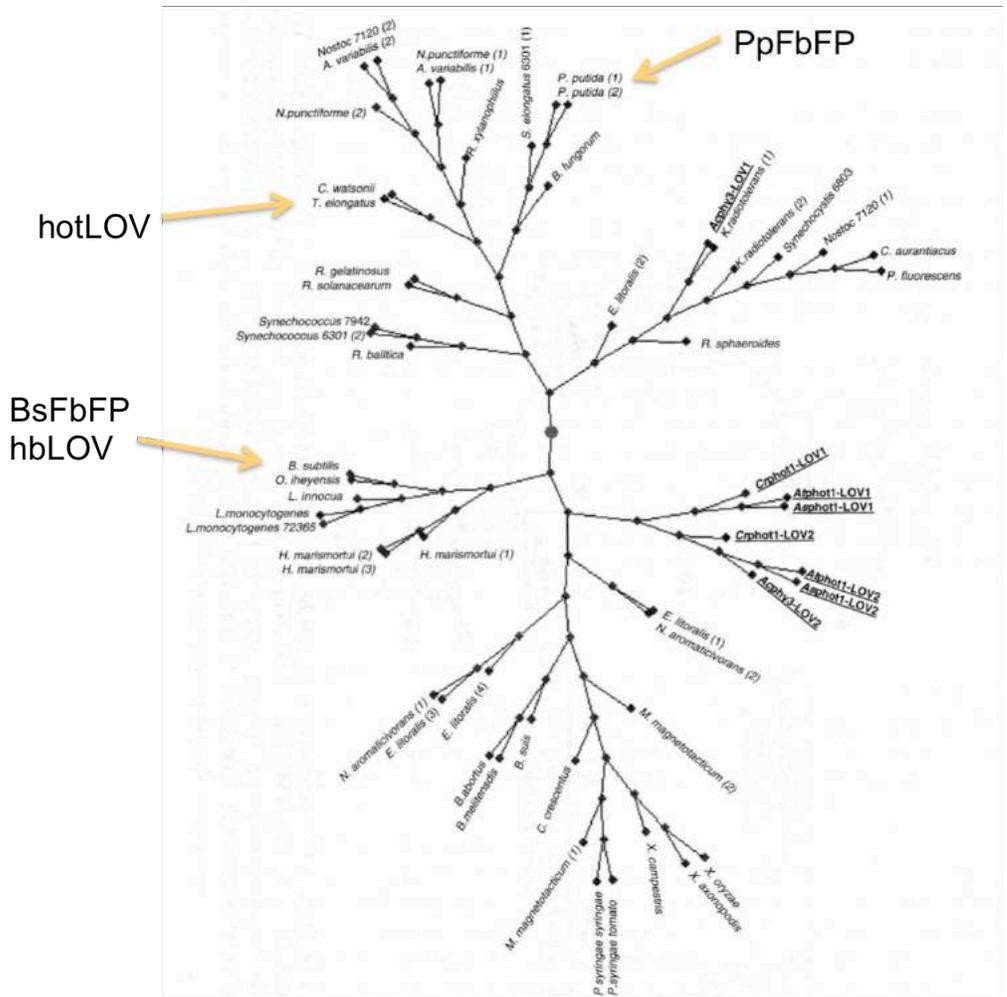


Figure 12. Phylogenetic tree of bacterial species that contain genes that encode for putative flavin-binding photosensors. The phylogeny was calculated on the basis of sequence alignment of LOV domains. *Avena sativa* and *Arabidopsis thaliana* were used as the outgroups. PpFbFP, from this work, corresponds to the LOV protein from *Pseudomonas putida*. hotLOV is the LOV protein from *Thermosynechococcus elongatus* that was codon optimized for *E. coli*. BsFbFP corresponds to the LOV protein from *Bacillus subtilis*, likewise, hbLOV corresponds to the same LOV protein as BsFbFP but contains several mutations. hbLOV was originally developed by Song *et al.* but was codon optimized for *B. subtilis* by B. Reeve (IC). *Figure taken and modified from Losi, 2006.*

The reason wild-type LOV proteins do not exhibit fluorescence but C62A mutants do, is that LOV domains bind non-covalently to the oxidized flavin mononucleotide (FMN) chromophore and when exposed to blue light, at 450 nm, they undergo a reversible photocycle that leads to the formation of an FMN-cysteine (Cys62 in YtvA) which exhibits

weak autofluorescence with a maximal emission wavelength of 495 nm [Salomon *et al.* 2000]. Blue-light activation of YtvA results in structural changes, *i.e.* the formation of a photoproduct. As confirmed by Drepper *et al.* in 2007, breaking the FMN-Cys62 bond results in a dramatic change in the recovery kinetics, which explains the increase in fluorescence of BsFbFP compared to YtvA [Losi 2006, Drepper *et al.* 2007, Krauss 2007].

FMN-based fluorescent proteins have been used as reporters in a variety of mesophilic hosts [Lobo *et al.* 2011, Tielker *et al.* 2009]. In this work, we propose to test FMN-based fluorescent proteins as reporters in *E. coli* and *G. thermoglucosidans*. However, it is possible that some degradation in *E. coli* and *G. thermoglucosidans* may occur, given that Drepper *et al.* (2007) reported that when BsFbFP was tested in *Rhodobacter capsulatus*, a Gram-negative mesophilic bacterium, the levels of fluorescence due to protein accumulation were very low, possibly due to protein degradation or inefficient gene expression.

3.1.2.2.3.2. HOTLOV

hotLOV is the third LOV-domain protein to be tested as a potential anaerobic fluorescent reporter for *G. thermoglucosidans* because:

- It is derived from a thermophile, therefore, it might be thermostable and thermophilic.
- It has been codon-optimized for *E. coli*

hotLOV is yet another LOV-domain protein variant, but it is from *Thermosynechococcus elongatus* (Accession number BAC08834.1). This thermophilic cyanobacterium also contains an FMN-binding blue-light sensing protein. It was developed by the Christie group at the University of Glasgow and was synthesized with codon optimization for *E. coli*, as seen in the alignment in **Figure 13**. Given that the source of the protein is a thermophilic host, it was assumed that the protein should be thermophilic and thermostable. The Christie group developed it based on the hypothesis that, as it was originated from a thermophilic host, it should have better folding properties than BsFbFP and therefore, might be more fluorescent in *E. coli* and might possibly also be thermophilic and thermostable. Taking this into consideration, it was decided that apart from BsFbFP and

PpFbFP, hotLOV should be tested as a potential anaerobic fluorescent reporter for *E. coli* and *G. thermoglucosidans*.

Thermosynechococcus elongatus BP-1 DNA, complete genome
 Sequence ID: [dbj|BA000039.2](#) Length: 2593857 Number of Matches: 1

Range 1: 1334352 to 1334681 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

| Score | Expect | Identities | Gaps | Strand |
|---------------|--|--------------|-----------|------------|
| 284 bits(314) | 6e-73 | 261/330(79%) | 0/330(0%) | Plus/Minus |
| Query 1 | ATCGCCAGCACCAACGGCATCGTCATTACGGACTATCGCCAACCGGACAACCCGGTCATC | 60 | | |
| Sbjct 1334681 | ATTGCTAGCACCAACGGCATTGTGATTACTGACTACCGCCAACCCGATAATCCAGTTATC | 1334622 | | |
| Query 61 | TACGTGAACCCGGCATTGTAACGCATGACCGGCTATCGTGAACGGAAGTCATTGGTAAA | 120 | | |
| Sbjct 1334621 | TATGTGAACCCGGCCTTTGAGCGGATGACGGGTACCGGCAACGGAGGTCAATTGGCAA | 1334562 | | |
| Query 121 | AACGCTCGTTTTCTGCAGGGCAGCGATCGCCATCAACCGGGTGCAGCCGTCATTGTAAT | 180 | | |
| Sbjct 1334561 | AATTGCCGCTTTTGCAGGGGAGCGATCGCCATCAACCGGGAGCTACAGCAATTCGTAAT | 1334502 | | |
| Query 181 | GCGATCAAAAAAGGCCAGTCTTGCCGCGTGGTTCGCGTAACTACCGTAAAAATGGTCAA | 240 | | |
| Sbjct 1334501 | GCCATTAAAAAAGGTCAATCCTGTCGTGTGGTTTTGCGGAACTATCGCAAAAACGGCCAG | 1334442 | | |
| Query 241 | CTGTTCTGGAACGAACTGGCAATTAGTCCGATCTACAATGAATTTGGCGAAATCACCCAC | 300 | | |
| Sbjct 1334441 | CTCTTCTGGAATGAACTTGCCATCTACCAATTTACAACGAGTTTGGTGAAATTACCCAC | 1334382 | | |
| Query 301 | TACATCGGCATCCAGTCGGACGTTACGGAA | 330 | | |
| Sbjct 1334381 | TACATTGGCATTCAAAGTGATGTTACTGAA | 1334352 | | |

Figure 13. Comparison of the genomic sequence of *Thermosynechococcus elongatus* corresponding to the LOV-domain protein with the codon optimized sequence of hotLOV. The alignment between the codon-optimized hotLOV gene (query) and the genomic sequence of *Thermosynechococcus elongatus* that was identified by NCBI BLAST as the closest match (subject). The mismatches seen between the Query and the Subject sequences can be attributed to codon optimization for *E.coli*.

3.1.2.2.3.3. HBLOV

hbLOV is a fourth variant of LOV-domain protein to be tested in this work:

- It is different from BsFbFP, PpFbFP and hotLOV in that it contains mutations that are predicted to confer thermostability based on a modelling approach.

Song et al. [2013] developed various different versions of the BsFbFP protein by modeling the protein structure and mutagenesis. They claim that their mutants should more thermostable than the original BsFbFP based on their modelling approach, where they selected mutations that are on the surface of the dimer interface because they were the most likely to produce salt bridges, thus making the interaction between both subunits stronger.

Figure 14 shows the eight sites that contributed to enhancing the fluorescence. The

mutations that conferred thermostability also produce stronger van der Waals interactions between neighbouring amino acids. Taking this into account, we believe that stronger chemical interactions would indeed result in a more robustly folded protein and it is therefore necessary to test this protein for thermophilicity and to corroborate its thermostability by expressing as a reporter in *G. thermoglucosidans*.

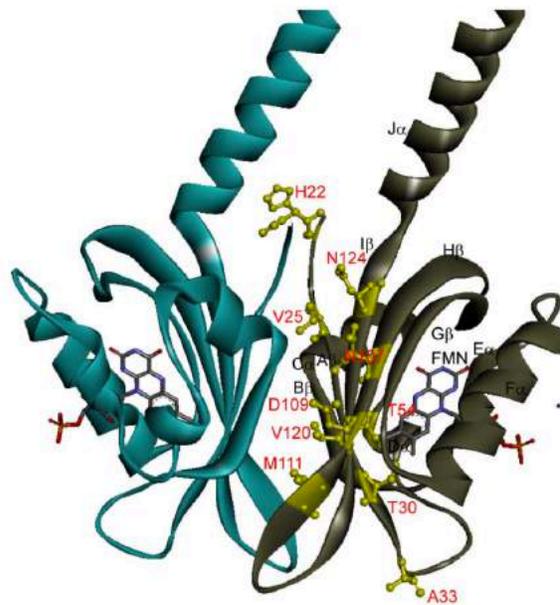


Figure 14. Three-dimensional structure of hbLOV. Highlighted in red are the eight residues which are predicted to confer thermostability to the hbLOV variants tested by Song *et al.* over the wild type BsFbFP. The gray domain corresponds to the LOV-domain identified because of its proximity to the FMN chromophore, whereas the cyan domain corresponds to the STAS-domain. It can be observed from this figure that the mutations that are predicted to confer thermostability occur mostly in the amino acids that are located where the two domains interact, and they result in stronger interactions between neighbouring amino acids which in turn has resulted in the prediction of more robust folding. *Figure taken from Song et al. 2013.*

Song *et al.* expressed several mutated variants of hbLOV and purified them. The purified variants were then heat-denatured and their fluorescence was measured at 450 nm to test for thermostability. The thermal denaturation was estimated for the mutants compared to the wild-type. The triple mutant N107Y-N124Y-M111F was suggested to be the most thermostable (by 32.1 °C compared to the wild-type: T_m of the wild-type is 42.8 °C, T_m of the triple mutant is 74.9 °C). According to the authors, the N124Y mutation forms extra hydrogen bonds, which result in closer contact between N124Y-V23',V25' and Q129.

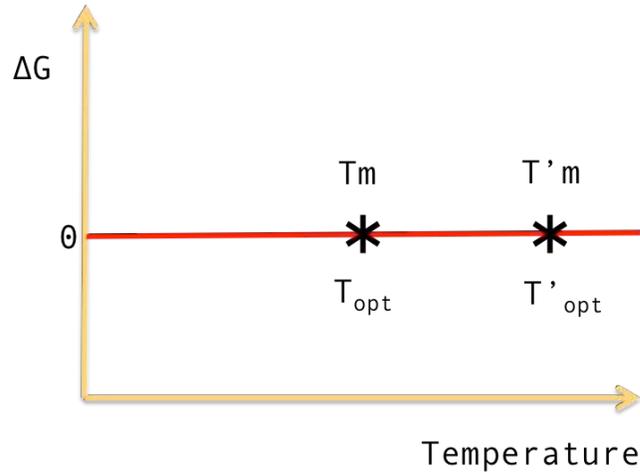


Figure 15. Hypothetical optimal melting temperatures for mesophilic and thermophilic proteins. Graph representing the hypothetical temperature profile of the free energy of mesophilic (T_m , T_{opt}) and thermophilic (T'_m , T'_{opt}) proteins, where ΔG represents the hypothetical difference in the free energy between native and denatured proteins. T_m and T'_m represent the hypothetical melting temperature for mesophilic and thermophilic proteins, respectively. T_{opt} and T'_{opt} represent the optimal growth temperatures of mesophilic and thermophilic organisms respectively. A higher melting temperature correlates to a higher growth temperature.
Figure modified from Jaenicke and Böhm [1998].

One of the most interesting facets about the research conducted by Song *et al.* was the strategy they used to predict what mutations would yield the most thermostable proteins. Their algorithm, FoldX, predicted free energy calculations for random mutants. Such calculations sought to find the difference in the free energies between the native and denatured proteins in order to obtain a configuration that is predicted to have a higher melting temperature (T_m) as seen in **Figure 15**. Experimentally it was confirmed that indeed those mutants did have improved thermostability. FoldX might to be a valuable tool for the rational engineering of thermostability, especially when the heterologous expression of mesophilic enzymes is required in a thermophile. It is relevant to this work because, in this work it is tested whether the prediction of FoldX for thermostability of hbLOV is tested experimentally in *G. thermoglucosidans*. If the prediction is correct, then FoldX would be a very valuable tool in the future for the design of thermostability in proteins as modelling starting point. The thermostability predictions are relevant to this work because, it will provide clarification about "thermostable" and "thermophilic" being two different properties. The "thermostable" variants of Song *et al.* were expressed in *E. coli*. Therefore, all variants folded in the host at 37 °C, so it is unknown if they are thermophilic, and expressing them in *G. thermoglucosidans* would be a way of finding out: if they fold correctly at high

temperatures, it means they are thermophilic, and if they remain folded and functional, they are thermostable.

3.1.3. SALIS RBS CALCULATOR

The Salis RBS Calculator is an essential tool for synthetic biology because it is an aid in designing RBSs that have predictable strengths and which do not form unwanted secondary structures depending on their combinatory context (promoter and CDS). It is relevant to this work because:

- The forward engineering function has been used for the *de novo in silico* design of RBS sequences that were then tested experimentally.
- The reverse engineering function has been used to calculate the strength of native RBSs of *G. thermoglucosidans* in both *E. coli* and *G. thermoglucosidans*.
- This work will test whether the Salis RBS Calculator's predictions can be applied also for *G. thermoglucosidans* given that the Salis RBS Calculator is optimized for *E. coli*, and it does not take into account the difference in the optimal temperatures of growth between *E. coli* (30-37 °C) and *G. thermoglucosidans*. Therefore, it will be interesting to see whether the predictions of the RBS calculator can also be applied to this host. If so, then the RBS Calculator could in the future become a valuable tool for predicting synthetic RBSs in *Geobacillus*.
- It is not the aim of this project to test the strength of the RBSs predicted or to characterize them. This work is focused exclusively on deciding whether they are functional or not based on protein expression.

Whilst promoters are transcriptional regulators, RBSs are bioparts that function as translational regulators, and there is an online modelling tool that permits to design RBS sequences *de novo* and to analyze existing ones. It is crucial for synthetic biology because the level of gene expression can be specified by designing the level of translation. The folding of mRNA near the RBS can strongly affect the expression levels of a gene. The mRNA that surrounds the RBS can interact with the CDS. It is therefore of the utmost necessity to be able to predict the interaction between RBS and CDS [Hoitz and Keasling 2010].

A thermodynamic model of translation initiation was developed by Salis *et al.* in 2009, called the Ribosome Binding Site (RBS) Calculator. It is an online tool⁴ designed to not only to develop parts *in silico*, but also to predict the rate of translation initiation in bacteria [Holtz and Keasling, 2010].

The RBS calculator offers two options: Forward and Reverse Engineering. In the Reverse Engineering mode, the user has the option of submitting an mRNA sequence (which already contains an RBS) and the output predicts how strong that particular RBS is. The Forward Engineering mode consists in predicting an entirely synthetic RBS. That is, an RBS designed nucleotide by nucleotide based on the predictions of the algorithms it runs with to achieve a desired translation initiation rate. The range of such rate can go from 0 to 100,000+ au [Salis, 2011].

The synthetic RBS sequences generated by this method are significantly longer (30-35 nucleotides) than the consensus “AGGAGG” RBS sequence. This means that the model must generate 4^{35} possible RBS sequences. All these sequences are then analyzed until one is found that matches as closely as possible the desired translation rate [Salis *et al.* 2009].

"K" and "NEQ" (“Kinetic trap” and “Not at equilibrium”, respectively) are warnings that indicate the folding of the mRNA transcript creates problems that violate the thermodynamic assumptions of the model [Salis, 2011].

The model requires at least the first 35 nucleotides of the protein’s CDS to make a prediction. The forward engineering option generally satisfies the model's assumptions, because the RBS is created from a non-existent one. The reverse engineering option, on the other hand, may result in K or NEQ outputs [Salis, 2011].

⁴ [<https://www.denovodna.com/software/>]

3.2. RESULTS

The purpose of this chapter is to show whether sfGFP, BsFbFP, PpFbFP, hbLOV and hotLOV are effective as fluorescent reporter proteins in *G. thermoglucosidans*. To test for expression, they are to be coupled to known *G. thermoglucosidans* promoters, namely *ldh*, *p43* and *RlpS*. First the plasmids containing the promoter-RBS-reporter cassettes had to be assembled and the results below explain in detail the strategy used for assembly. The plasmids were assembled in *E. coli* and for this reason *E. coli* was tested first for fluorescence. After the correct expression of all the reporters in *E. coli* was confirmed, the plasmids were transformed into *G. thermoglucosidans* and the expression of the reporters was determined by analysis of fluorescence and protein purification.

3.2.1. SUPERFOLDER GREEN FLUORESCENT PROTEIN (sfGFP)

STEP 1: ASSEMBLY OF PLASMID PUCG18+LDH+sfGFP

Primers emk001 and emk002 were initially designed to amplify sfGFP. All primers sequences are included in the **Appendix Table A12**. The PCR yielded a 774 bp fragment, which was run through a 0.8% agarose gel to determine the fragment's correct size and gel-extracted.

The forward primer (emk001) encoded an *XbaI* site and the same RBS as the one that had already been used in the pGR002 (*i.e. pheB-RBS*) construct at the 5'-end (see **Figure A137**). The reverse primer (emk002) added a rho-independent transcriptional terminator at the 3'-end.

Due to the lack of a restriction site at the 3'-end, a third primer (emk003) was subsequently designed to add an *EcoRI* site so that sfGFP could be ligated as an *XbaI/EcoRI* flanked DNA fragment. Thus, 10 ng of the 774 bp PCR-product obtained with the primer pair emk001 and emk002 were used as template for a second PCR-amplification with primers emk001 and emk003, yielding a 794 bp fragment, which was subsequently digested with *XbaI* and *EcoRI*, to generate a 782 bp cohesive-end fragment.

The vector into which sfGFP was integrated was pGR002 [Bartosiak-Jentys, 2009]. This plasmid was digested with *XbaI* and *EcoRI* to remove pheB and replace it for sfGFP. This way, sfGFP is expressed under the control of the *ldh* promoter.

The *XbaI/EcoRI* pGR002 vector and the *XbaI/EcoRI* sfGFP insert were mixed in a 3:1 molar ratio (insert:vector) ligation reaction, which was incubated overnight at 4 °C and subsequently was used to transform chemically competent *E. coli* JM109.

In order to confirm the presence of sfGFP in *E. coli*, an observational assay would not have picked up any green fluorescent colonies because the NCA1503 promoter does not work in *E. coli*. This was chosen because it was the strongest known promoter available for *G. thermoglucosidans* under aerobic conditions at the time. An equivalent result was also observed by Bartosiak-Jentys when the aforementioned *ldh* promoter was placed downstream from the pheB gene (*xylE* reporter).

Therefore, to select for the correct transformant colonies, a colony PCR was done to screen for the presence of the sfGFP gene using the primer pair emk001 and emk002. The result can be observed in **Figure 16**. Eleven out of twelve colonies produced a band of the expected size (774 bp), thus confirming the presence of the sfGFP insert in the plasmid backbone.

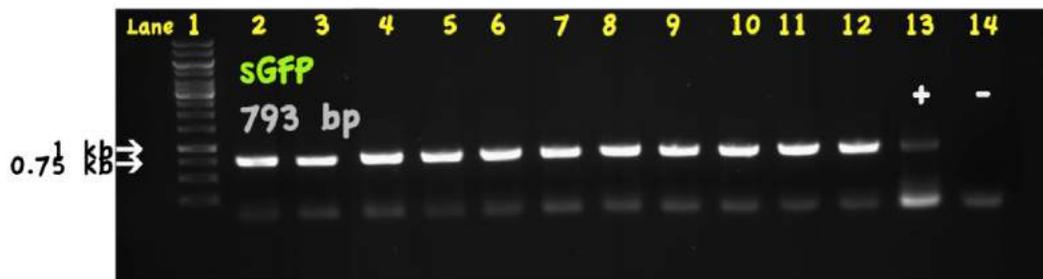


Figure 16. Colony-PCR products looking for the presence of sfGFP in *E. coli* transformants analyzed on a 0.8% agarose gel. Lane 1 contained the molecular size marker (1 kb GeneRuler, Fermentas). Relevant band sizes indicated. The size expected for sfGFP was 793 bp. Lanes 2-12 contained 20 μ M of the PCR-product obtained from screening 11 transformant DH10B colonies using the primer pair emk001 and emk003. A positive and negative control were included in the PCR, in lanes 13 and 14, respectively.

STEP 2: TESTING OF PLASMID pUCG18+LDH+sfGFP IN *G. THERMOGLUCOSIDANS*

Once confirmed, the resulting plasmid, which hereafter will be referred to as "pUCG18+ldh+sfGFP", depicted in **Figure 17**, was isolated from the correct transformant *E. coli* colonies and 100 ng were used to transform *G. thermoglucosidans* DL33 and DL44. The detailed sequence can be found in **Appendix Figure A115**.

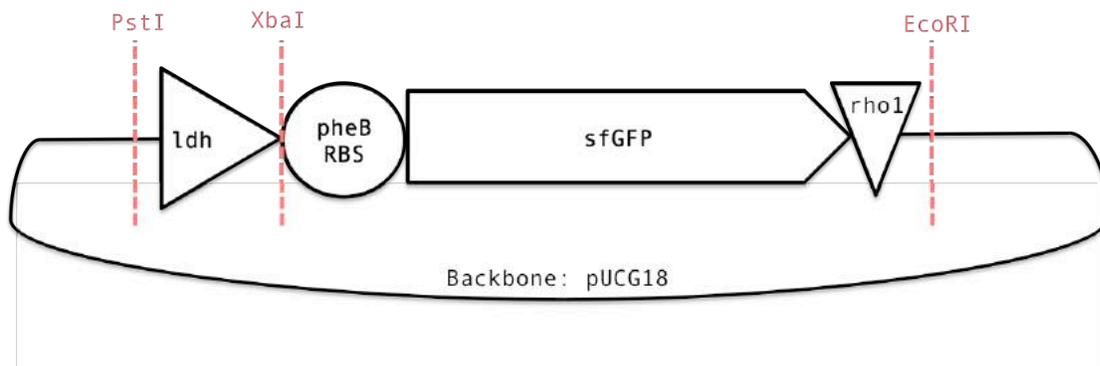


Figure 17. Cassette used to express sfGFP under the control of the *ldh* promoter, in *G. thermoglucosidans* DL44. Restriction sites marked in red. Detailed sequence in **Figure A115**.

Two of the transformant colonies obtained were restreaked on a fresh TGP/kanamycin plate and the green fluorescence seen in **Figure 18** confirmed the expression of sfGFP under the control of the NCA1503 *Ldh* promoter.



Figure 18. Expression of sfGFP by *G. thermoglucosidans*. Photograph of two colonies of *G. thermoglucosidans* transformed with pUCG18+ldh+sfGFP expressing sfGFP under the control of the ldh promoter and the pheB-RBS were restreaked on a TGP-kanamycin plate and grown overnight at 55 °C. The colony at the left corresponds to the wild-type DL33 strain of *G. thermoglucosidans* and the colony to the right corresponds to DL44, an ldh knockout strain of DL33. The fluorescence of sfGFP in the plate was viewed on a table-top UV viewer with orange filter and irradiated with UV light.

3.2.2. THE *LDH* PROMOTER IN *E. COLI*

E. coli JM109 did not show any green fluorescence when transformed with a plasmid containing the *ldh* promoter upstream from sfGFP. There are two possible explanations for this: (1) the *ldh* promoter doesn't work in *E. coli*, perhaps due to the absence of a necessary transcription factor or sigma factor in *E. coli*, or (2) the RBS located downstream from the *ldh* promoter is not compatible with the sequence of sfGFP or pheB.

To investigate the latter, the sequence was analyzed by the reverse engineering function of the Salis RBS calculator⁵. The results are shown in **Appendix Figures A116 and A117**. Two RBSs were analyzed using this function: the native RBS found upstream from the start codon of the *ldh* gene of *G. stearothermophilus* and the RBS developed by Bartosiak-Jentys [2009] to express pheB, which was also used in the "pUCG18+ldh+sfGFP".

⁵ [<https://www.denovodna.com/software/>]

The RBS Calculator could not perform a definitive analysis of the strength of either of the RBSs reporting both "OLS" and "NEQ", which are warnings that indicate that the criteria for the model could not be satisfied. Further research is necessary to investigate whether changing the RBS would result in an *ldh* promoter that works in both *E. coli* and *G. thermoglucosidans*.

3.2.3. sfGFP IN THE ABSENCE OF OXYGEN

When *G. thermoglucosidans* DL44 containing plasmid pUCG18+*ldh*+sfGFP were grown anaerobically or microanaerobically, no fluorescence was observed. It was assumed sfGFP was being expressed because the culture was inoculated with the cells containing the pUCG18-*ldh*+sfGFP plasmid that showed fluorescence aerobically and kanamycin was supplemented to the growth medium to ensure retention of the plasmid. In an attempt to refold the misfolded proteins, DL44 pUCG18+*ldh*+sfGFP was grown anaerobically for 16 hours. The cells were harvested by centrifugation and were allowed to aerate by transferring them to a baffled flask containing 50 mL of 2TY medium supplemented with 30 µg/ml chloramphenicol (to halt protein production) and shaken at 250 rpm for two hours. After which, they were harvested once more by centrifugation and observed under blue light. No fluorescence was observed. Furthermore, in a second attempt to refold the misfolded proteins, the cell pellet was heated to 100 °C for 20 minutes in order to denature all proteins, after which, they were transferred to a baffled flask containing 50 mL of 2TY supplemented with chloramphenicol and they were allowed to aerate for two hours, in the hope that sfGFP can be refolded after being denatured, but once again, no fluorescence was observed.

The absence of fluorescence when sfGFP was grown anaerobically was unlikely to be a sign that the *ldh* promoter was not active anaerobically, as refuted by N. Crowhurst [2010], but rather, because oxygen is required for the correct formation of the sfGFP chromophore, the lack of oxygen in these conditions prevent fluorescence [Drepper *et al.* 2007]. Therefore, taking into consideration that the *ldh* promoter is active anaerobically according to Bartosiak-Jentys [2009], it was assumed that DL44 is expressing sfGFP, but seeing as there are no oxygen molecules for the chromophore to form, the proteins misfold and thus are not fluorescent. This indicates that once sfGFP in the cells has folded incorrectly or has not been expressed, it is irreversible and a functional formation cannot be

achieved. A Western Blot should have been carried out to confirm that sfGFP was effectively being expressed under anaerobic conditions under the control of the *ldh* promoter.

3.2.4. BsFBFP IN *E. COLI* AND *G. THERMOGLUCOSIDANS*

STEP 1: ASSEMBLY OF PLASMID PUCG18+P43+BSFBFP

The *ytvA* gene was amplified as a 843 bp fragment from the genome of *B. subtilis* 168 using the pair of primers emk004 and emk005. As seen in **Figure 19**, eight primers were designed to amplify *YtvA* and convert it to BsFbFP. The PCR products obtained with each pair are shown in **Figure 20**. Primers emk006 and emk007 were designed to carry out the directed mutagenesis required to transform *YtvA* to BsFbFP. The sequences are included in **Appendix Figures A119** and **A120**.

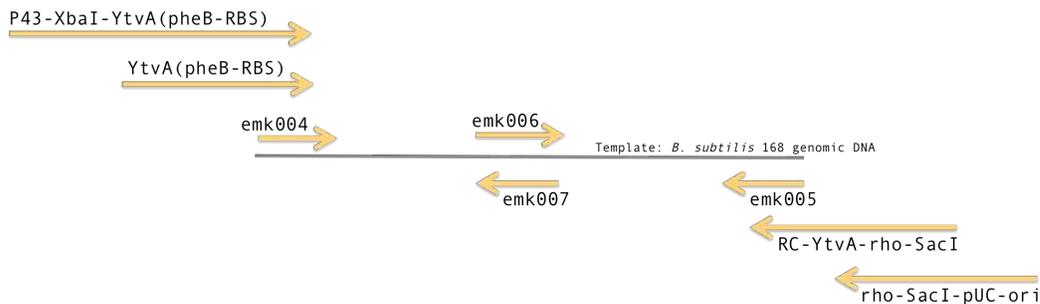


Figure 19. Schematic representation of the primers used to amplify *ytvA*.

As seen in **Figure 20**, the mutation to change the cysteine residue in position 62 for an alanine residue was inserted by PCR-amplification. The PCR-amplified *ytvA* was used as a template for two subsequent PCR amplifications: the primer pair emk004 and emk007 added the mutation to one of the PCR products, and the primer pair emk006 and emk005 added the mutation to the second PCR product.

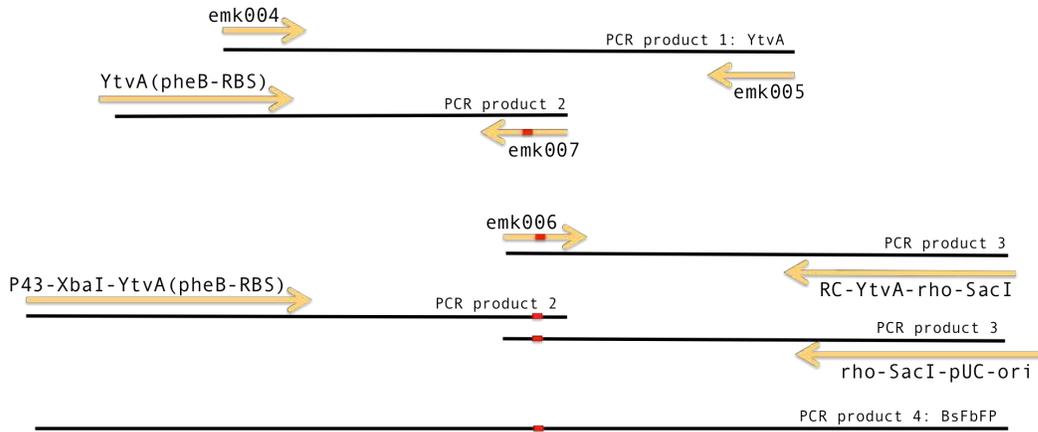


Figure 20. Primer pairs and PCR products used to amplify *ytvA* from the genome of *B. subtilis* 168 and convert it to BsFbFP.

The parental *ytvA* was digested with *DpnI* to eliminate it, and PCR products 2 and 3 were joined together by PCR-amplification to yield BsFbFP as a fourth PCR product, amplified with the primer pair p43-XbaI-YtvA(pheB-RBS) and rho-SacI-pUC_ori. This final pair of primers added overhangs for the P43 promoter in the 5'-end and for pUC_ori in the 3'-end. The p43 promoter was PCR amplified from pNCBS5 supplied by N. Crowhurst, using the primer pair pUC_ori-PacI-p43 and RC-p43-XbaI-YtvA(pheB-RBS) (see **Figure 21**). This primer pair also added 30 bp of overhang to the pUCG18 vector at the 5'-end end to *ytvA* at the 3'-end. The vector backbone, pUCG18, was amplified to add overhangs for both, the p43 promoter and BsFbFP using primers rho-SacI-pUC_ori and RC-p43-XbaI-YtvA(pheB-RBS).

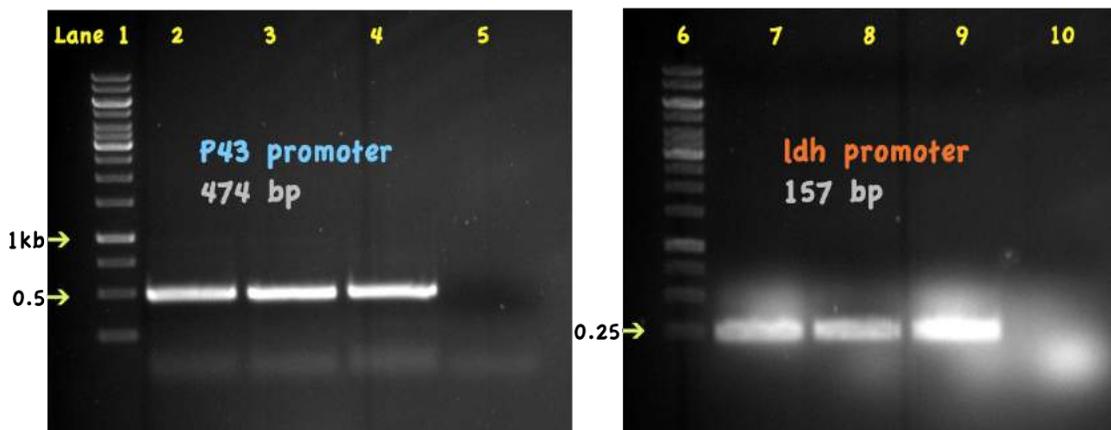


Figure 21. Analysis of PCR-amplification of the p43 and *ldh* promoters in an 0.8% agarose gel. Lanes 2-4 are the PCR product of the amplification of the *P43* promoter using primers pUC_ori-PacI-P43 and RC-P43-XbaI-YtvA. Lanes 7-9 are the PCR product of the amplification of the *ldh* promoter using primers pUC_ori-PstI-*ldh* and RC-*ldh*-XbaI-YtvA. Lanes 1 and 6 correspond to the molecular size marker (Fermentas, 1 kb-ladder). The sizes expected were 0.474 Kb for the *p43* promoter and 0.157 Kb for the *ldh* promoter.

The three fragments: p43, BsFbFP and the pUCG18 backbone were assembled together using Gibson assembly, thus resulting in the plasmid depicted in **Figure 22**, which will be named "pUCG18+p43+BsFbFP". The detailed sequence can be found in **Appendix Figure A121**.

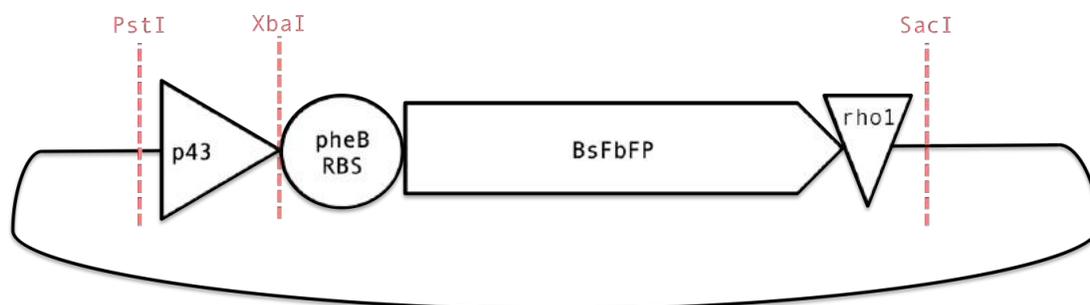


Figure 22. Cassette used to express BsFbFP under the control of the p43 promoter, in both *E. coli* JM109 and *G. thermoglucosidans* DL44. Restriction sites shown in red. Detailed sequence is in **Figure A145**.

The correct *E. coli* transformant colonies containing pUCG18+p43+BsFbFP were selected by the presence of cyan-green fluorescence (449 nm maximal emission at 495 nm upon excitation), observed under a blue light transilluminator. Unlike the *ldh* promoter, the p43 promoter is active in *E. coli*, so fluorescent colonies were easy to detect. The fluorescence of BsFbFP is shown in **Figure 23**.



Figure 23. Fluorescence of BsFbFP in *E. coli*. Six *E. coli* colonies, labeled 1-6, transformed with plasmid pUCG18+p43+BsFbFP were restreaked on an LB-ampicillin plate and grown overnight at 37 °C. The cyan-green fluorescence is due to the expression of BsFbFP under the control of the p43 promoter. The fluorescence was seen when observing the plate under a table-top UV light transilluminator with an orange filter.

STEP 2: TESTING OF PLASMID pUCG18+LDH+SFGFP IN *G. THERMOGLUCOSIDANS*

The plasmid pUCG18+p43+BsFbFP was isolated from the fluorescent *E. coli* colonies and 300 ng of DNA were used to transform *G. thermoglucosidans* DL44. The transformant colonies were selected on TGP/kanamycin plates at 55 °C. No cyan-green fluorescence was observed by inspection under a blue-light transilluminator of the transformant DL44 colonies at this temperature.

The absence of fluorescence in *G. thermoglucosidans* had three possible explanations: either the protein was not stable at temperatures as high as 55 °C, or, the p43 promoter was not strong enough to allow the detection of fluorescence or the protein was not made (a Western blot was not carried out to confirm this). Given the lack of experimental confirmation about the expression of BsFbFP in thermophiles, the thermostability and thermophilicity of this protein were unknown. To test the former hypothesis, the DL44 cells containing pUCG18+p43+BsFbFP were grown at 45 °C, which is the lowest temperature at which *G. thermoglucosidans* will grow. No fluorescence was observed at 45 °C either, so the attention was focused on replacing the p43 promoter for a different promoter that's stronger in DL44.

STEP 3: ASSEMBLY OF PLASMID pUCG18+LDH+BSFBFP

The p43 promoter in the pUCG18+p43+BsFbFP construct was excised using PstI and XbaI. The primer pair pUC_ori-PstI-ldh and RC-ldh-XbaI-(pheB-RBS) was used to amplify the *ldh* promoter from the pUCG18+ldh+sfGFP plasmid. The PCR product was digested with PstI and XbaI and a 157 bp band was extracted from a 1.2% agarose gel (**Figure 39**). The PstI/XbaI *ldh* promoter was ligated to the promoterless PstI/XbaI pUCG18+BsFbFP backbone to yield pUCG18+ldh+BsFbFP, shown in **Figure 24**.

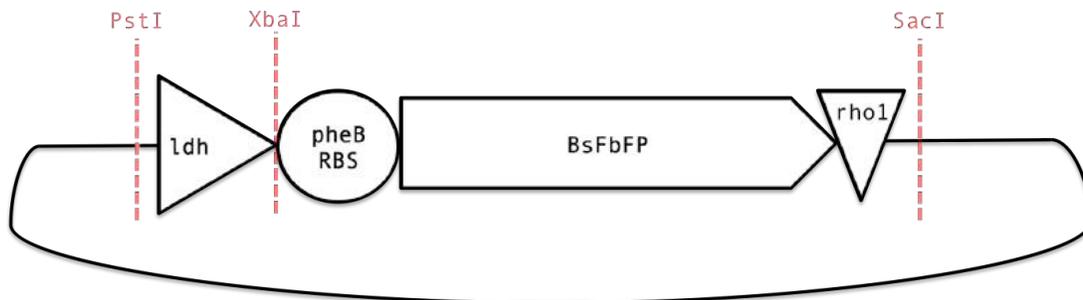


Figure 24. Cassette used to express BsFbFP under the control of the *Ldh* promoter, in *G. thermoglucosidans* DL44. Restriction sites shown in red. Detailed sequence is in Figure A145.

Since the *ldh* promoter does not work in *E. coli*, a colony PCR was performed to screen for the correct transformant colonies. The result is shown in Figure 25. Four correct transformants were selected and grown in 5 mL LB/ampicillin. The pUCG18+*ldh*+BsFbFP plasmid was isolated and 300 ng of it were used to transform *G. thermoglucosidans* DL44. The sequence of the *ldh* promoter and BsFbFP can be found in the Appendix Figure A122.

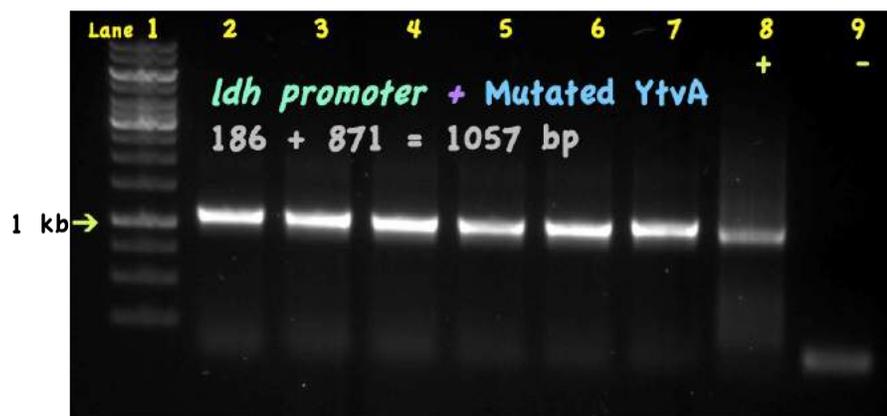


Figure 25. Colony-PCR of *ldh*+BsFbFP analyzed on a 0.8% agarose gel. Lane 1 is the 1kb GeneRuler molecular size marker. Lanes 2-7 are the PCR-products obtained by screening for the *ldh* promoter plus the mutated form of *ytva* (BsFbFP) from six colonies of *E. coli*. Lane 8 is the positive control. Lane 9 is the negative control.

STEP 4: TESTING OF PLASMID pUCG18+LDH+BsFbFP IN *G. THERMOGLUCOSIDANS*

By visual inspection under a blue-light transilluminator of DL44 transformed with pUCG18+ldh+BsFbFP or pUCG18+p43+BsFbFP and grown at 45 °C, no differences were seen between the BsFbFP cells and the negative controls. A Western Blot to confirm protein expression was not carried out because the sequence of the protein did not include a tag. However, to test whether DL44 were transcribing BsFbFP, RNA was extracted from a culture that had been growing at 47 °C and converted to cDNA, after which reverse transcription reaction was performed using the primer pair emk004 and emk005.

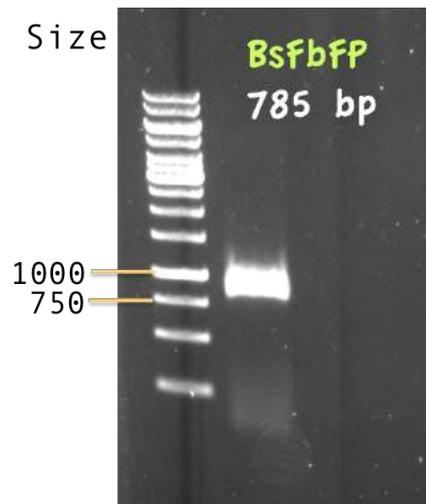
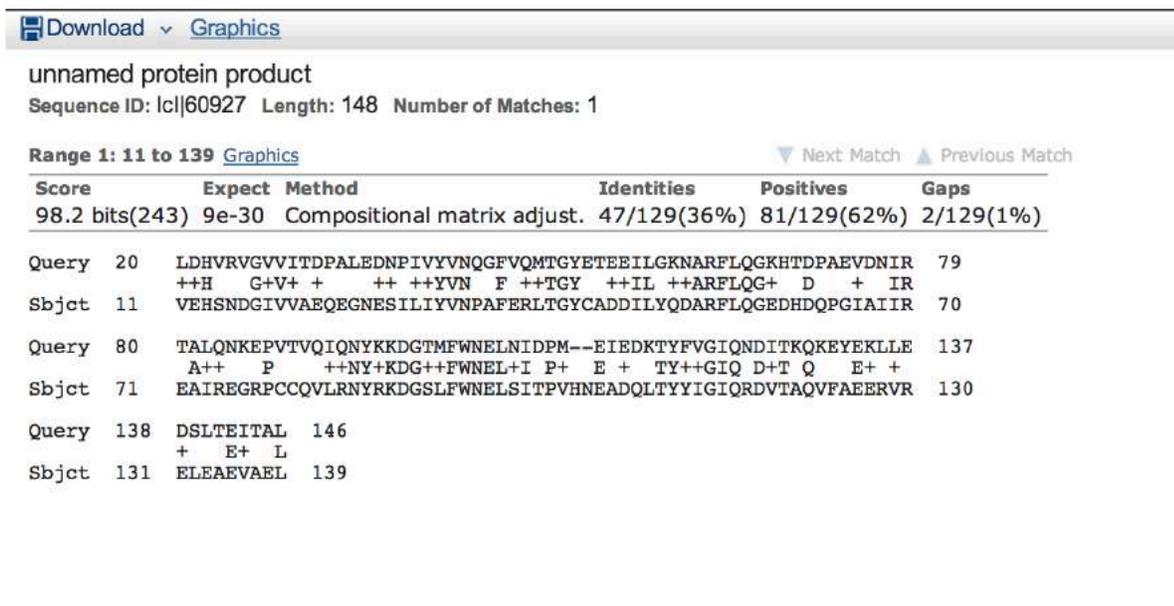


Figure 26. 0.8% agarose gel showing the RT_PCR of BsFbFP. Lane 1 is the 1kb GeneRuler molecular size marker. Lane 2 is the result of the PCR amplification, using primers emk004 and emk005, of the cDNA obtained from the reverse transcription of the RNA extracted from DL44 cells containing the plasmid pUCG18+p43+BsFbFP. The size expected for BsFbFP was 785 bp.

As it can be seen in **Figure 26**, a band that appears to be the correct size was seen, thus confirming the presence of mRNA transcript of BsFbFP suggesting that the absence of fluorescence might be due to the protein not being made. The primers should not have amplified the plasmid DNA because DNA removal was carried out when total RNA was extracted by DNaseI treatment.

3.2.5. PpFbFP

A sequence comparison study among BsFbFP and PpFbFP (**Figure 27**) revealed that there is only 36 % similarity between the amino acid sequences of BsFbFP and PpFbFP, indicating that structurally, even though both have the LOV domains, their amino acid sequences are somewhat different and therefore, their folding patterns might be different and perhaps PpFbFP might be more thermostable or less prone to degradation as suggested by Drepper *et al.* (2007) who reported that *Rhodobacter capsulatus* expressed PpFbFP much more efficiently than BsFbFP.



The screenshot shows a BLAST search interface. At the top, there are buttons for 'Download' and 'Graphics'. Below that, the search parameters are listed: 'unnamed protein product', 'Sequence ID: lcl|60927', 'Length: 148', and 'Number of Matches: 1'. The search results are displayed in a table with columns for 'Score', 'Expect', 'Method', 'Identities', 'Positives', and 'Gaps'. The top result shows a score of 98.2 bits (243), an expectation value of 9e-30, and a method of 'Compositional matrix adjust.'. The identity is 47/129 (36%), with 81/129 (62%) positives and 2/129 (1%) gaps. The query sequence is 'LDHVRVGVVITDPALEDNPIVYVNVQGFVQMTGYETEEILGKNARFLQGKHTDPAEVDNIR' and the subject sequence is 'VEHSNDGIIVVAEQEGNESILYVNPAPERLTGYCADDILYQDARFLQGEDHDQPGIAIIR'. The alignment shows a high degree of similarity between the two sequences.

| Score | Expect | Method | Identities | Positives | Gaps |
|----------------|---|------------------------------|-------------|-------------|-----------|
| 98.2 bits(243) | 9e-30 | Compositional matrix adjust. | 47/129(36%) | 81/129(62%) | 2/129(1%) |
| Query 20 | LDHVRVGVVITDPALEDNPIVYVNVQGFVQMTGYETEEILGKNARFLQGKHTDPAEVDNIR | | | | 79 |
| | ++H G+V+ + ++ ++YVN F ++TGY ++IL ++ARFLQG+ D + IR | | | | |
| Sbjct 11 | VEHSNDGIIVVAEQEGNESILYVNPAPERLTGYCADDILYQDARFLQGEDHDQPGIAIIR | | | | 70 |
| Query 80 | TALQNKEPVTVQIQNYKKDGTMFWNELNIDPM--EIEDKTYFVGIQNDITKQKEYEKLE | | | | 137 |
| | A++ P ++NY+KDG++FWNEL+I P+ E + TY++GIQ D+T Q E+ + | | | | |
| Sbjct 71 | EAIREGRPCQVLRNRYRKDGSLFWNELSITPVHNEADQLTYIIGIQRDVTAQVFAEERVR | | | | 130 |
| Query 138 | DSLTEITAL | 146 | | | |
| | + E+ L | | | | |
| Sbjct 131 | ELEAEVAEL | 139 | | | |

Figure 27. Comparison of the amino acid sequences among BsFbFP and PpFbFP using protein BLAST⁶. The subject sequence is BsFbFP and the query sequence is PpFbFP. The similarity between them is 36%, so it was hypothesized that the amino acid sequence was different enough to justify testing PsFbFP for thermostability.

STEP 1: ASSEMBLY OF PLASMID P11AK+RPLS+PPFBFP

To test PpFbFP, the gene encoding it, as reported in Drepper *et al.* (2007), was synthesized with codon optimization for *G. stearothermophilus* by GeneART (Life Technologies). A 6X-Histidine tag was added before the stop codon for purification purposes. An RBS was designed using the forward engineering function of the RBS calculator. An

⁶ [http://www.ncbi.nlm.nih.gov/].

XbaI site was included in the 5'-end and a *SacI* site was added to the 3'-end of the gene as well. This entire sequence was synthesized by GeneART. A map of the construct is shown in **Figure 28**. The gene encoding PpFbFP together with the synthetic RBS were placed downstream from the wild-type RplS promoter developed by Benjamin Reeve (IC).

The gene string was amplified with the primer pair PpFbFP and RC-PpFbFP and digested using *XbaI* and *SacI*. Plasmid p11AK (discussed in Chapter 2 of this work) was also digested with *XbaI* and *SacI*. Both parts were ligated together overnight and the ligation was transformed into chemically competent DH10B cells.

The correct transformants were selected from plates by their weak cyan-green fluorescence under UV light. Four correct transformant colonies were grown in 5 mL LB/ampicillin overnight and the plasmid was isolated. 300 ng of it were used to transform *G. thermoglucosidans* DL44 electrocompetent cells.

STEP 2: TESTING OF PLASMID P11AK+RPLS+PPFBFP IN *G. THERMOGLUCOSIDANS*

The presence of the gene encoding the 6X-His PpFbFP as well as the synthetic RBS were confirmed by sequencing of DNA from plasmid isolated from *G. thermoglucosidans* DL44.

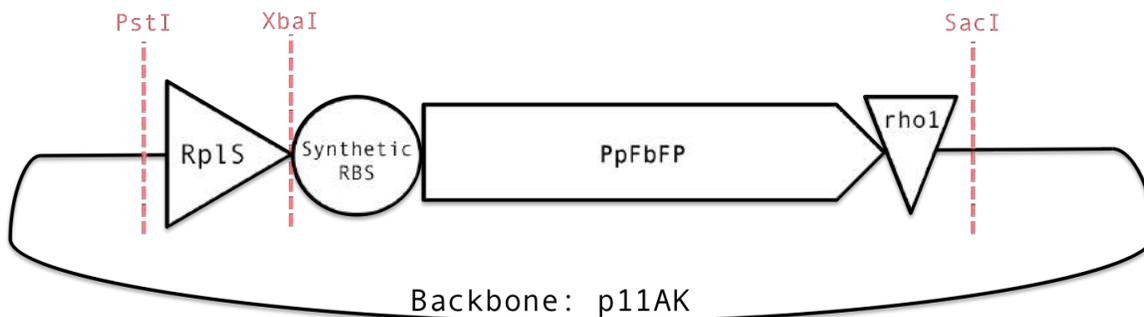


Figure 28. Cassette used to express PpFbFP under the control of the RplS promoter, in both *E. coli* DH10B and *G. thermoglucosidans* DL44. Restriction sites shown in red. Detailed sequence is in **Figure A146**.

As shown in **Figures 31 and 32**, *E. coli* DH10B and *G. thermoglucosidans* DL44 cells containing the plasmid p11AK+RplS+PpFbFP showed no fluorescence. So, to confirm the expression of PpFbFP by *G. thermoglucosidans* DL44, the protein was purified using a Western blot using the 6X-His tag added to the 3'-end. The result is shown in **Figure 29**. The band observed is closer to 30 kDa, whereas the size of the band expected was 17.8 kDa.

Given that flavin-photosensors, when heterologously expressed, could potentially take up too much of the FMN pool from the host [Losi, 2006], LB and 2TY plates, containing the appropriate antibiotic, were prepared supplemented with 100 mg FMN (flavin mononucleotide). BsFbFP and PpFbFP *G. thermoglucosidans* transformants were restreaked on these plates. No fluorescence was observed on these plates either, indicating that the unavailability of FMN is not the reason for the lack of fluorescence of either BsFbFP or PpFbFP.

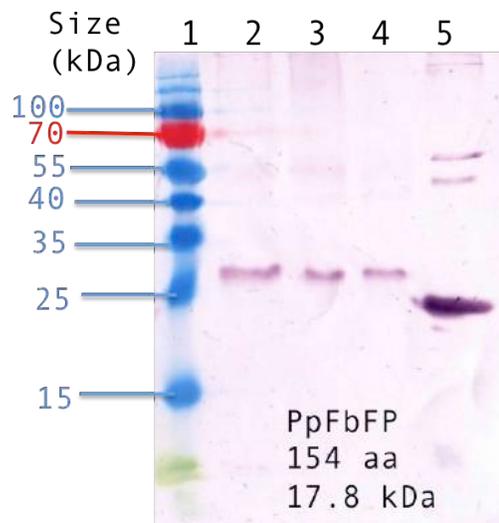


Figure 29. Western blot detecting the PpFbFP protein from cultures of *G. thermoglucosidans* DL44 containing the p11AK+RplS+PpFbFP construct. Lane 1, prestained protein size marker. Lanes 2-4, cell lysate from colonies 1-3. Lane 5, positive control (purified 6X his-tagged Kdc).

The possibility that the 6X histidine tag might be affecting the folding of the protein, seeing as no fluorescence was observed in *E. coli* either, which should have been fluorescent according to the results obtained by Drepper *et al.* in 2007. This is unlikely, however, considering that Gawthorne *et al.* in 2012, reported on the expression of a C-terminal 8X His-tagged shortened version of BsFbFP in *E. coli* and it did not affect the fluorescence.

3.2.6. HOTLOV

Benjamin Reeve (Imperial College London) expressed hotLOV under the control of the *ldh* promoter and using the pheB-RBS. The fluorescence was measured in both *E. coli* and *G. thermoglucosidans*, and results show that while it shows the expected cyan-green fluorescence when expressed in *E. coli*, in *G. thermoglucosidans* it exhibits no fluorescence [Benjamin Reeve, *personal communication*].

3.2.7. HBLOV

As shown by Drepper et al. in 2007, codon optimization for *E. coli* of BsFbFP was one of the factors that produced a BsFbFP version with enhanced fluorescence, so Benjamin Reeve (IC) optimized the sequence of the Song et al. triple-mutant (N107Y-N124Y-M111F) that showed the highest melting temperature (T_m).

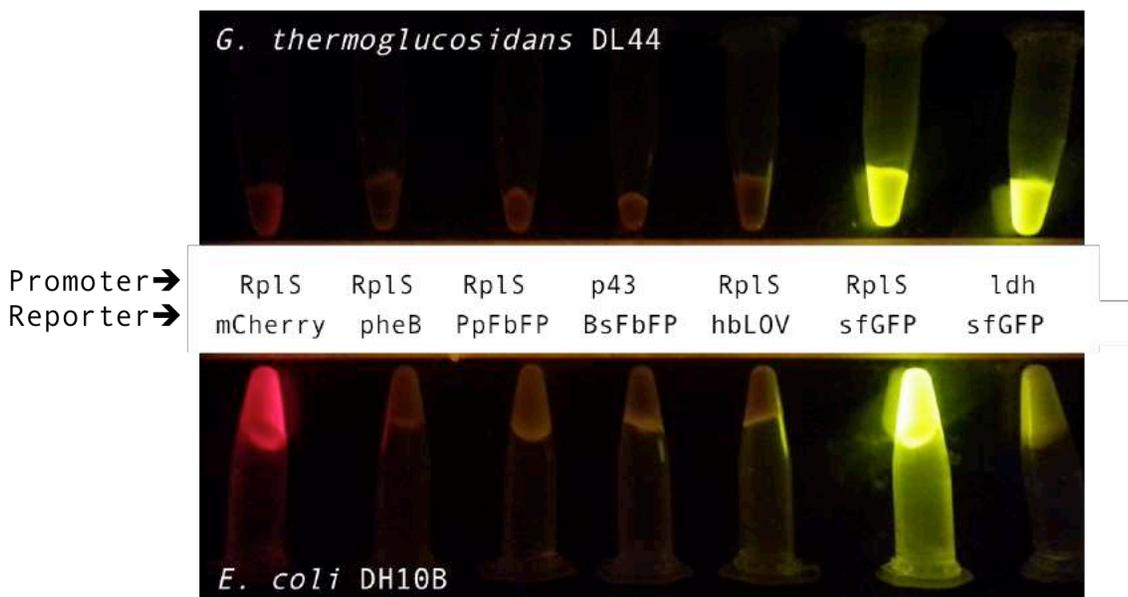


Figure 30. Fluorescence observed for each of the reporter proteins tested in this study when exposing pellets to blue light in a blue-light transilluminator. At the top are pellets corresponding to *G. thermoglucosidans* DL44 aerobic cultures and at the bottom are cell pellets from *E. coli* DH10B cultures expressing, from left to right, mCherry under the Rp1S promoter, pheB under the Rp1S promoter, PpFbFP under the Rp1S promoter, BsFbFP under the p43 promoter, hbLOV under the Rp1S promoter, sfGFP under the Rp1S promoter and sfGFP under the *ldh* promoter. All *Geobacillus* cultures were grown at 45 °C, except for the sfGFP ones, which were grown at 55 °C. All *E. coli* cultures were grown at 37 °C.

3.2.8. FLOW CYTOMETRY

Flow cytometry readings were carried out to test whether the cells shown in **Figure 30** had no fluorescence because (1) the excitation wavelength produced by the table-top blue light transilluminator might have not been the most appropriate to obtain the maximum emission and (2) to accurately discern if the cells show the same amount of green fluorescence as non-fluorescent negative controls. Results indicate that there is weak fluorescence shown by BsFbFP that can only be detected by flow cytometry.

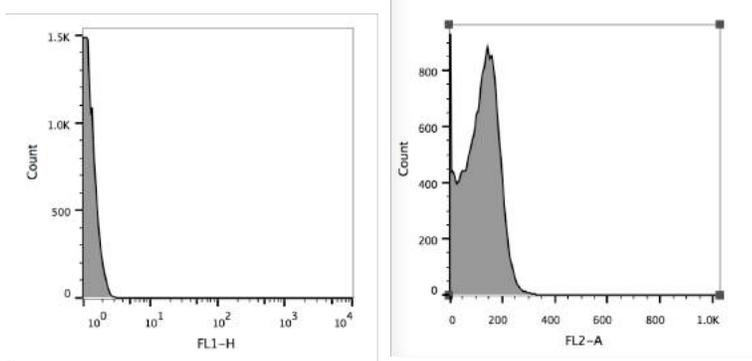
Figure Appendix A124 shows a comparison between the size and agglomeration of *E. coli* and *G. thermoglucosidans* when compared by flow cytometry. *G. thermoglucosidans* cells are bigger in size and less agglomerated than *E. coli* but the same gating settings were applied to *G. thermoglucosidans* cells as for *E. coli* cells.

Figure 31 shows the results obtained from analyzing the same cultures of *E. coli* (A-G) and *G. thermoglucosidans* (I-K) from **Figure 30** by flow cytometry.

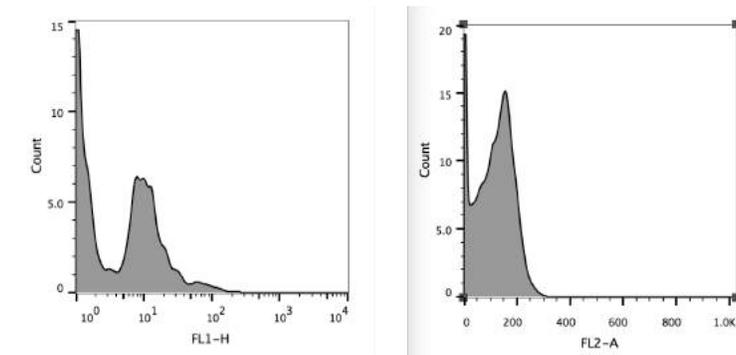
Histogram B shows that BsFbFP exhibits green fluorescence in *E. coli* due to the right shift of the peak in FL1-H. PpFbFP (C) also exhibits very weak fluorescence. The large amount of cells that exhibit no fluorescence can be explained by degradation of the FbFP proteins during stationary phase. hbLOV (D) shows a much more significant green shift and no loss of fluorescence, evidenced by the one peak in FL1-H. The green fluorescence exhibited by any of the flavin proteins is nowhere nearly as bright as sfGFP, for instance, the FL1-H median for hbLOV is 10.9 and for sfGFP its 334. Remarkably, histogram G shows that there is a trace amount activity of the Ldh promoter in *E. coli*. *E. coli* expressing mCherry shows a significant shift of the peak to the right in FL2-A.

In the case of *G. thermoglucosidans*, on the other hand, histogram I shows that there is fluorescence of BsFbFP in this host, both under the control of the p43 promoter and under the *ldh* promoter (J). It is interesting that such fluorescence could not be detected visually, but it could be detected by the flow cytometer. Also, it appears to be significantly strong (FL1-H median of 16.4 in I compared to 89.8 for Ldh+sfGFP in M). hbLOV (K) is significantly weaker than BsFbFP. PpFbFP (L) is definitely not fluorescent in this thermophile. The

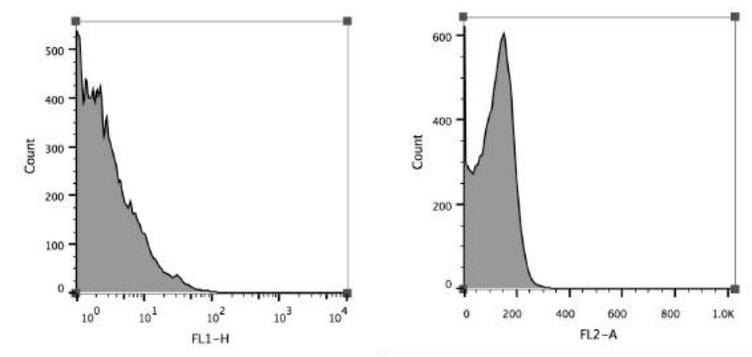
comparison between *ldh*+sfGFP (M) and *RpIS*+sfGFP (N) reveals that the wild-type *RpIS* promoter appears to be much stronger than the *ldh* promoter. Histogram O confirms that the red shift by mCherry is very weak, the median in FL2-A was 219 compared to 134 for the negative control (H).



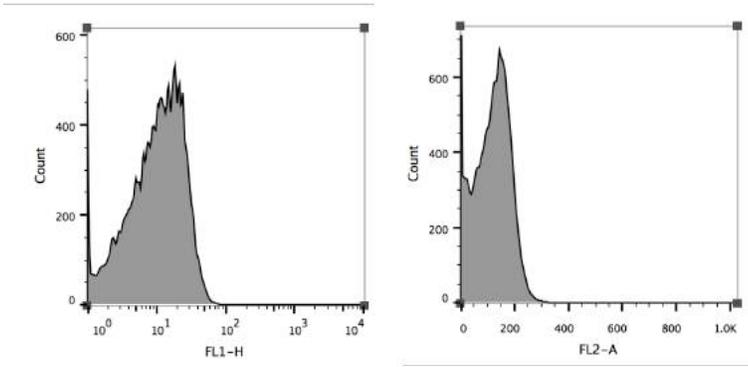
A. *E. coli* DH10B (negative control). Median FL1-H= 1.03, FL2-A= 111



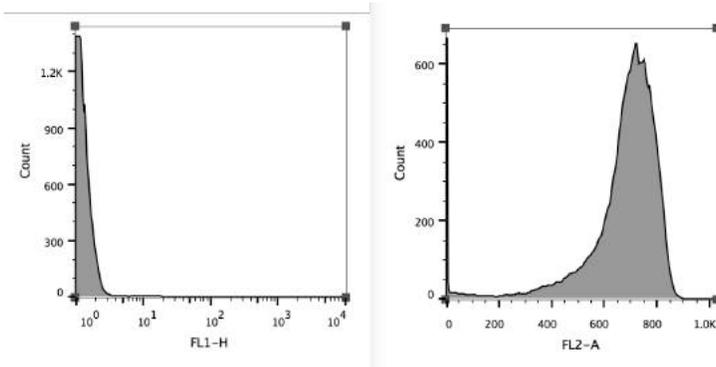
B. *E. coli* DH10B *p43+BsFbFP*. Median FL1-H= 1.32, FL2-A= 111



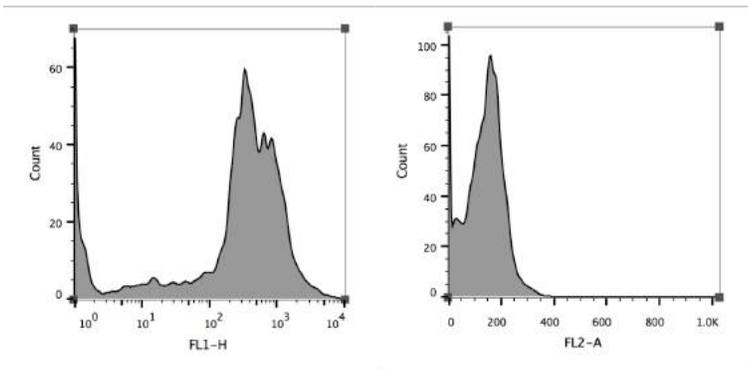
C. *E. coli* DH10B *RpIS+PpFbFP*. Median FL1-H= 2.09, FL2-A= 110



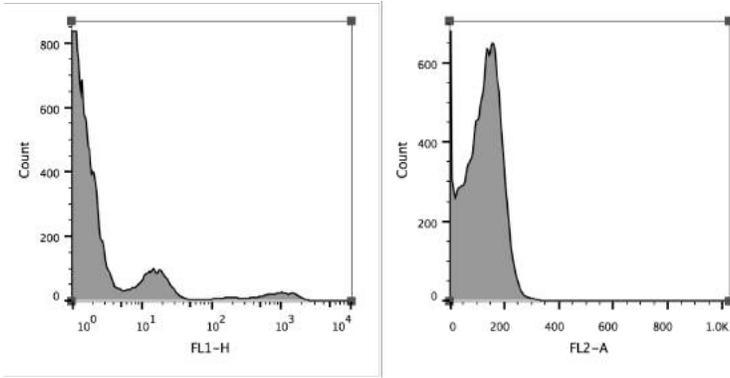
D. *E. coli* DH10B *RpIS+hbLOV*. Median FL1-H= 10.9, FL2-A= 110



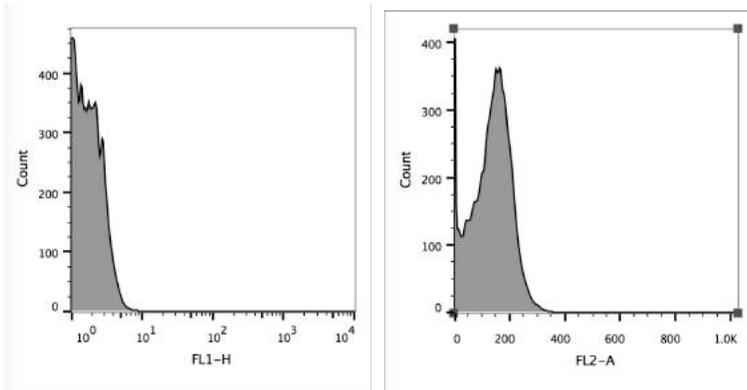
E. *E. coli* DH10B *RpIS+mCherry*. Median FL1-H= 1.04, FL2-A= 686



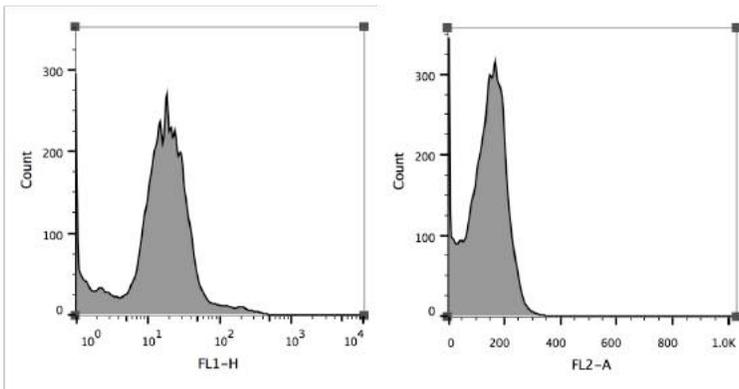
F. *E. coli* DH10B *RpIS+sfGFP*. Median FL1-H= 334, FL2-A= 140



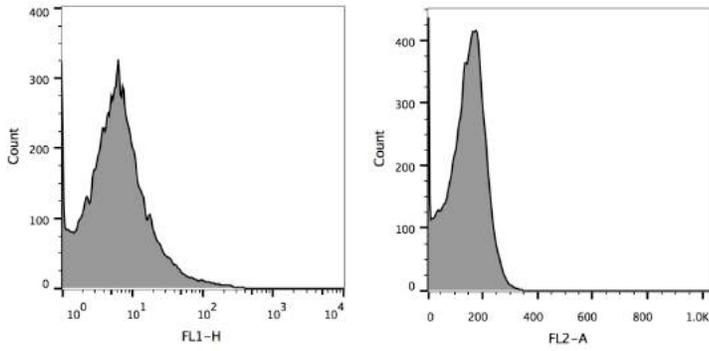
G. *E. coli* DH10B *ldh+sfGFP*. Median FL1-H= 1.39, FL2-A= 117



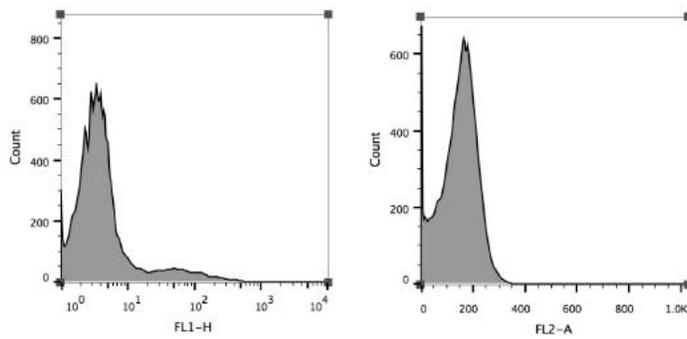
H. *G. thermoglucosidans* DL44 (negative control). Median FL1-H= 1.54, FL2-A= 134



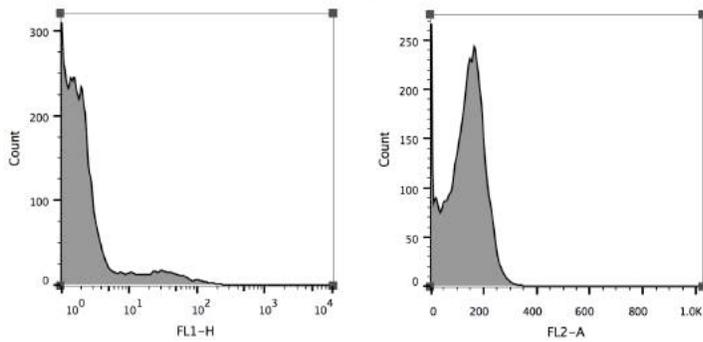
I. *G. thermoglucosidans* DL44 *p43+BsFbFP*. Median FL1-H= 16.4, FL2-A= 141



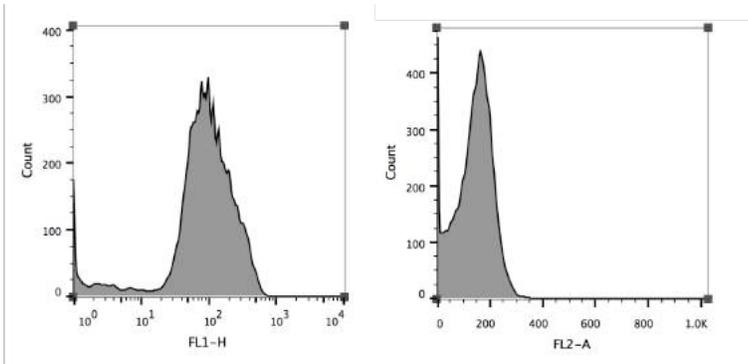
J. *G. thermoglucosidans* DL44 *Idh+BsFbFP*. Median FL1-H= 5.94, FL2-A= 142



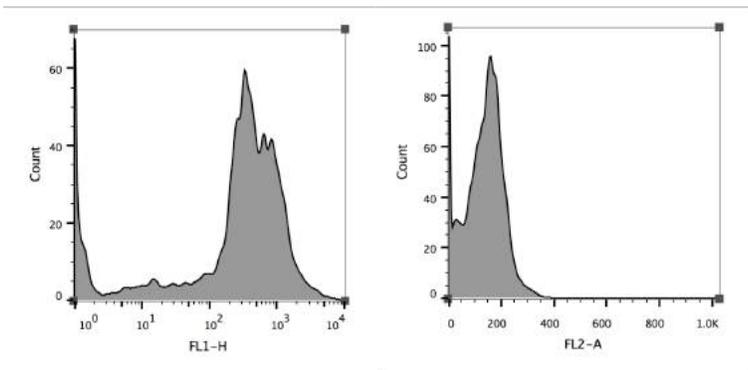
K. *G. thermoglucosidans* DL44 *RpIS+hbLOV*. Median FL1-H= 1.72, FL2-A= 136



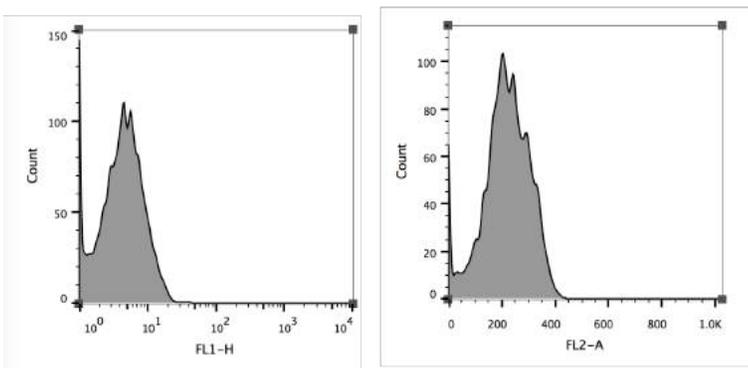
L. *G. thermoglucosidans* DL44 *RpIS+PpFbFP*. Median FL1-H= 1.72, FL2-A= 135



M. *G. thermoglucosidans* DL44 *Idh+sfGFP*. Median FL1-H= 89.8, FL2-A= 144



N. *G. thermoglucosidans* DL44 *RpIS+sfGFP*. Median FL1-H= 334, FL2-A= 140



O. *G. thermoglucosidans* DL44 *RpIS+mCherry*. Median FL1-H= 4.37, FL2-A= 219

Figure 31. Histograms A-O obtained from flow cytometry analysis of the various promoter /reporter protein combinations. FL1-H indicates green fluorescence. FL2-A indicates red fluorescence.

3.3. DISCUSSION

The advantages/disadvantages of each of the promoters and reporters used in this work will be discussed in detail below.

3.3.1. PROMOTERS FOR *G. THERMOGLUCOSIDANS*

In an effort to try to understand why the *ldh* promoter is not functional in *E. coli*, the analysis of the sequence of the RBS using the reverse engineering function produced inconclusive results and due to the uncertainty regarding the strength of the RBSs, the next possible explanation would be that the *ldh* promoter from *G. stearothermophilus* NCA1503 doesn't work in *E. coli* due to *E. coli*'s metabolism. *E. coli* represses the *ldh* promoter according to Toyoda *et al.* [2009]. Therefore, growing *E. coli* aerobically means that the *ldh* promoter is repressed because the production of NADH is not needed. This theory was not pursued in the present work due to the availability of other promoters such as p43 or RpIS that have been shown to be functional in both *E. coli* and *G. thermoglucosidans*.

Since the formation of lactate is a fermentative pathway, the *ldh* promoter in theory should be induced only by low aeration. However, at 55 °C, the solubility of oxygen is significantly lower than at lower temperatures. For instance, at 35 °C, the solubility of oxygen in fresh water (0% salinity) is 7 mg/L and in sea water (3.5% salinity) it's 5.6 mg/L. Whereas at 50 °C, the solubility of oxygen in fresh water (0% salinity) is 5.6 mg/L and in sea water (3.5% salinity) it's 4.6 mg/L⁷, therefore a fully aerobic culture cannot be achieved when growing *G. thermoglucosidans* at 55 °C. For this reason, it acts as a constitutive promoter in *Geobacillus* spp. [Bartosiak-Jentys 2012]. It has been used in previous publications [Bartosiak-Jentys, 2012 and Crowhurst, 2010] to express the reporter protein pheB. The combination of *ldh* promoter with pheB_RBS appears not to be context dependent since expression of both sfGFP and pheB [Bartosiak-Jentys, 2009] were obtained when these two bioparts were placed in combination. The *ldh* promoter does not have activity on *E. coli*, and depending on the application, this might either be an advantage. For instance, when *E. coli* is used exclusively as a cloning host and it is not required to express any protein in *E. coli*, or if the plasmid cannot be assembled because the protein is toxic for *E. coli*, then the *ldh*

⁷ http://www.engineeringtoolbox.com/gases-solubility-water-d_1148.html

promoter should be the promoter of choice. On the other hand, the RplS promoter has activity in both *E. coli* and *G. thermoglucosidans*, so when paired with a reporter protein, it facilitates quick selection of fluorescent transformant *E. coli* and *G. thermoglucosidans* colonies and the same applies for the p43 promoter. It was not an aim of this work to characterize the activity of promoters, so their strengths were not compared.

In addition, it is possible to tell from the results obtained so far that the p43 and RplS promoters, coupled to the pheB-RBS will successfully drive transcription and translation of potentially any coding sequence placed downstream, given that they have successfully directed expression of proteins of diverse sequences (pheB, sfGFP, mCherry, BsFbFP), so it can be assumed that these parts in combination create minimal secondary structures that interact with the CDS and therefore can be appropriate to be used for future heterologous expression or as references for characterization efforts.

3.3.2. REPORTERS FOR *G. THERMOGLUCOSIDANS*

3.3.2.1. SUPERFOLDER GREEN FLUORESCENT PROTEIN (sfGFP)

sfGFP is an ideal reporter to characterize promoters aerobically in *Geobacillus spp.* This work confirms that sfGFP is expressed efficiently and that the protein is stable at 55 °C, therefore it will be included in the *E. coli/G. thermoglucosidans* toolbox. Flow cytometry should be used to measure sfGFP output when characterizing promoters aerobically, however, it cannot be used to characterize promoters anaerobically because the correct formation of the fluorophore and thus the emission of green fluorescence, is dependent on molecular oxygen.

According to Andrews *et al.* [2007], the de novo folding of sfGFP occurs before chromophore formation but the refolding of GFP after it has been denatured by heat and urea, occurs in the presence of the already cyclized chromophore. In this work, in the experiments that attempted to refold GFP from anaerobic cultures, the absence of fluorescence after reintroducing oxygen could mean that the chromophore has formed without oxygen and did not mature, so the de novo folding of sfGFP occurred without the presence of oxygen and the chromophore didn't form due to the absence of molecular

oxygen. The attempted refolding of sfGFP could not occur due to absence of a cyclized chromophore.

3.3.2.2. RED FLUORESCENT PROTEIN (mCherry)

The presence of mCherry in *E. coli* can be detected by eye because the colonies look pink/purple, just as sfGFP can be detected by eye because *E. coli* colonies look light green, at least when expressed under the RplS promoter. The presence of the reporters can be confirmed by looking at the plates under a table-top blue light transilluminator, which will excite the fluorophores and the difference in coloration compared to non-fluorescent cells is greater. This is an advantage because colonies on a plate can be screened instantly without having to run every colony on a flow cytometer or plate reader just to check whether red or green (or both) transformant colonies have been obtained and to easily separate them from background non-fluorescent colonies, an experiment which was carried out in Chapter 5.

3.3.2.3. ANAEROBIC FLUORESCENT PROTEINS: LOV-DOMAIN REPORTERS

The LOV-domain reporters tested in this work were isolated from three different bacterial species: *Bacillus subtilis*, *Pseudomonas putida* and *Thermosynechococcus elongatus*. As indicated by Losi *et al.* [2006] it appears that LOV-domain proteins have a very ancient origin and their sporadic distribution among bacterial species could be explained by several independent endosymbiotic events or perhaps instances of horizontal gene transfer.

Of the four LOV-domain proteins tested, only BsFbFP could potentially be used as an anaerobic fluorescent reporter in *G. thermoglucosidans* as long as the cultures are grown at 45 °C and fluorescence of the cells is measured by flow cytometry, but this would be inappropriate considering that the downstream long term use for *G. thermoglucosidans* is in industrial fermenters that are maintained at 60 °C.

The absence or weak fluorescence observed for LOV-domain proteins could probably be due to incorrect folding during expression and/or the inability to retain a functional structure due to high temperatures. However, they could be good candidates for reporter genes in *E. coli*, since, during the course of the experiments with them, it was observed that

the proteins degrade and the fluorescence is lost in cultures that have been in stationary phase for more than 12 hours. For some applications, this could be an advantage. sfGFP does not degrade, so the accumulation of sfGFP over time does not allow to quantify promoter activity in exponential phase alone or in stationary phase alone. Or for example, in the case of biosensors which exhibit fluorescence as an output, using sfGFP makes the biosensor a one-time use application because the reporter doesn't degrade, but having a reporter that does degrade could make a biosensor reusable.

Flow cytometry confirmed that the Song *et al.* hbLOV does not fold at high temperatures, highlighting the fact that a thermostable protein isn't necessarily thermophilic, and it needs to be both to function in a thermophile. *FoldX's predictions for thermostability of hbLOV did not seem to be correct because the models only seems to take into account the stability of the structure once it has folded, but it does not take into account that kinetics during folding are different for thermophiles due to the high temperatures.*

It is difficult to understand why hotLOV showed no fluorescence in *G. thermoglucosidans*. Especially when taking into consideration that hotLOV is derived from a host whose optimal growth temperature is 55 °C. It is difficult to accept the explanation that the absence of fluorescence is due to the protein not being thermophilic. Folding should not have been a problem.

Endogenous FMN depletion cannot be the reason the proteins are not fluorescent, given that there was no fluorescence after FMN was supplemented to the media, unless the extracellular FMN was not incorporated by the cells. However, sequencing confirmed the presence of the cysteine codon mutation, so it is not possible that the absence of fluorescence is due to the FMN binding to the cysteine residue. And if FMN doesn't bind, it is improbable FMN is becoming sequestered so the presence or absence of FMN in *Geobacillus* is irrelevant to fluorescence. This is supported by the findings of Gaidenko *et al.* in 2009, who confirmed that the cysteine 62 is required for YtvA to function as a regulatory element involved in the stress response. Which indicates that without the cysteine, the protein is inactive. Therefore, it is not possible either that *G. thermoglucosidans* is not fluorescent because the LOV-proteins have unwanted interactions with the stress signalling pathway in this host.

A second option to explain the lack of fluorescence could be due to the intracellular environment of *G. thermoglucosidans*. It is possible that the internal pH is causing misfolding. The most likely explanation is that none of the LOV proteins work in *G. thermoglucosidans* for the same reason but more research is needed to understand it.

In the Song paper, it is evident that the formation of new van der Waals interactions as well as hydrogen bonds helped confer thermostability to the protein. It was very interesting to see how new interactions between amino acids formed after mutations were introduced and while in some cases they resulted in the reduction of fluorescence, as it was to be expected, in other cases, such as for the triple mutant N107Y-N124Y-M111F, the fluorescence remained, indicating that strong binding between the two domains is required. The stronger the binding, the more stable. The misfolding at 55 °C could be explained because high temperatures make difficult the permanence of van der Waals interactions and hydrogen bonds given that there is higher mobility among the components of proteins, which is why chemical reactions catalyzed by enzymes occur faster, according to the Arrhenius equation.

The Western Blot included as **Figure 29** clearly shows that the purified PpFbFP is bigger (about 30 KDa) than the expected 18 KDa. The incorrect size obtained could be explained due to misfolding, which might have caused unwanted interactions between neighbouring amino acids that generated a three dimensional structure with the incorporation of additional unwanted domains from either other PpFbFP proteins or from other proteins within the cell.

A weakness of this work is that a tag was not added to BsFbFP, hotLOV and hbLOV so it was not possible to confirm the presence of the protein in cell lysates from *G. thermoglucosidans* by Western blotting.

Throughout this work it has been believed that it is not so important that the LOV-domain proteins are involved in a signalling response in their respective bacterial hosts because in this work we are expressing the non-active mutated form of the protein and as Gaidenko [2006] points out, the signalling response cannot be triggered without the interaction between FMN and the cysteine 62 of the LOV domain. However, it might be possible that these proteins are triggering a signalling response nonetheless, or it might be possible

that they are sequestering too much of the cell's supply of FMN. There must be a reason that explains why hosts like *E. coli*, as reported by Drepper *et al.* [2007] are degrading the BsFbFP proteins when heterologously expressed by them. A possible explanation might be that the expression of these proteins might be having a negative effect on the cell's physiology and one of the ways of surviving would be by degrading this protein that might be very detrimental to their health or that might be interfering too much with their own signalling pathways and which is completely non-essential or them. That is an unlikely explanation however, because the strains of *E. coli* used (DH10B) have mutations that make them less likely to generate a response to a stressful input, such as a toxic heterologously expressed protein. If the burden is too much, then the cells might not survive or they might grow too slowly, in which case, colonies would not have been seen on a plate after overnight incubation. Therefore, the most likely explanation for the absence of fluorescence is that the proteins are misfolding when expressed at temperatures of 45 °C or higher. 45 °C is the lowest temperature at which *G. thermoglucosidans* will grow, and therefore, it was not possible to try a lower incubation temperature. This hypothesis might explain why some variants like BsFbFP are slightly more fluorescent than others, *i.e.*, it has to do with their structure. All these variants were chosen to be tested as reporters in *G. thermoglucosidans* because they all have a different amino acid sequence and therefore their three-dimensional structures would be very different, so it was possible that one of these structures might have had a higher degree of interactions among neighbouring amino acids that conferred it more robust folding kinetics. This seems to have been the case for BsFbFP. Even though it is not very robust, it seems to fold better at higher temperatures than the PpFbFP, hotLOV or hbLOV counterparts. It is possible that the three-dimensional structure of BsFbFP, which was originally derived from *Bacillus subtilis*, is more agreeable with *G. thermoglucosidans* due to their common evolutionary ancestry.

Protein degradation might be an explanation for low fluorescence, given that it might be possible that some degradation in *E. coli* and *G. thermoglucosidans* occurs, seeing as Drepper *et al.* (2007) reported that when BsFbFP was tested in *Rhodobacter capsulatus*, a Gram-negative mesophilic bacterium, the levels of fluorescence due to protein accumulation were very low, possibly due to protein degradation or inefficient gene expression.

The next Chapter (Chapter 5) focuses on the development of parts of the toolbox that are not directly involved in gene expression, but are vital if *G. thermoglucosidans* is to

become a chassis for synthetic biology, and those bioparts are constituents of the plasmid backbone required for any genetic modification in this thermophile.

3.4. CONCLUSIONS

The main contributions of this chapter are:

- sfGFP can be used as a reporter in *G. thermoglucosidans* under aerobic growth conditions but not anaerobic.
- mCherry can be used as a reporter in *G. thermoglucosidans* under aerobic growth conditions but not anaerobic.
- sfGFP and mCherry have excitation/emission wavelengths that do not overlap, so can be used when two reporter need to be expressed simultaneously.
- mCherry could be used as a reporter in *G. thermoglucosidans* under anaerobic growth conditions as long as the fluorescence is measured by flow cytometry and growth occurs at 45 °C.
- BsFbFP was the only LOV-domain protein that exhibited weak fluorescence in *G. thermoglucosidans*.

3.5. FUTURE WORK

The main work that needs to be carried out in the future should be the characterization of the reporters that this work has incorporated in the *E. coli/G. thermoglucosidans* toolbox, namely:

- Growth must be carried out in baffled flasks to achieve better aeration and healthier (faster growing) cultures.
- Reporter stability must be measured at a range of temperatures starting at 45 °C and ending at 70 °C.
- Different media types need to be tested to determine whether salt content might have an effect on oxygen solubility and therefore fluorescence of the reporters.

More research needs to be carried out to determine how important oxygen, salt content and temperature really are for the expression of fluorescent reporters. sfGFP requires molecular oxygen for the formation of the chromophore. mCherry also requires molecular oxygen for the formation of the chromophore. Throughout this project, reporters for

G. thermoglucosidans were routinely tested by growing in 5 mL of rich media (TGP) in falcon tubes and incubated with shaking at 200 rpm (with 45 ° inclination angle) at 55 °C. These are standard conditions for growing *G. thermoglucosidans* and they are considered "aerobic". However, throughout this project it has become increasingly evident that more aerobic conditions and healthier and faster growing cells have been obtained when growing *G. thermoglucosidans* in 50 mL of media in 250 mL baffled conical flasks, because they permit better aeration of the cultures and aeration is a problem due to the lower solubility of oxygen at high temperatures. In particular, cells expressing mCherry should be grown in baffled flasks and their fluorescence compared to cells grown in falcon tubes. It is possible that the higher permeation of oxygen might result in increased fluorescence.

4. CHAPTER 4: SHUTTLE VECTORS FOR *E. COLI* / *G. THERMOGLUCOSIDANS*

The key contributions and results of this chapter are:

1. Two sets of *E. coli* / *G. thermoglucosidans* shuttle vectors have been constructed.
2. The architectonic configuration new vectors is entirely modular.
3. The issue of host-specificity for the use of thermostable kanamycin nucleotidyltransferase in *E. coli* has been resolved.
4. The new vectors have higher transformation efficiencies than previously existing ones.
5. An entirely synthetic Multiple Cloning Site was incorporated into the shuttle vectors, which includes entirely unique restriction sites and is limited by two transcriptional terminators.

4.1. INTRODUCTION

It is urgently necessary to develop a new set of shuttle vectors for *E. coli* / *G. thermoglucosidans* because the existing vectors are extremely inefficient due to:

- Their extremely large size, which reduces transformation efficiencies.
- The lack of knowledge regarding the sequences composing them.
- The large amount of unnecessary DNA they contain.
- Their lack of modularity.
- The lack of suitable selection markers.
- The inadequacy of the current antibiotic selection markers.

This work will focus on the creation of a new set of shuttle vectors for *E. coli* / *G. thermoglucosidans* with the following aims:

- To carefully study the sequences existing plasmids for *G. thermoglucosidans* are composed of to determine the useful DNA from the unnecessary sequences.
- Create vectors that are minimal in size by identifying the minimal functional bioparts required to achieve full functionality (promoter, RBS, CDS, terminator).
- Confirm whether a smaller size of shuttle vector results in higher transformation efficiencies in *E. coli* and *G. thermoglucosidans*.
- Investigate online software tools that could be used to analyze plasmid sequences.

- Create vectors that are modular to follow the ethos of synthetic biology.
- Investigate the incorporation of new selection markers to the toolbox, namely chloramphenicol resistance and uracil auxotrophy.
- Use the knowledge generated from the abstract study of the minimal bioparts required for the construction of synthetic biology devices to develop an adequate and efficient selection marker.
- Propose potential uses for the vectors included in the toolbox.

The majority of molecular tools constructed for the genetic modification of *Geobacillus* spp. involve shuttle vectors [Kananavičiūtė and Čitavičius 2015] but the number of shuttle vectors that can be used in *G. thermoglucosidans* is limited mainly because research in this area is still in its infancy. The lack of efficient vectors hampers the exploration of the full potential of this thermophile and it's a subject regularly mentioned in *Geobacillus*-related publications that involve genetic engineering [Blanchard *et al.* 2014, Lin *et al.* 2014, Bartosiak-Jentys *et al.* 2013] because there's a fundamental need for efficient vectors.

pUCG3.8 [Bartosiak-Jentys *et al.* 2013] is the most used *E. coli*/*G. thermoglucosidans* shuttle vector by the Leak group now. Before the construction of this vector, the most commonly used vector was pUCG18 [Kananavičiūtė and Čitavičius 2015]. Both plasmids are very similar in terms of their components but they differ in size. pUCG18 is relevant to this work because it is the starting point for the creation of the new shuttle vectors and pUCG3.8 is important because it will be the reference the new plasmids are compared against. In the following section, these two plasmids will be introduced.

4.1.1. EXISTING SHUTTLE VECTORS FOR *E. COLI*/*G. THERMOGLUCOSIDANS*: PUCG18 AND PUCG3.8

The plasmid map of pUCG18 is shown in **Figure 32**. pUCG18 is composed of five essential parts detailed below:

1. **ColE1** is the pUC18 origin of replication for the well-known *E. coli* pUC18 plasmid. It is a high copy number origin of replication in *E. coli* (300-400 copies per cell). It was derived from a naturally occurring plasmid of *E. coli* and the high copy number is attained because its replication is independent from that of the chromosome

[Yanisch-Perron *et al.* 1985]. *E. coli* is used because it is a well-established transformation system.

2. ***lacZ***, encoding for β -galactosidase, allows for blue/white screening selection in *E. coli*, and was also derived from pUC18. It is located between *thermostable_kanR* and pUC_ori. In the middle of *lacZ* is the Multiple Cloning Site (MCS) for pUC18.
3. ***ampR*** is the gene that encodes for a β -lactamase, which confers resistance to ampicillin in mesophiles such as *E. coli*.
4. ***repBST1*** is the origin of replication for *Geobacillus*. It was derived from a plasmid created for *G. stearothermophilus* called pBST22 [Liao and Kanikula, 1990]. The exact copy number of plasmids with this origin within cells has not yet been determined.
5. ***Thermostable_kanR*** encodes a thermostable kanamycin nucleotidyltransferase, which confers resistance to kanamycin in thermophiles and was also derived from pBST22. The mutations that confer thermostability are Asp-80-Tyr and Thr-130-Lys and the protein becomes denatured at 65 °C [Liao and Kanikula, 1990].

As the name indicates, half of pUCG18 actually corresponds to the pUC18 plasmid from *E. coli*. The vector is essentially a fusion between pUC18 and pBST22 so the plasmid is significantly big, which is its main disadvantage: 6332 bp is significantly large for an empty vector and its size can only increase if cargo is added to the MCS. This is a problem because the larger the plasmid, the lower the transformation efficiency [Hanahan, 1983; Szostkova, 1998]. In an attempt to solve this problem, Bartosiak-Jentys *et al.* [2013] created a smaller version of pUCG18 and it was named pUCG3.8. The plasmid map for pUCG3.8 is included in **Figure 33**. Small plasmids for the genetic transformation of *Geobacillus* spp. are of the utmost necessity because their thick peptidoglycan layer and low permeability of the plasma membrane are the main hindrances for DNA transfer to cells (Lian and Xu, 2013).

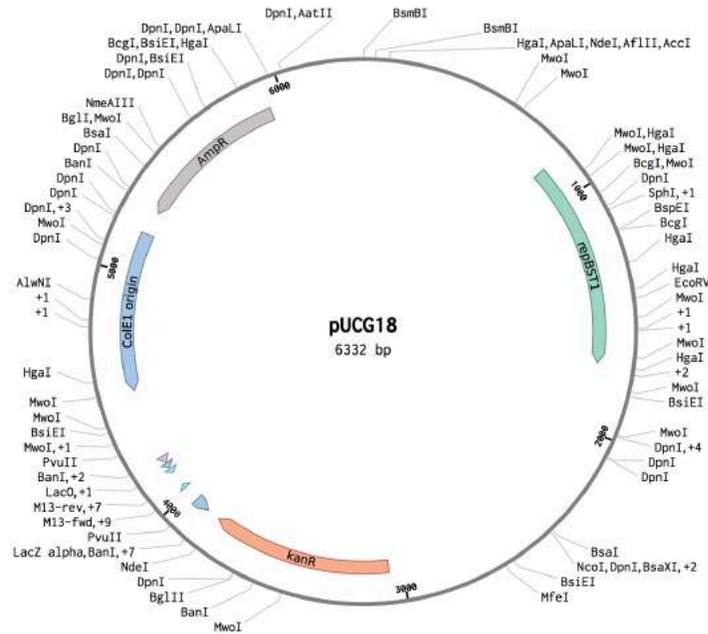


Figure 32. Shuttle vector pUCG18. The plasmid is composed of the same four essential parts: *ColE1* is an origin of replication for *E. coli*, *repBST1* is an origin of replication for *G. thermoglucosidans*, *AmpR* is a mesophilic ampicillin resistance gene and *kanR* is a thermostable kanamycin resistance gene. The multiple cloning site (MCS) and *lacZ* are located between *kanR* and pUC18_ori and they are derived from pUC18. The size of the empty plasmid is 6332 bp. Recognition sites for common restriction enzymes have been included.

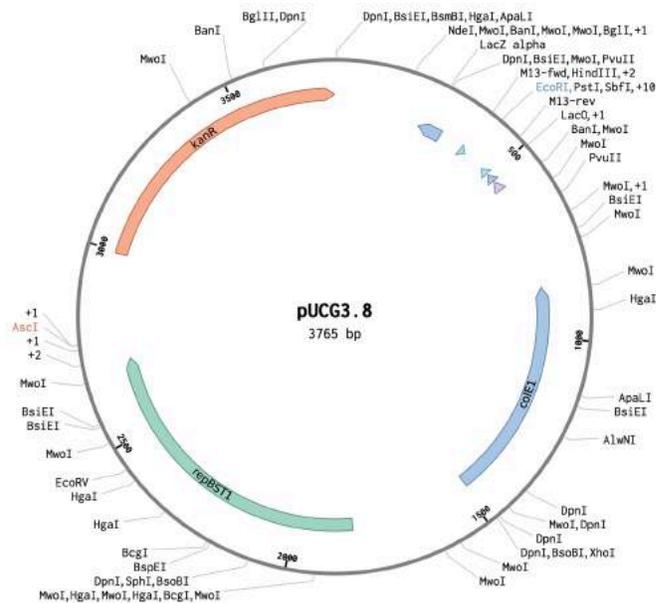


Figure 33. Plasmid map of pUCG3.8. The shuttle vector is composed of the same four essential parts as pUCG18: *ColE1* is an origin of replication for *E. coli*, *repBST1* is an origin of replication for *G. thermoglucosidans* and *kanR* encodes a thermostable kanamycin nucleotidyltransferase. The multiple cloning site (MCS) and *lacZ* are located between *kanR* and pUC18_ori and they are from pUC18. The difference between pUCG18 and pUCG3.8 them is the absence of *AmpR*, the mesophilic ampicillin resistance gene, in the latter. Recognition sites for common restriction enzymes have been included.

pUCG3.8 was **constructed** because an even smaller vector was necessary to express cellulases in *G. thermoglucosidans*, which often are extremely large genes [Bartosiak-Jentys, *personal communication*]. pUCG3.8 is composed of same parts as pUCG18: *ColE1*, *repBST1*, *kanR*, and *lac(MCS)Z* located after *kanR*. Ampicillin selection was left out. *However*, this vector incorporated some modularity: an *AsiI* restriction site was placed between *kanR* and pUC18_ori, *XhoI* was placed between pUC18_ori and *repBST1* and *AscI* separated *repBST1* from *kanR*. But pUCG3.8 has three problems: (1) new parts cannot be easily incorporated and (2) the MCS contains many restriction sites that are common and which cut parts that could be placed within it like promoters, reporters or genes encoding cellulases. And (3) even though high transformation efficiencies were reported (2.8×10^5 c.f.u./ μg), these transformations efficiencies are for *G. thermoglucosidans*, not for *E. coli*. Kanamycin selection in *E. coli* with this vector is actually extremely troublesome. Very low transformation efficiencies are achieved **as it will be evidenced in the results section of this chapter**. For applications where high transformation efficiency is required, for example when **creating a library of synthetic bioparts**, it would definitely present a problem. The Leak group has tried to overcome this problem by using LB medium supplemented with glucose to recover cells but low transformation efficiencies in *E. coli* are still an issue.

4.1.2. ARE TWO ANTIBIOTIC RESISTANCE GENES REALLY REQUIRED?

- Ideally, only one resistance marker should be useful for both hosts, but that cannot always be the case.
- Shuttle vectors typically carry two origins of replication and either one or two antibiotic resistance selection markers because the expression or resistance profile of these selection markers is host-specific [Kananavičiūtė and Čitavičius 2015].
- Kanamycin resistance would be an ideal selection marker for both *E. coli* and *G. thermoglucosidans* but for the moment, it is functional only in *G. thermoglucosidans*. The efficiencies obtained with *E. coli* are poor (as it will be demonstrated in the results section of this chapter).
- This project will attempt to use the knowledge generated by the engineering side of synthetic biology to attempt to break the host-specificity barrier of *kanR*.

Throughout the development of plasmids for thermophiles, it has been understood that the standard *E. coli* antibiotic selection markers like β -lactamase (AmpR) are unsuitable as markers for thermophilic organisms because either the protein or the antibiotic itself are thermolabile. That's not the problem for kanamycin resistance though: kanamycin is the most thermostable of the commonly-used antibiotics and the KanR used in pUCG18 is a double mutant more thermostable than its parent [Liao and Kanikula, 1990].

When *E. coli* is transformed with pUCG18 it is typical to find fewer colonies when selecting on kanamycin than on ampicillin. The present work seeks to answer if a synthetic biology approach could elucidate the reason *E. coli* cannot be selected on kanamycin when transformed with either pUCG18 or pUCG3.8. The section below contains a glimpse of the problem.

4.1.3. CHLORAMPHENICOL RESISTANCE IN THERMOPHILES

The use of chloramphenicol resistance selection in thermophiles is problematic but not impossible. For instance, Turner *et al.* [1992] mention that the chloramphenicol acetyltransferase gene encoded in pMK4 was able to confer resistance to chloramphenicol in *G. stearothermophilus* at 70 °C and that this enzyme retained 90% of its activity after being heated at 70 °C for one hour, which suggests that this chloramphenicol acetyltransferase should be tested in *G. thermoglucosidans*.

4.1.4. PMK3 AND PMK4

pMK3 and pMK4 are *B. subtilis/E. coli* shuttle vectors (NCBI accession numbers EU549779 and EU549778, respectively). The plasmids maps for both vectors have been included below (Figures 34 and 35). They are relevant for this work because three parts will be isolated from them and tested for inclusion in the toolbox. These three parts are:

- **repB**- temperature sensitive origin of replication useful for doing knock-outs in *G. thermoglucosidans* as explained in Chapter 6 of this work.
- **repC**- an origin of replication which will be tested for functionality in *G. thermoglucosidans*.
- **catE**- chloramphenicol acetyltransferase which will be tested for functionality in *G. thermoglucosidans*. It is not functional in *E. coli* [Gryczan *et al.* 1978].

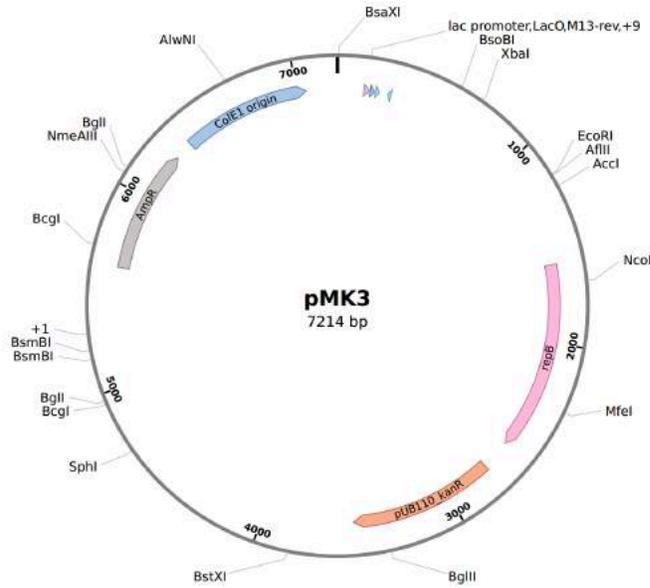


Figure 34. Plasmid map of pMK3. pMK3 is an *E. coli/B. subtilis* shuttle vector [Sullivan *et al.* 1984] containing an origin of replication for *E. coli* (*ColE1*), β -lactamase (*ampR*) for selection in *E. coli*. For selection in *B. subtilis*, the plasmid encodes a kanamycin nucleotidyltransferase (*kanR*). The plasmid also contains an origin of replication (*repB*) derived from a *Staphylococcus aureus* cryptic plasmid. The sequence can be found under NCBI accession number EU549779.

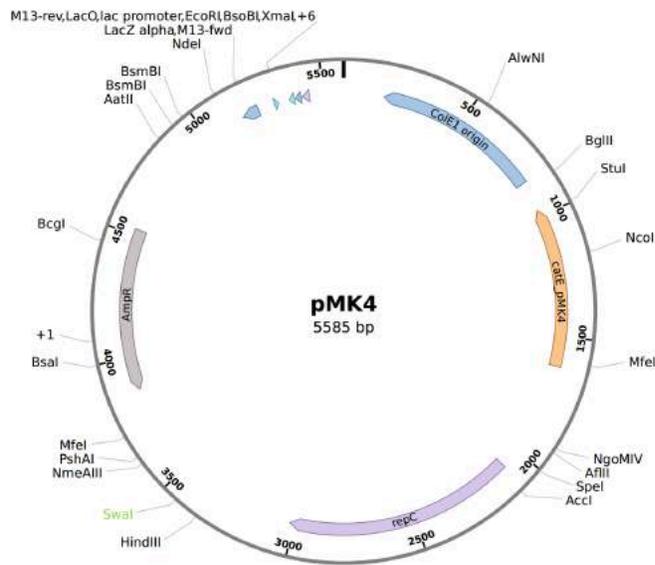


Figure 35. Plasmid map of pMK4. pMK4 is an *E. coli/B. subtilis* shuttle vector [Sullivan *et al.* 1984] containing an origin of replication for *E. coli* (*ColE1*), β -lactamase (*ampR*) for selection in *E. coli*. For selection in *B. subtilis*, the plasmid encodes a chloramphenicol acetyltransferase (*catE*), which is not a functional in *E. coli*. The plasmid also contains an origin of replication (*repC*) derived from a *Staphylococcus aureus* cryptic plasmid. The plasmid sequence can be found under NCBI accession number EU549778.

The first requirement for the development of *G. thermoglucosidans* as a chassis for synthetic biology is the construction of a toolbox of genetic components and the starting point for that is the creation of shuttle vectors. Their design and the method used for their assembly are discussed in the results section of this chapter, found below.

4.2. RESULTS

The results obtained in this work can be divided into the following subjects:

1. Analysis of DNA sequence of existing plasmids to determine the minimal composition.
2. Design of the shuttle vectors based on:
 - How will modularity for each vector be maintained?
 - What are components that will be included in the toolbox?
 - How will these components be split into shuttle vectors?
3. Assembly of the shuttle vectors.
4. Testing of the shuttle vectors.

4.2.1. ANALYSIS OF THE DNA SEQUENCE OF EXISTING PLASMIDS

- Existing plasmids were analyzed to identify the functional units comprising them (*i.e.* promoters, RBSs, CDS, terminators) in order to identify the minimal sequences required to assemble a minimal plasmid.

It was decided that the first modular plasmid to be constructed for *Geobacillus* would be an improved version of pUCG18. This plasmid was selected as a reference because it was the most commonly used shuttle vector plasmid in the Leak Group at the time when this research project began. Initial work was undertaken to carefully analyze the entire sequence of pUCG18 in order to be able to draw a delimiting line between each of the parts it is composed of, and to identify important non-coding regions, such as promoters, transcription factors and transcriptional terminators, that should not be removed unless necessary.

Annotations for the coding sequences (hereinafter CDS) were found using the NCBI database. In particular, accession number EU547236 refers to the sequence of pUCG18 [Taylor *et al.* 2008]. Potential promoter regions were identified using BPPROM [Softberry, Solovyev and Salamov, 2011⁸], an online tool that identifies, by comparison against a database, consensus -35 and -10 transcription site boxes for bacterial promoters. Common transcription factors are also highlighted by BPPROM. Rho-independent transcriptional terminator regions were identified by analyzing the sequences located directly after the stop codons with IDT's online Oligoanalyzer tool v.3.1⁹ specifically looking for the presence of hairpins that may arise and the results are depicted in **Figure 59**. The results of the terminator search are highlighted in **Figures A148** for *repBST1*, **A149** for *kanR* and **A150** for *ampR*.

4.2.2. THE PROBLEM WITH *KANR*

- Kanamycin selection doesn't work in *E. coli* due to the lack of a transcriptional terminator at the 3'-end of the gene in the pUCG18 and pUCG3.8 plasmids.

When analyzing the sequences of the parts that comprise pUCG18 and pUCG3.8 and trying to identify the functional units that form each part, it was discovered that perhaps the reason the kanamycin resistance gene has poor functionality in *E. coli* is because a synthetic biology approach was not applied when both plasmids were assembled.

It is not surprising that even though pUCG3.8 is smaller than pUCG18, the transformation efficiencies when selecting *E. coli* on kanamycin are similarly poor. The problem is that both vectors have the same architecture and the same basic design element is missing: there is no transcriptional terminator in either vector at the 3'-end of *kanR*. This detail was omitted in both the pUCG18 paper [Taylor *et al.* 2008] and Mark Taylor's thesis [2007].

The reason *kanR* does not have a transcriptional terminator is that when joining the two plasmids pUC18 and pBST22, the original transcriptional terminator of *thermostable_kanR* was eliminated and replaced by an *NdeI* site used to join both plasmid

⁸ [<http://linux1.sof tberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>]

⁹ [<http://eu.idtdna.com/analyzer /Applications/OligoAnalyzer/Default.aspx>]

vectors together. Therefore, there is a need for a sequence downstream from the stop codon of thermostable *kanR* that can potentially act as its transcriptional terminator. As seen in **Figure 36**, the DNA located immediately after *kanR* is that corresponding to the *lacZ* gene. The hairpin formed seems to indicate that might be a transcriptional terminator, but, given that the orientation of the *lacZ* gene is opposite to the orientation of the *kanR* gene, it's most likely that it is a unidirectional terminator for *lacZ* and it does not function as a bidirectional terminator for *kanR*. **Figure 59** shows the hairpin structure associated with a transcriptional terminator when the sequence of the *lacZ* terminator is analyzed by the IDT OligoAnalyzer tool¹⁰.

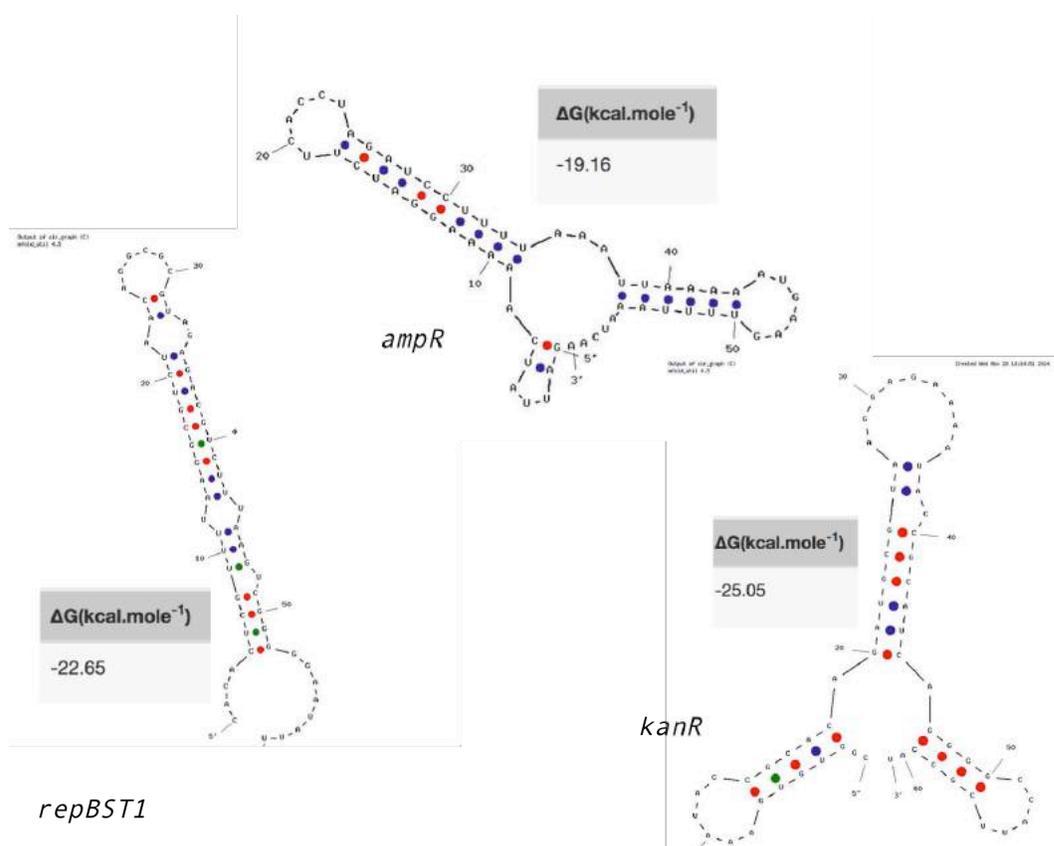


Figure 36. Secondary structures of transcriptional terminators. The RNA sequences of the terminator regions of *repBST1*, *ampR* and *kanR* as in pUCG18 and pUCG3.8 were analyzed for the formation of secondary structures by the software OligoAnalyzer¹¹. The hairpin-shaped secondary structures obtained from the modelling analysis suggest that the sequences act as transcriptional terminators by creating a physical block to the RNA polymerase. For sequence detail, visit **Figures A148-A150**.

¹⁰ [<http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/Default.aspx>]

¹¹ <http://eu.idtdna.com/analyzer/applications/oligoanalyzer/>

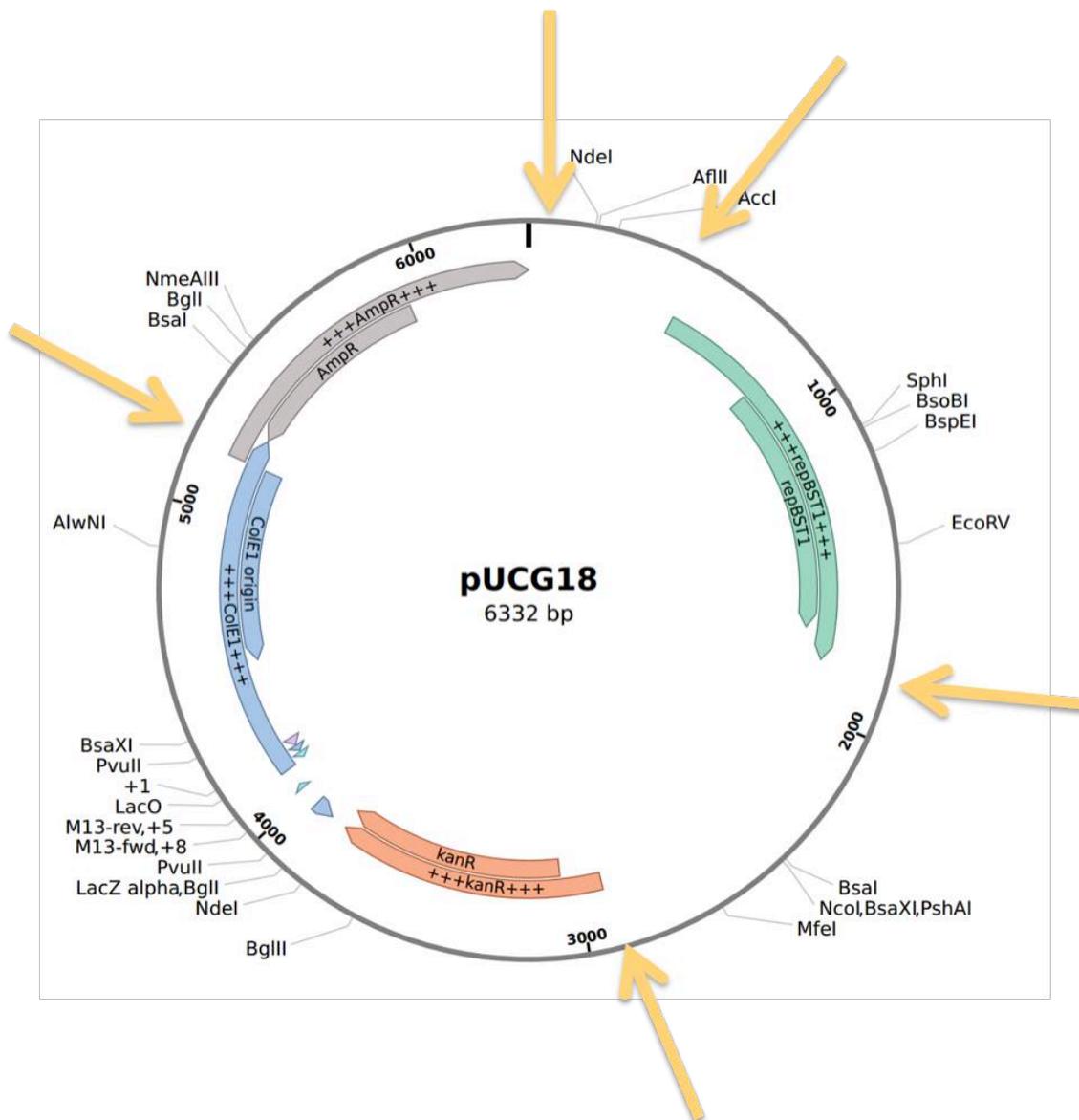


Figure 37. Plasmid map of pUCG18. Two kinds of annotations are displayed: small annotations indicate exclusively the location of the CDS for each of the parts. Larger annotations, *i.e.* annotations that have six plus signs included in the name, indicate the minimal sequence for each of the parts including promoter, CDS and terminator. Big yellow arrows indicate where restriction sites should be to permit creating a minimal vector that would allow swapping parts. Size in base pairs (bp).

The "minimal" sequence necessary to compose each of the parts was selected based on the information encoded before and after the coding sequence so that promoters and terminators for each coding sequence were also included. As it can be observed in **Figure 37**, it was determined that pUCG18 contains a lot of unnecessary DNA, especially surrounding *repBST1*.

Furthermore, from looking at **Figure 37**, it is evident that in a vector like pUCG18, replacing any one part for its “minimal” version would not be a straightforward process given the absence of existing common restriction enzyme cutting sites in key locations separating some of these parts. Replacing any part would result in an unpractical and unnecessary accumulation of one-time use plasmids each having various degrees of modification. **Therefore building modular plasmids that reduce the amount of work needed to replace one part can be a** long-term cost-effective solution and fits with the ethos of synthetic biology [Keasling, 2012].

4.2.3. DESIGN OF THE SHUTTLE VECTORS

The construction of a new set of *G. thermoglucosidans*/*E. coli* shuttle vectors was proposed with one principle in mind: modularity. Having a standardized plasmid system can provide efficiency, ease of use and will facilitate sharing parts between different research groups. The creation of a new set of plasmids needs to start by having one modular backbone to begin with, not only because the possibility of exchanging parts within it then exists, but also because it sets the rules for standardizing what all the following parts must conform to if they are to contribute to the expanding toolkit.

In order to design this first modular backbone, attention was focused on existing design norms for plasmids suggested in the literature. One such example is SEVA, described below.

4.2.3.1. PSEVA

pSEVA is a design for a plasmid for one host, not for a plasmid like a shuttle vector that involves two hosts, so the pSEVA design cannot be implemented to the letter in this work, however:

- The Multiple Cloning Site MCS of the new modular vectors must be compatible with the pSEVA architecture because it will make easier the sharing of cargo bioparts.**
- Ease of sharing cargo bioparts may in turn facilitate the adoption of *G.thermoglucosidans* as a thermophile chassis.**

The Standard European Vector Architecture (SEVA) was published in 2013 by Silva-Rocha *et al.* and updated in 2014 by Martínez-García *et al.* and it is a guideline scaffold for the design of plasmids which intends to facilitate the swapping of parts. The basic design of a pSEVA vector is shown in **Figure 38**. The ideal of SEVA is that all plasmids use the same and a specific standard nomenclature. However, the original pSEVA design allows for one antibiotic marker and one origin of replication. For the toolbox shuttle vectors, two origins and two antibiotic resistance markers need to be incorporated so the pSEVA designed will not be followed except for the Multiple Cloning Site because it should facilitate the sharing and testing of bioparts between the *Geobacillus* and the wider synthetic biology community. Therefore, the design elements of SEVA incorporated into the toolbox design were the *AscI* and *SwaI* restriction enzyme recognition sites to delimit the MCS.

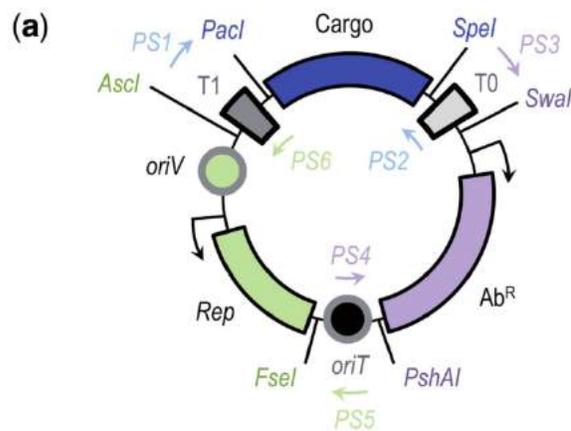


Figure 38. The basic design of a pSEVA plasmid. In pSEVA plasmids, the functional parts are separated by the rare-cutter restriction sites *AscI*, *Pacl*, *SpeI*, *SwaI*, *PshAI* and *FseI*. Two rho-independent transcriptional terminators T1 and T0 delimit the cargo region. *oriT* permits the conjugative transfer of DNA between *E. coli* and alternative hosts. The new modular plasmids will also contain an MCS delimited by the restriction enzyme *AscI* and *SwaI* recognition sites.

Figure taken from Silva-Rocha *et al.* [2013].

Rare-cutting restriction enzyme recognition sites *NotI*, *PmeI*, *SwaI*, *AscI*, *FseI* were identified as appropriate enzymes to cut at sites at the boundary of parts because their 8 bp recognition sequences were absent from all parts to be included in the vector set. In addition, they are commercially available and with the exception of *SwaI*, are all compatible with the same buffer (New England Biolabs).

The architecture of the first modular plasmid was designed to resemble that of pUCG18 as closely as possible, so that its efficacy could be compared to this existing internal standard.

4.2.3.2. FIRST MODULAR PLASMID DESIGN

The first modular plasmid follows the design seen in **Figure 39**. The design of this first plasmid was to have the sequences *ColE1* delimited by the *SwaI* and *FseI* sites, *ampR* by *FseI* and *NotI*, *repBST1* by *NotI* and *PmeI* and *kanR* between *PmeI* and *AscI*. A fifth building block would also be added between these main four plasmid parts. This fifth block would be a Multiple Cloning Site (MCS) containing many restriction enzyme sites to allow the plasmids created to be easy to use as shuttle vectors. The obvious way to include DNA sequences into shuttle vectors is to design a zone designated for this purpose, with plenty of unique restriction sites to choose from. The MCS included in the set is represented in **Figure 40**. The MCS was designed entirely *de novo* to contain a variety of restriction sites that don't inconveniently cut any of the parts expected to be parts of the *toolbox*. The MCS was designed to be placed between *kanR* and *ColE1*. Additionally, the wild type RplS promoter as well as the Song *et al.* hbLOV fluorescent protein were included so that it could be immediately assayed for being able to host an expression cassette. The entire sequence was synthesized by GeneART (Life Technologies).

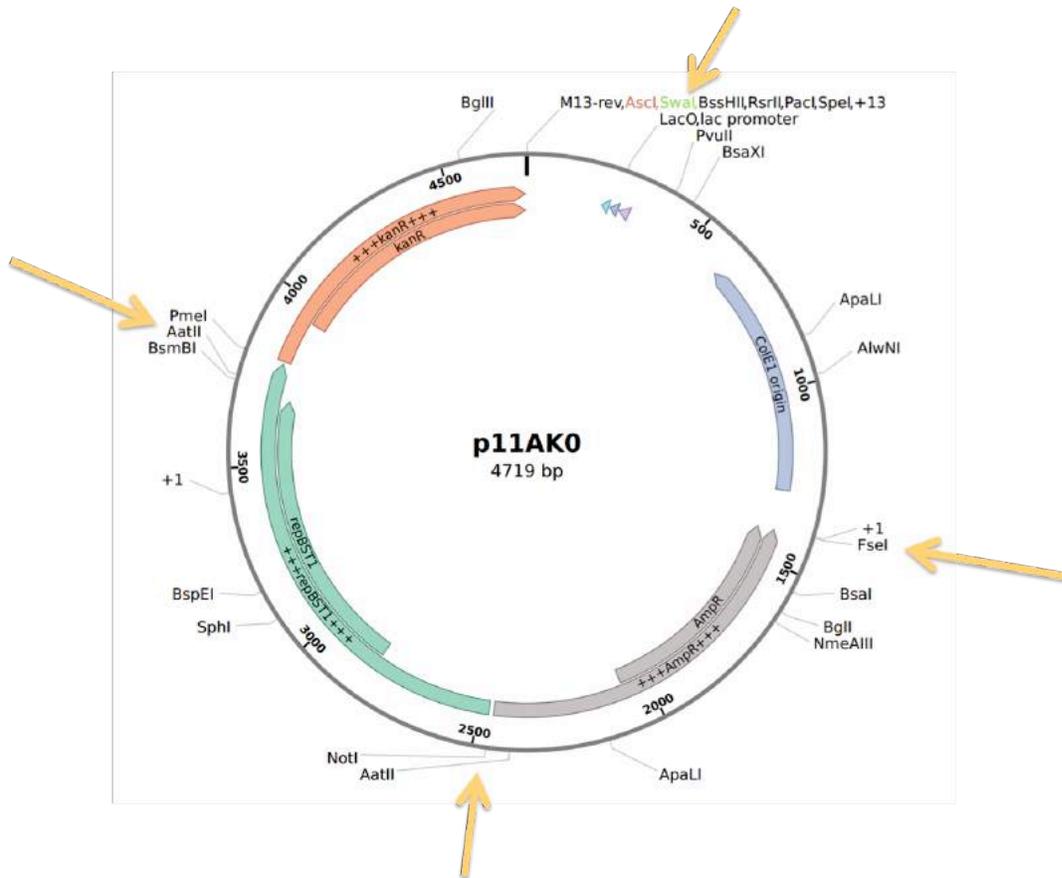


Figure 39. Plasmid map of the first modular plasmid. The first modular plasmid was designed to resemble the architecture of pUCG18 as closely as possible. Two kinds of annotations are displayed: small annotations indicate exclusively the location of the CDS for each of the parts. Larger annotations, *i.e.* annotations that have six plus signs included in the name, indicate the minimal sequence for each of the parts including promoter, CDS and terminator. Big yellow arrows indicate the location of the rare-cutter restriction sites that confer modularity to the vector, namely, PmeI, Ascl, SwaI, FseI and NotI. Size in base pairs (bp).

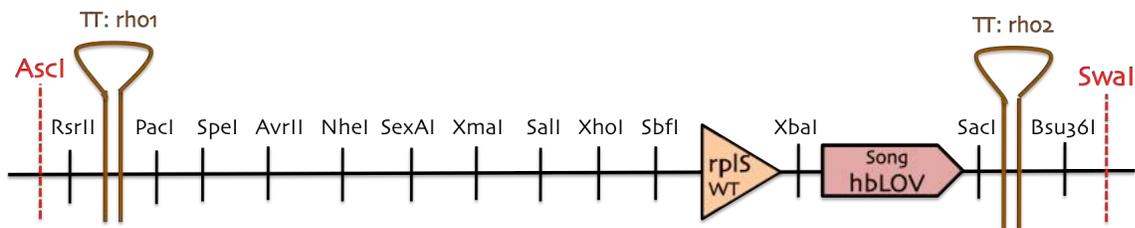


Figure 40. Design of the synthetic MCS to be included in the modular plasmids. The wild-type RplS promoter was placed between the *SbfI* and *XbaI* sites and the sequence for the Song *et al.* hbLOV was placed between the *XbaI* and *SacI* sites. Two rho-independent transcriptional terminators (rho1 and rho2) were placed delimiting the MCS. Other restriction sites included within the transcriptional terminators are *PacI*, *SpeI*, *AvrII*, *NheI*, *SexAI*, *XmaI*, *Sall*, *XhoI*, *SbfI*, *XbaI* and *SacI*. In addition, the *RsrII* and *Bsu36I* restriction sites will permit placing parts outside of the transcriptional terminators.

Within the synthesized part, two rho-independent transcriptional terminators designated “rho1” and “rho2” delimit the central region of the MCS. Their specific location in the MCS region can be seen in **Figure 40**. The hairpins they form can be seen in **Figure 41**. The first one, “rho1”, has the purpose of terminating transcription of the *kanR* gene. The second one, “rho2”, terminates transcription of genes included between the *PacI* and *SacI* sites if they are in forward orientation.

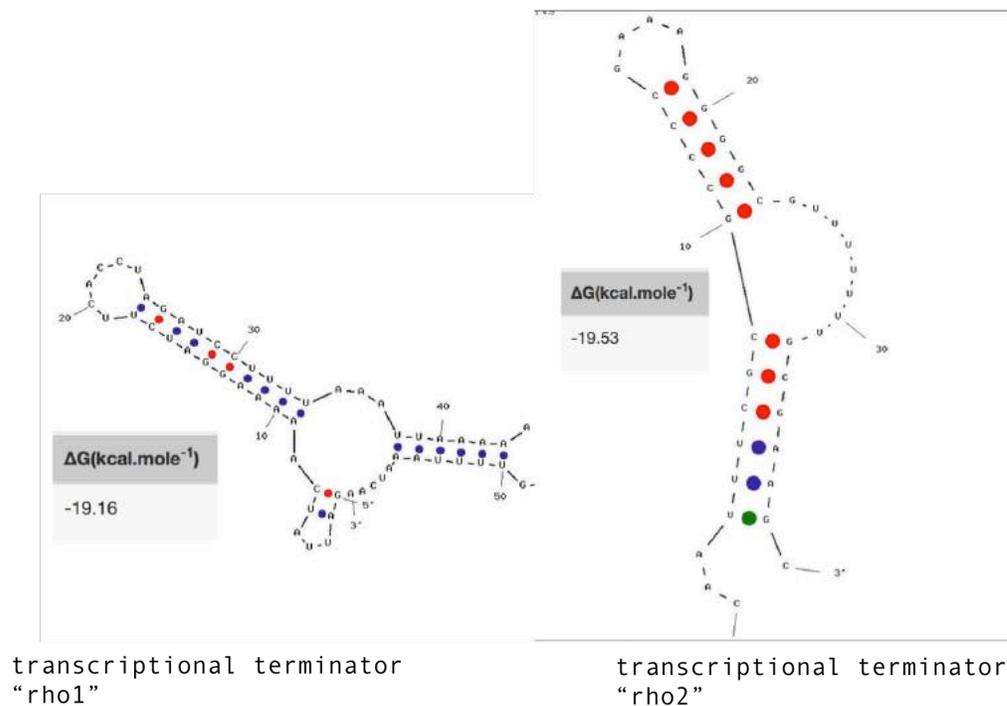


Figure 41. Secondary structures of transcriptional terminators. The RNA sequences of the two rho-independent transcriptional terminators included in the MCS were analyzed for the formation of secondary structures by the software OligoAnalyzer¹². The hairpin-shaped secondary structures obtained from the modelling analysis suggest that the sequences act as transcriptional terminators by creating a physical block to the RNA polymerase.

4.2.3.3. ASSEMBLY OF THE PLASMIDS OR DNA SYNTHESIS?

For the creation of a set of modular *Geobacillus* plasmids, various new DNA assembly standards and techniques, such as MODAL (Modular Overlap Directed Assembly with Linkers), Golden Gate and Gibson Assembly were considered to construct the modular

¹² <http://eu.idtdna.com/analyzer/applications/oligoanalyzer/>

plasmids. DNA synthesis of the entire plasmid was not contemplated because the cost would have been significantly higher than the price paid for the necessary oligonucleotides required for PCR amplification plus reagents. It was concluded that the easiest and most accessible way to allow plasmids to not only be built but to allow plasmid parts to be quickly swapped in and out would be traditional digestion and ligation that leaves standard restriction cutting sites between different parts. Therefore, in order to base the modularity of this vector on restriction cutting, suitable restriction sites needed to be chosen. The appropriateness of rare-cutters is accentuated in this situation as their target sequences are longer than the usual 6 bp of common type II restriction endonucleases, and are therefore much less often encountered within existing DNA sequences.

4.2.4. CONSTRUCTION OF THE FIRST MODULAR PLASMID

To construct the plasmid, an approach was taken to build up a modular equivalent of pUCG18 in one go from all the parts necessary. All parts were PCR amplified from pUCG18 (except the synthesized MCS) and the primers used to amplify each part are included in **Table A12**. Primers and PCRs were designed to remove unnecessary DNA between pUCG18 parts and add in the necessary restriction sites between each region. The approach required a 5-part ligation to work to give any plasmid that could be further modified. Despite trying repeated times; this always resulted in no colonies, probably due to the complexity of a 5-part assembly (data not shown). So in order to construct the vector, a second approach was taken where pUCG18 would be iteratively modified by PCR mutation until it became compatible with the intended modular design and contained the necessary restriction sites in the right places.

Figures 42-50 illustrate the step-by-step process used to create the first modular plasmid, hereafter referred to as "p11AK1". First, as shown in **Figure 42**, pUCG18 was PCR-amplified with primers *Ascl-Swal-pUC_ori* and *RC-kanR-Ascl*. The forward primer encoded *Ascl* and *Swal* sites. The reverse primer encoded an *Ascl* site.

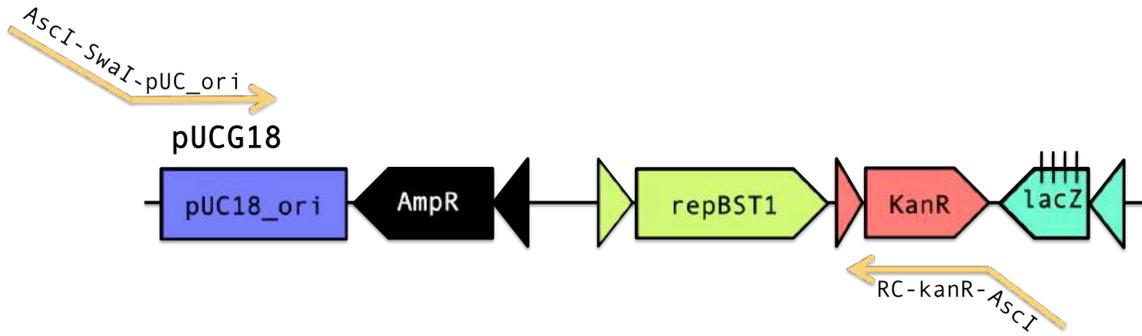


Figure 42. Primers *AscI-SwaI-pUC_ori* and *RC-kanR-AscI* were used to amplify pUCG18. The forward primer encoded *AscI* and *SwaI* sites. The reverse primer encoded an *AscI* site.

The PCR product was digested with *AscI* and self-ligated to circularize. The resulting plasmid is shown in **Figure 43**. Subsequently, it was digested with *AscI* and *SwaI*, as shown in **Figure 67**.

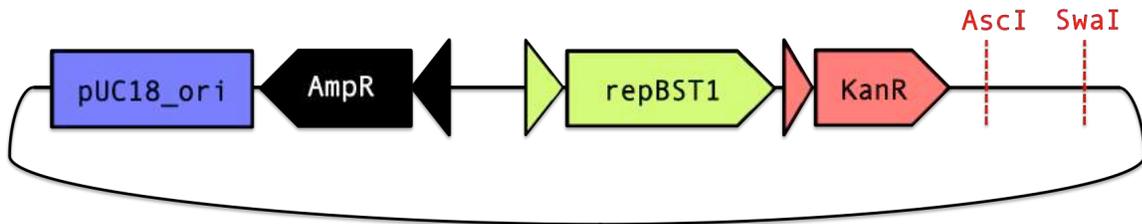


Figure 43. Plasmid resulting from the *AscI* self-ligation of the PCR product from **Figure 64**.

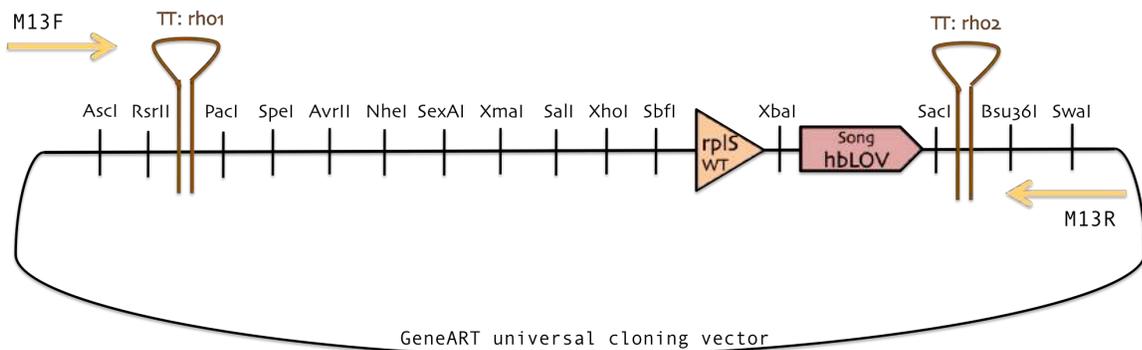


Figure 44. M13F and M13R primers were used to PCR-amplify the sequence corresponding to the synthetic MCS that included the wild-type *RplS* promoter as well as the *hbLOV* which was codon optimized for *B. subtilis*.

As seen in **Figure 44**, primers M13F and M13R were used to amplify the synthetic MCS sequence that also included the wild-type *RplS* promoter and the *Song_hbLOV*

reporter. The PCR product was digested with *AscI* and *SwaI*. These restriction sites were already included in the sequence that was synthesized. The *AscI/SwaI* plasmid and the *AscI/SwaI* MCS were ligated and transformed. The resulting plasmid is shown in **Figure 45**. *E. coli* colonies were selected based on the presence of fluorescence produced by *hbLOV* being expressed.

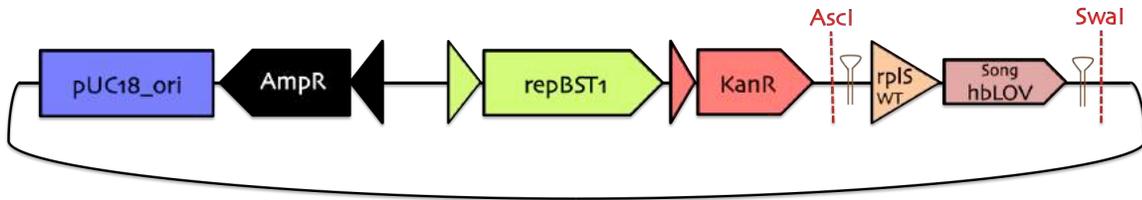


Figure 45. Plasmid resulting from the ligation of the plasmid shown in **Figure 43** digested with *AscI* and *SwaI* with the *AscI/SwaI* MCS which also included the wild-type *RplS* promoter as well as *hbLOV*.

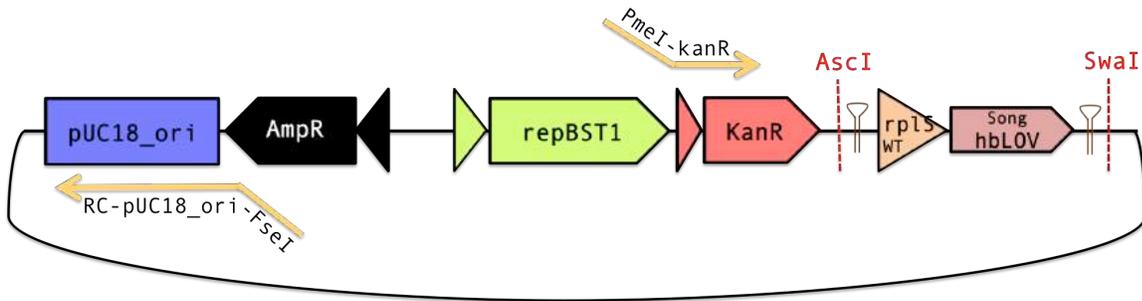


Figure 46. PCR-amplification of the plasmid from **Figure 45** with primers *PmeI*-kanR and RC-pUC18_ori-*FseI*.

The plasmid from **Figure 45** was PCR amplified to add *PmeI* and *FseI* restriction sites with primers *PmeI*-kanR and RC-pUC18_ori-*FseI* (**Figure 46**). Likewise, the same plasmid was also amplified with primers *FseI*-ampR and RC-repBST1-*PmeI* (**Figure 47**) to also add *FseI* and *PmeI*.

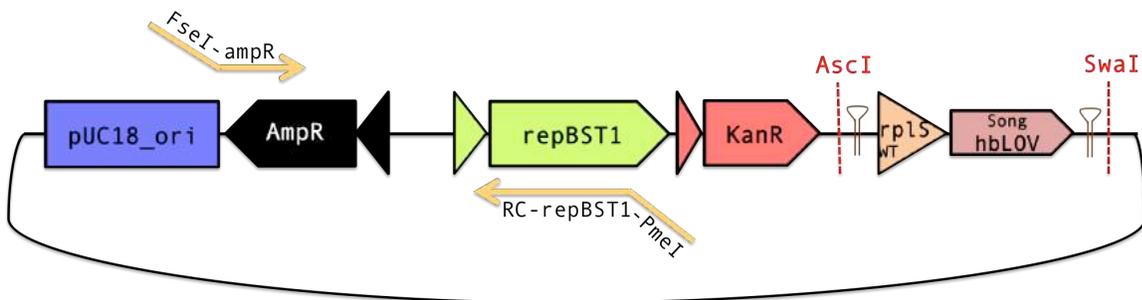


Figure 47. PCR-amplification of the plasmid from **Figure 46** with primers *FseI*-ampR and RC-repBST1-*PmeI*.

The PCR products from **Figures 46** and **47** were digested with *FseI* and *PmeI* and ligated together. The ligation was transformed into *E. coli* and the resulting plasmid, shown in **Figure 48**, was isolated and test-digested to confirm the presence of the *FseI* and *PmeI* restriction sites.

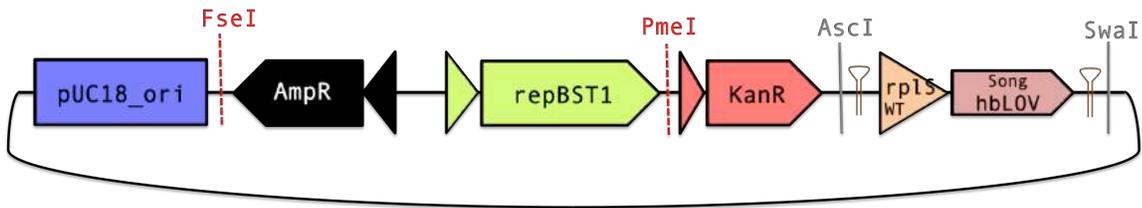


Figure 48. Plasmid resulting from the ligation of the PCR products from **Figures 46** and **47** digested with *FseI* and *PmeI*.

Subsequently, the plasmid from **Figure 48** was PCR-amplified this time with primers *NotI*-repBST1 and RC-AmpR-*NotI* (**Figure 49**).

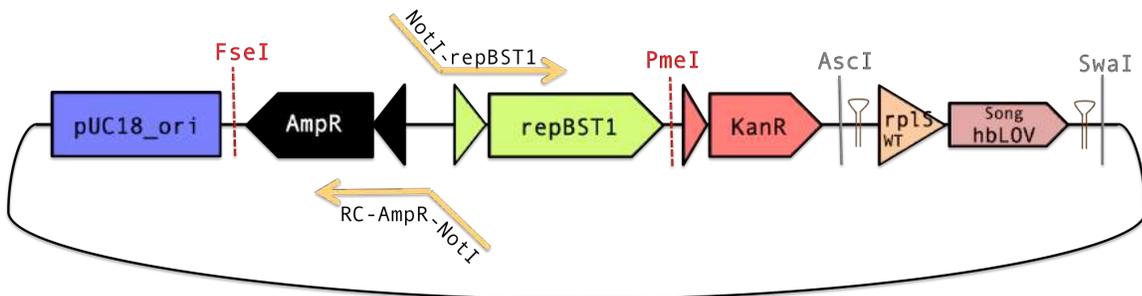


Figure 49. Primers *NotI*-repBST1 and RC-AmpR-*NotI* were used to amplify the entire vector shown in **Figure 48** to add *NotI* at either end of the PCR product.

The PCR-product from **Figure 49** was digested with *NotI* and self-ligated. The resulting minimal modular plasmid, p11AK1 is shown in **Figure 50** and the detailed sequence can be found in **Figure A129**. This was the first plasmid created to match our standard modular design.

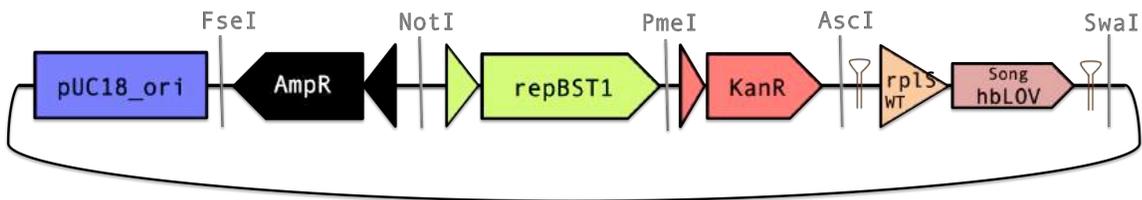


Figure 50. Schematic view of the new minimal modular plasmid p11AK1. Rare cutter restriction enzyme recognition sites (*FseI*, *NotI*, *PmeI*, *AscI* and *SwaI*) delimit each of the modules of the new shuttle vector. This vector is composed of the same essential elements as pUCG18, with the addition of the synthetic MCS, which contains the *RpIS* promoter and *hbLOV*.

The functionality of the boundary enzymes in the plasmid was confirmed after a series of test-digests. Examples of the results from this are given in **Figures 51 and 52**.

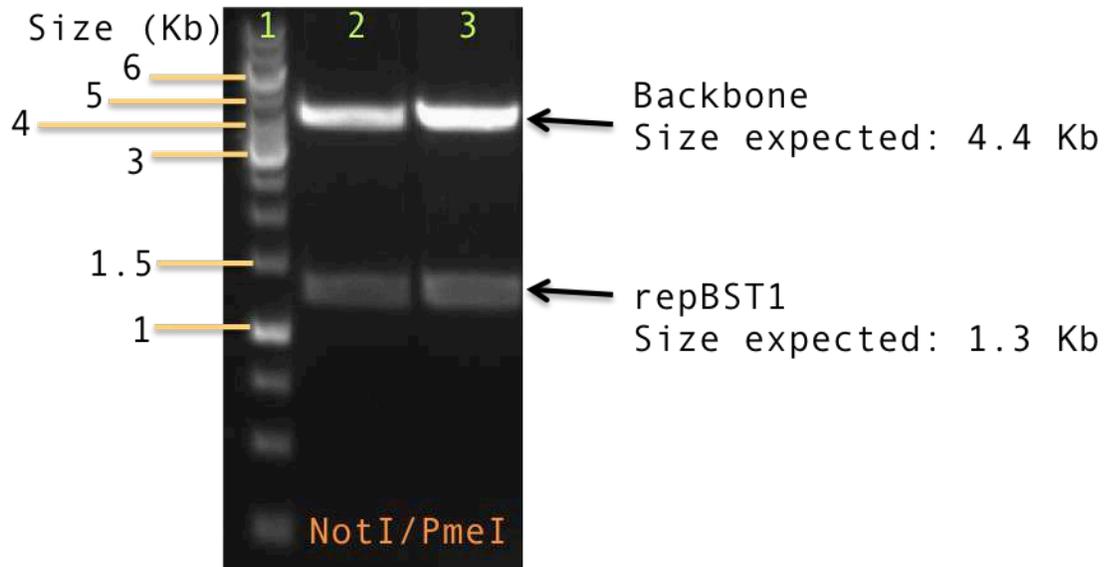


Figure 51. 1% agarose gel of test-digest with NotI/PmeI to demonstrate the functionality of two of the boundary restriction sites in p11AK1. Lane 1: ladder (Fermentas 1 kb). Lanes 2,3: DNA extracted from *E. coli* colonies 1 and 2. Size represented in Kilo base pairs (Kb).

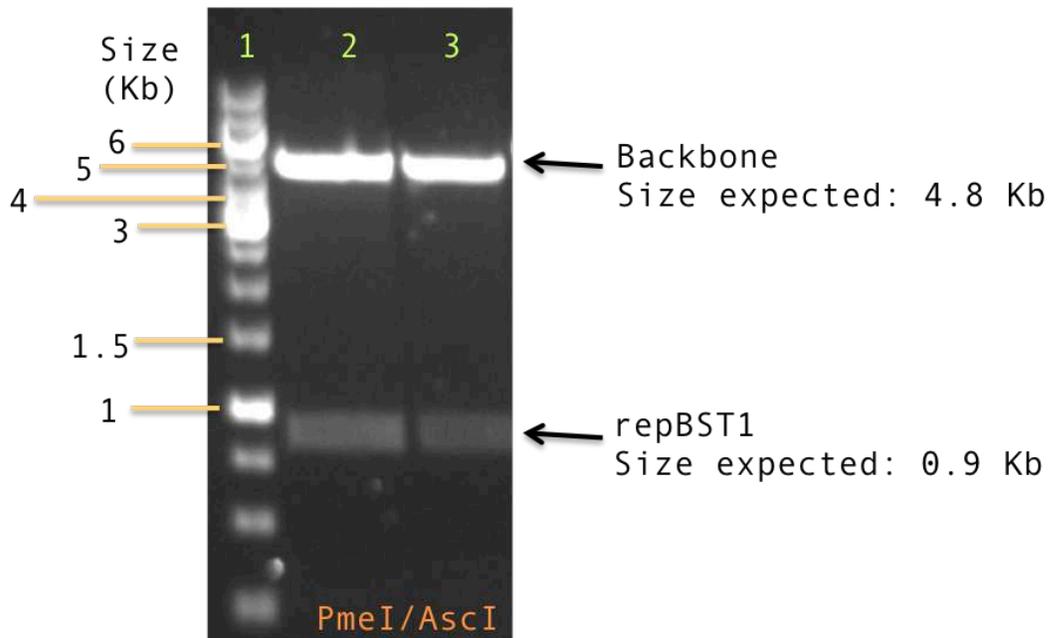


Figure 52. 1% agarose gel of test-digest with PmeI/AscI to demonstrate the functionality of these two boundary restriction sites in p11AK1. Lane 1: ladder (Fermentas, 1 kb). Lanes 2,3: DNA extracted from *E. coli* colonies 1 and 2. Size represented in Kilo base pairs (Kb).

4.2.5. CONSTRUCTION OF A SECOND MODULAR PLASMID

In plasmid p11AK1, the region between *AscI* and *SwaI* contains an MCS occupied by the *RpIS* promoter and *hbLOV* between the *SbfI* and *SacI* sites. To remove them to leave an empty MCS, the plasmid was digested with *SbfI* and *SacI*. Primers BR068 and BR069 were designed (by Benjamin Reeve) to anneal to each other to form a small dsDNA part that completes the MCS that can easily ligate in place of the promoter and *hbLOV* when digested with *SbfI* and *SacI* and mixed with p11AK1. Both plasmid and dsDNA parts were ligated together to produce the shuttle vector with an empty MCS as shown in **Figure 53**.

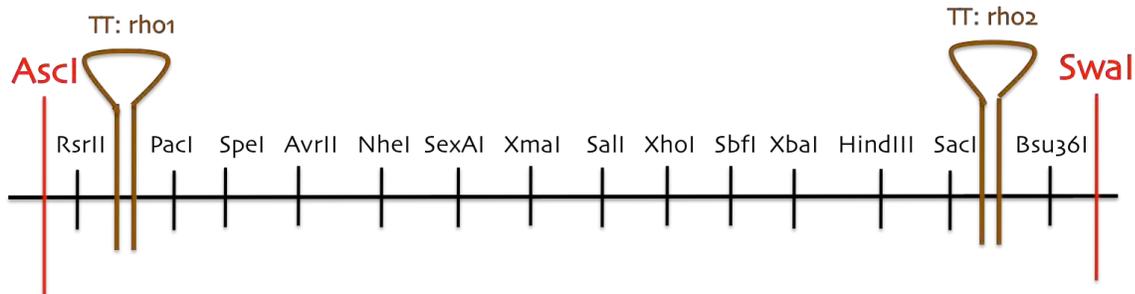


Figure 53. Restriction map of the empty Multiple Cloning Site (MCS) and location of the “rho1” and “rho2” transcriptional terminators. The *RpIS* promoter and *hbLOV* have been replaced by a *HindIII* site.

The resulting plasmid was named p11AK0 following the naming convention given in **Table 3**.

| Table 8. Naming convention for modular toolbox plasmids based on parts composition. | | | |
|--|---|---------------------------------|--|
| <i>E. coli</i> origin of replication | <i>G. thermoglucosidans</i> origin of replication | Antibiotic resistance | MCS |
| <i>pUC18_ori</i> (<i>ColE1</i> =1) | <i>repBST1</i> =1 | Ampicillin ^R =A | Empty=0 |
| | <i>repB</i> =2 | Kanamycin ^R =K | <i>rplS_{WT}</i> + <i>hbLOV</i> =1 |
| | <i>repC</i> =3 | Chloramphenicol ^R =C | <i>rplS_{WT}</i> + <i>sfGFP</i> =2 |
| | | | <i>rplS_{WT}</i> + <i>mCherry</i> =3 |
| | | | <i>rplS_{WT}</i> + <i>pheB</i> =4 |

| Table 9. Naming given to modular toolbox plasmids according to naming convention. | | | | | |
|--|--|-----------------------|-----|-------------------|--|
| <i>E. coli</i> origin of replication | <i>G. thermoglucosidans</i> origin of replication | Antibiotic resistance | MCS | Resulting plasmid | Components |
| 1 | 1 | AK | 1 | p11AK1 | <i>ColE1, repBST1, ampR, kanR, RplS+hbLOV</i> |
| 1 | 1 | AK | 0 | p11AK0 | <i>ColE1, repBST1, ampR, kanR, empty MCS</i> |
| 1 | 1 | AK | 2 | p11AK2 | <i>ColE1, repBST1, ampR, kanR, RplS+sfGFP</i> |
| 1 | 1 | AK | 3 | p11AK3 | <i>ColE1, repBST1, ampR, kanR, RplS+mCherry</i> |
| 1 | 1 | AK | 4 | p11AK4 | <i>ColE1, repBST1, ampR, kanR, RplS+pheB</i> |
| 1 | 1 | AC | 1 | p11AC1 | <i>ColE1, repBST1, ampR, pMK4_catE, RplS+hbLOV</i> |
| 1 | 2 | AK | 1 | p12AK1 | <i>ColE1, repB, ampR, kanR, RplS+hbLOV</i> |
| 1 | 3 | AK | 1 | p13AK1 | <i>ColE1, repC, ampR, kanR, RplS+hbLOV</i> |
| 1 | 2 | AK | 1 | p12AK1 | <i>ColE1, repB, kanR, empty MCS</i> |
| 1 | 2 | AC | 1 | p12AC1 | <i>ColE1, repB, ampR, kanR, RplS+hbLOV</i> |
| 1 | 1* | AC | 2 | p11*AC1 | <i>ColE1, mutant_repBST1, ampR, catE, RplS+sfGFP</i> |
| 1 | 2 | AK | 3 | p12AK3 | <i>ColE1, repB, ampR, kanR, RplS+mCherry</i> |
| 1 | 1 | K | 0 | p11K0 | <i>ColE1, repBST1, kanR, empty MCS</i> |
| 1 | 2 | K | 0 | p12K0 | <i>ColE1, repB, kanR, empty MCS</i> |

The plasmids constructed to conform the early *E.coli*/*G. thermoglucosidans* toolbox are listed in **Table 4** and schematic views of the two plasmids sets are included in **Figures 54** and **55**.

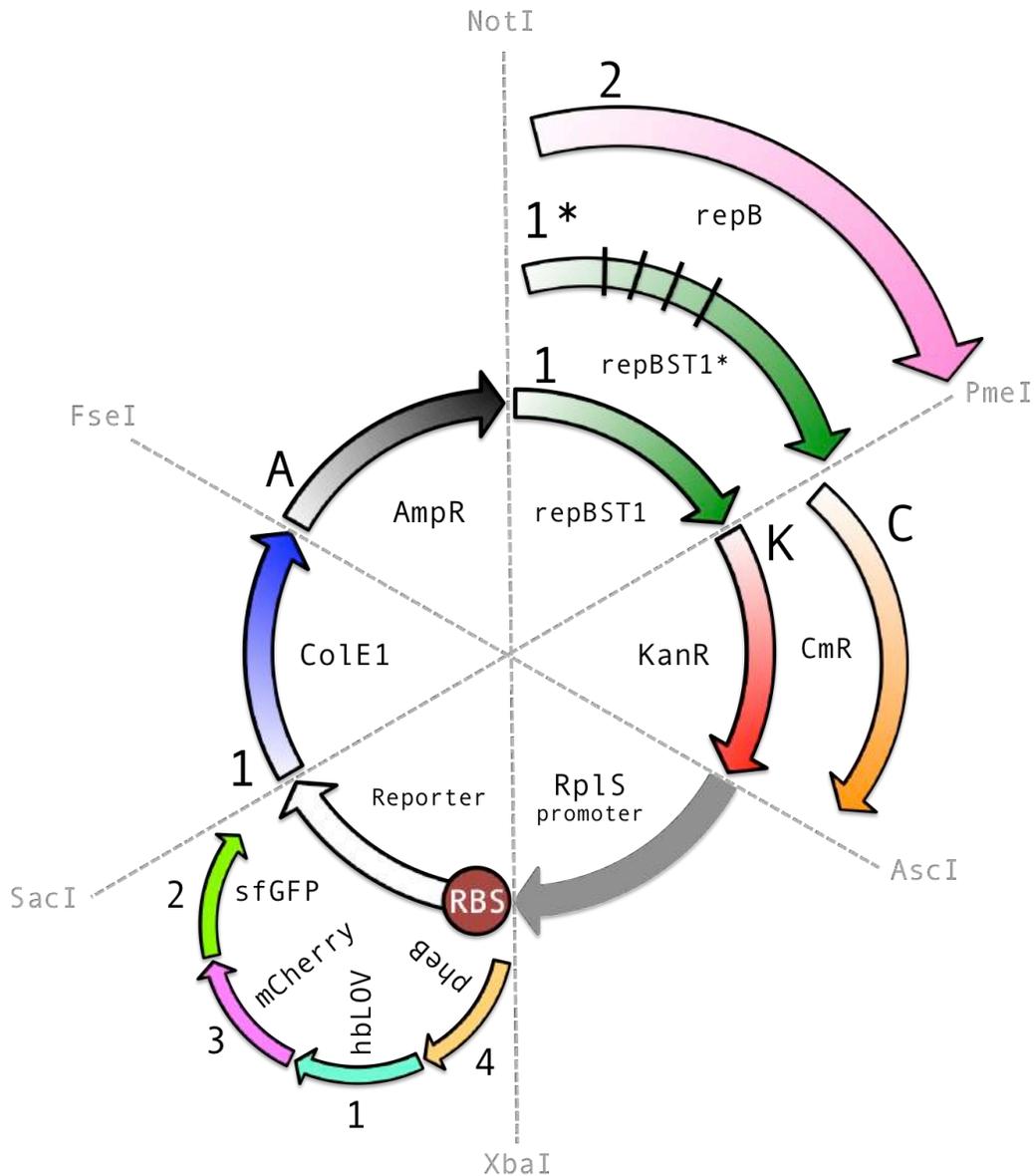


Figure 54. Schematic representation of the larger vector set constructed to be part of the *E. coli*/*G. thermoglucosidans* toolbox. Two sets of shuttle vectors were constructed to constitute an early *E. coli*/*G. thermoglucosidans* toolbox. The larger vectors include two antibiotic resistance genes: one for *E. coli* (*ampR*) and one for *G. thermoglucosidans* (*kanR* or *catE*) although *kanR* can be used to select both *E. coli* and *G. thermoglucosidans*. Also included within the toolbox are three origins of replication for *G. thermoglucosidans* (*repBST1*, the mutant *repBST1** or the temperature-sensitive *repB*), and an origin of replication for *E. coli* (*ColE1*). The suitability of this larger set as an expression vector is demonstrated by the expression of the four reporter proteins (*pheB*, *hbLOV*, *mCherry* and *sfGFP*) discussed in chapter 4 of this work. The transcription of each reporter is driven by the *RplS* promoter and the translation is driven by the *pheB_RBS*. The number next to each part is the one used in the naming convention shown in **Table 3**.

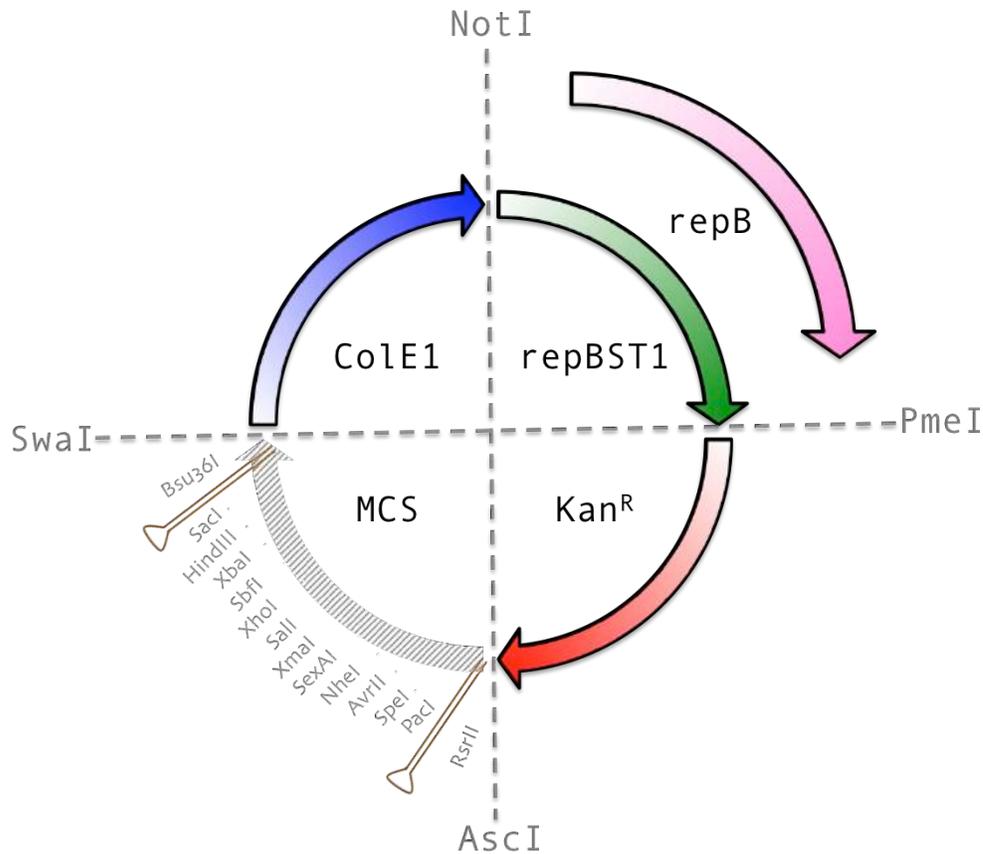


Figure 55. Schematic representation of the smaller vector set constructed to be part of the *E. coli*/*G. thermoglucosidans* toolbox. The smaller, more minimal, vectors include only one antibiotic resistance gene, *kanR* because it can be used to select both *E. coli* and *G. thermoglucosidans*. In addition, included within the toolbox are two origins of replication for *G. thermoglucosidans* (*repBST1* or the temperature-sensitive *repB*), and an origin of replication for *E. coli* (*ColE1*). This smaller vector set is also supplemented with an empty MCS. The number next to each part is the one used in the naming convention shown in **Table 3**.

4.2.6. CONSTRUCTION OF P11AC1

In order to replace the kanamycin resistance gene of p11AK1, two chloramphenicol resistance genes were tested for thermostability in *G. thermoglucosidans* DL44. The first tested was *catE* from the *E. coli* plasmid pLysE, a protein known to confer resistance to chloramphenicol in this mesophile at concentrations as high as 33 µg/mL. The second one was *catE* from pMK4 [Sullivan *et al.* 1984] - a vector commonly-used for *B. subtilis* [Yao *et al.*

2003, Ozcelik *et al.* 2004]. These two genes were selected due to the availability of the plasmids within the research group.

p11AK1 was digested with *PmeI* and *AscI* to extract *kanR*. *pLysE_catE* was amplified using primers *PmeI-catE_ATG* and *RC-catE_ATG-AscI*. *pMK4_catE* was PCR-amplified using primers *PmeI-pMK4_catE_JUN* and *RC-pMK4_catE_JUN_AscI*. The PCR-products were also digested with *PmeI* and *AscI*. Each gene was ligated to the *PmeI/AscI* p11AK1.

4.2.7. CHLORAMPHENICOL RESISTANCE IN *G. THERMOGLUCOSIDANS*

pLysE_catE was observed to confer resistance to chloramphenicol in *E. coli* DH10B at concentrations of up to 33 µg/mL but when the plasmid was transformed into *G. thermoglucosidans* DL44, no colonies were observed at 55 °C or at 45 °C. This was determined by transforming *G. thermoglucosidans* DL44 with 300 ng of the plasmid and allowing the cells to recover in 2TY medium at either 45 °C or 50 °C for 90 minutes, and either plating them on 2TY chloramphenicol plates supplemented with 8 or 12 µg/mL chloramphenicol.

pMK4_catE was observed to not be able confer resistance to chloramphenicol in *E. coli* DH10B, not even at concentrations as low as 6 µg/mL, but when the plasmid was transformed into *G. thermoglucosidans* DL44, colonies were retrieved when incubating at 55 °C for 48 hours in 2TY plates supplemented with 12 µg/mL chloramphenicol.

An artificial selection method was then applied to attempt to obtain a much more thermostable version of *pMK4_catE* by subculturing a colony of *G. thermoglucosidans* DL44 at 50 °C in liquid 2TY media and adding ever increasing amounts of chloramphenicol. The first culture of the transformed colony was obtained when supplementing the medium with 12 µg/mL and growth was observed only after 48 hours. 500 µL of this culture were used to inoculate fresh 2TY medium supplemented with 16 µg/mL. Growth was observed after 48 hours. Once again, 500 µL of this culture were used to inoculate fresh medium, this time supplemented with 20 µg/mL chloramphenicol. After 48 hours, 500 µL of the resulting culture were inoculated to fresh medium containing 24 µg/mL chloramphenicol. Surprisingly, this time, growth was observed after 24 hours. 500 µL of this were inoculated and selected on 28 µg/mL. Once again, growth was detected after 24 hours. Finally, 500 µL of this culture were

used to inoculate medium that was supplemented with 32 µg/mL chloramphenicol. It was decided to stop the selection at this point because 32 µg/mL is a sufficiently high enough concentration of antibiotic. The resulting plasmid was isolated and sequenced.

Interestingly, sequencing revealed four mutations in the *repBST1* gene, but none in the *catE* region. These mutations are marked in the sequence in **Figure A153**. This plasmid will be hereafter referred to as p11*AC1 because *repBST1* is mutated.

4.2.8. CONSTRUCTION OF P12AC1

In order to replace the origin of replication in p11*AC1 with *repB*, plasmid p11*AC1 was digested with *PmeI* and *NotI* to extract *repBST1*. *repB* was PCR-amplified using primers NotI-*repB* and RC-*repB*-*PmeI*. The PCR-product was also digested with *PmeI* and *NotI*. Each gene was ligated to the *PmeI/NotI* p11AK1. The ligation was transformed into *E. coli* DH10B and the correct colonies were confirmed by sequencing. Then this plasmid was transformed into *G. thermoglucosidans* DL44. Colonies appeared on the 2TY/chloramphenicol (12 µg/mL) plate after 48 hours when incubated at 50 °C.

One colony was subjected to the same selection pressure as that used to increase the tolerance to chloramphenicol, which resulted in p11*AC1, but growth was not observed above 16 µg/mL, suggesting that this origin of replication is not as resilient to mutation as *repBST1*.

4.2.9. CONSTRUCTION OF P12AK1

repB is another origin of replication that has been shown to have activity in *G. thermoglucosidans* by Cripps *et al.* (2009). To determine the minimal sequence needed to obtain a functional gene, the promoter region was screened with BPRM for the -10 and -35 boxes and with the IDT OligoAnalyzer tool to find transcriptional terminators. The detailed sequence has been included in **Figure A128**. The promoter -35 and -10 boxes are highlighted in this sequence, and the terminator has been marked as well.

Plasmid p11AK1 was digested with *NotI* and *PmeI* to remove *repBST1*. *repB* was PCR-amplified from pUB190 [Cripps *et al.*, 2009] using primers *NotI*-*repB3* and *RC*-*repB*-*PmeI*. The PCR-product was digested with *NotI* and *PmeI* and ligated to the *NotI/PmeI* vector. The ligation was transformed into *E. coli* DH10B and the correct transformants were confirmed by sequencing. The resulting plasmid was transformed into *G. thermoglucosidans* DL44. Colonies were obtained when selecting on 2TY/kanamycin plates at 55 °C. The correct transformants were confirmed by test-digesting the plasmids isolated from them.

4.2.10. CONSTRUCTION OF P13AK1

Originally, one of the aims of this work was to test whether the components of pMK4, a plasmid shown in **Figure 35**, would work in *G. thermoglucosidans*. When pMK4 was transformed into *G. thermoglucosidans* DL44, no colonies were retrieved when the plates were incubated at 50 °C and supplemented with 12 µg/mL chloramphenicol. pMK4 is composed of *repC* and *pMK4_catE*. It has been demonstrated as part of this work that *pMK4_catE* does work in *G. thermoglucosidans* at 50 °C (see plasmid p11AC1). It was assumed that *repC* and *catE* were perhaps not the best combination of parts and it was possible that the absence of colonies could be due to *repC* being a very low copy plasmid and not producing enough copies to produce enough chloramphenicol acetyltransferase. Decreasing the amount of chloramphenicol was not recommendable because when incubating 2TY plates supplemented with chloramphenicol at 50 °C, the antibiotic starts degrading and the longer they are incubated and the lower the amount of chloramphenicol used, the more background can be observed. 12 µg/mL is actually the lowest recommendable amount for chloramphenicol supplemented to plates, although 8 µg/mL will also be acceptable in the cases where colonies can be retrieved after 24 hours (see plasmid p11*AC1).

repC is an origin of replication for *B. subtilis* also found in pMK4 [Sullivan *et al.* 1984]. Since *B. subtilis* is a close relative to *G. thermoglucosidans*, the functionality of this origin of replication in the thermophile seemed probable. So instead of decreasing the amount of chloramphenicol, it was decided to exploit the modularity and ease of cloning new parts into the toolbox plasmids and replace *repBST1* in p11AK1 with *repC*, thus creating p13AK1. So plasmid p11AK1 was digested with *NotI* and *PmeI* to remove *repBST1*. *repC* was PCR-amplified using primers *NotI*-*repC* and *RC*-*repC*-*PmeI*. The PCR-product was digested with

NotI and *PmeI* and ligated to the *NotI/PmeI* vector. The ligation was transformed into *E. coli* DH10B and the correct colonies were confirmed by sequencing. However, when this plasmid was transformed into *G. thermoglucosidans* DL44, no colonies were observed, at either 55 °C or 45 °C, suggesting that this origin of replication is not suitable for *G. thermoglucosidans*. It was therefore not pursued any further.

4.2.11. SMALLER SET OF MINIMAL THREE-PART MODULAR PLASMIDS

Two minimal plasmids were also constructed as part of the toolbox: p11K0 and p12K0. Plasmid p11K0 was constructed by PCR-amplifying plasmid p11AK0 with primers *NotI*-repBST1 and RC-pUC_ori-*NotI*. Plasmid p12K0 was constructed by PCR-amplifying plasmid p11AK0 with primers *NotI*-repB3 and RC-pUC_ori-*NotI*. The PCR-products were digested with *NotI* and self-ligated. The ligation was transformed into *E. coli* DH10B cells and the correct colonies were confirmed by sequencing. These two plasmids are smaller because they only have one antibiotic resistance gene: *kanR* as well as an empty MCS.

4.2.12. CHARACTERIZATION OF PARTS: KANR

Table 10 shows a comparison of the transformation efficiencies obtained with some of the plasmids of the modular toolbox and pUCG18 and pUCG3.8 for comparison.

| Table 10. Average number of colonies obtained after transforming with 15 ng of DNA. N=3 | | | | |
|--|------------------------|----------------------|-----------------------------|----------------------------|
| Organism | <i>E. coli</i> | | <i>G. thermoglucosidans</i> | |
| Incubation temperature | 37 °C | 37 °C | 55 °C | 55 °C |
| Antibiotic | Ampicillin (100 µg/mL) | Kanamycin (50 µg/mL) | Kanamycin (25 µg/mL) | Chloramphenicol (12 µg/mL) |
| p11AK0 | >1000 | >1000 | 547 | - |
| p11K0 | - | >1000 | >1000 | - |
| p12AK1 | >1000 | 355 | >1000 | - |
| p12AC1 | 734 | - | - | 23.5 |
| p11*AC1 | >1000 | - | - | 18 |
| pUCG18 | 2.5 | 9* | 0.3 | - |
| pUCG3.8 | N/A | 6* | >1000 | - |

*Observed after 48 hours of incubation at 37 °C.

In **Figures 56** and **57**, it can be seen that the transformation efficiencies, when selecting on ampicillin and kanamycin, are much higher for the modular plasmids than for the reference pUCG18 and pUCG3.8. In *E. coli*, for the modular plasmids the transformation efficiencies obtained when selecting on ampicillin were just as high as when selecting on kanamycin.

For DL44, it can be clearly observed that the number of colonies obtained when selecting on chloramphenicol is much lower than when selecting on kanamycin. Also, the efficiencies between pUCG3.8 and the modular plasmids appear to be similar. The size differences between the vectors that have two antibiotic markers and the ones that have only one does appear to have an effect in *G. thermoglucosidasius* transformation efficiency, confirming previous anecdotal data from the Leak Group.

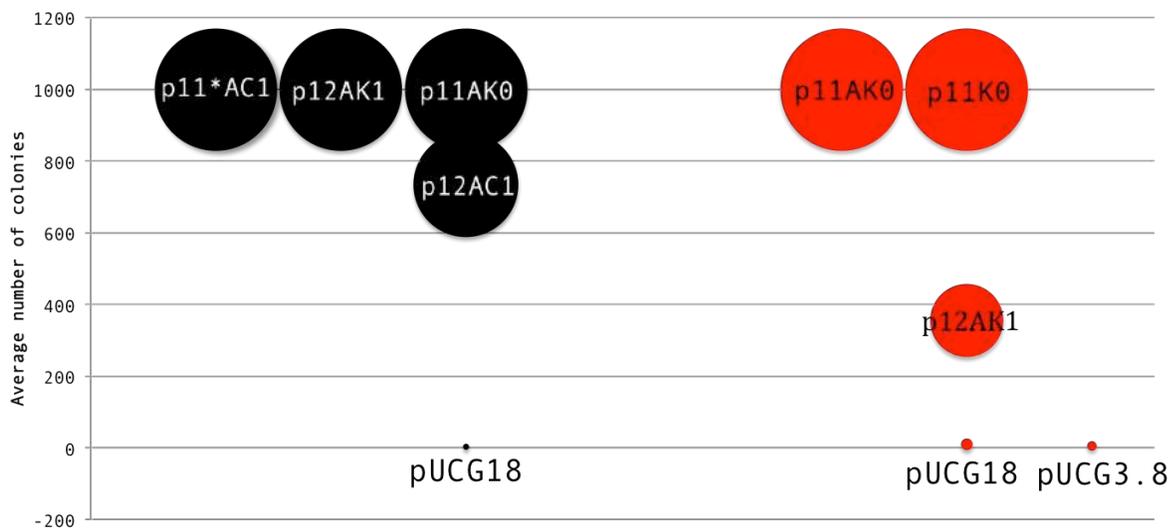


Figure 56. Transformation efficiencies of *E. coli* obtained with the modular vectors. Graphic representation of the transformation efficiencies reported in **Table XX** of chemically competent *E. coli* DH10B obtained with the plasmids of the toolkit compared to the pre-existing reference plasmids pUCG18 and pUCG3.8. The size of the circle correlates to the average number of colonies obtained after three transformations using 15 ng of DNA. The black colour scheme indicates that the selection of colonies was carried out in the presence of ampicillin (100 µg/mL), whereas red represents kanamycin (50 µg/mL) selection.

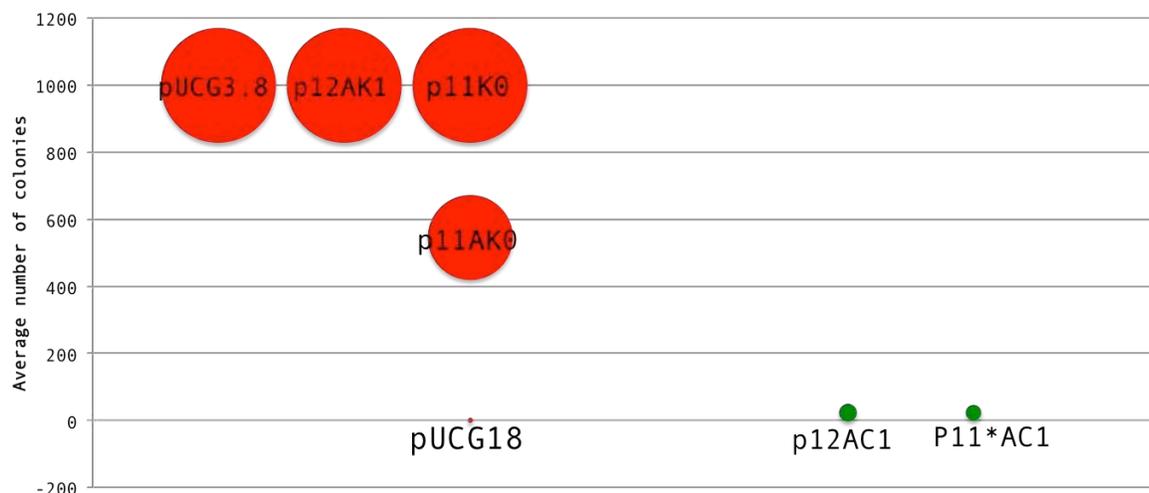


Figure 57. Transformation efficiencies of *G. thermoglucosidans* obtained with the modular vectors. Graphic representation of the transformation efficiencies reported in Table 3 of electrocompetent *G. thermoglucosidans* DL44 obtained with the plasmids of the toolkit compared to the pre-existing reference plasmids pUCG18 and pUCG3.8. The size of the circle correlates to the average number of colonies obtained after three transformations using 15 ng of DNA. The red color scheme indicates that the selection of colonies was carried out in the presence of kanamycin (f.c. 25 µg/mL), whereas green represents chloramphenicol (f.c. 12 µg/mL) selection.

4.2.13. PLASMID COMPATIBILITY

To determine whether it is possible for *G. thermoglucosidans* to maintain two plasmids simultaneously, two new plasmids were created: (*mutant_repBST1**, *pMK4_catE* and *RpIS+mCherry*) as well as p12AK2 (*repB*, *kanR*, *RpIS+sfGFP*). These were chosen **because** they have different origins of replication and different antibiotic selections, therefore these two plasmids could potentially replicate and propagate independently within a single cell. p11*AC3 was transformed into *G. thermoglucosidans* DL44 first. The cells were grown at 50 °C and the correct colonies were confirmed by screening for red fluorescence. These cells were made electrocompetent and transformed once more, this time with p12AK2.

To test for plasmid incompatibility, the cells were grown in kanamycin and chloramphenicol. When these cells were seen under a microscope to screen for either green or red fluorescence of each individual cell, as seen in **Figure 58**, it was evident that the majority of the cells express either sfGFP or mCherry. Only in a very small number of cells, can weak expression of both reporters be seen.

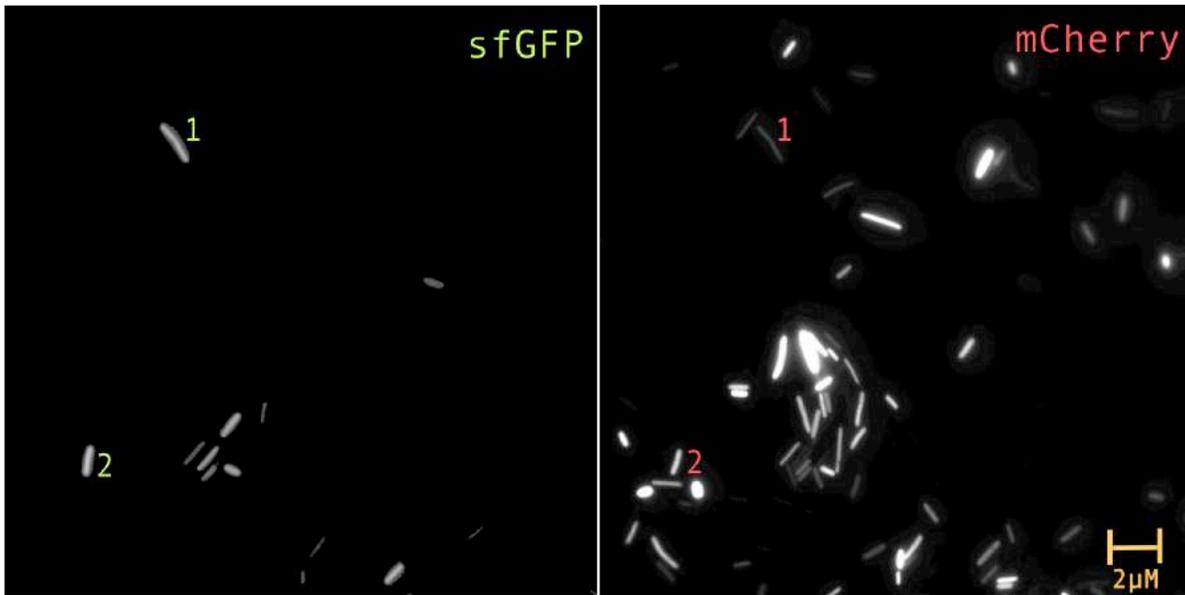


Figure 58. Microscopy imaging detection of sfGFP and mCherry fluorescence to test for plasmid (in)compatibility. *G. thermoglucosidans* was transformed with p11*AC3 (containing the mutant *repBST1**, *catE*, and *mCherry*) and p12AK2 (containing *repB*, *kanR* and *sfGFP*) and selected simultaneously on chloramphenicol (f.c. 32 $\mu\text{g}/\text{mL}$) and kanamycin (f. c. 12 $\mu\text{g}/\text{ml}$). Cells were restreaked on a microscope slide and were viewed under a microscope to detect fluorescence of sfGFP and mCherry. Labeled as "1" and "2" are the cells that appear to show fluorescence of both reporters.

4.2.14. CHARACTERIZATION OF PARTS: *E. COLI* ORIGIN OF REPLICATION

The several rounds of PCR required in DNA assembly resulted in two mutations in *pUC18_ori (ColE1)* (shown in **Figure A129**), but these weren't corrected because the yields obtained, when compared to the parent pUCG18, were still significantly high, as seen in **Figure 59**. Therefore, the effect of these mutations on the copy number was not regarded as problematic.

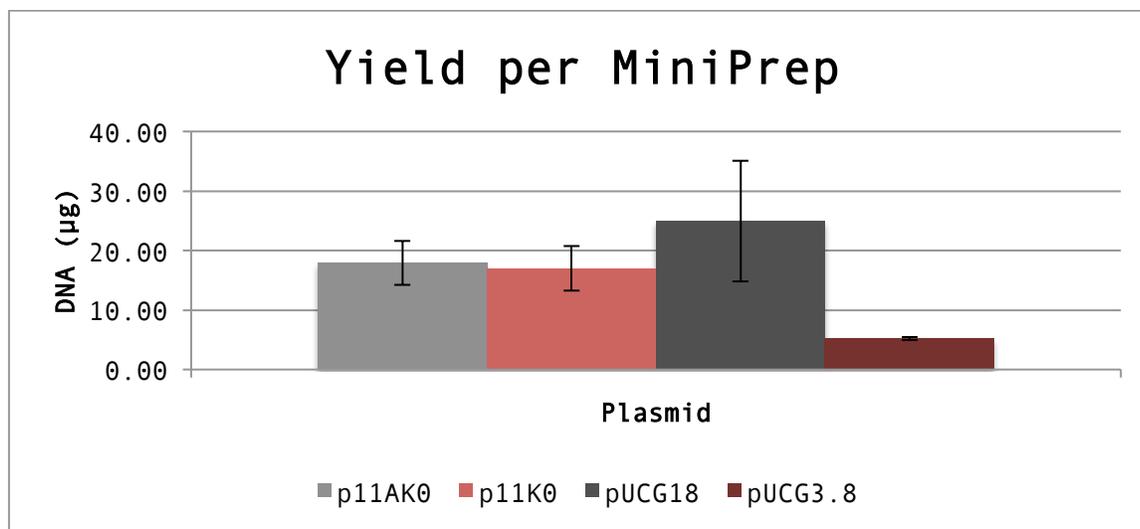


Figure 59. Average DNA concentration obtained after extracting each plasmid from overnight *E. coli*. Plasmid DNA was isolated from 5 mL cultures of *E. coli* growing in LB/ampicillin in the case of p11AK0 and pUCG18 and LB/kanamycin for p11K0 and pUCG3.8 at 37 °C. Error bars indicate standard deviation of a population of at least 3 samples per plasmid.

4.2.15. CHARACTERIZATION OF PARTS: *GEOBACILLUS* ORIGINS OF REPLICATION

In an attempt to determine what is the difference between *repB*, *repBST1* and the mutagenic *repBST1**, *G. thermoglucosidans* DL44 cells transformed with p11AK1, p11*AC1 and p12AC1 were inoculated to 5 mL of 2TY supplemented with the appropriate antibiotic and were grown at 50 °C. After overnight growth, the cells were harvested by centrifugation and their plasmid DNA was extracted. The DNA was quantified using a NanoDrop 1000 to compare yields. The results are shown in **Figure 60**.

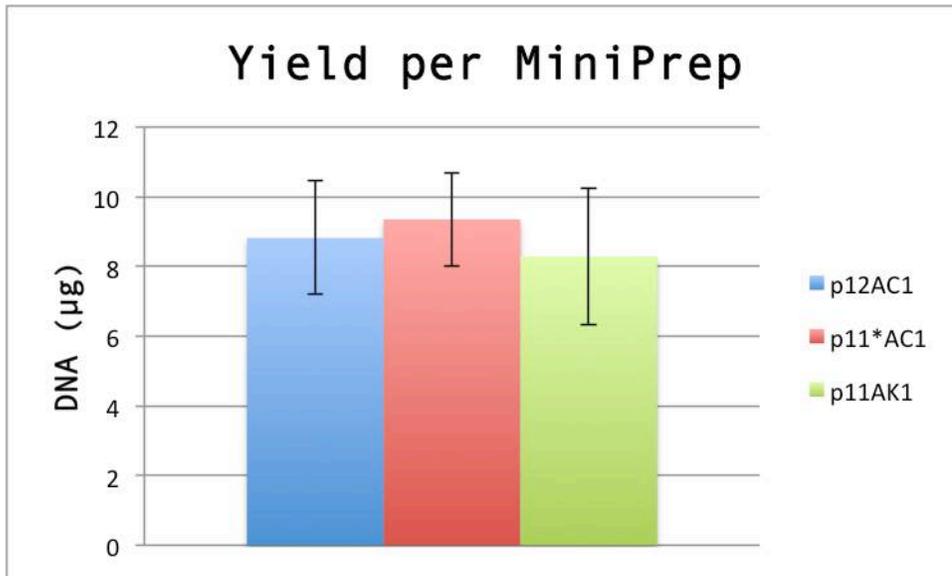


Figure 60. Average DNA concentration obtained after extracting each plasmid from overnight *G. thermoglucosidans*. Plasmid DNA was isolated from 5 mL cultures of *G. thermoglucosidans* DL44 grown in 2TY/chloramphenicol (f.c.12 µg/mL) at 50 °C with shaking at 200 rpm. Error bars indicate standard deviation of three measurements.

As it can be seen in **Figure 60**, it's not possible to make a clear distinction of the copy numbers of each plasmid based on the plasmid yield alone, since all three produced very similar yields.

In another experiment, *G. thermoglucosidans* DL44 modified with p12AC1 and p11*AC1 were grown in parallel in a baffled flask containing 50 mL of 2TY supplemented with 32 µL of chloramphenicol at 50 °C to determine whether there is any difference in their growth, but, as it can be seen in **Figure 61**, the differences observed do not allow to clearly establish if these two origins of replication have any effect on growth. It's interesting to see that the mutations of *repBST1* do not appear to help the cells grow significantly faster.

Tracking the growth on chloramphenicol was selected for this experiment because cells take longer to grow than on kanamycin so any difference should have been made much more obvious by this.

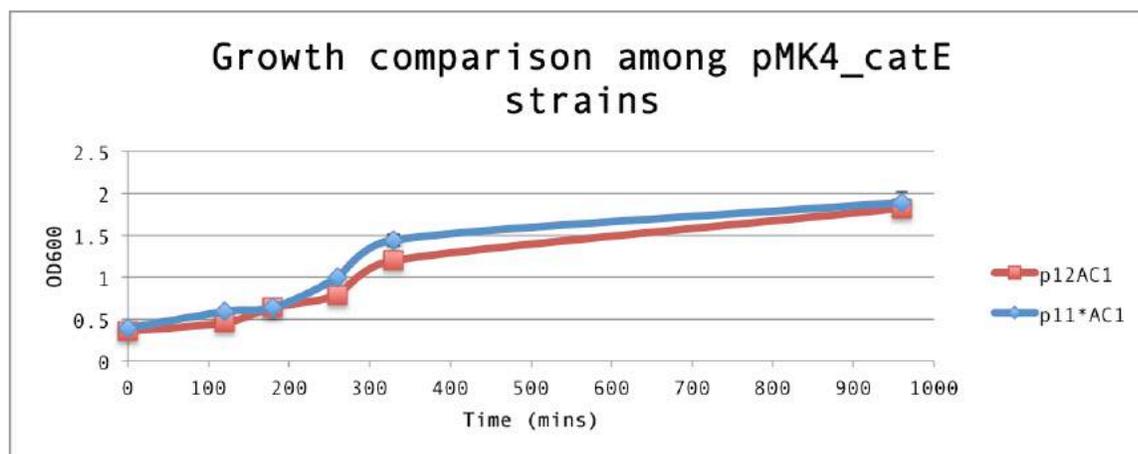


Figure 61. Growth comparison among strain p12AC1 (*repB pMK4_catE*) and strain p11*AC1 (*repBST1 pMK4_catE*). Fresh 50 mL 2TY medium was supplemented with 32 µg/mL (f.c.) chloramphenicol. 500 µL of *G. thermoglucosidans* DL44 transformed with plasmids p12AC1 and p11*AC1 were inoculated from an overnight culture and grown in baffled 250 mL flasks at 55 °C. Growth was followed by taking OD₆₀₀ measurements at selected time points. Error bars represent the standard deviation of two measurements.

4.3. DISCUSSION

- The transformation efficiency yields demonstrate the positive effect the new transcriptional terminator in the synthetic MCS are having on the transcriptional regulation of the gene and it's clearly a significant improvement over the previously existing vectors.
- Chloramphenicol resistance with *catE_pMK4* in *G. thermoglucosidans* is inconclusive from the experiments carried out so far. Oxygenation needs to be investigated because the enzyme is dependent on acetyl-CoA.
- *repBST1* and *repB* are not compatible origins of replication.

4.3.1. CHLORAMPHENICOL RESISTANCE IN *G. THERMOGLUCOSIDANS*

Of the two versions of the chloramphenicol resistance protein tested, *pLysE_catE* proved to not be suitable for thermophiles. To the author's best knowledge, there are no reports in literature of this protein being used in a thermophile, thus suggesting that it is not thermostable. *pMK4_catE* on the other hand, was shown to be able to withstand much higher temperatures, but it is not recommendable selecting at a temperature above 50 °C because chloramphenicol itself degrades at these temperatures and the result observed on plates is a

high level of background; making it impossible to distinguish positive transformants. Chloramphenicol selection at 50 °C is also not recommendable because 50 °C is actually a low temperature for *Geobacillus*, and is far from the ideal 60 °C. Thus cells grows at a much slower rate and transformants cannot be observed on plates until after 48 hours incubation and by then, a large amount of the chloramphenicol on the plates will have degraded, and individual colonies can no longer be differentiated since there is a thin layer of cells on the plate due to the bacteriostatic nature of chloramphenicol.

The mutations seen in *repBST1* might help confer chloramphenicol resistance at 55 °C. It may be that the mutations lead to a higher copy number of *repBST1* and therefore enough *catE* copy-number to produce enough of the resistance enzyme to allow growth in the presence of chloramphenicol. Although, one would think that it would be easier for the cells to mutate *catE* instead of *repBST1*. Since this wasn't the case, it suggests that the only thermostable configuration of the gene is the one it presently has, and that if any other mutations arose, they probably did not permit the survival of the cells.

In conclusion, *pMK4_catE* can be used as an alternative selection marker in *G. thermoglucosidans* as long as the cells are selected at 55 °C for no more than 24 hours and the plasmid being used has the mutated *repBST1*. Unfortunately transformation efficiencies of this will be very low compared to kanamycin selection (**Figure 57**).

In summary, the results obtained from this experiment are inconclusive until it can be demonstrated that the mutant *repBST1**-chloramphenicol resistance plasmid can be transformed back into *G. thermoglucosidans* with the obtention of transformants at 12 µg/mL at 55 °C overnight. Otherwise, the mutagenic origin of replication does not actually confer any improved resistance to chloramphenicol. Also, taking into consideration that chloramphenicol resistance is dependent on acetyl-CoA and there is more production of acetyl-CoA in aerobic fermentation, then it is most likely that the reason more colonies are obtained when selecting at lower temperatures (*i.e.* 45 °C) is that oxygen solubility is lower at those temperatures and therefore there is more cofactor available for the enzyme to interact with, and the reason is not so much that the actual enzyme is not temperature sensitive.

4.3.2. KANAMYCIN RESISTANCE IN *G. THERMOGLUCOSIDANS*

The thermostable kanamycin nucleotidyltransferase produced by Liao *et al.* (1990) appears to be a significantly better selection marker for thermophiles than chloramphenicol. On top of this, it works well in *E. coli* too, which means that the shuttle plasmid can be smaller in size if the *kanR* gene is used by both organisms. Previously, thermostable KanR expression in *E. coli* had been problematic, meaning that three-part vectors (ones that use thermostable KanR for selection in both organisms) gave very low numbers of colonies when transformed into *E. coli*. In this work, it was identified that a transcriptional terminator was missing from previously constructed plasmids containing a thermostable KanR. Both pUCG18 and by extension, pUCG3.8, lack a transcriptional terminator and have low efficiencies when transformed into *E. coli* and selected by kanamycin. In the case of pUCG18, the low efficiencies are due to two factors: the lack of the terminator and the very large size of the empty vector. In the modular plasmids, the transcriptional terminator "rho2" was placed at the 5'-end of the MCS. This acts as a transcriptional terminator for *kanR* and could explain the improved transformation efficiency of the modular plasmids when transformed into *E. coli* and selected by kanamycin. The comparison of the transformation efficiencies obtained when selecting *E. coli* on kanamycin shows how significant having an effective transcriptional terminator after *kanR* is. Likely this terminator is positively affecting the expression of this gene. While this seems to have little impact in the case of *G. thermoglucosidans*, it does make a substantial difference for *E. coli*. And with this design, having a minimal plasmid that has only *kanR* for selection in both hosts is very promising.

It would be interesting to find out why the absence of a terminator at the end of *kanR* has had no effect on *G. thermoglucosidans*, whereas in *E. coli* it does have a strong impact. Understanding this could be useful for future synthetic biology device designs and characterization of parts.

The transformation efficiencies obtained between the modular plasmids that have two antibiotic resistance genes and those that have one seem to have an important effect in *G. thermoglucosidans*. This highlights the need for an efficient minimal vector that can be used on both hosts effectively.

Interestingly, during the course of this work it was found that no matter what concentration of kanamycin is used in *E. coli* (50 µg/mL or 100 µg/mL), false positives were occasionally encountered when plating on LB. This has previously been reported by Studier [50], who discovered that *E. coli* can develop natural resistance to kanamycin when there is a high concentration of phosphate in the medium combined with amino acids in rich media. During the course of this investigation, it was also found that the amount of salt in LB media is too high to effectively select on kanamycin. The amount of salt in media is directly linked to the number of false positives seen on plates, which can be discerned from real transformants because the colonies are significantly smaller and won't grow when transferred to liquid media with kanamycin. The best way to avoid this is to add only half the amount of salt that LB normally requires.

4.3.3. PLASMID (IN)COMPATIBILITY

repB and *repBST1* appear from the results to be incompatible origins of replication. It is possible also that having the two plasmids simultaneously is putting the cells under too much stress. It is remarkable that within one colony there are three different phenotypes: mCherry, sfGFP or weak sfGFP/mCherry. And more remarkable yet is the fact that the cells are able to survive the presence of two antibiotics. The loss of one of the two plasmids would explain why cells are either green or red, but they couldn't possibly survive if they lost both *kanR* and *catE*. This indicates the possibility that the plasmids are being recombined into the genome (even though they lack homology regions) in order to guarantee the survival of the cell. This result hints towards the powerful recombination capabilities of the thermophilic host cell.

4.3.4. MULTIPLE CLONING SITE (MCS) DESIGN

The MCS design is one of the strongest attributes contained in the modular plasmids that form part of the toolbox. It contains restriction sites for enzymes that do **cut** not any of the fragments that are included in the toolbox, and it offers the user a significant number of **unique** enzymes to choose from. None of the restriction sites included in the MCS are blunt cutters. *PacI* produces a two-nucleotide overhang, for example, while *RsrII* leaves four nucleotides. Most of the enzymes chosen however, leave at least four nucleotides of overhang, such as *SpeI*, *AvrII*, *NheI*, *XmaI*, *Sall*, *XhoI*, *SbfI*, *XbaI*, *HindIII* and *SacI*. In the case of *SexAI*, a five-nucleotide overhang remains after cutting. Most of these enzymes are compatible with the New England Biolabs CutSmart Buffer and they are available commercially. These restriction sites were selected because longer overhangs are much more likely to be more efficient in a ligation than an overhang of only two base pairs.

4.4. CONCLUSION

In conclusion, the shuttle vectors created as part of this **project** are **minimal, modular** and offer **several choices regarding biopart composition**, for instance, it includes **two origins of replication for *G. thermoglucosidans***, one of it is **temperature sensitive** and it allows **doing genomic recombinations in this thermophile**. Both origins of replication appear have **similar copy numbers and can be isolated comparable to *E. coli***, although it appears that the two origins of replication are not compatible, so it is not possible to have these two plasmids simultaneously in *G. thermoglucosidans*.

The toolbox also offers three antibiotic resistance markers that permit **selecting on ampicillin or kanamycin (*E. coli*) and either kanamycin or chloramphenicol for *G. thermoglucosidans***, depending on the temperature of growth. Kanamycin selection is much more recommendable at any temperature than chloramphenicol selection because it offers higher transformation efficiencies, but in the case that another origin of replication is to be tested for compatibility then it will be possible to do so with the chloramphenicol marker.

The high copy number origin of replication for *E. coli* ensures that large amounts of the plasmid will be produced, which can subsequently be used to transform *G. thermoglucosidans*. If need be, the *E. coli* origin can be easily replaced by a lower copy number origin if the vector are being used to express a protein that's toxic or that creates significant stress. **And it's not just** the *E. coli* origin can be easily swapped. All other parts can be readily exchanged **due to the modular nature of the vector**. The use of rare cutters as border enzymes is particularly advantageous because it is highly unlikely they will be present in any additional part to be included in the future in **the** toolbox. With that in mind, the Multiple Cloning Site offers many unique restriction sites, **to accommodate new parts**. The MCS can be transferred to a SEVA vector, thus making sharing parts easier among the SEVA community.

The minimal three part vectors are smaller in size than any of the previous **vectors** for *G. thermoglucosidans* and thus offer the possibility to clone large genes, such as genes for lignocellulosic degrading enzymes, without compromising transformation efficiency, which is definitely a desirable property for any plasmid, and the shuttle vectors created as part of this toolbox offer much higher transformation efficiencies than any of the previously existing vectors.

The elements included in this toolbox and assembled as functional shuttle vectors should facilitate, enhance and accelerate the transfer of bioparts between *G. thermoglucosidans* and *E. coli*.

4.5. FUTURE WORK

The shuttle vector have not been tested for stability. Additional experimental work needs to be carried out to determine whether the plasmids have:

- 1. segregational stability to ensure that all the cells in the culture have the plasmid.**
- 2. consistent copy number- to ensure product consistency.**
- 3. ability to carry large sequences of DNA.**

More origins of replication from other classes or orders also classified under Firmicutes could be tested, especially those from other thermophiles such as *Clostridium thermocellum* for example, which **might show activity in *G. thermoglucosidans*. It would be**

ideal to build a toolkit that contains multiple parts that have different functions. For instance, other origins of replication that are compatible with each other could be used to express two cassettes simultaneously carried by two vectors, or lower/higher copy number origins of replication could be swapped depending on the effect they have in a particular application, especially when too many/too little copies of the plasmid are a drawback to achieve optimal results. It would also be interesting to determine whether the origins of replication work in other *Geobacillus* species, apart from *G. thermoglucosidans* because the shuttle vectors could be used by the *Geobacillus* community and not only by the *G. thermoglucosidans* community.

Other selection markers that could be included in the toolbox are for instance, hygromycin resistance, which has been shown to work on other thermophiles such as *Thermus thermophilus* (Cava *et al.* 2008) or *Sulfolobus solfataricus* (Cannio *et al.* 2001). But not all selection markers need to be antibiotic selection markers. For instance, alcohol dehydrogenases have been used for selection in hyperthermophilic archaea that have sensitivity to alcohols, by catalyzing the conversion of the alcohol to a non-toxic derivative (Aravalli and Garrett, 1997). Yet a third potential selectable marker could be an auxotrophic marker: by eliminating the pathway for production of an essential compound, selection can be achieved by supplementing the necessary compound. Cultures may develop resistance to antibiotics but auxotrophy markers might be more difficult to evolve resistance to because the cells would have to acquire the pathway *de novo*.

The antibiotic selection resistance markers need to be characterized under various oxygen and salinity conditions. For kanamycin resistance it is known that the concentration of kanamycin affects translational processes and the general cell physiology to different extents [Gorochofski *et al.* 2013] and it was observed in the course of this investigation that salt concentration has an effect on the number of *E. coli* false positives. Therefore, it is necessary to test whether salt concentrations have an effect on the transformation efficiencies of *G. thermoglucosidans*. For chloramphenicol resistance, because the protein (chloramphenicol acetyl-transferase) is dependent on acetyl-CoA, it is necessary to determine whether oxygen availability has an effect on the amount of acetyl-CoA produced and therefore and the amount of cofactor available. It is possible that the growth conditions need to be made more aerobic to permit better functionality of the marker.

CHAPTER 5:

GENOMIC KNOCK-IN/OUTS IN *G. THERMOGLUCOSIDANS*

The key contributions and results of this chapter are:

- A shuttle vector containing the combination of parts *repB+thermostable_kanR* was constructed.
- The shuttle vector was used in an attempt to create a knock-out of *ldh* by simultaneously doing a knock-in of *sfGFP* in its locus.
- The strategy resulted too ambitious: it was not possible to do a knock-out of *ldh* in the wild-type strain DL33, but it was possible to do a knock-in of *sfGFP* in the *ldh* locus of the *ldh⁻* strain DL44.
- Due to the lack of a selection marker for the second recombination event, it was not possible to select for colonies that did not revert their phenotype back to wild-type.
- Uracil auxotrophy is proposed as a selection marker for the second recombination event.
- The genome of *G. thermoglucosidans* was screened to determine the feasibility of knocking-out the uracil biosynthesis pathway by targeting the gene *pyrE* for a knock-out.

5.1. INTRODUCTION

Genomic knock-in/outs in *G. thermoglucosidans* are relevant to this work because:

- The temperature-sensitive shuttle vectors from this toolbox can be used to do genomic knock-ins or knock-outs.
- The shuttle vectors included in this toolbox that contain the *repB/thermostable_kanR* combination of parts should be more effective for the knock-in or knock-out process in *G. thermoglucosidans* than previously existing vectors.
- *repB* is a temperature sensitive origin of replication and *thermostable_kanR* differs from the mesophilic *kanR* by two mutations.
- The vectors previously used for knock-outs were not ideal because they did not contain the correct parts to make them efficient, *i.e.* they contained the mesophilic *kanR* instead of the thermophilic one.

- The hypothesis that the parts in this toolbox are better for doing a knock-out will be experimentally verified in this work.
- The knock-out process in *G. thermoglucosidans*, based on reports found in the literature, consists of two recombination events.
- While kanamycin resistance is the marker that permits selecting transformants from the first recombination event, currently there is no selection marker for the second recombination event, which limits the strategy and level of modifications to the genome of *G. thermoglucosidans*.
- It is necessary to find a marker that permits selection of transformants from the second recombination event.
- The tools provided in the *G. thermoglucosidans* toolbox are the first step required for the creation of a better and improved method for genome scale modification of this thermophile.

5.1.1. THE KNOWN KNOCK-OUT PROCEDURE

Cripps *et al.* (2009) created a vector called pUB190, with the purpose of doing knock outs in *Geobacillus* spp. This plasmid contains a temperature sensitive origin of replication called "**repB**", which becomes non-functional at temperatures above 65 °C. It also contains a kanamycin resistance gene, but it is different from that of Liao and Kanikula, [1990] in that it does not contain the two mutations that add thermostability at temperatures between 55 and 68 °C. The mesophilic kanamycin resistance gene, works poorly at temperatures above 55 °C, as demonstrated in an earlier report [Martinez-Klimova, 2011], in which the mesophilic *kanR* resistance gene was paired with *repBST1*.

Taylor [2007] was the first report of a knock-out in *G. thermoglucosidans*. The *ldh* gene of wild-type DL33 was knocked out to create DL44, which has now become the strain of choice as the *G. thermoglucosidans* model strain because it is the one that has the highest resilience and produces the highest transformation efficiencies [Bartosiak-Jentys, *personal communication*]. The Taylor strategy for doing the knock-out, as well as the strategy used by Cripps *et al.* [2009] and by Crowhurst in *G. thermodenitrificans* [2010] to replicate the *ldh* knockout of Taylor [2007] is illustrated in **Figures 62 and 63**. In order for the inactive version of *ldh* to permanently replace the functional version of *ldh* in the genome, two cross-over events are needed, each of which depends on two regions of homology.

Figure 62 shows the first cross-over between the *Geobacillus* spp. genomic DNA and the plasmid. First (**A**, **Figure 62**) a plasmid was created that contained *repB*, the temperature sensitive origin of replication as well as a kanamycin resistance gene. In Taylor [2007], Crowhurst [2010] and Cripps *et al.* [2009], the kanamycin resistance gene used was the mesophilic one, because these authors were under the assumption that it was thermophilic. In addition, the plasmid had two 500 bp regions that were identical to the *ldh* gene. In the middle, 50 bp were omitted and replaced instead by a NotI restriction site. The first region of homology from the plasmid pairs with the corresponding identical region in the chromosome (**B**, **Figure 62**) and the first cross-over yields a genotype that has incorporated the entire plasmid into the chromosome (**C**, **Figure 62**).

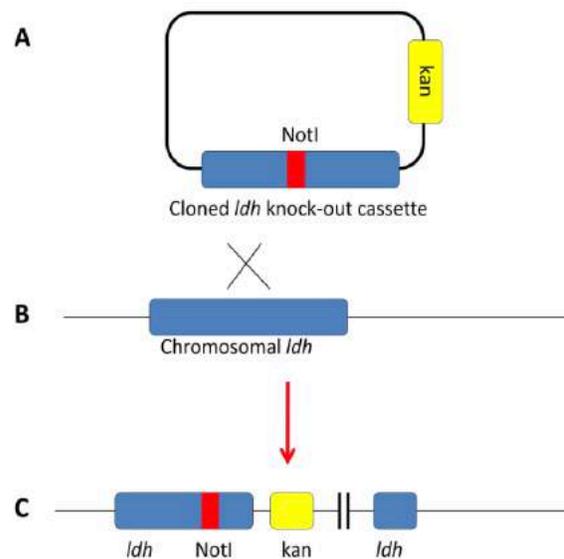


Figure 62. Illustration of the sequence of events leading to the first cross-over event according to Crowhurst [2010], Taylor [2007] and Cripps *et al.* [2009]. (A) shows the plasmid containing two regions identical in sequence to the chromosomal *ldh* gene (B), but in the plasmid-bound version, the *ldh* gene has had 50 bp in the middle deleted and replaced by a NotI site. In (C) the genomic incorporation of the entire plasmid has taken place due to the recognition of the two identical plasmid-bound and genomic *ldh* regions.

Figure taken from Crowhurst [2010].

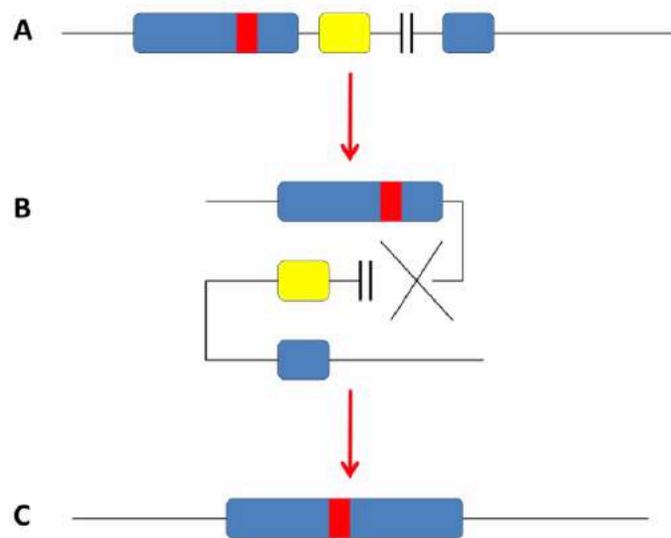


Figure 63. Illustration of the sequence of events leading to the second cross-over event according to Crowhurst [2010], Taylor [2007] and Cripps *et al.* [2009]. (A) shows the genomic incorporation of the entire plasmid has taken place due to the recognition of the two identical plasmid-bound and genomic *ldh* regions. (B) shows the loop-in event between the second homology regions in the plasmid and in the genome. (C) shows the remaining non-functional copy in the genome and the absence of the functional copy as well as of the rest of the plasmid components, such as *kanR*. Figure taken from Crowhurst [2010].

In order for the genome to incorporate the plasmid, the temperature of the cultures needs to be raised to 68 °C. At this temperature, the plasmids can no longer replicate autonomously because *repB* has lost functionality. The presence of kanamycin in the media prompts the cells to incorporate the plasmid-bound *kanR* resistance cassette into the genome in order to survive, and the homology region of the plasmid facilitates finding the target where the genome recombination is to take place.

After the first recombination event, the genome contains the non-functional *ldh* as well as the remaining part of the functional *ldh*, *i.e.*, the second homology region that didn't recombine, remained in the genome after the first cross-over event.

During the second cross over event, shown in **Figure 63**, the part of the genome that contains both copies of *ldh* loops-in, and thus the second homology region of the non-functional copy is aligned to the second homology region of the functional copy. The non-functional copy replaces the functional copy in its entirety and the plasmid components that

were in the genome are lost in this process. The outcome is therefore a strain that is Δldh , and has lost resistance to kanamycin.

5.1.2. PYRE COUNTER-SELECTION FOR SECOND CROSS-OVERS

The problem with the approach that has been used so far by Crowhurst [2010], Taylor [2007] and Cripps *et al.* [2009] is that, while it is easy to select for the first recombination event, there is no selection for the second recombination event, so hundreds of colonies have to be restreaked everyday on kanamycin and non-kanamycin containing plates in the hope that within ten days one colony will appear that has lost resistance to kanamycin. This is extremely labor-intensive and time consuming and shouldn't be necessary in an era of synthetic biology. Therefore, this work investigates uracil auxotrophy as a second selection marker that could be used:

- As a selection marker alternative to *kanR*.
- For the quick selection of the second cross-over event in knock-outs.

Heap *et al.* in 2012 reported a system to do genomic recombinations in *Clostridium acetobutylicum* that involved creating a uracil-auxotrophic strain as a baseline to do subsequent genomic modifications.

pyrE in *C. acetobutylicum* is the equivalent of the well-known *ura3* marker of *Saccharomyces cerevisiae*. A uracil-auxotrophic strain is created when *pyrE* is disrupted by a genomic recombination event and similarly requires two homology regions that are identical in their DNA sequence to the part of *pyrE* that requires to be knocked out, as shown in **Figure 64**.

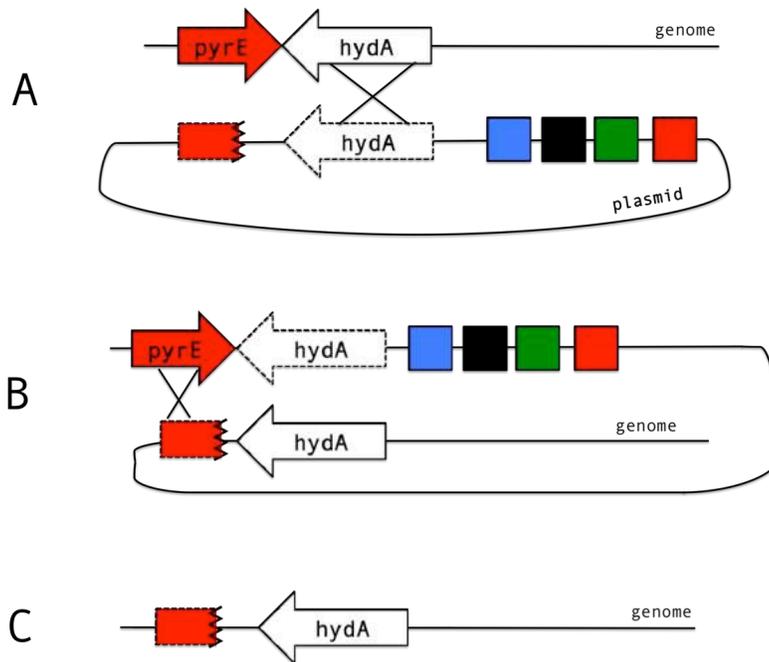


Figure 64. Creation of a *pyrE* knock-out strain. (A) A plasmid needs to contain two homology regions identical to the genomic region it needs to recombine into. In the case of Heap *et al.* [2012], these regions were 1200 bp and 300 bp. (B) the first cross-over event occurs between the 1200 bp region corresponding to *hydA*, which in the gene proximal to *pyrE* in the *C. acetobutylicum* chromosome, and the 1200 bp region. Its larger size means that this is the region that has the highest likelihood of recombining first. (C) The second cross-over occurs between the smaller 300 bp region and *pyrE*, which encodes an orotate phosphoribosyltransferase (EC 2.4.2.10), thus yielding a uracil auxotrophic strain when *pyrE* becomes fractured after losing half of its 3'-end sequence in the second cross-over event.

Figure modified from Heap *et al.* [2012].

A uracil auxotrophic strain can be selected in two ways: (1) by supplementing it with uracil and (2) by using 5-FOA (5-fluoroacetic acid). The orotate phosphoribosyltransferase encoded by *pyrE* has the capacity to convert 5-FOA into the compound 5-fluorouracil that will cause cell death. Therefore, supplementing the growth medium with 5-FOA results in colonies that do not encode a functional *pyrE* and which therefore should be uracil-auxotrophic. 5-FOA is a thermostable chemical compound [Suzuki *et al.* 2012] so it can be used for selection at high temperatures.

Using a *pyrE* knock-out would solve the problem of lack of selection for the second cross-overs in *G. thermoglucosidans* knock-in procedures. The first cross-over in this thermophile is forced by incubating at 68 °C. The second cross-over could be forced by the addition of 5-FOA to a $\Delta pyrE$ strain. A knock-in into the *G. thermoglucosidans* genome could be achieved by following the procedure illustrated in Figure 65. However, using such approach would mean that genes can only be knocked-in in the *pyrE* locus, because having

a functional *pyrE* confers an advantage over **uracil auxotrophy**, and the recombination is more likely to occur if the cells can take up a function they need.

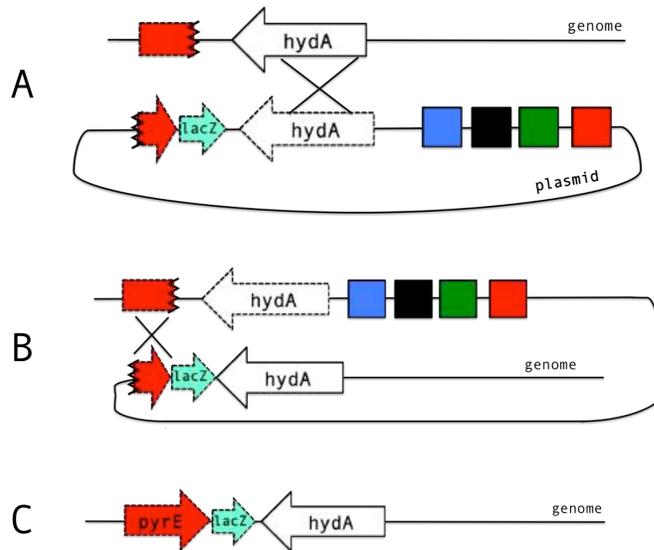


Figure 65. Knock-in of *lacZ* into the chromosome of *Clostridium acetobutylicum* as described by Heap *et al.* [2012]. (A) A plasmid contains a 1200 bp and a 300 bp identical regions to the *pyrE* region. The 300 bp region corresponds to the 3'-end of the fractured sequence of *pyrE* produced at the end of the second cross-over shown in **Figure 64**. In the middle of these two homology regions is a sequence corresponding to any gene of interest that is to be knocked-in, in this case, the authors used *lacZ*. The first cross-over occurs between the 1200 bp region and its homologous counter-part *hydA*. (B) The second cross over occurs between the region corresponding to the 3'-end deleted part of the *pyrE* gene together with a *pyrE* homology region. (C) The functionality of the *pyrE* region has been restored after the fracture was repaired after the second recombination.

Figure modified from Heap et al. [2012].

5.2. RESULTS

5.2.1. KNOCK-IN OF SFGFP IN THE *G. THERMOGLUCOSIDANS* GENOME

The design for the genomic recombination experiment followed a modified **version of** the procedure used by Crowhurst [2010], Taylor [2007] and Cripps [2009] but with slight modifications. As shown in **Figure 65**, the first step needs to be the creation of a vector that contains a temperature-sensitive origin of replication, such as *repB*, which becomes inactive at a temperature lower than the upper temperature limit at which *G. thermoglucosidans* can grow. This vector also needs to contain a thermostable antibiotic resistance marker for selection.

Two recombination events are required. The sequence of events that should lead to the first recombination event are illustrated in **Figure 66**. The second cross-over is illustrated in **Figure 67**. It was decided that the *ldh* gene should be the target for inactivation because it has been demonstrated that Δldh cells are viable (*G. thermoglucosidans* DL44).

In order to select for the first cross-over, the culture of *G. thermoglucosidans* needs to be incubated at 68 °C to inactivate *repB* and promote recombination via one of the two homology regions. The presence of kanamycin acts as a selection marker to ensure that the cells incorporate the entire plasmid into the genome.

The selection of the second cross-over is much more difficult because it cannot be forced to happen. There is no selection available. The cells need to loop-out the *kanR* gene together with the rest of the plasmid components on their own. The way Crowhurst [2010], Taylor [2007] and Cripps [2009] approached this problem was by continuously replica-planting hundreds of colonies in both 2TY and 2TY/kanamycin plates until one colony is found that can no longer survive in the presence of kanamycin.

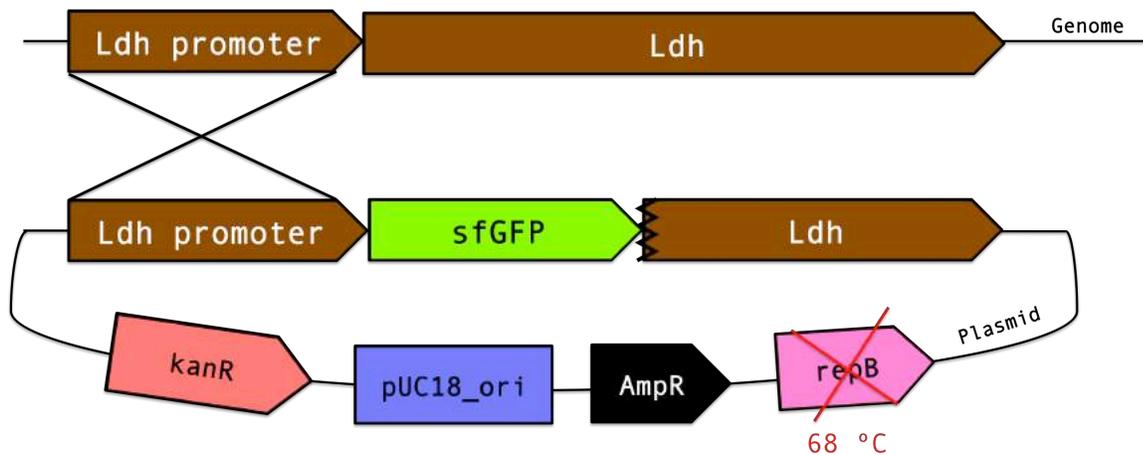


Figure 66. Sequence of events leading to the first cross-over event. The temperature-sensitive origin of replication *repB* becomes inactive at 68 °C, so at this temperature the plasmid can no longer replicate autonomously. Selection with kanamycin prompts the incorporation of the entire plasmid into the genome by recombination between the *ldh* promoter regions of the plasmid and of the genome.

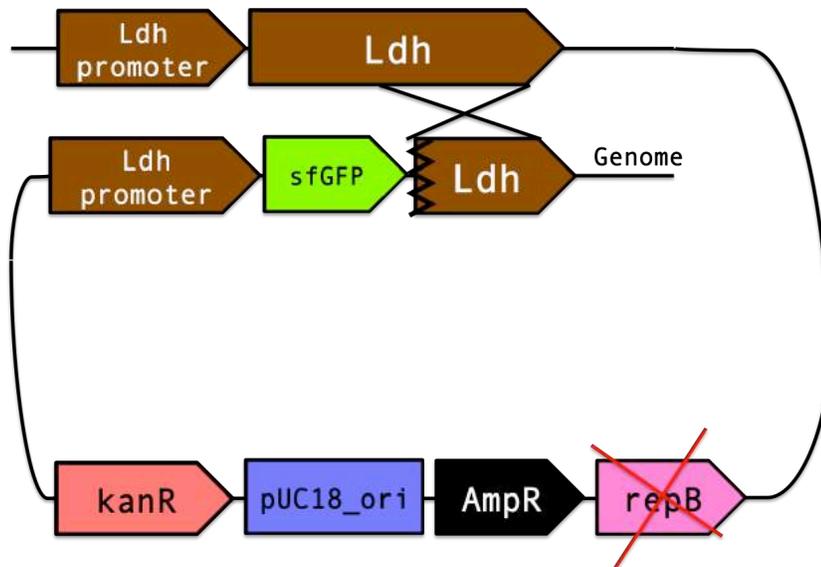


Figure 67. Sequence of events leading to the second cross-over event. The entire plasmid has been incorporated into the genome. Now there's two copies of the *ldh* gene, the genomic one and the one interrupted by *sfGFP*. The second recombination event occurs when the genome loops-in and both copies anneal, after which only one copy will remain and the plasmid components, including *kanR*, are looped-out as a result.

5.2.1.1. CONSTRUCTION OF THE KNOCK-OUT/IN PLASMID

In order to demonstrate the functionality of a plasmid that was created earlier as part of this project [Martinez-Klimova, 2011] which contained a longer version of the temperature sensitive origin of replication "*repB*" than the one included in p12AK0, as well as the thermostable kanamycin resistance marker, an experiment was designed to attempt to do a knock-out of *ldh* in DL33 by interrupting it with the sequence of sfGFP.

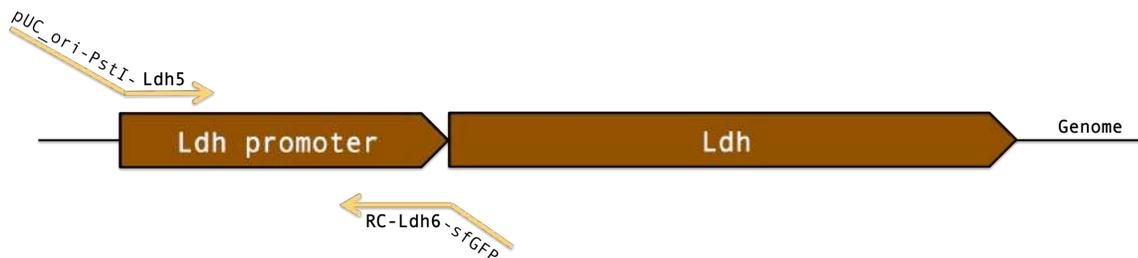


Figure 68. PCR-amplification of the first homology region using primer pair pUC_ori-PstI-Ldh5 and RC-Ldh6-sfGFP were used to amplify a 500 bp region corresponding to the *ldh* promoter of the genome of *G. thermoglucosidans* DL33.

As shown in **Figure 68**, primers pUC_ori-PstI-Ldh5 and RC-Ldh6-sfGFP were used to amplify a 500 bp region corresponding to the *ldh* promoter of the genome of *G. thermoglucosidans* DL33. The forward primer in turn added the *PstI* site and a 35 bp homology region to the 3'-end of pUC18_ori and the reverse primer added 35 bp of homology to *sfGFP*.

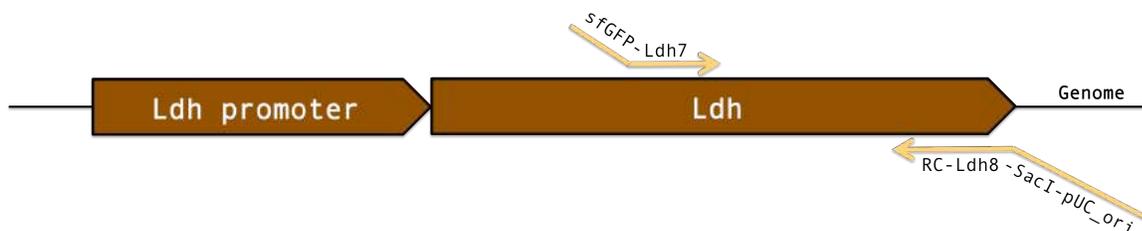


Figure 69. PCR-amplification of the second homology region using primers sfGFP-Ldh7 and RC-Ldh8-SacI-pUC_ori, which amplified a 500 bp region at the 3'-end of the *ldh* gene of *G. thermoglucosidans* DL33.

Figure 69 shows the amplification of the second homology region of 500 bp at the 3'-end of the *G. thermoglucosidans* DL33 gene with primers sfGFP-Ldh7 and RC-Ldh8-SacI-pUC_ori. The forward primer in turn added 35 bp of homology to sfGFP and the reverse primer added a *SacI* site as well as 35 bp of homology to pUC_ori.

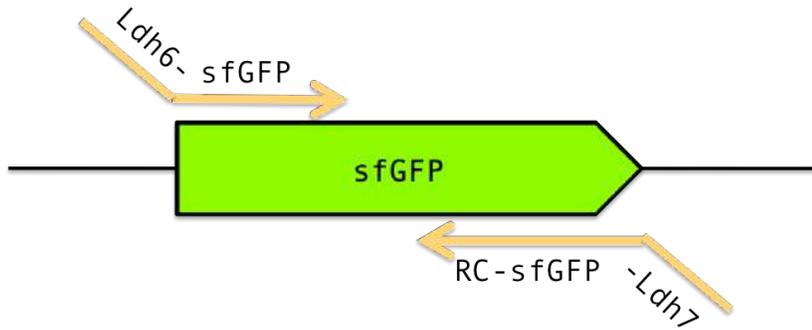


Figure 70. PCR-amplification of the gene encoding for sfGFP using primers Ldh6-sfGFP and RC-sfGFP-Ldh7.

Figure 70 shows the amplification of the gene encoding sfGFP using primers Ldh6-sfGFP and RC-sfGFP-Ldh7. The forward primer in turn added 35 of homology to the promoter region of *ldh* PCR-amplified from **Figure 68**, and the reverse primer added 35 of homology to the 3'-region of *ldh* PCR-amplified from **Figure 69**. The three fragments from **Figures 68, 69** and **70** were joined together using Gibson isothermal assembly and the product was PCR-amplified using primers pUC_ori-PstI-Ldh5 and RC-Ldh8-SacI-pUC_ori (see **Figure 71**) and digested with *PstI* and *SacI*.

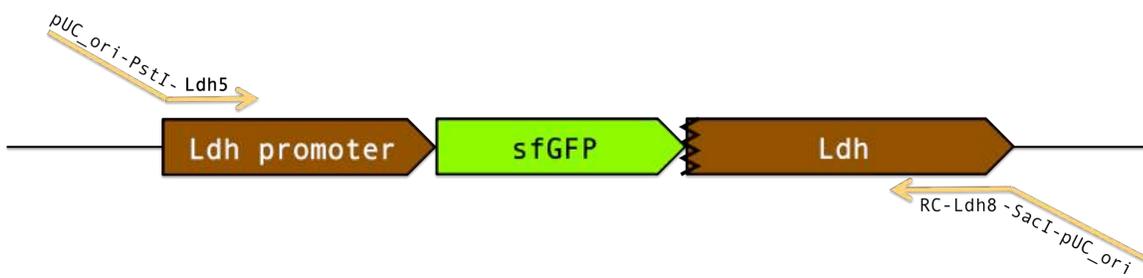


Figure 71. PCR-amplification of the Gibson isothermal assembly reaction product using primers pUC_ori-PstI-Ldh5 and RC-Ldh8-SacI-pUC_ori.

The plasmid containing the temperature sensitive origin and the thermostable *kanR* was digested with *PstI* and *SacI* and the *PstI/SacI* PCR-amplification of the Gibson assembly product were ligated together. The plasmid resulting from the ligation is shown in **Figure 72**.

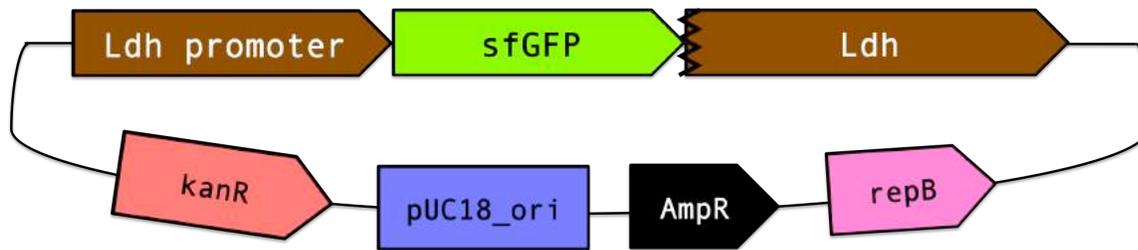


Figure 72. Plasmid resulting from the *Pst*/*Sac*I ligation of the PCR-amplification of the Gibson isothermal assembly reaction product with the vector containing the temperature-sensitive origin of replication *repB* and the thermostable *kanR* gene.

The plasmid containing the temperature sensitive origin and the thermostable *kanR* as well as the two homology regions to the *ldh* promoter and to the 3'-end of *ldh* was transformed into *G. thermoglucosidans* DL33, but no colonies were retrieved when selecting the transformed cells on 2TY/kanamycin plates at 55 °C.

The same plasmid was transformed into *G. thermoglucosidans* DL44. In this case, colonies were retrieved after plating the transformation on 2TY/kanamycin plates at 55 °C but the colonies lacked green fluorescence even though the gene encoding *sfGFP* was placed in frame with the *ldh* promoter: *ldh*, since the start codon, was replaced by *sfGFP*.

An analysis carried out using the reverse function of the RBS Calculator¹³ revealed that the RBS of the native *ldh* promoter of DL33 has the very low transcription initiation rate of 48.66 au (**Figure 73**), possibly due to low compatibility of the RBS with the sequence of *sfGFP*, which might explain the absence of fluorescence.

¹³ [[https://www.denovodna.com/ software/](https://www.denovodna.com/software/)]

| Method | Title | Summary |
|---------------|--------------------|---|
| RBS For. Eng. | Gth_pldh_sfGFP_FWD | After 400 iterations, a synthetic RBS with 4114.39 T.I.R ACCACCTCAAATACGAAGGACAATAATTAATTA |
| RBS Rev. Eng. | Gth_pldh_sfGFP_REV | 10 start codons from 2.40 to 3842.30 T.I.R |

Submitted: 05/12/2013 Status: Finished CPU Time: 0:02:12.5 Organism: Geobacillus sp. C56-T3 (ACCTCCTTT)
 Version: v1.1 Unique Identifier: nBr7Qlih

mRNA Sequence
 CATT TTTTGAAAGGATGACAGACAGCGATGCGTAAAGGCGAAGAGCTGTTCAC TGGTGTCGTCCTATTCTGGTGGAAGCTGGATG

| Start Position | Translation Initiation Rate (au) | ΔG_{total} | $\Delta G_{mRNA-rRNA}$ | $\Delta G_{spacing}$ | $\Delta G_{standby}$ | ΔG_{start} | ΔG_{mRNA} | Structure | Accuracy[?] |
|----------------|----------------------------------|--------------------|------------------------|----------------------|----------------------|--------------------|-------------------|------------|-------------|
| 6 | 140.26 | 6.4 | 1.32 | 1.52 | -0.0 | -0.04 | -3.6 | Click here | OK |
| 15 | 144.77 | 6.33 | -4.28 | 0.01 | 0.0 | -1.19 | -11.8 | Click here | NEQ |
| 28 | 48.66 | 8.75 | -6.98 | 1.73 | 0.0 | -1.19 | -15.2 | Click here | NEQ |
| 46 | 2.43 | 15.41 | -6.78 | 1.52 | 0.0 | -0.03 | -20.7 | Click here | NEQ |
| 53 | 23.34 | 10.38 | -5.28 | 0.0 | 0.0 | -0.03 | -15.7 | Click here | CDS OLS |
| 56 | 5.18 | 13.73 | -2.18 | 0.29 | 0.0 | -0.07 | -15.7 | Click here | CDS NEQ OLS |
| 70 | 136.75 | 6.46 | -2.88 | 0.67 | 0.0 | -0.03 | -8.7 | Click here | CDS NEQ OLS |
| 73 | 269.49 | 4.95 | -1.68 | 0.01 | 0.0 | -0.07 | -6.7 | Click here | CDS OLS |
| 79 | 967.13 | 2.11 | -3.38 | 0.01 | 0.0 | -0.03 | -5.52 | Click here | CDS NEQ OLS |
| 83 | 3842.35 | -0.96 | -5.28 | 0.0 | 0.0 | -1.19 | -5.52 | Click here | CDS NEQ OLS |

All Gibbs free energies (ΔG) are reported in units of kcal/mol.

Figure 73. RBS calculations to determine the suitability of the *ldh* native RBS with sfGFP. The Reverse Engineering function of the Salis RBS Calculator was used to analyze whether the native RBS of the genomic *ldh* of *G. thermoglucosidans* is compatible with sfGFP. The analysis identifies several potential RBS sites and measures translation initiation rates. The result of the analysis of the native RBS of *ldh* is highlighted in yellow. The term NEQ suggests that the parameters of the analysis could not be met and therefore it is likely that the native RBS of *ldh* is not compatible with the sequence of sfGFP, which would explain the absence of fluorescence.

Even though the absence of green fluorescence was certainly a desirable property, it was not crucial for the experiment. The first recombination event was enforced after *G. thermoglucosidans* DL44 cells transformed with the plasmid that contained the knock-out cassette was incubated in liquid 2TY medium, supplemented with 12 $\mu\text{g}/\text{mL}$ kanamycin at 68 $^{\circ}\text{C}$ while shaking overnight, after which, 300 μL of the culture were plated on a 2TY plate supplemented with kanamycin and left to grow overnight and also, 300 μL of the same culture were used to inoculate fresh 2TY medium supplemented with 12 $\mu\text{g}/\text{mL}$ kanamycin.

No colonies were observed on the plate after overnight incubations, but growth was observed in the liquid 2TY. Therefore, it was decided that instead of restreaking hundreds of colonies on plates over the course of the next several days, as per the procedure followed by Crowhurst [2010], Taylor [2007] and Cripps [2009], a variation using liquid 2TY medium would be employed: 300 μL of the liquid 2TY culture in which growth was observed at 55 $^{\circ}\text{C}$ were used to inoculate fresh liquid 2TY medium and while not supplementing with antibiotic, the cells were left to grow overnight. The following day, 300 μL of this new culture were used

to once more inoculate 300 μ L of liquid 2TY medium, without antibiotic, and were left to grow overnight. The same produce was repeated for another five days, after which 300 μ L of the culture were plated on a 2TY plate and left to incubate overnight at 55 $^{\circ}$ C. Thirteen of the resulting colonies were replica plated on two fresh 2TY plates with and without 12 μ g/mL kanamycin. All colonies were able to grow on 2TY as well as on 2TY kanamycin. None of them lost the kanamycin resistance gene.

These thirteen colonies were subjected to colony PCR with primers SEQ-Ldh-KO and SEQ-RC-Ldh-KO. These primers align in the genome of DL44, in locations adjacent to the *ldh* gene, as illustrated in **Figure 74**.

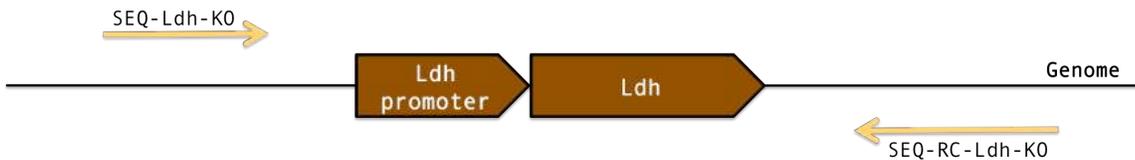


Figure 74. Primers SEQ-Ldh-KO and SEQ-RC-Ldh-KO anneal to regions adjacent to the *Ldh* locus in the genome of *G. thermoglucosidans* DL44. This primer pair does not anneal anywhere in the plasmid.

The results of the PCR amplification are shown in **Figure 75**. Only three colonies produced bands. Based on the size, it can be deduced that two of them reverted back to wild-type and did not incorporate the *sfGFP* gene, whereas one colony appears to have both phenotypes within the same colony, *i.e.*, some of the cells within the same colony reverted back to wild-type and some of the cells within the same colony did incorporate the knock-in of *sfGFP* in the *ldh* locus.

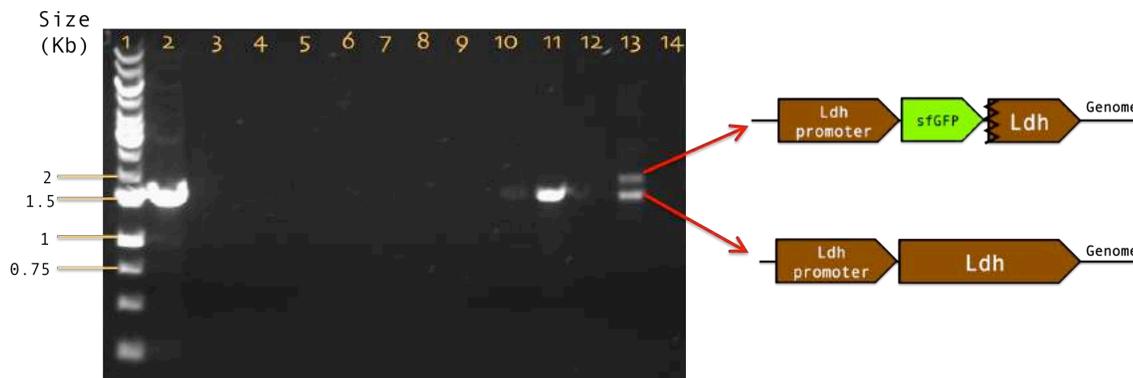


Figure 75. Colony-PCR screening results of the *Ldh* locus using primers SEQ-Ldh-KO and SEQ-RC-Ldh-KO. Primer pair SEQ-Ldh-KO and SEQ-RC-Ldh-KO are primers that anneal in the genomic region flanking the *Ldh* gene and do not anneal anywhere in the plasmid. Therefore, this primer pair was used to PCR-amplify the genomic content of the colonies that underwent several days of subculturing in order to determine if the plasmid is still present (*i.e.* if the genetic content of the cells is still that obtained after the first recombination event) or if the second recombination event had occurred. Lane 1: molecular size marker (Fermentas, 1 kb). Lanes 2-14: thirteen colonies were screened. Size band expected for the wild-type gene was 1.5 kb, and for the knock-in, the size expected was 2 kb. The colonies from lanes 2 and 11 have either (1) reverted back to wild-type or (2) never incorporated the plasmid so didn't undergo the first recombination event. The colonies from lanes 3-10, 12 and 14 probably still contain the entire plasmid in the genome and no bands are seen because the fragment was too large to amplify (nearly 6 kb). Lane 13 contains two bands which can be attributed to both phenotype states: some of the cells within the same colony have reverted back to wild-type and some of the cells have incorporated the knock-in of *sfGFP* in the *Ldh* locus.

The knock-in experiment demonstrates one thing: it is of the utmost necessity to find a method to force the second recombination event to happen, as it is not ideal to screen hundreds of colonies and wait for the desired phenotype to occur.

5.2.2. *PYRE* KNOCK-OUT IN *G. THERMOGLUCOSIDANS*

The presence of *pyrE* in the genome of *G. thermoglucosidans* was investigated using the KEGG database¹⁴. The results are shown in **Figure 76**. *Geobacillus* sp. Y4.1MC1 was selected as the species of reference because, as shown in **Figure 5**, it is the closest relative to *G. thermoglucosidans*.

¹⁴ [<http://www.genome.jp/kegg/kegg2.html>]

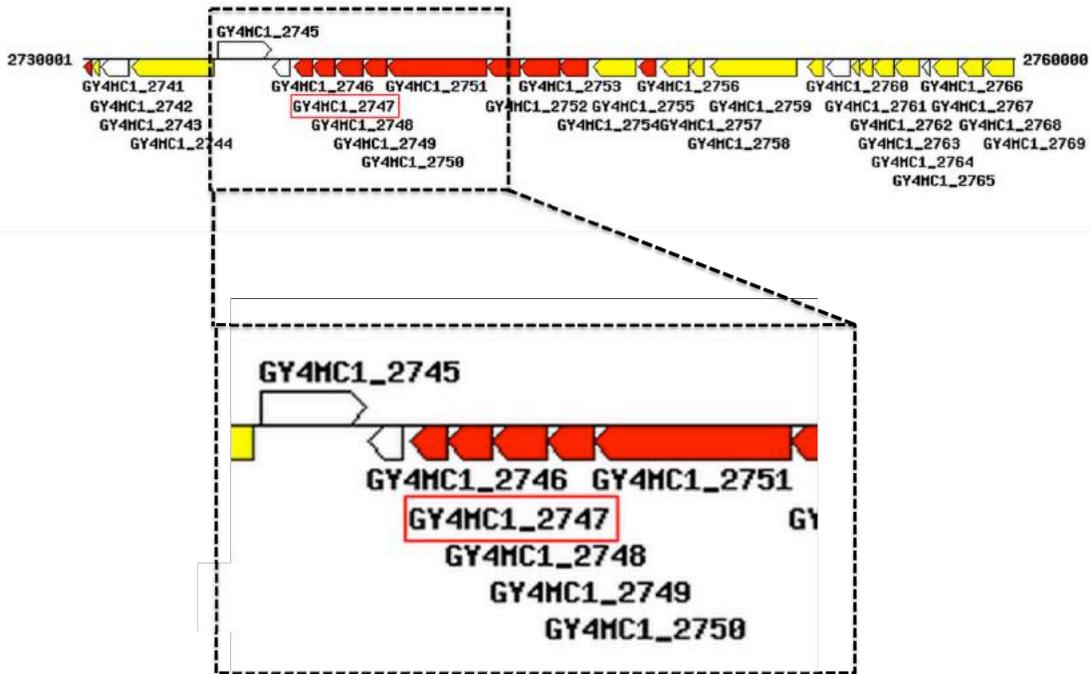


Figure 76. Genome map of *Geobacillus* sp. Y4.1MC1 highlighting (red box) the location of the *pyrE* gene in the genome. *pyrE* (orotate phosphoribosyltransferase) corresponds to gene GY4MC1_2747. Gene GY4MC1_2745 corresponds to a fibronectin-binding A domain-containing protein, GY4MC1_2746 is a pseudogene, GY4MC1_2748 corresponds to *pyrF* (orotidine 5'-phosphate) decarboxylase, GY4MC1_2749 is a dihydroorotate dehydrogenase and GY4MC1_2750 is a dihydroorotate dehydrogenase, electron transfer subunit, iron-sulfur cluster binding domain-containing protein.

The sequence corresponding to *pyrE*, as well as the sequence downstream from *pyrE* was compared between *Geobacillus* sp Y4.1MC1 and the genome sequence of *G. thermoglucosidans* DL33, provided by Ward [2014]. After it was confirmed that they match, the DL33 genome sequence was used as template to design primers to amplify the two homology regions that would result in a *pyrE* knock-out following the procedure by Heap *et al.* [2012]. The primers anneal to the regions shown in **Figure 77**.

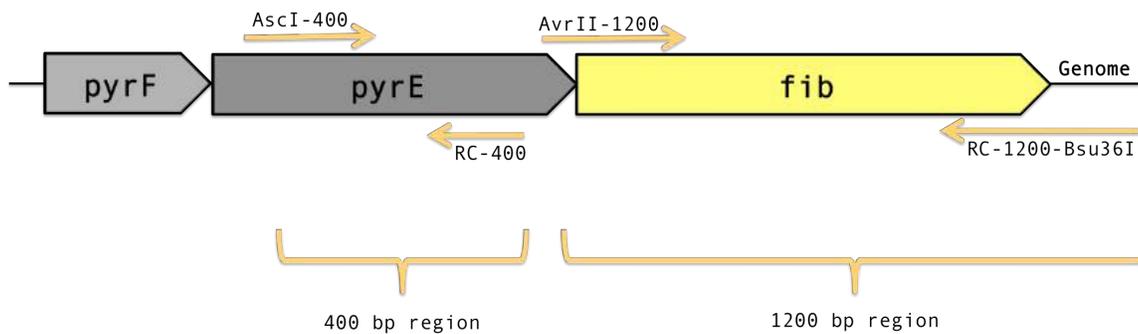


Figure 77. Primers AscI-400 and RC-400 were designed to amplify the first 400 bp homology region. Primers AvrII-1200 and RC-1200-Bsu36I were designed to amplify the second 1200 bp homology region.

As it can be seen in **Figure 78**, bands of the correct size were obtained when the 400 bp region was amplified with primers AscI-400 and RC-400. This region is present in DL33, DL44 and TM89. This last strain is also an *ldh* knock-out of a wild-type strain.

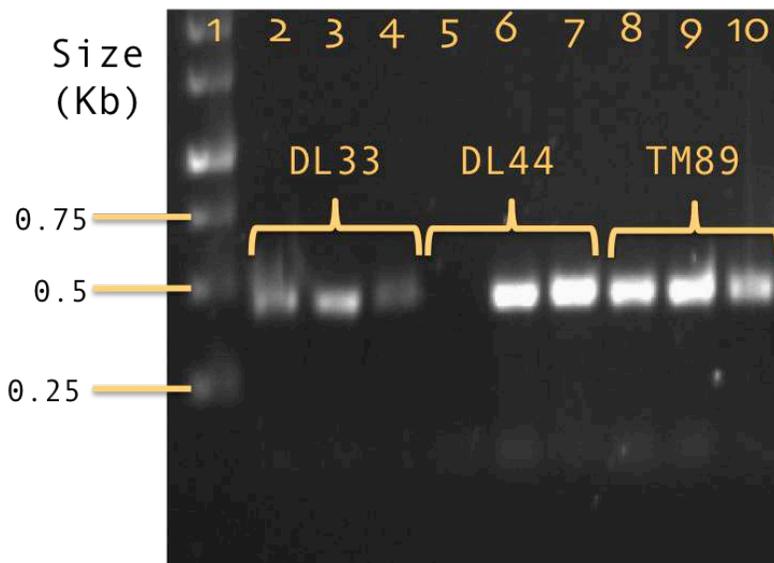


Figure 78. PCR-amplification of the 400 bp region using primers AscI-400 and RC-400. Lane 1: molecular size ladder. Lanes 2-4: DL33 genomic DNA was used as template. Lanes 5-7: DL44 genomic DNA was used as a template. Lanes 8-10: genomic DNA from TM89 was used as template. DL33 is the wild-type strain of *G. thermoglucosidans*. Both DL44 and TM89 are Δldh strains.

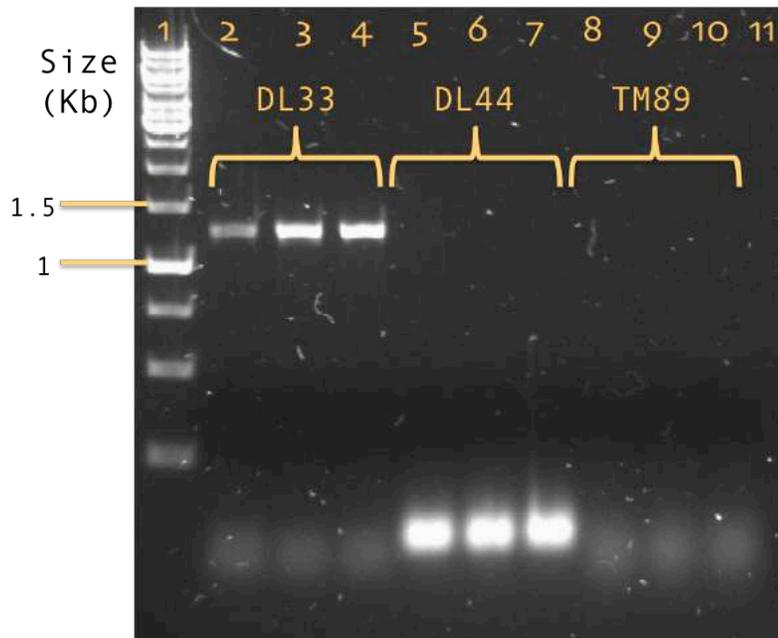


Figure 79. PCR-amplification of the 1200 bp region using primers AvrII-1200 and RC-1200-Bsu36I. Lane 1: molecular size ladder. Lanes 2-4: DL33 genomic DNA was used as a template. Lanes 5-7: DL44 genomic DNA was used as a template. Lanes 8-10: genomic DNA from TM89 was used as template.

As it can be seen in **Figure 79**, when the 1200 bp second homology fragment was PCR-amplified with primers AvrII-1200 and RC-1200-Bsu36I, a band of the correct size was produced from DL33, but not from DL44 or TM89.

5.3. DISCUSSION

5.3.1. WHY IS IT IMPORTANT TO KNOW THE PARTS OF THE VECTOR IN USE?

In the knock out experiments of Taylor [2007], Cripps *et al.* [2009] and Crowhurst [2010], it is evident that the lack of knowledge about the parts composing the vector used to knock-out *ldh* in *G. thermoglucosidans* resulted in an unnecessarily inefficient process: the selection marker being used was the wrong one. The fact that Cripps *et al.* used the mesophilic KanR (derived from pUB190) instead of the thermophilic one (derived from pBST22) indicates that, in order for the procedure to have worked, there must have been some residual activity of the mesophilic KanR at 65 °C. However, this could only have been fortuitous and less than optimal. In subsequent discussions with Dr. B. Rudd [*personal communication*], who developed pUB190, the use of a mesophilic KanR would explain why

the colonies that were selected had the pUB190_kanR incorporated multiple times in the targeted gene. So the only colonies that were able to tolerate the presence of kanamycin were those that were expressing a significant number of copies of the mesophilic kanamycin nucleotidyl transferase. This also would explain why Taylor [2007], Cripps *et al.* [2009] and Crowhurst [2010] had to screen hundreds of colonies by restreaking in kanamycin and non-kanamycin plates for ten days to find a single colony that has lost kanamycin resistance that might indicate that the second recombination has taken place.

- In the future, knock-outs should be done using the vector included this toolbox that contains the temperature-sensitive origin of replication repB and the thermophilic kanR to avoid multiple genomic copies of *kanR*.

5.3.2. KNOCK-OUT/IN IN *G. THERMOGLUCOSIDANS*

It was not possible to knock-out the *ldh* gene in *G. thermoglucosidans* DL33, which is the wild-type strain, perhaps because the size of *sfGFP* was too large to be incorporated in the genome. To the best of the author's knowledge, it had never been attempted before to do a knock-in in *G. thermoglucosidans*, especially of a fragment the size of *sfGFP*. In addition, it is most likely that the plan to knock a gene out by knocking a second gene in its place was extremely ambitious and the DL33 strain, which does not appear to be as resilient to stress as DL44, could not cope with the stress caused by having to recombine the plasmid to the genome and rearranging its genome and its metabolic pathways fast enough to survive. Second, it appears that trying to avoid replica-plating hundreds of colonies by growing them in liquid medium instead is not a good strategy to obtain the second cross-over strain. There is more control over the entire process when each colony is identified and marked and followed through. Third, the fact that the second recombination cannot be forced to happen due to the lack of a counter-selection mechanism is a severe limitation to the creation of knock-out or knock-in strains in this thermophile.

5.3.3. PYRE KNOCK-OUT IN *G. THERMOGLUCOSIDANS* DL33

The lack of a band for the 1200 bp second homology region indicates that DL44 is not genetically identical to DL33 except for the *ldh* gene. And the same seems to be the case for TM89, which also happens to be an *ldh* knock-out strain. It appears that knocking-out the *ldh* gene brought significant stress that resulted in a rearrangement of the genome, and a genome sequence of DL44 is required if a *pyrE* knock-out is to ever be created using this strain. Unfortunately, due to lack of time, the *pyrE* knock-out of DL33 was not pursued any further.

5.4. CONCLUSION

The creation of the *pyrE* knock-out strain would permit having uracil auxotrophy as a second selection marker and it would put in evidence that the vectors in the toolbox paired with uracil auxotrophy can be used in conjunction as an improved strategy to do knock-outs, which is a requirement for *G. thermoglucosidans* to become a chassis so its potential for various downstream applications can be explored.

5.5. FUTURE WORK

It is necessary to create a *pyrE* or a *pyrF/pyrR* knock-out strain to test if 5-FOA can be used as a selection marker. This would significantly speed up genomic recombinations. However, it seems that the genomes among strains are significantly variable, at least in the *pyrE* region, so it will be necessary to know the genome sequence, at least of *pyrE* and its surroundings for DL44 and for any other strain that's more used than DL33.

6. CHAPTER 6:

ISOBUTANOL PRODUCTION BY *GEOBACILLUS THERMOGLUCOSIDANS*

The key contributions and results of this chapter are:

1. A pathway for isobutanol production was introduced in *G. thermoglucosidans* as a proof-of-principle demonstration that the new modular shuttle vectors can be efficiently used for the metabolic engineering of *G. thermoglucosidans* to achieve the production of valuable compounds that are not ethanol.
2. The production of 50 mg/L of isobutanol was achieved under aerobic growth conditions.
3. It was discovered that a native enzyme (acylating aldehyde dehydrogenase, or ALDH) of *G. thermoglucosidans* has the potential to produce isobutanol.
4. This work presents the first experimental confirmation that ALDH has the capacity to produce isobutanol from 2-ketoisovalerate as a precursor.
5. This work also confirms the findings of Lin *et al.* [2014] regarding the poor thermostability of Kdc.

6.1. INTRODUCTION

The potential of *G. thermoglucosidans* for useful applications needs to be explored in order to justify the effort invested into developing tools that permit the efficient metabolic modification of this thermophile because:

- Demonstrating the utility of *G. thermoglucosidans* would promote its candidacy as a chassis for synthetic biology.
- The use of *G. thermoglucosidans* so far has only been limited to the industrial production of ethanol (TMO Renewables), but that doesn't have to be the case.
- *G. thermoglucosidans* could be used for the production of thermostable pharmaceuticals, fuels or chemical precursors from cheap feedstocks.
- *G. thermoglucosidans* attracted attention from the biotechnological research community because (1) it is a natural producer of ethanol, (2) it has tolerance to elevated ethanol concentrations (up to 10%), (3) it is capable of degrading cheap

lignocellulosic feedstocks to use them as a carbon source and (4) the several advantages that running industrial application at high temperatures brings. These four properties made it an environmentally friendly host for the industrial production of bioethanol.

- However, *G. thermoglucosidans* could be the producer of other fuel substituents as an example, to promote the potential uses of this thermophile.
- Ethanol is an alcohol, so the present work explores the possibility to engineer pathways for the production of other alcohols, in particular, higher chain alcohols, which are being investigated for their potential to become fuel substituents and replace ethanol.
- Isobutanol is one of the most researched higher chain alcohols due to its high energy content and octane number that are very similar to gasoline.
- If this work demonstrates that it is possible to modify *G. thermoglucosidans* to introduce a pathway for isobutanol production, it could open the doors to the scientific exploration of:
 1. The use of the toolbox for the metabolic engineering of *G. thermoglucosidans*, which in turn would result in further contributions to the toolbox.
 2. The use of *G. thermoglucosidans* or other *Geobacillus* species for the production of useful compounds from cheap feedstocks.
 3. The enhanced study of the biology of thermophiles brought by a change in perspective achieved when thermophiles are seen as platforms of production instead of sources of thermostable enzymes.

6.1.1. BUTANOLS AS FUELS

Butanols have been proposed as substitutes or supplements for gasoline, since they are particularly advantageous as biofuels [Atsumi and Liao, 2008, Chen *et al.* 2011]. They produce less sulphur and nitrogen oxide emissions in comparison to gasoline and, given that they have more carbon-hydrogen bonds than ethanol, their energy content is higher [Altun *et al.* 2011]. n-Butanol, for instance, has an energy content of 27 MJ/L, which is similar to the 32 MJ/L of gasoline [Atsumi and Liao, 2008]. Isobutanol, on the other hand, is an isomer of 1-butanol. It has similar physical and chemical properties as n-butanol, but it has a higher octane number due to the branched CH₃ group [Atsumi and Liao, 2008]. **Table 11** shows the physical and chemical properties of isobutanol.

| | |
|---------------------|----------------------------------|
| IUPAC nomenclature | 2-methylpropan-1-ol |
| Molecular formula | C ₄ H ₁₀ O |
| Molar mass | 74.122 g/mol |
| Density | 0.802 g/cm ³ |
| Melting point | -101.9 °C |
| Boiling point | 107.9 °C |
| Solubility in water | 8.7 mL/100 mL |
| Refractive index | (n _D) 1.3959 |
| Viscosity | 3.95 cP at 20°C |
| Appearance | Colorless liquid |
| Solubility in water | 8.5 % at 20 °C and 7.5% at 30 °C |

Ramey [2007] demonstrated that n-butanol can completely replace gasoline, with no modifications to the engine but there hasn't been a report of a car that has been tested with 100% isobutanol in the tank. Testing performed regarding the use of butanols as gasoline substituents show that engines can operate when up to 20% isobutanol is mixed with gasoline [Karabektas and Hosoz, 2009]. So the 20% vs. 100% butanol/isobutanol-in-the-tank uncertainty still remains to be solved, but in the meantime, both butanol and isobutanol could be added to gasoline as additives that constitute less than 20% of the mixture [Bruno *et al.* 2009]. Butanols are also attractive as fossil-fuel substituents because water-containing butanol blends with diesel result in reduced emissions of persistent organic pollutants¹⁵ (POP) [Chang *et al.* 2014].

6.1.2. TOXICITY OF BUTANOLS

The successful biosynthesis of butanols by a non-natural producer is still a challenge because heterologous pathways interfere with the native metabolism of the cell [Connor and Liao, 2009] and because they are toxic for the cells [Brynildsen and Liao, 2009]. For instance, *Saccharomyces cerevisiae* cannot tolerate more than 4% butanol [Zaki *et al.*

¹⁵ including polycyclic aromatic hydrocarbons, polychlorinated dibenzo-p-dioxins and dibenzofurans, polychlorinated biphenyls, polychlorinated diphenyl ethers, polybrominated dibenzo-p-dioxins and dibenzofurans, polybrominated biphenyls and polybrominated diphenyl ethers.

2014], *Zymomonas mobilis* and *Clostridium acetobutylicum*, *E. coli*. Some industrial yeast strains can grow under 25% (v/v) ethanol [Shi *et al.* 2009]. In contrast, *E. coli* is inhibited by 3% ethanol and the cyanobacteria *Synochocystis* sp. is inhibited by 1.5% ethanol (v/v), 0.2% butanol. *Lactobacillus* spp. have tolerance to up to 18% (v/v) ethanol and up to 3% butanol. *Lactobacillus* spp. are considered to be the species with the most biofuel tolerance capacity by Liu *et al.* [2008].

| Table 12. Tolerance to alcohols by natural and non-natural producers of ethanol and butanol. According to [Zaki <i>et al.</i> 2014, Shi <i>et al.</i> 2009, Liu <i>et al.</i> 2008, Wu <i>et al.</i> 2010]. | | |
|---|-------------------------|-------------------------|
| Organism | Ethanol tolerance (v/v) | Butanol tolerance (v/v) |
| <i>E. coli</i> | 3 % | 1 % |
| <i>S. cerevisiae</i> | 25 % | 4 % |
| <i>Lactobacillus</i> spp. | 18 % | 3 % |
| <i>Synochocystis</i> spp. (cyanobacteria) | 1.5 % | 0.2 % |

Butanols are toxic for the cells because they can accumulate in the membrane, altering its fluidity. They have also been found to decrease intracellular pH and ATP concentration, inhibit the uptake of glucose and, common to other stress responses, intracellular reactive oxygen species levels are increased. Isobutanol is also known to cause quinone depletion, which leads to respiratory distress. Solvents partition into the membrane and disrupt its fluidity, which in turn results in ion leakage [Fischer *et al.* 2008].

E. coli cannot tolerate isobutanol concentrations of more than 0.84% (v/v) under aerobic conditions [Blombach and Eikmanns, 2011]. Wu *et al.* in 2010 developed a strain of *E. coli* that can tolerate 1% (v/v). This was achieved by selection on ever-increasing amounts of isobutanol. When the resulting strain was sequenced, it was revealed that the cells were overexpressing *glmS*¹⁶, which participates in the synthesis of peptidoglycans and cell wall lipopolysaccharides. The low tolerance of isobutanol suggests that isobutanol has undesirable effects on the cell wall. Wu *et al.* discovered that the 1% (v/v) isobutanol-tolerant strain does not produce higher amounts of isobutanol than the less tolerant one, which

¹⁶ *glmS*: Glucosamine-6-phosphate synthase; glucosamine--fructose-6-phosphate aminotransferase; C-terminal F6P-binding domain has isomerase activity (EG10382)) [Zhou and Rudd, 2012. <http://www.ecogene.org/>]

means that when isobutanol is added to the growth medium, it does not permeate the cells. So, the strain developed tolerance to exogenous isobutanol but not to endogenous isobutanol, which is very possibly the reason why there was no increase in isobutanol production by this strain. Endogenously produced isobutanol is still toxic, possibly because it accumulates inside the cells.

6.1.3. NATURAL PRODUCERS OF BUTANOLS

Clostridium acetobutylicum produces butanol naturally as part of its characteristic ABE fermentation (Acetone-Butanol-Ethanol) [Kaminski, 2011], but in low yields and low maximum concentrations due to limited tolerance, which makes it an unsuitable host to produce butanol at an industrial scale [Alper and Stephanopoulos, 2009]. *Saccharomyces cerevisiae* [Chen *et al.* 2011] and *Lactococcus lactis* [Connor and Liao, 2009] produce isobutanol naturally, but in trace amounts.

6.1.4. HETEROLOGOUS PRODUCERS OF BUTANOLS

Using metabolic engineering, the genes for natural pathways that lead to the production of higher alcohols as metabolic by-products have been moved to bacterial user-friendly hosts, in particular *Escherichia coli*. One of the most influential works on the area of biofuels is that by Atsumi and Liao. In 2008, the heterologous production of butanol, isobutanol and other higher alcohols such as 1-propanol, 2-methyl-1-butanol and 3-methyl-1-butanol using *E. coli* as a production platform, **was published**. It was the first report of a successful non-native strategy for the production of isobutanol, **which** resulted in a yield of 86% of the theoretical maximum. The pathway proposed and patented by Atsumi and Liao [2008] **for isobutanol production** is summarized in **Figure 80**.

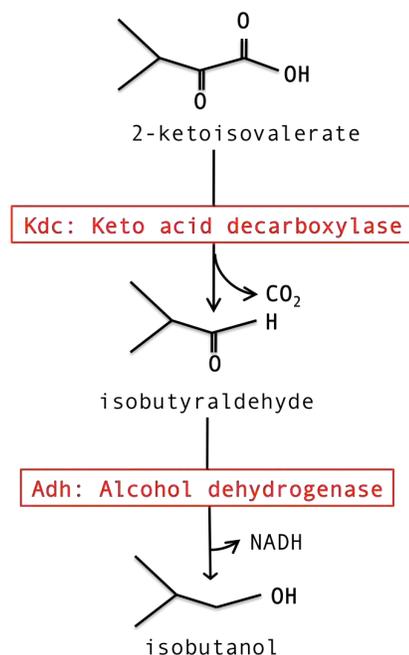


Figure 80. Atsumi and Liao's 2008 strategy to produce isobutanol in *E. coli*. 2-ketoisovalerate is the precursor to valine. A keto-acid decarboxylase (Kdc) converts 2-ketoisovalerate into isobutyraldehyde and then an alcohol dehydrogenase (Adh) converts isobutyraldehyde into isobutanol.

6.1.5. SOURCES OF 2-KETO ACID DECARBOXYLASES (KDC)

| Table 13. Kdc and Adh enzymes screened by Atsumi and Liao [2008] | | |
|--|---------------------------------------|--|
| Kdc | Plus Adh | Equals: Final isobutanol production by <i>E. coli</i> (μM) |
| <i>Pdc6</i> from <i>S. cerevisiae</i> | <i>adh2</i> from <i>S. cerevisiae</i> | Not detectable |
| <i>Aro10</i> from <i>S. cerevisiae</i> | <i>adh2</i> from <i>S. cerevisiae</i> | 2,094 |
| <i>Thi3</i> from <i>S. cerevisiae</i> | <i>adh2</i> from <i>S. cerevisiae</i> | Not detectable |
| <i>Kdc</i> from <i>L. lactis</i> | <i>adh2</i> from <i>S. cerevisiae</i> | 5,242 |
| <i>Pdc</i> from <i>C. acetobutylicum</i> | <i>adh2</i> from <i>S. cerevisiae</i> | 260 |

Keto acid decarboxylases are common in plants, yeasts and fungi, but not in bacteria [3]. Atsumi and Liao screened five enzymes in the hope to find one able to decarboxylate ketoisovalerate to isobutyraldehyde. Their results are shown in **Table 13**.

After screening a number of potential candidates, it was evident that the best results were obtained with the *Lactococcus lactis* **Kdc**. In fact, it seems that *L. lactis* has several enzymes that can transform α -keto acids into aldehydes [Plaza *et al.* 2004, Smit *et al.* 2005]. Pyruvate decarboxylases (PDC), on the other hand, seem to be highly specific for pyruvate and show no activity against non-pyruvate α -keto acids [Plaza *et al.* 2007].

Atsumi and Liao [2008] therefore opted to use the broad range 2-ketoacid decarboxylase (Kdc) from *Lactococcus lactis*, to convert the direct precursor of valine, 2-ketoisovalerate, to isobutyraldehyde in combination with the *adh2* gene from *S. cerevisiae*, encoding an alcohol dehydrogenase, which converts isobutyraldehyde into isobutanol. Isobutanol production was further increased by overexpressing *ilvIHCD*, encoding enzymes involved in the production of 2-ketoisovalerate and by deleting competing pyruvate consuming pathways. The result was a strain that produced, under aerobic growth conditions, 0.86 mol of isobutanol per mol of glucose. The same strategy has subsequently been implemented and perfected in *Bacillus subtilis* and *Corynebacterium glutamicum*, as summarized in **Table 14**.

6.1.6. ISOBUTANOL PRODUCTION IN *B. SUBTILIS*

When Li and Wen, in 2011, implemented this strategy in *B. subtilis*, the production of 0.2 mol of isobutanol per mol of glucose was achieved after the integration of *kivd*, *adh2* and *ilvIHCD* in the genome, under the control of the strong constitutive p43 *B. subtilis* promoter. This experiment was carried out under oxygen-limited conditions in order to create a pseudo-fermentative process, but *B. subtilis* is only weakly fermentative, so the low yield could be explained by a redox imbalance. Also, the expression of the enzymes from the genome, instead of plasmid-bound, might have resulted in less protein being produced and so limited pathway flux. The low yield might also be explained by toxicity because *B. subtilis* is not tolerant to concentrations higher than 2% butanol [Fisher *et al.* 2008].

6.1.7. ISOBUTANOL PRODUCTION IN *C. GLUTAMICUM*

Corynebacterium glutamicum is the workhorse of industrial amino acid production [Smith *et al.* 2010]. Since the pathways for 2-ketoisovalerate overproduction have already

been engineered in *C. glutamicum* for the production of valine, a straightforward alternative application for this strain was the production of isobutanol.

Blombach *et al.* [2011] reported the production of 0.77 mol of isobutanol, per mol of glucose, under oxygen-limited conditions by *C. glutamicum*. Competing pathways that consume pyruvate were knocked-out and subsequently, *ilvBNCD*, *kivd* and *adhA* were expressed plasmid-bound together with *pntAB*. The purpose of including *pntAB*, which encodes a membrane-bound transhydrogenase, was to increase the production of NADPH to counteract redox imbalance.

| Organism | Relevant genomic characteristics of the strain | Plus plasmid-bound genes | Isobutanol production (mol isobutanol/mol glucose) |
|----------------------|--|---|---|
| <i>E. coli</i> | $\Delta adhE$, $\Delta ldhE$, $\Delta frdBC$, Δfnr , Δpta , $\Delta pflB$ | <i>kivd</i> from <i>L. lactis</i> , <i>adh2</i> from <i>S. cerevisiae</i> , <i>ilvIH</i> from <i>B. subtilis</i> and <i>ilvCD</i> from <i>E. coli</i> | 0.86 Aerobic conditions (20 g/L) [Atsumi and Liao, 2008] |
| <i>E. coli</i> | $\Delta ldhA-fnr::FRT$, $\Delta adhE::FRT$, $\Delta frd::FRT$, $\Delta pflB::FRT$, $F'(lacIq)$, $\Delta ilvC::PLLacO1::LI_kivd1::Ec_ilvD_coEc::FRT$, $\Delta pta::PLLacO1::Bs_alsS1$, $FRT::KAN::FRT$ | <i>ilvC_NADH_dependent_mutant</i> and <i>adhA_mutant</i> from <i>L. lactis</i> | 1.03 Anaerobic conditions (13.4 g/L) [Bastian et al. 2011] |
| <i>C. glutamicum</i> | $\Delta aceE$ Δpqo $\Delta ilvE$ $\Delta ldhA$ Δmdh | <i>ilvBNCD</i> , <i>pntAB</i> , <i>kivd</i> , <i>adhA</i> | 0.77 (Oxygen-limited conditions) [Blombach, 2011] |
| <i>B. subtilis</i> | $\Delta amyE::(P43_kivd-adh2-SPC')$, $P43::ilvD-ilvC-alsS-SPC'$ | | 0.2 (Oxygen-limited conditions) [Li et al. 2011] |

Bastian *et al.* in 2011 reported a yield of 1.03 mol of isobutanol per mol of glucose (*i.e.* the theoretical stoichiometric maximum) achieved under anaerobic conditions using a strain of *E. coli* that had competing pathways inactivated, as well as integration of *kivd* from *L. lactis* under the control of P_{LacO1} . The issue of NADPH/NADH redox imbalance when trying to produce isobutanol anaerobically was addressed by two procedures:

- (1) Over-expressing a transhydrogenase (PntAB) and
- (2) By mutating the cofactor binding site of IlvC in order to change its cofactor dependence from NADPH to NADH.

In the *E. coli* strain of Bastian *et al.* [2011], the natural chromosomal *ilvC* encoding the NADPH dependent enzyme was knocked out and replaced with a plasmid-bound copy of the mutant *ilvC* encoding an NADH dependent enzyme. Therefore, given that *E. coli* generates two mol of NADH per mol of glucose during glycolysis, the use of an NADH dependent IlvC and an NADH dependent alcohol dehydrogenase turned isobutanol production into a sustainable fermentation pathway.

6.1.8. BIOSYNTHESIS OF 2-KETOISOVALERATE IN *E. COLI* AND *C. GLUTAMICUM*

- Since 2-ketoisovalerate is the precursor for isobutanol, it is important to explore the possibility of upregulating the biosynthetic pathway for this compound.
- 2-Ketoisovalerate is the direct precursor of valine.
- So far only two organisms, *Corynebacterium glutamicum* and *E. coli*, have been metabolically engineered to overproduce valine [Park and Lee, 2010]. The biosynthesis of valine in *E. coli* is shown in **Figure 81**.
- It starts with a sugar, generally glucose, being converted to pyruvate via glycolysis.

The first step in the biosynthesis of valine is the condensation between two molecules of pyruvate, catalyzed by *ilvIH* (encoding an acetohydroxy acid synthase) to form 2-acetolactate. *ilvC* (acetohydroxy acid reductoisomerase) converts this to 2,3-dihydroxyisovalerate. Subsequently, *ilvD* (dihydroxy-acid dehydratase) will transform 2,3-dihydroxyisovalerate into 2-ketoisovalerate. And finally, *ilvE* (branched chain amino acid aminotransferase) converts 2-ketoisovalerate into valine. There are two other fates for 2-

ketoisovalerate: one is the formation of pantothenate by *panB* (3-methyl-2-oxobutanoate hydroxymethyl transferase) and the other is the formation of leucine by *leuA* (2-isopropylmalate synthase). Bacterial *IlvHC* proteins are composed of two subunits: *IlvH* is the large catalytic one, sized about 60 kDa and *IlvI* is the small regulatory one of 9-18 kDa [Porat *et al.* 2004].

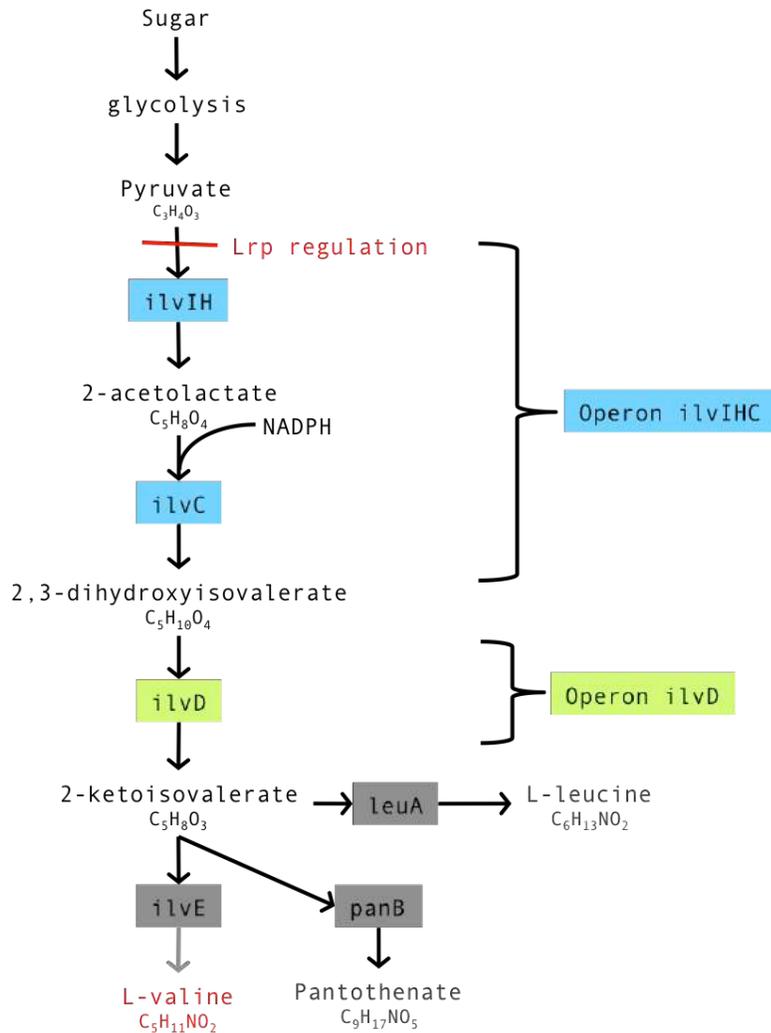


Figure 81. Map of the biosynthesis of the branched-chain amino acid valine in *E. coli*. The highlight in blue indicates that genes *ilvIH* and *ilvC* are together in the genome of *E. coli* in an operon. The green highlight indicates that *ilvD* is in a separate operon. Pyruvate is the starting point for valine biosynthesis. Lrp regulation activates *ilvIH*. 2-ketoisovalerate is the direct precursor of valine, leucine and pantothenate.

6.1.9. KETO ACID DECARBOXYLASES (KDC) FOR *G. THERMOGLUCOSIDANS*

There are two keto acid decarboxylases from *Lactococcus lactis* subsp. *lactis* that have decarboxylase activity on 2-ketoisovalerate and catalyze the formation of isobutyraldehyde.

These two enzymes are:

1. kdc - Accession number CAB34226
2. kdc- Accession number LLKF_1386
 - The present work is focused on kdc_CAB34226 exclusively.
 -

6.1.9.1. KETO ACID DECARBOXYLASE KDC_CAB34226

- The keto acid decarboxylase kdc_CAB34226 was sourced from a mesophile.
- Therefore, it is necessary to test whether it is (1) thermophilic and (2) thermostable before expressing it in *G. thermoglucosidans*.

The gene encoding the *Lactococcus lactis* Kdc used by Atsumi and Liao in 2008 was originally reported by Plaza *et al.* in 2004. It belongs to the *Lactococcus lactis* subsp. *lactis* strain IFPL730 *kivd* gene encoding an alpha-ketoisovalerate decarboxylase. The sequence of the gene can be found on NCBI, under the accession number AJ746364 (protein sequence CAB34226).

Keto acid decarboxylases are TPP-dependent enzymes. Thiamine pyrophosphate (TPP) is a cofactor they require for correct three-dimensional folding. Pyruvate decarboxylases (PDC), just like keto acid decarboxylases, are another member of the family of TPP-dependent enzymes that have attracted interest for their ethanol production capabilities. It has been reported in a work by Davidson (2009) that PDC is unable to fold properly at high temperatures, suggesting that the folding of the protein around the TPP cofactor is impaired at high temperatures resulting in a misfolded inactive enzyme. This might also be the case for Kdc, so testing is required to determine whether incorrect folding will affect the activity of Kdc when expressed at high temperatures by *G. thermoglucosidans*.

6.1.10. TESTING KETO ACID DECARBOXYLASES FOR ACTIVITY

As observed by Davidson, two tests need to be carried out when expressing mesophilic proteins in a thermophile in order to determine whether the protein will function correctly at high temperatures:

1. It is necessary to determine whether Kdc_CAB34226 is thermostable as well as thermophilic. A thermostable protein remains correctly folded and retains activity at high temperatures. However, in order to express a mesophilic protein, such as Kdc_CAB34226 in a thermophile.
2. It is necessary to determine whether the protein is also thermophilic. If the protein is thermophilic, it will be able to fold correctly at high temperatures and the three-dimensional structure will have activity.

Gocke *et al.* reported in 2007 that this enzyme showed decarboxylase activity on 2-ketoisovalerate at an optimal temperature of 50 °C. The authors determined the thermostability of Kdc by two assays:

1. A *direct decarboxylase assay*, which consisted of spectrophotometrically measuring the decay of 2-ketoisovalerate when mixing 2-ketoisovalerate with purified Kdc.
 2. A *coupled assay*, which consisted in pairing the purified kdc with a wide-range Adh (Alcohol dehydrogenase) and measuring spectrophotometrically the NADH output, since Adh is NAD⁺ dependent and an increase in NADH could be linked to an increase in isobutanol production.
- Based on the results obtained from both assays, the authors suggested that the Kdc CAB34226 might be thermostable, which was remarkable given that the source of the Kdc was the mesophile *Lactococcus lactis* and the probability of Kdc to have activity at high temperatures was low.
 - But isobutanol production at high temperatures wasn't tested.

Taking the work by Gocke *et al.* into account, it was decided that this work would focus on the expression of Kdc CAB34226 with the aim of testing its thermostability by:

1. Corroborating the results obtained from the *coupled assay* of Gocke *et al.* [2007].
2. Corroborating by HPLC that the protein is capable of producing isobutanol at 50 °C.

6.1.11. ISOBUTANOL PRODUCTION IN *G. THERMOGLUCOSIDANS*

At the time of the realization of this work, a paper by Lin *et al.* [2014] was published that reported three main findings:

1. The keto acid decarboxylase Kdc_CAB34226 from *L. lactis* is not thermostable.
2. The keto acid decarboxylase Kdc_LLKF_1386 from *L. lactis* is thermostable.
3. Kdc_LLKF_1386 was used to produce isobutanol in *G. thermoglucosidans*.

Lin *et al.* [2014] compared the thermostability of two keto acid decarboxylases from *L. lactis*. They were named CAB34226 and LLKF_1386 according to their amino acid sequence NCBI accession numbers. Kdc_CAB34226 has no activity at 50 °C, whereas Kdc_LLKF_1386 starts losing activity at 55 °C. The two proteins differ by only seven amino acids. Kdc_LLKF_1386 was used to produce isobutanol in this thermophile at 50 °C. The authors, however, report that no isobutanol could be produced at 55 °C due to "instability of the protein".

The yields achieved by Lin *et al.* are shown in **Table 15**. The authors overcame the issue of *ilvIH* feedback inhibition by overexpressing *alsS*, which is the *ilvIH* gene of *B. subtilis*. **AlsS** is not inhibited by end-products and it has a higher specificity towards pyruvate. Also, by replacing the native *ilvIH* promoter by the strong, constitutively expressed *Ldh* promoter, transcriptional attenuation of *ilvIH* was simultaneously eliminated.

Table 15. Isobutanol production titres achieved by Lin *et al.* [2014] in *G. thermoglucosidans* using the Kdc_LLKF_1386. *alsS* is the *ilvIH* gene from *B. subtilis* that has been overexpressed due to its higher affinity for pyruvate (see Figure XX). *ilvC* and *ilvD* further contribute to the overproduction of 2-ketoisovalerate and hence, isobutanol. Table modified from Lin *et al.* 2014

| | Genes expressed | Titer (g/L) |
|--|-----------------|-------------|
| | | N/D |
| | | 0.2 ± 0.0 |
| | | 3.3 ± 0.4 |
| | | 2.8 ± 0.5 |

Based on the contradictory findings regarding the thermostability of Kdc_CAB34226, it was decided as an aim of this work to compare the results obtained from the work by Lin *et al.* [2014] to the results obtained by Gocke *et al.* [2007] to determine whether Kdc_CAB34226 can be used for the production of isobutanol in *G. thermoglucosidans*.

6.1.12. ALTERNATIVE PATHWAY FOR ISOBUTANOL PRODUCTION

Two other options have been reported in the literature about enzymes that have the capacity to produce isobutyraldehyde:

1. AlsS, the acetolactate synthase from *B. subtilis* has 2-ketoisovalerate decarboxylase activity as well, following a report by Atsumi *et al.* in 2009. However, the yield of isobutanol obtained was very low (0.17 mol of isobutanol/mol glucose) compared to all the strains which use a Kdc.

2. The second alternative is an acylating aldehyde dehydrogenase (ALDH). ALDH is encoded by gene *ald* in *Clostridium beijerinckii*. As a report by Toth *et al.* suggests, this enzyme is capable of catalyzing the reduction of acetyl-CoA and butyryl-CoA to acetaldehyde and butyraldehyde in this host. These aldehydes are then reduced to ethanol and butanol by an alcohol dehydrogenase (ADH). It is NADH-dependent and in *C. beijerinckii*, it has a preference for butyryl-CoA as a substrate. There is, however, no information in the literature about its capacity to produce isobutyraldehyde, apart from a United States Patent by Donaldson *et al.* from 2010, which suggests that isobutyraldehyde could be produced from isobutyryl-CoA by this enzyme, using NADH as an electron donor. They are known by the EC numbers 1.2.1.10 and 1.2.1.57 and they are also available from *Clostridium acetobutylicum*, *Pseudomonas putida* and *Thermus thermophilus*. The proposed pathway is represented in **Figure 82**.

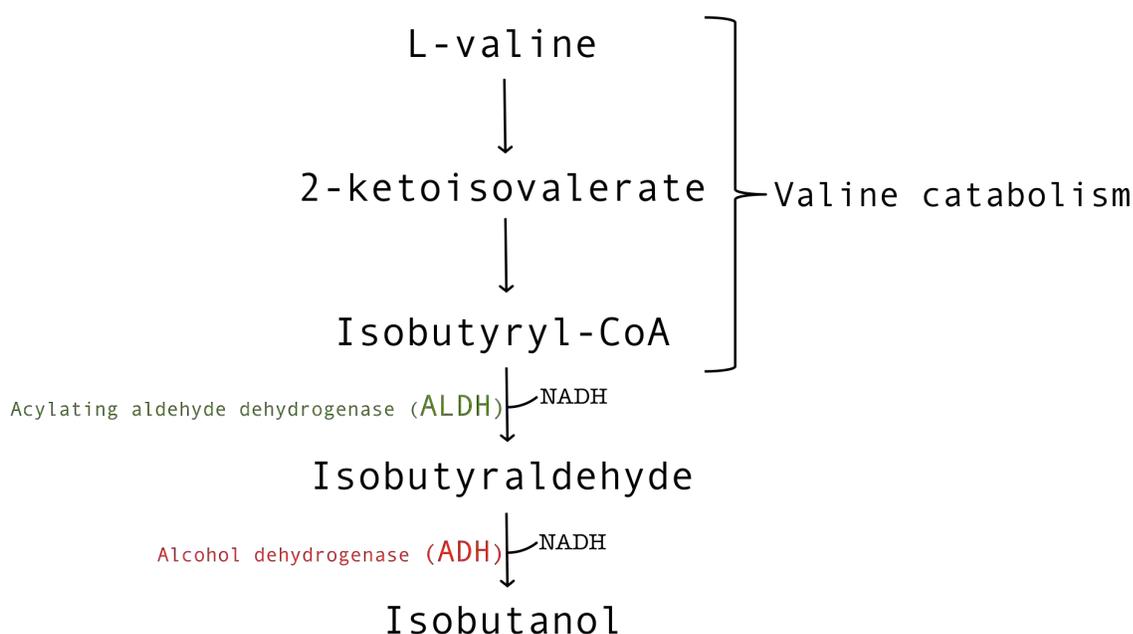


Figure 82. Pathway for isobutanol production as proposed by Donaldson *et al.* [2010], mediated by an acylating aldehyde dehydrogenase (ALDH) and an alcohol dehydrogenase (Adh). The difference between the approach proposed using ALDH and the approach that uses Kdc is that ALDH converts a valine catabolism intermediate, isobutyryl-CoA, to isobutyraldehyde, whereas Kdc converts a valine anabolism intermediate, 2-ketoisovalerate, to isobutyraldehyde. The final step in both approaches is the same: an alcohol dehydrogenase (Adh) will convert isobutyraldehyde into isobutanol.

The fact that it is present in the thermophile *T. thermophilus*, suggests that ALDH is thermostable and that it might also be present among other thermophiles, namely, *G. thermoglucosidans*.

6.1.13. SOURCES OF ALCOHOL DEHYDROGENASES (ADH)

Alcohol dehydrogenases are oxidoreductases. They catalyze, reversibly, the reduction of aldehydes to alcohols/ketones. Generally, they are versatile enzymes, and different families with different substrate specificities exist [Savrasova *et al.* 2011].

Atsumi and Liao, in 2008, demonstrated that the alcohol dehydrogenase from *Saccharomyces cerevisiae* (***adh2***) had the capacity to convert isobutyraldehyde into isobutanol. The authors knew that ***adh2*** was a good choice because *S. cerevisiae* naturally produces isobutanol in low yields [Dickinson *et al.* 1998]. Subsequently, it was reported that ***adh2*** is not the only enzyme that can do this, one of the native alcohol dehydrogenases of *E. coli*, also has the capacity to convert isobutyraldehyde into isobutanol very efficiently [Atsumi *et al.* 2010]. This implies that, in the 2008 paper, it is impossible to tell how much of the isobutanol produced was due to the heterologously expressed ***adh2*** from *S. cerevisiae* and how much of it was the work of the native ***yqhD***. In terms of cofactor dependence, ***yqhD*** requires NADPH while ***adh2*** requires NADH [Savrasova *et al.* 2011], but their experiment was carried out aerobically, so redox imbalance did not present a problem.

The aim of the work in the present chapter was to use the toolbox created using a synthetic biology approach to demonstrate, as proof-of-principle, its capacity to express the heterologous proteins Kdc from *Lactococcus lactis* and ALDH from *Thermus thermophilus*, as well as to overexpress an intrinsic ALDH from *G. thermoglucosidans*. In addition, a secondary aim of this chapter was to determine whether the addition of 2-ketoisovalerate to the growth medium would result in a higher production of isobutanol, which, if it indeed is achieved, would suggest that it is possible to bypass the need to overexpress precursor pathways before testing the thermostability of an enzyme.

6.2. RESULTS

6.2.1. TOLERANCE OF *G. THERMOGLUCOSIDANS* TO ISOBUTANOL

An experiment was performed to determine how tolerant *G. thermoglucosidans* is to isobutanol. Several concentrations of isobutanol (v/v) were tested. 10 μ L of an overnight culture that had been growing in 2TY at 55 $^{\circ}$ C (250 rpm) were inoculated to 5 mL of 2TY and supplemented with 0.5%, 1%, 1.5% and 2% (v/v) isobutanol and were left to grow overnight (aerobically and anaerobically) in a rotary shaker at 55 $^{\circ}$ C and 250 rpm.

It was found that growth occurs at concentrations of up to 1.5% (1.2 g/L) in 2TY. The growth is shown in **Figure 83**.

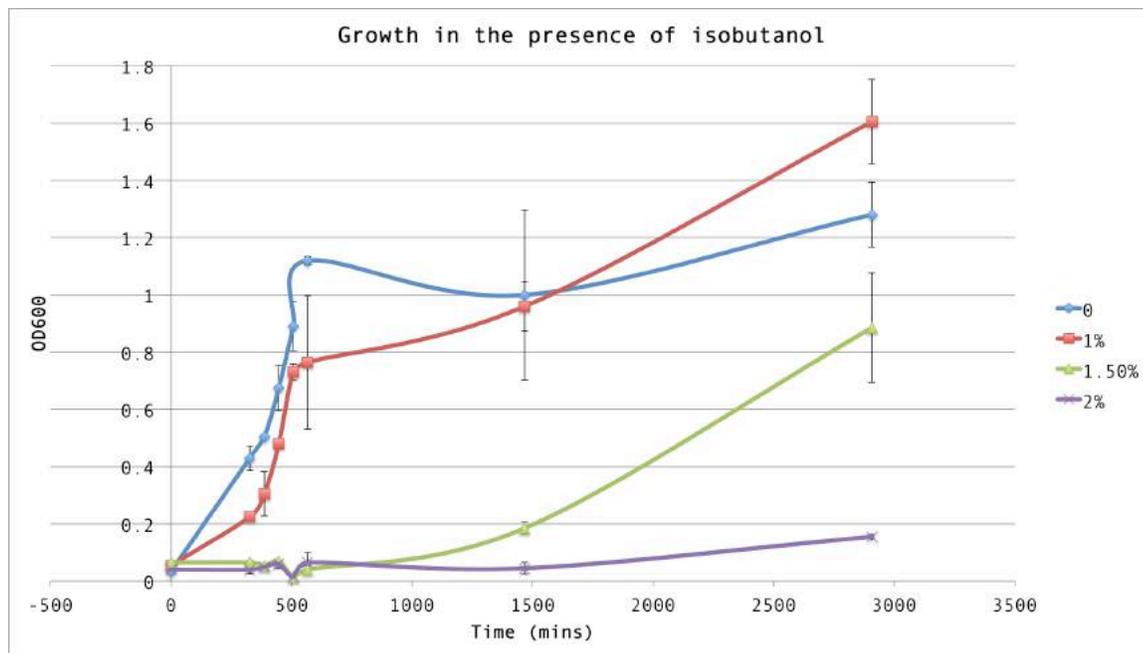


Figure 83. Growth of DL44 in the presence of 0, 1, 1.5 and 2% isobutanol monitored over the course of 49 hours. The error bars represent the standard deviation of two replicates (n=2). DL44 cells were grown in 50 mL of 2TY in a 250 mL baffled flask, and grown shaking at 55 $^{\circ}$ C.

It's possible to see in **Figure 83** that the culture supplemented with 1% isobutanol exhibited a higher OD₆₀₀ after 3 days, which could be attributed to isobutanol being used as a carbon source by *G. thermoglucosidans*. Therefore, another experiment was set up to verify if *G. thermoglucosidans* grows on isobutanol as a sole carbon source. The cells were grown

on the minimal ASM medium, but all carbon sources were removed, except for 1% isobutanol. No growth was observed after 48 hours (data not shown), indicating that even in the rare case that isobutanol was oxidized, it was converted to a by-product that the cells cannot use.

6.2.2. EXPRESSION OF *LACTOCOCCUS LACTIS* KDC

Considering that the *L. lactis Kdc* CAB34226 was the Kdc that produced the best result for Atsumi and Liao in 2008, it was decided in 2009 that the *L. lactis Kdc* CAB34226 should be tested in *G. thermoglucosidans* for thermostability. In order to do so, the sequence of the gene encoding this protein was codon-optimized by GeneART (Life Technologies) using the *G. stearothermophilus* codon usage tables available in the online codon usage database¹⁷. Both the non codon-optimized and the codon-optimized sequences have been included in **Appendix Figures A132 and A133**, respectively.

The synthetic codon optimized *kdc* was included in a GeneART commercial vector, placed downstream from the T7 promoter. A close inspection using the reverse engineering function of the Salis RBS Calculator revealed that, in order for the *kdc* to be transcribed under the control of the T7 promoter in *E. coli*, an RBS needed to be placed upstream from *kdc*.

An RBS was designed for this particular CDS using the forward engineering function of the Salis RBS Calculator. The synthetic RBS had a predicted strength of 22373 AU for *E. coli* and 12464 AU for *G. thermoglucosidans*. The RBS was included in the construct encoded in primer rbs1-kdc. Together with the 5'-phosphorylated reverse primer RC-T7-prom, this pair was used to PCR-amplify the entire vector, which was then re-circularized by incubation with T4 ligase.

E. coli BL21 cells were transformed with this ligation, but no transformant colonies were obtained. While the T7 promoter is inducible, it is however, well-known that this promoter has basal activity even in the absence of the inducer IPTG. Therefore, to inquire into the possibility that the absence of transformants was caused by toxicity of Kdc

¹⁷ [<http://www.kazusa.or.jp/codon/>].

expression, BL21 (DE3) pLysE (Bioline) cells were transformed instead. Encoded in pLysE lies the T7 lysozyme, which is constitutively expressed, and which inhibits basal T7 RNA Polymerase.

When BL21 (DE3) pLysE cells were transformed, and confirmed by sequencing. It was necessary to promote the loss of pLysE in order to stop the T7 lysozyme from inhibiting the expression of *kdc*, which was under the control of the T7 promoter. To make the cells lose pLysE, which has a chloramphenicol resistance marker, and keep the *kdc* plasmid, which had a kanamycin resistance gene, individual colonies were serially restreaked on LB/kanamycin plates as well as on LB/kanamycin+chloramphenicol plates for five days, which was when colonies were found that grew exclusively on kanamycin but not on chloramphenicol. To confirm the loss of the pLysE plasmid, a test-digest with PstI and Accl was carried out, before the serial restreaking and after it. The results of these digests are shown in **Figures 84** and **85**, respectively.

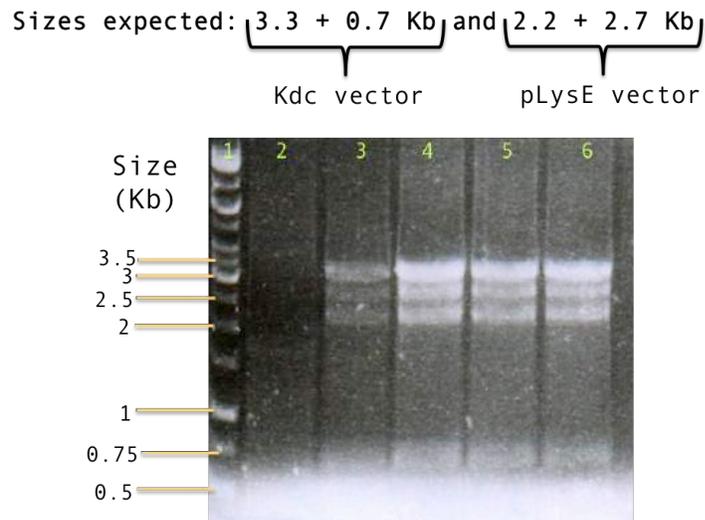


Figure 84. Accl/PstI test-digest to determine the presence of the pLysE and the *kdc* plasmids analyzed on a 1% agarose gel. Lane 1, 1 kb GeneRuler molecular size marker. Lanes 3-6, digested plasmid from four BL21 (DE3) pLysE colonies grown on LB/kanamycin on the first day of serial restreaking.



Figure 85. *AccI/PstI* test-digest to determine the loss of the pLysE plasmid analyzed on a 1% agarose gel. Lane 1, 1 kb GeneRuler molecular size marker. Lanes 2-4, digested plasmid from four BL21 (DE3) pLysE colonies grown on LB/kanamycin on the fifth day of serial restreaking.

Figure 85 confirms that after five days, the BL21 (DE3) pLysE cells lost the pLysE plasmid and retained the *kdc* plasmid, thus becoming solely BL21 cells rather than BL21 (DE3). To confirm that no mutations appeared on the *kdc* gene after successive days of subculturing, the plasmid was extracted and sequenced. No mutations on the *kdc* gene, or on the RBS or on the T7 promoter were seen. The sequencing results have been added to the **Appendix Figure A135**.

BL21 colonies containing the T7+RBS+*kdc* construct were grown overnight and 500 μ L of the culture were used to inoculate 50 mL of fresh LB/kanamycin medium in a baffled 250 mL conical flask to allow for optimal aeration of the cells. Once the cells reached an OD_{600} of 0.5, they were induced with 1mM IPTG for 3 hours, after which the cells were harvested, the pellet was resuspended in 5 mL KOH-MES buffer and the cells were lysed using a sonicator. 30 μ L of the lysate was analyzed on an SDS-PAGE gel, shown in **Figure 86**. The band expected for *kdc* was of 61.8 kDa according to the Protein Calculator v3.4. From **Figure 86** it is difficult to determine whether the protein is indeed being expressed. An overexpression band is seen in the induced cells, but runs with a molecular weight comparable to 55 kDa.

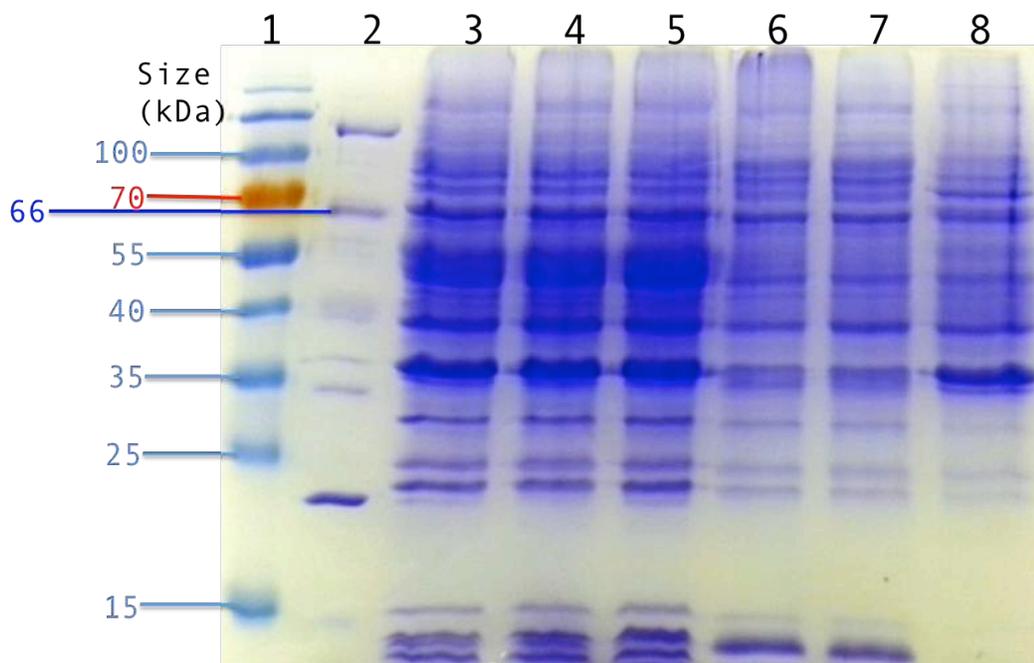


Figure 86. SDS-PAGE gel electrophoresis of intracellular protein from cultures of BL21 containing the T7+RBS+kdc construct. Lane 1, prestained protein molecular marker. Lane 2, unstained protein molecular marker. Lanes 3-5, protein extracted from three separate colonies. Lanes 6-8, protein extracted from three separate untransformed BL21. All cultures were induced by supplementing the growth media with 1mM IPTG.

To test for **thermostability of Kdc_CAB34226** it was attempted to determine Kdc activity by carrying out both the direct and coupled decarboxylation assays at a variety of temperatures with cell lysate from *E. coli* BL21. Both assays consist of measuring, spectrophotometrically at 340 nm, the decay of either 2-ketoisovalerate for the direct assay, or the decay of NADH for the coupled one.

In order to do so, it was necessary to incubate the cell lysate from BL21 cells transformed with the T7+RBS+kdc construct at 55 °C, because that is the temperature most commonly used to grow *G. thermoglucosidans*. It was found that incubating BL21 cell lysate at this temperature results in the lysate becoming white and cloudy, due to the misfolding effect the temperature has on mesophilic proteins. This made any spectrophotometric reading for the direct and coupled assays impossible. Following the work carried out by Gocke *et al.* in 2007, it was decided that it would be necessary to purify the protein in order to separate from the rest of the mesophilic proteins of BL21. A 6X histidine tag was added to the C-terminus of the protein encoded in the 5'-phosphorylated primer RC-His-tag_kdc. Together with the forward primer His-tag_kdc, this pair was used to PCR-amplify the entire

T7+RBS+kdc plasmid. The PCR-product was self-ligated using T4 ligase and BL21 (DE3) pLysE cells were transformed with the ligation. These cells were, as previously, consecutively restreaked for five days until the loss of the pLysE plasmid was confirmed. The resulting plasmid was named "T7+RBS+kdc+his-tag".

BL21 colonies containing the T7+RBS+kdc+his-tag construct were grown overnight and 500 μ L of the culture were used to inoculate 50 mL of fresh LB/kanamycin medium in a baffled 250 mL conical flask to allow for optimal aeration of the cells. Once the cells reached an OD_{600} of 0.5, gene expression was induced with 1mM IPTG for 3 hours, after which the cells were harvested, the pellet was resuspended in 5 mL KOH-MES buffer and the cells were lysed using a sonicator.

The 6XHis-tagged proteins were purified under native conditions using the immobilized nickel chelate chromatography using the Ni-NTA Spin kit by Qiagen. The purification consisted in two wash steps and four elution steps. The protein yield of each step was analyzed on a Western Blot, shown in **Figure 87**.

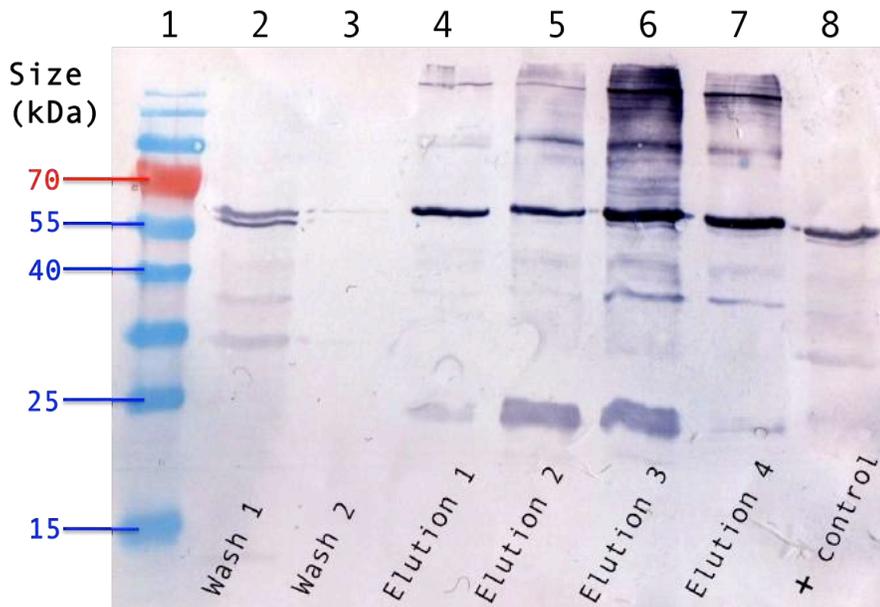


Figure 87. Western blot analysis of the yields of each of the two wash steps and four elution steps carried out to purify the 6XHis-tagged proteins. Lane 1, prestained protein molecular marker. Lanes 2-3, protein yield of the first wash and second washes, respectively. Lanes 4-7, protein yields of each of the four elution steps. Lane 8, contained a 6X-His-tag positive control.

As it can be seen in the Western blot in **Figure 87**, a strong band of the correct size (61.7 KDa) was observed in each of the elution steps, confirming the expression and purification of the 6X-His-tagged Kdc by BL21 cells. The smaller bands seen at 25 KDa correspond to *E. coli* heat shock proteins.

The direct and coupled assays described above were done with 50 μL of the purified Kdc obtained from elution with 250 mM imidazole. The results of the coupled assay are shown in **Figure 88**. Two temperatures were tested: 37 and 55 $^{\circ}\text{C}$.

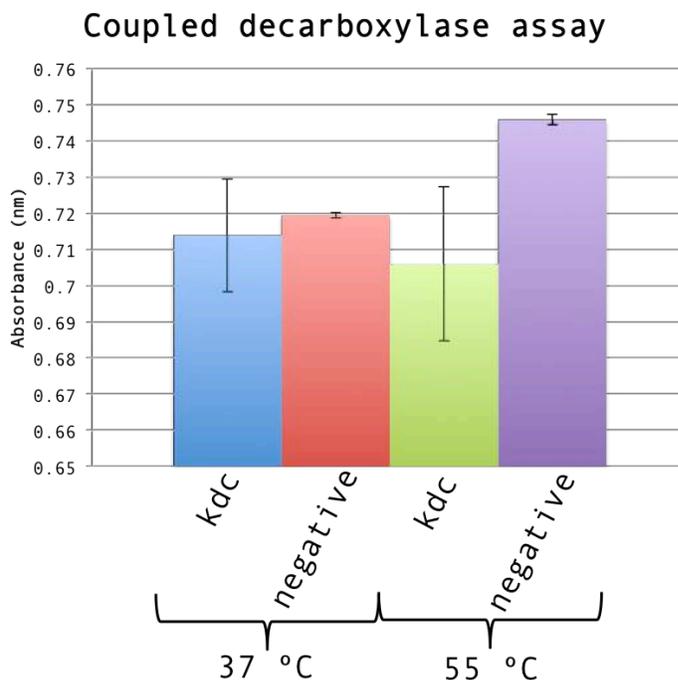


Figure 88. Absorbance readings of the coupled decarboxylase assay. The excitation wavelength was 340 nm. The reaction was carried out in a final volume of 1 mL. 50 μL of purified Kdc was incubated potassium phosphate buffer pH 6.8, 2.5 MgSO_4 and 0.1 mM TPP, 60 mM 2-ketoisovalerate, 0.25 mM NADH. Negative controls ("No Kdc") contained potassium phosphate buffer pH 6.8 with the same amounts of MgSO_4 , TPP, 2-ketoisovalerate, NADH, but no purified Kdc was added. The incubation was carried out at 37 $^{\circ}\text{C}$ or 55 $^{\circ}\text{C}$ for an hour, after which and 0.25 U/mL of broad range HL-ADH were added and the reaction was allowed to proceed at room temperature for 1 hour. Error bars indicate the standard deviation of two replicates N=2.

As it can be observed in **Figure 88**, the absorbance readings obtained with Kdc are much more variable than the results obtained with the negative controls. The fact that the variability was observed at both 55 $^{\circ}\text{C}$ and 37 $^{\circ}\text{C}$ suggests that Kdc has activity at both these temperatures, but the high variability observed in the absorbance readings makes it difficult to quantify the NADH oxidizing activity with certainty. It is evident, nonetheless, that the

absorbance observed with Kdc, at either 37 °C or 55 °C, is on average lower than that observed for the reference reactions without Kdc, which indicates that some oxidation of NADH is taking place, and this is caused by Kdc activity, and not by NADH degradation under the experimental conditions used. The degree of NADH oxidation compared to the control is greater at 55 °C than at 37 °C.

To test the Kdc activity in *G. thermoglucosidans*, the gene encoding Kdc was amplified using primers XbaI-RBS-kdc and RC-kdc-SacI. Plasmid T7+RBS+kdc+6XHis-tag was used as template for the PCR reaction. Encoded in the forward primer, downstream from the XbaI site was the RBS (previously denoted as "pheB-RBS") used to express reporter proteins in plasmids p11AK1, p11AK2, p11AK3, p11AK4 and p11AK5, detailed in Chapter 1. The expression cassette, incorporated into plasmid p11AK1 following digestion and ligation with XbaI and SacI, will hereafter be referred to as p11AK+kdc, and it is shown in **Figure 89**.

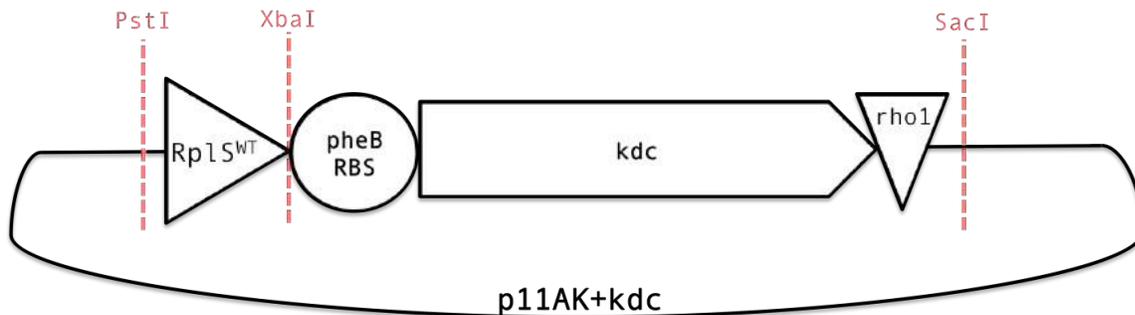


Figure 89. Design of the cassette used to express Kdc by *G. thermoglucosidans* DL44. The RplS^{WT} promoter together with the pheB-RBS direct transcription of *kdc* in the thermophile. "rho1" is the rho-independent transcriptional terminator placed downstream from *kdc*.

To confirm the expression of the Kdc under the control of the wild-type RplS promoter and the pheB-RBS in plasmid p11AK-kdc, cell lysate from transformant *G. thermoglucosidans* DL44 was analyzed in a Western blot, which is shown in **Figure 90**. The Western blot shows that Kdc is being expressed because a band of the right size was observed at about 62 kDa from the two colonies screened, as shown in lanes 4 and 5. It is much more evident from the colony in lane 4, the colony in lane 5 shows weaker expression, but the band is the right size nonetheless. *G. thermoglucosidans* DL44 cells transformed with p11AK+kdc were routinely grown at 55 °C. Cells containing this construct did not grow at 50 °C. This is probably due to toxicity of Kdc. It becomes denatured at about 50 °C according to

Lin et al. [2014] (Kdc_CAB34226). Therefore, while it's still in a more active form, it's too toxic for the cells because it sequesters 2-ketoisovalerate. In a more inactive form, achieved by thermal denaturation, it does not sequester as much ketoisovalerate, thus allowing the cells to survive.

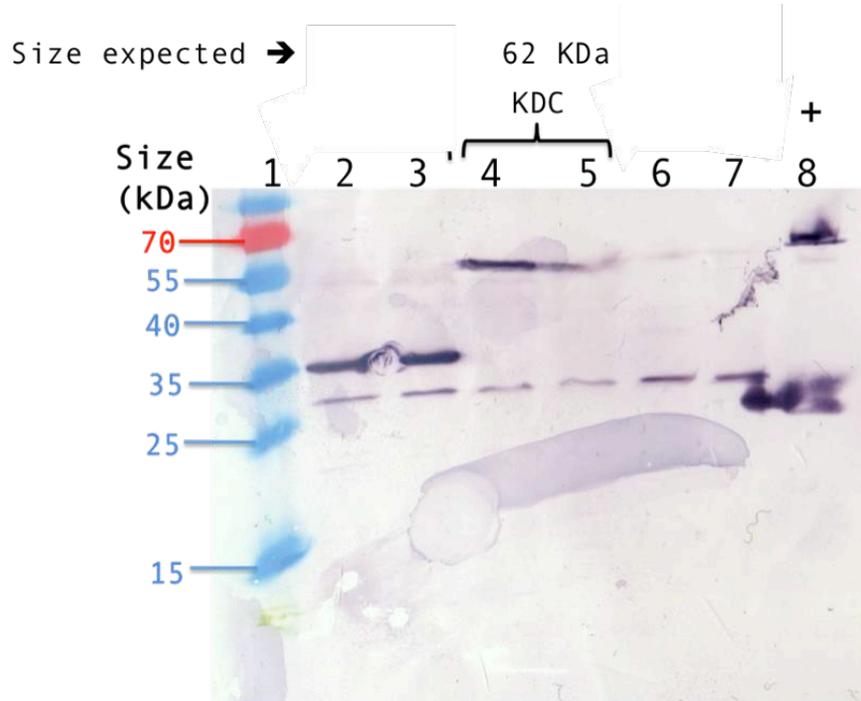


Figure 90. Cell lysate of *G. thermoglucosidasius* DL44 analyzed on a Western blot. Lane 1, prestained protein molecular marker (Fermentas). Lanes 2-3 and 6-7 correspond to a sample described in Figure XX and are not relevant in this section. Lanes 4-5, cell lysate from two colonies transformed with p11AK+kdc. Lane 8, positive 6XHis-tag control corresponding to a sample from a previous unrelated experiment. The sizes expected were: 34 KDa for Tt_ADHL.

As experiments were being carried out to determine whether the *L. lactis* Kdc CAB34226 used in the 2008 Atsumi paper was thermostable, a paper was published in May 2014 that contradicted the work by Gocke *et al.* [2007] and which confirmed that Kdc_CAB34226 is not thermostable. The publication of the paper by Lin *et al.* entitled "Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*" confirmed that it was possible to produce isobutanol at 50 °C by this thermophilic host. They used the keto acid decarboxylase approach to convert 2-ketoisovalerate to isobutyraldehyde, which was then reduced to isobutanol by an alcohol dehydrogenase. It was decided that an alternative pathway to produce isobutanol should be sought because the production isobutanol had to be used as evidence that the toolbox is functional for the expression of heterologous genes and that *G. thermoglucosidans* should be developed as a chassis for the production of useful compounds. It was known that several

Clostridium spp. are able to produce butanol using an ALDH (acetylating aldehyde dehydrogenase), and it is known that such protein naturally exists in *Thermus thermophilus*, which strongly implies that it is both thermophilic and thermostable, unlike Kdc.

6.2.3. FINDING A THERMOSTABLE ACYLATING ALDEHYDE DEHYDROGENASE (ALDH)

In order to determine whether it is a feasible possibility to implement a pathway for isobutanol production that mirrored the native *Clostridium beijerinckii* butanol production pathway, the metabolic pathways of *G. thermoglucosidans* were inspected in the search for the production of isobutyryl-CoA. It was discovered that, according to KEGG, an online database of genome information as well as metabolic pathway data, *G. thermoglucosidans* possesses a branched-chain keto acid dehydrogenase, which catalyzes the conversion of 2-ketoisovalerate to isobutyryl-CoA, using NADH as an electron acceptor (see **Figure 91**). In *Geobacillus* spp., this enzyme is also known as 3-methyl-2-oxobutanoate dehydrogenase. It is comprised of 331 amino acids, and the sequences can be found on NCBI (accession number YP_003988619).

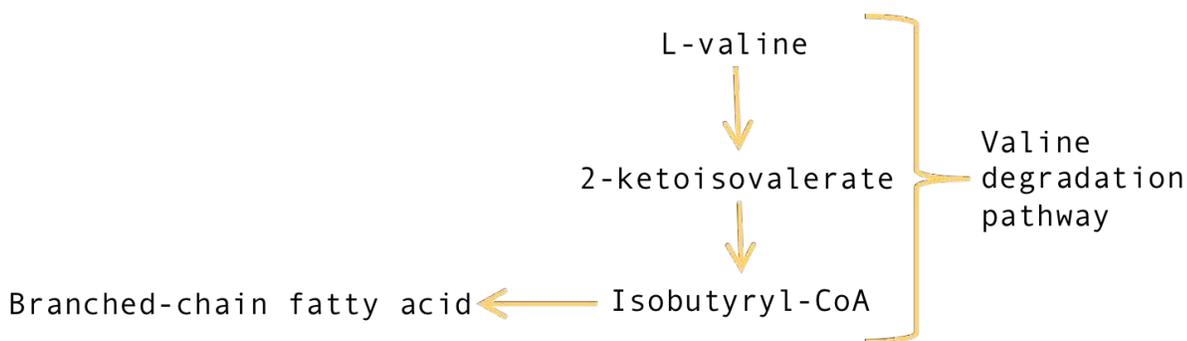


Figure 91. Degradation pathway of valine in *G. thermoglucosidans*. L-valine is degraded to 2-ketoisovalerate which is then converted to isobutyryl-CoA. Isobutyryl-CoA normally is involved in the biosynthesis of branched chain fatty-acids.

A Concise BLAST search against the *adl* gene of *Thermus thermophilus*, which can be found on NCBI, under the accession number (YP_145486) revealed, as it can be seen in **Figure 92**, that a close-match to the ALDH protein is also present in *Geobacillus* spp.

| Organism | Protein Name | Accession L |
|---|----------------------------|--------------|
| Thermomicrobium roseum DSM 5159 | acetaldehyde dehydrogenase | YP_002522598 |
| Roseiflexus castenholzii DSM 13941 | acetaldehyde dehydrogenase | YP_001432509 |
| Chloroflexus aggregans DSM 9485 | acetaldehyde dehydrogenase | YP_002463357 |
| Roseiflexus sp. RS-1 | acetaldehyde dehydrogenase | YP_001276898 |
| Chloroflexus aurantiacus J-10-fl | acetaldehyde dehydrogenase | YP_001634967 |
| Geobacillus sp. Y4.1MC1 | acetaldehyde dehydrogenase | YP_003989256 |
| Pelotomaculum thermopropionicum SI | acetaldehyde dehydrogenase | YP_001211032 |
| Sphingomonas wittichii RW1 | acetaldehyde dehydrogenase | YP_001262141 |
| Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446 | acetaldehyde dehydrogenase | YP_003185021 |
| Bacillus tusciae DSM 2912 | acetaldehyde dehydrogenase | YP_003588851 |
| Brevibacillus brevis NBRC 100599 | acetaldehyde dehydrogenase | YP_002772458 |
| Geobacillus thermodenitrificans NG80-2 | acetaldehyde dehydrogenase | YP_001127242 |
| Geobacillus sp. Y412MC61 | acetaldehyde dehydrogenase | YP_003253384 |
| Thermomonospora curvata DSM 43183 | acetaldehyde dehydrogenase | YP_003301047 |
| Bacillus pseudofirmus OF4 | acetaldehyde dehydrogenase | YP_003427464 |
| Bacillus megaterium QM B1551 | acetaldehyde dehydrogenase | YP_003564035 |
| Bacillus megaterium DSM 319 | acetaldehyde dehydrogenase | YP_003598756 |
| Mycobacterium tuberculosis H37Rv | acetaldehyde dehydrogenase | NP_218052 |
| Mycobacterium sp. JLS | acetaldehyde dehydrogenase | YP_001072486 |
| Geobacillus sp. Y4.1MC1 | acetaldehyde dehydrogenase | YP_003988983 |
| Mycobacterium abscessus ATCC 19977 | acetaldehyde dehydrogenase | YP_001701377 |

Figure 92. Concise BLAST results of the search against the NCBI database of the *T. thermophilus* ALDH. Several hit candidates were observed; their accession numbers are included in the third column. *Geobacillus* sp. Y4.1MC1, *Geobacillus thermodenitrificans* and *Geobacillus* sp. Y412MC61 were among the candidates. This result prompted the search for the ALDH gene in the wild-type DL33 strain of *G. thermoglucosidans* available in our laboratory.

Two conserved domains were identified for these proteins, which suggested they'd be able to react with aldehydes. As it can be seen in **Figure 93**, these domains are: (1) Semialdehyde dehydrogenase, NAD binding domain and (2) a prokaryotic acetaldehyde dehydrogenase, dimerization; members of this family are found in prokaryotic acetaldehyde dehydrogenase (acylating). They mediate dimerization of the protein [Marchler-Bauer, 2011].

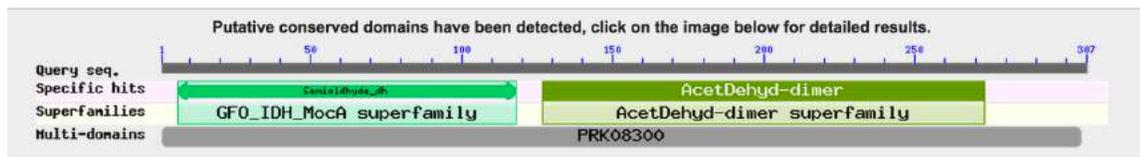


Figure 93. Conserved domains identified by NCBI Blast for the acylating aldehyde dehydrogenase of *T. thermophilus*. The presence of conserved domains involved in the dehydrogenation of acetaldehyde was a positive result, which indicated the potential for reaction with aldehydes.

An alignment between the two ALDH amino acid sequences, shown in **Figure 94**, revealed that there is 66% similarity among them.

Range 1: 3 to 298 [Graphics](#) ▼ Next Match ▲ Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|----------------|--|------------------------------|--------------|--------------|-----------|
| 403 bits(1035) | 1e-144 | Compositional matrix adjust. | 195/296(66%) | 238/296(80%) | 1/296(0%) |
| Query 30 | SKVKVIGILGSGNIGTDLMYKILNNDGHMELSI IAGIVPESEGLMRARSMGIRASHEGIAA | | | | 89 |
| | +VKV ILGSGNIGTDLMYK+L N GHMEL + GI P+SEGL RAR++G+ ASHEGIA | | | | |
| Sbjct 3 | ERVKVAIILGSGNIGTDLMYKLLKNP GHMELVAVVGIDPKSEGLARARALGLEASHEGIAY | | | | 62 |
| Query 90 | ILEDPPVRIVFDATSAKAHRQNAEALRAAGKIVIDLTPAAIGPFVPPINLNEHINAWNV | | | | 149 |
| | IIE P+++IVFDATSAKAH ++A+ LR AGKI IDLTPAA GP+VVPP+NL EH++ NV | | | | |
| Sbjct 63 | ILERPEIKIVFDATSAKAHVRHAKLLREAGKIAIDLTPAARGPYVPPVNLKEHLDKDNV | | | | 122 |
| Query 150 | NMISCGGQATIPLVYAVSRIVSVQYAEVITTTASRSIGPGTRQNMDEFTYTTAQGLQSIG | | | | 209 |
| | N+I+CGGQATIPLVYAV R+ V YAE+++T ASRS GPGTRQ+DEFT+TTA+GL++IG | | | | |
| Sbjct 123 | NLITCGGQATIPLVYAVHRVAPVLYAEMVSTVASRSAGPQTRQNI DEFTFTTARGLEAIG | | | | 182 |
| Query 210 | GAKIARAIPVINSADPTIIMTNTVYAAVE-EDFDEKAIILSIKDMVDEVSKYVPGYRLKA | | | | 268 |
| | GAK +AI ++N A+P I+MTNTV E E FD +A++ S++ M EV YVPGYRLKA | | | | |
| Sbjct 183 | GAKKGKAI IILNPAEPPILMTNTVRCIPEDEGFDREAVVASVRAMEREVQAYVPGYRLKA | | | | 242 |
| Query 269 | TPFIDYRDTFPWGRLPVVVILNEVEGSEDFLPAYAGNLDIMTASARRVGESYAQHLL | | | | 324 |
| | P + TPWG VV +L EVEG+ D+LP YAGNLDIMTASARRVGE +AQHLL | | | | |
| Sbjct 243 | DPVFERLPTPWGERTVVSMLLEVEGAGDYLPKYAGNLDIMTASARRVGEVFAQHLL | | | | 298 |

Figure 94. Alignment between the ALDH amino acid sequences of *T. thermophilus* ALDH (Subject) and the *G. thermoglucosidans* ALDH (Query). There is 66% similarity between both sequences, which indicates that both might have similar dehydrogenase activity.

The *T. thermophilus* *ald* gene sequence was sent to GeneART (Invitrogen) to be synthesised with codon-optimization for *G. thermoglucosidans*. The codon-optimized version was compared to the original version, and the alignment can be seen in **Figure 95**.

Range 1: 6 to 945 [Graphics](#) ▼ Next Match ▲

| Score | Expect | Identities | Gaps | Strand |
|---------------|---|--------------|-----------|-----------|
| 540 bits(598) | 2e-157 | 686/941(73%) | 2/941(0%) | Plus/Plus |
| Query 6 | CGAACGCGTCAAAGTCGCCATTTTGGGCAGCGGCAACATTGGCACGGATTGATGTATAA | 65 | | |
| Sbjct 6 | CGAAAGGGTTAAGGTAGCCATCCTGGGCTCCGGCAACATCGGGACGGACCTGATGTACAA | 65 | | |
| Query 66 | ATTGTTGAAAAATCCGGGACATATGGAATTGGTCGCGTGGTCGGCATTGATCCGAAAAG | 125 | | |
| Sbjct 66 | GCTCCTGAAGAACCCTGGGCCACATGGAGCTTGTGGCGGTGGTGGGGATAGACCCCAAGTC | 125 | | |
| Query 126 | CGAAGGCTTGGCTCGCGCTCGCGCTTTGGGCTTGAAGCCAGCCATGAAGGCATTGCCTA | 185 | | |
| Sbjct 126 | CGAGGGCTTGGCCCGGGCGGGCCCTTAGGGTTAGAGGCGAGCCACGAAGGGATCGCCTA | 185 | | |
| Query 186 | TATTTTGAACGCCCAGAAATCAAATTTGCTTTTGTATGCCACGAGCGCAAAGCCCATGT | 245 | | |
| Sbjct 186 | CATCCTGGAGAGGCCGGAGATCAAGATCGTCTTTGACGCCACCAGCGCCAAGGCCACGT | 245 | | |
| Query 246 | CCGCCATGCCAAATTTGTTGCGCGAAGCCGGAAAAATTGCCATTGATTTGACACCGGCTGC | 305 | | |
| Sbjct 246 | GCGCCACGCCAAGCTCCTGAGGGAGCGGGGAAGATGCCATAGACCTCACGCCGGCGGC | 305 | | |
| Query 306 | TCGGGGACCGTATGTCGTTCCACCGGTCAACTTGAAAGAACATTTGGATAAAGACAACGT | 365 | | |
| Sbjct 306 | CCGGGGCCCTTACGTGGTGCCCGGTGAACCTGAAGGAACACTGGACAAGGACAACGT | 365 | | |
| Query 366 | CAACTTGATTACGTGCGGAGGCCAAGCCACGATTCCGTTGGTCTATGCCGTCCATCGGGT | 425 | | |
| Sbjct 366 | GAACCTCATCACCTGCGGGGGCAGGCCACCATCCCCCTGGTCTACGCGGTGCACCGGGT | 425 | | |
| Query 426 | CGTCCGGTCTTGTATGCCGAAATGGTCAGCACGGTCGCTTCGCGCAGCGCTGGACCGGG | 485 | | |
| Sbjct 426 | GGCCCCGTGCTCTACGCGGAGATGGTCTCCACGGTGGCCTCCCGCTCCGCGGGCCCGG | 485 | | |
| Query 486 | AACACGCCAAAACATTGATGAATTTACGTTTACGACGGCTCGCGGATTGGAAGCCATTGG | 545 | | |
| Sbjct 486 | CACCCGGCAGAACATCGACGAGTTCACCTTACCACCGCCCGGGGCTGGAGGCCATCGG | 545 | | |
| Query 546 | CGGAGCCAAAAAGGCAAAGCCATTATTATCTTGAATCCGGCTGAACCACCGATTTTGAT | 605 | | |
| Sbjct 546 | GGGGGCCAAGAAGGGGAAGGCCATCATCTCTGAACCCGGCGGAACCCCCATCCTCAT | 605 | | |
| Query 606 | GACGAACACGGTGCCTGCATTCGGGAAGATGAAGGCTTTGATCGCGAAGCCGTCGTCGC | 665 | | |
| Sbjct 606 | GACCAACACCGTGCCTGCATCCCGGAGGACGAGGGCTTTGACCGGGAGGCGCTGGTGGC | 665 | | |
| Query 666 | CAGCGTTCGCGCTATGGAACCGGAAGTCCAAGCCTATGTTCCGGGATATCGCTTGAAGC | 725 | | |
| Sbjct 666 | GAGCGTCCGGGCCATGGAGCGGGAGGTCCAGGCCTACGTGCCCGGCTACCGCCTGAAGGC | 725 | | |
| Query 726 | CGATCCGGTCTTTGA-ACGCTTGCCGACGCCGTGGGAGAAGCGCACGGTTCGTCAGCATGT | 784 | | |
| Sbjct 726 | GGACCCGGTGTGAGAGGCTT-CCCACCCCTGGGGGAGCGCACCGTGGTCTCCATGC | 784 | | |
| Query 785 | TGTTGGAAGTCAAGGCGCTGGCGATTATTTGCCGAAATATGCCGAAACTTGGATATTA | 844 | | |
| Sbjct 785 | TCTTGGAGGTGGAGGGGGCGGGGACTATTTGCCCAAATACGCCGGAACCTGGACATCA | 844 | | |
| Query 845 | TGACGGCCAGCGCACGCCGCGTCCGGCGAAGTCTTTGCCAACATTTGCTGGGCAAACCGG | 904 | | |
| Sbjct 845 | TGACGGCTTCTGCCCGAGGGTGGGGAGGTCTTCGCCAGCACCTCCTGGGGAAGCCG | 904 | | |
| Query 905 | TCGAAGAAGTCGTCGCCACCATCACCATCATCATTAATAA | 945 | | |
| Sbjct 905 | TGGAGGAGGTGGTGGCGCATCACCATCATCACCCTGATAA | 945 | | |

Figure 95. Sequence comparison among the codon-optimized *T. thermophilus* Ald gene (Query) and the non-codon optimized version of the same gene (Subject). The differences observed are due to codon optimization for *Geobacillus stearothermophilus* carried out by GeneART using specialized algorithms that take into account the formation of secondary structures.

6.2.4. EXPRESSION ACYLATING ALDEHYDE DEHYDROGENASE (ALDH)

A synthetic RBS was designed for this gene using the Forward Engineering function of the Salis RBS Calculator. The resulting synthetic RBS had a predicted strength of 28442 AU for *E. coli* and 51056 AU for *G. thermoglucosidans*. The codon-optimized *Ald* gene was synthesized as a Gene String (Invitrogen) and so included in the sequence made were the synthetic RBS located upstream from the start codon and an XbaI site located upstream from the RBS. A 6X histidine tag was placed before the stop codon and a SacI site was placed downstream from it.

The entire synthetic *Ald* sequence was thus amplified with primers Tt_ALDH and RC-Tt_ALDH. The PCR-product was digested with XbaI and SacI, and was ligated to an XbaI/SacI p11AK1 plasmid, resulting in the construct depicted in **Figure 96**.

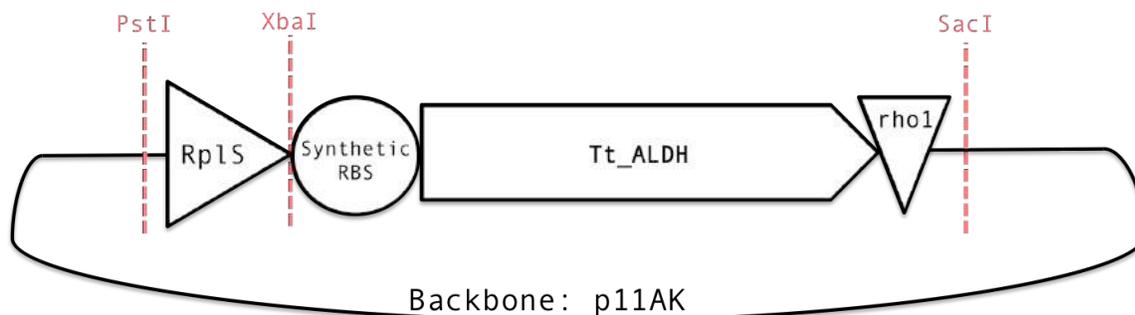


Figure 96. Design of the cassette used to express Tt_Aldh by *G. thermoglucosidans* DL44. The RplS promoter, together with a synthetic RBS direct transcription of Tt_Aldh in the thermophile. "rho1" is the rho-independent transcriptional terminator placed downstream from Tt_Aldh.

E. coli DH10B cells were transformed with this ligation and the p11AK+Tt_ALDH plasmid was isolated and used to transform *G. thermoglucosidans* DL44. The resulting transformant colonies were screened by colony-PCR to identify the correct transformants, which were afterwards confirmed by sequencing (the sequencing results have been included in **Appendix Figure A139**). Two of these were used to inoculate 5 mL of 2TY/kanamycin and were grown overnight. 500 μ L of this culture were used to inoculate 50 mL of 2TY/kanamycin and the cells were grown to an OD₆₀₀ of 2. The cells were harvested by centrifugation and the pellet was resuspended in 5 mL of KOH-MES buffer. The cells were lysed by sonication and the cell lysate was analyzed on a Western blot to confirm the expression of Tt_ALDH, which is shown in **Figure 100**. Bands that appear to be the correct

size (34 KDa) were seen in the Western blot, together with weaker bands seen at about 30 kDa, which seem to belong to a protein from DL44, which is perhaps being targeted by the histidine tag antibodies as well due to its histidine content.

As stated above, the Concise BLAST search also revealed that, apart from the *Thermus thermophilus Ald* gene, there is an equivalent version of the same gene in *Geobacillus* sp. Y4.1MC1 and *G. thermodenitrificans*. This provided a second option to be overexpressed. Given that *Geobacillus* sp. Y4.1MC1 is more closely related to *G. thermoglucosidans*, the sequence of the Y4.1MC1 *Ald* gene, which can be found under the accession number YP_003989256, was compared against the genome of *G. thermoglucosidans* DL33, the wild-type strain. It was found that DL33 also contains the same *Ald* gene, and the DL33 gene is 99% similar that of *Geobacillus* sp. Y4.1MC1. *G. thermoglucosidans* DL33 was selected because it's genome sequence has been made available by Ward [2014] .

Primers *Geoth_ALDH* and *RC-Geoth_ALDH* were designed according to the sequence of *Geobacillus* sp. Y4.1MC1 *Adl* gene, and they were used to PCR-amplify *Adl* from *G. thermoglucosidans* DL33. Bands of the correct size were observed, as it can be seen in **Figure 97**.



Figure 97. PCR-amplification of gene *Ald* from *G. thermoglucosidans* DL33. Lane 1, molecular size ladder (Fermentas, 1 kb ladder). Lanes 2-10, eight DL33 colonies were screened with primers *Geoth_ALDH* and *RC-Geoth_ALDH*. The size of the bands expected was 1.017 Kb.

A synthetic RBS was designed for *Geoth_ALDH* using the Forward Engineering function of the Salis RBS Calculator and the sequence for this was included into amplification

primers. The predicted strength of the RBS was 21389 AU for *E. coli* and 32069 AU for *G. thermoglucosidans*.

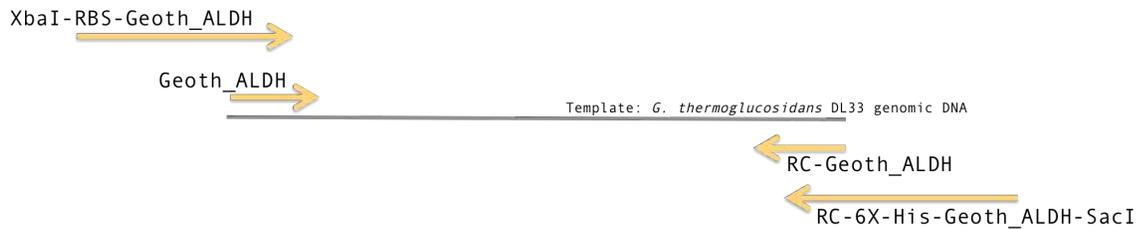


Figure 98. Primers used to amplify *Geoth_ALDH* from *G. thermoglucosidans* DL33. The first PCR to amplify the gene from the genome was carried out using primer pair *Geoth_ALDH* and *RC-Geoth_ALDH*. The second PCR was carried out using primer pair *XbaI-RBS-Geoth_ALDH* and *RC-6XHis-Geoth_ALDH-SacI*.

The PCR-product obtained with primers *Geoth_ALDH* and *RC-Geoth_ALDH* (**Figure 98**) was gel extracted and used as template for a second PCR-amplification with primers *XbaI-RBS-Geoth_ALDH* and *RC-6XHis-Geoth_ALDH*. This forward primer also encoded the synthetic RBS upstream from the start codon of *Ald* and an *XbaI* site upstream from the RBS. The reverse primer also encoded a 6X histidine tag that was located upstream from the stop codon of *Ald*, and a *SacI* site located downstream from the stop codon.

The PCR-product of the second amplification was digested with *XbaI* and *SacI*, and was ligated to an *XbaI/SacI* p11AK1 plasmid, resulting in the construct depicted in **Figure 99**.

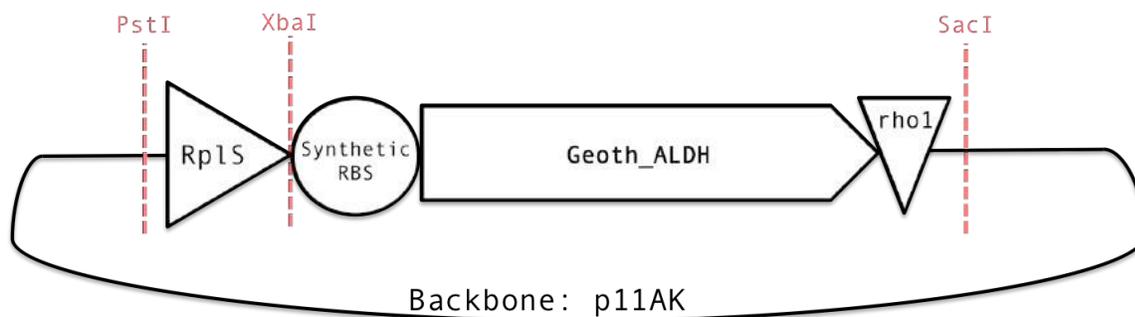


Figure 99. Design of the cassette used to express *Geoth_Aldh* by *G. thermoglucosidans* DL44. The *RplS* promoter together with a synthetic RBS direct transcription of *Geoth_Aldh* in the thermophile. "rho1" is the rho-independent transcriptional terminator placed downstream from *Geoth_Aldh*.

E. coli DH10B cells were transformed with this ligation and the p11AK+*Geoth_ALDH* plasmid was isolated and used to transform *G. thermoglucosidans* DL44. Two transformants

of these were used to inoculate 5 mL of 2TY/kanamycin and were grown overnight. 500 μ L of this culture were used to inoculate 50 mL of 2TY/kanamycin and the cells were grown to an OD_{600} of 2. The cells were harvested by centrifugation and the pellet was resuspended in 5 mL of KOH-MES buffer. The cells were lysed by sonication and the cell lysate was analyzed on a Western blot to confirm the expression of Geoth_ALDH, which is shown in **Figure 100**.

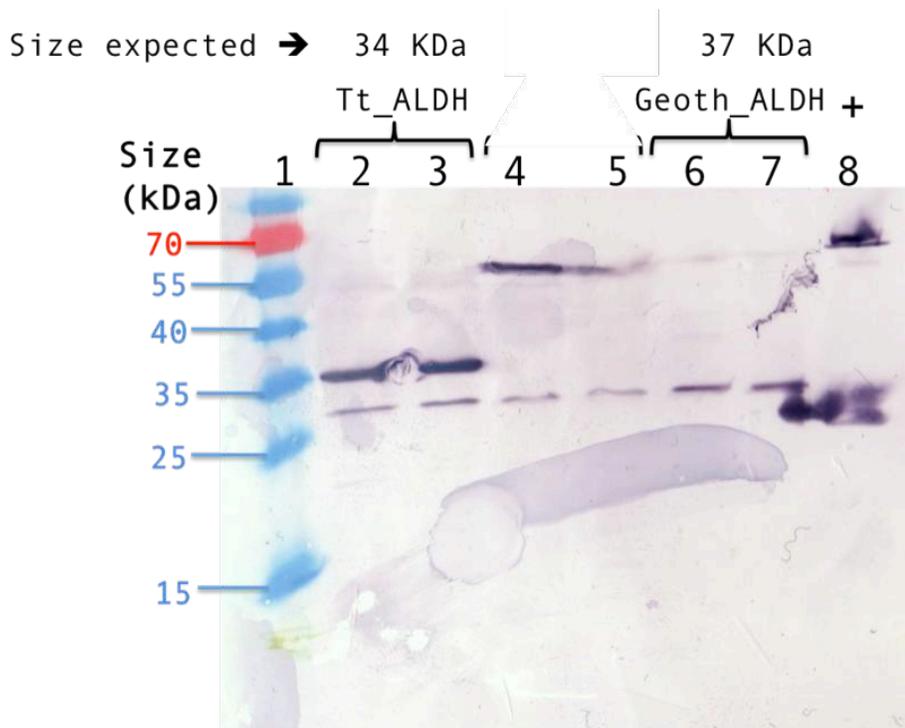


Figure 100. Cell lysate of *G. thermoglucosidarius* DL44 analyzed on a Western blot. Lane 1, prestained protein molecular marker (Fermentas). Lanes 2-3, cell lysate from two colonies transformed with plasmid p11AK+Tt_ALDH. Lanes 4-5 correspond to a sample described in **Figure XX** and are not relevant in this section. Lanes 6-7, cell lysate from two colonies transformed with plasmid p11AK+Geoth_ALDH. Lane 8, positive 6XHis-tag control corresponding to a sample from a previous unrelated experiment. The sizes expected were: 34 KDa for Tt_ALDH.

Bands of the correct size, expected to be around 37 KDa were not seen on the Western Blot, indicating that this protein could not be overexpressed.

The sequencing results (**Appendix Figure A140**) revealed that only the first 31 bp of Geoth_ALDH remained in the construct after the plasmid was extracted from *G. thermoglucosidans* DL44. The presence of those 31 bp confirm that the protein was initially incorporated into the plasmid but then it was lost, possibly due to digestion and ligation of the plasmid by the cells. The gel pictures showing the colony PCRs that demonstrate the presence of bands for Geoth_ALDH from the *E. coli* plasmid isolations used to transform *G.*

*thermogluco*sidans are shown in the **Appendix Figure 156**. The same colony PCR gel pictures, but with the *G. thermogluco*sidans colonies have been included in **Appendix Figure A157** and **A158**. While most of the *E. coli* plasmid isolations do have the entire Geoth_ALDH, none of the *G. thermogluco*sidans colonies have it, which means that they all must have undergone the same digestion and ligation, possibly due to the presence of *G. thermogluco*sidans restriction sites for native type II restriction enzymes.

6.2.5. HPLC ANALYSIS TO DETECT ISOBUTANOL PRODUCTION

Figure 101 shows the HPLC chromatogram obtained after analyzing the sterile ASM medium used to grow DL44. The peaks for glucose and 2-ketoisovalerate supplemented to the medium are conspicuous as expected. The chromatograms for the standards of glucose and 2-ketoisovalerate are included in **Figures A141** and **A142**, respectively.

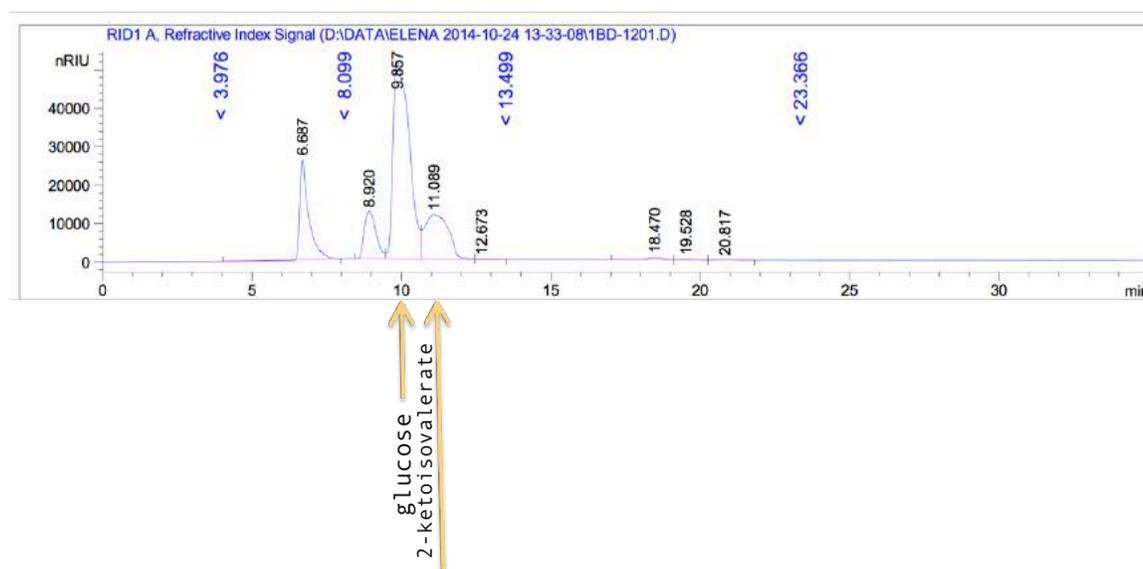


Figure 101. HPLC chromatogram of the sterile ASM medium used to grow *G. thermogluco*sidans. The peak for 0.1 M glucose is found at a retention time of 9.857 minutes and the peak for 2-ketoisovalerate is found at 11.089 minutes. No peak for isobutanol is seen, which demonstrated that the production of isobutanol was obtained after growth in this medium.

In order to determine whether the HPLC column is capable of detecting isobutanol, a 0.1M isobutanol standard was prepared in water and analyzed. The resulting peak is observed at a retention time of 32.693 minutes (**Figure 102**).

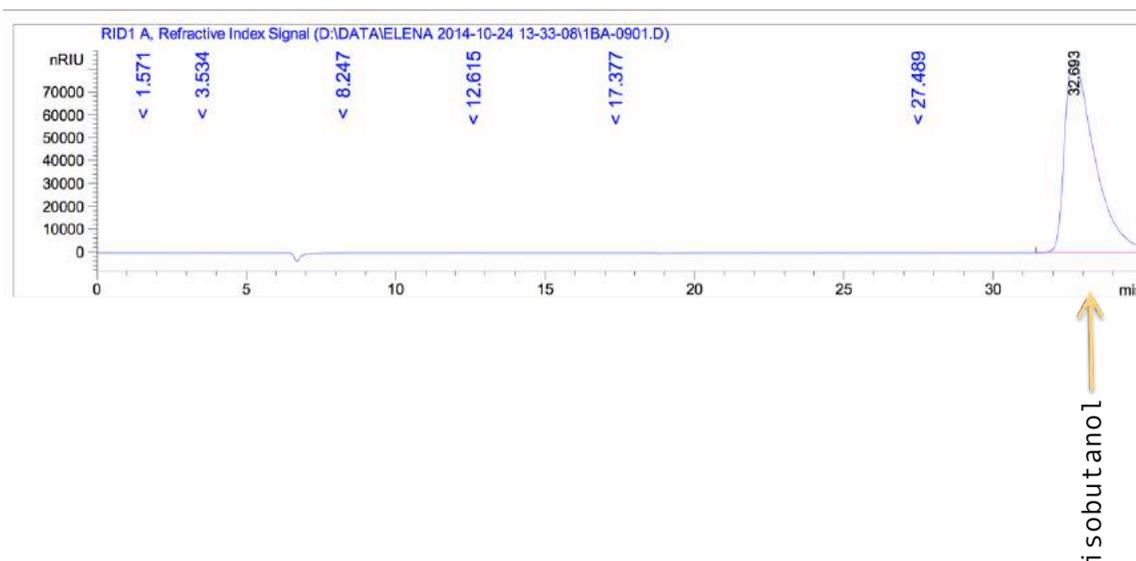


Figure 102. HPLC chromatogram of the isobutanol 0.1 M standard. The peak is observed at a retention time of 32.693 minutes.

Once it was confirmed that isobutanol can indeed be detected by the equipment, *G. thermoglucosidans* DL44 cells transformed with p11AK+kdc, p11AK+Geoth_ALDH and p11AK+Tt_ALDH were grown aerobically overnight at 55 °C, in 5 mL ASM supplemented with 0.1 M glucose, 0.05 M 2-ketoisovalerate and 1% tryptone. The cultures were centrifuged and 1 mL of the supernatant was filtered through a 0.2 µM filter to remove cell debris and was analyzed on the HPLC.

No isobutanol was detected from the Kdc cultures that were grown at 55 °C (see **Figure 103**). As Lin *et al.* suggested, the reason is most likely due to Kdc being unstable at 55 °C.

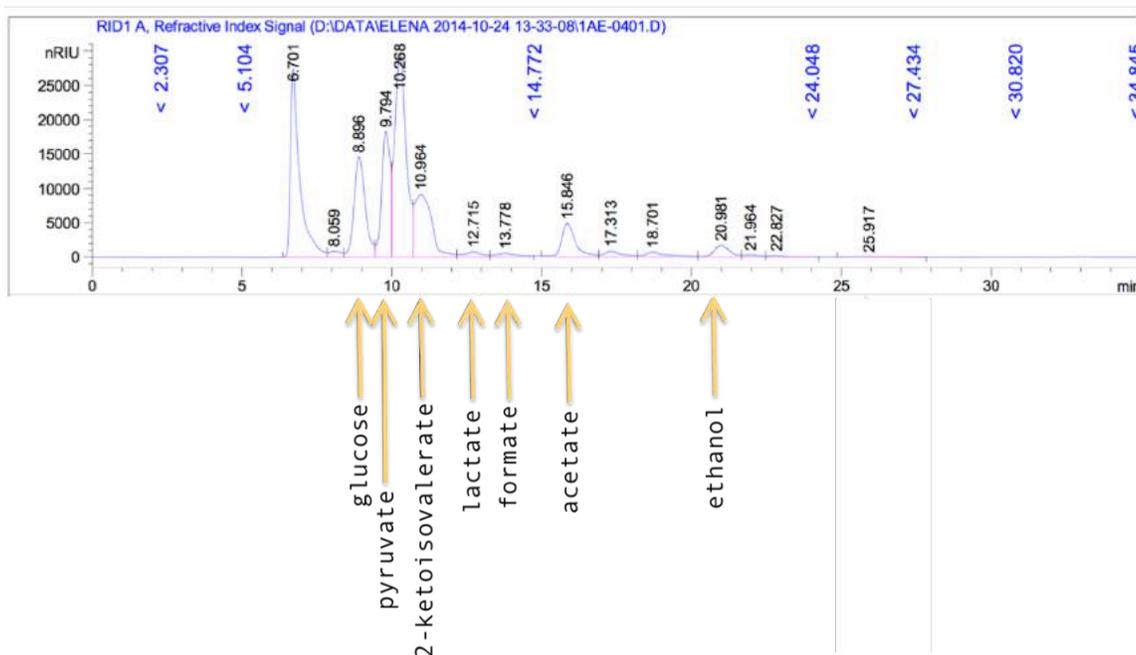


Figure 103. HPLC chromatogram of the supernatant of DL44 p11AK+kdc grown in ASM supplemented with 0.05 M ketoisovalerate. Peaks have been annotated according to the known standards. No isobutanol peak was detected.

The presence of isobutanol was detected in the culture of DL44 transformed with p11AK+Tt_ALDH, as it can be seen in **Figure 104**. Isobutanol was also observed in the DL44 culture transformed with p11AK+Geoth_ALDH (**Figure 105**), even though no bands were retrieved for Geoth_ALDH in the Western blot. This can be observed in **Figure 100**.

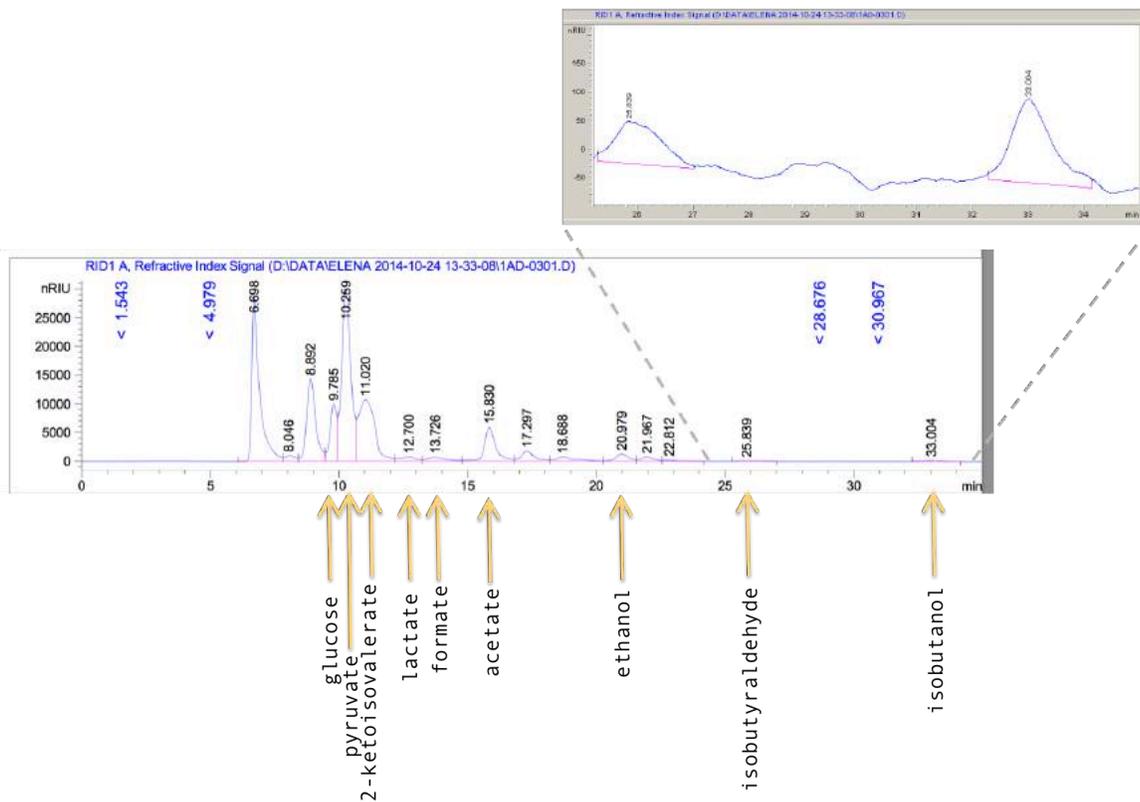


Figure 104. HPLC chromatogram of the supernatant of DL44 p11AK+Tt_ALDH grown in ASM supplemented with 0.05 M ketoisovalerate. The isobutanol peak is seen at 33.004 minutes. Peaks have been annotated according to the known standards. The area of the peak can be found in [Table 16](#).

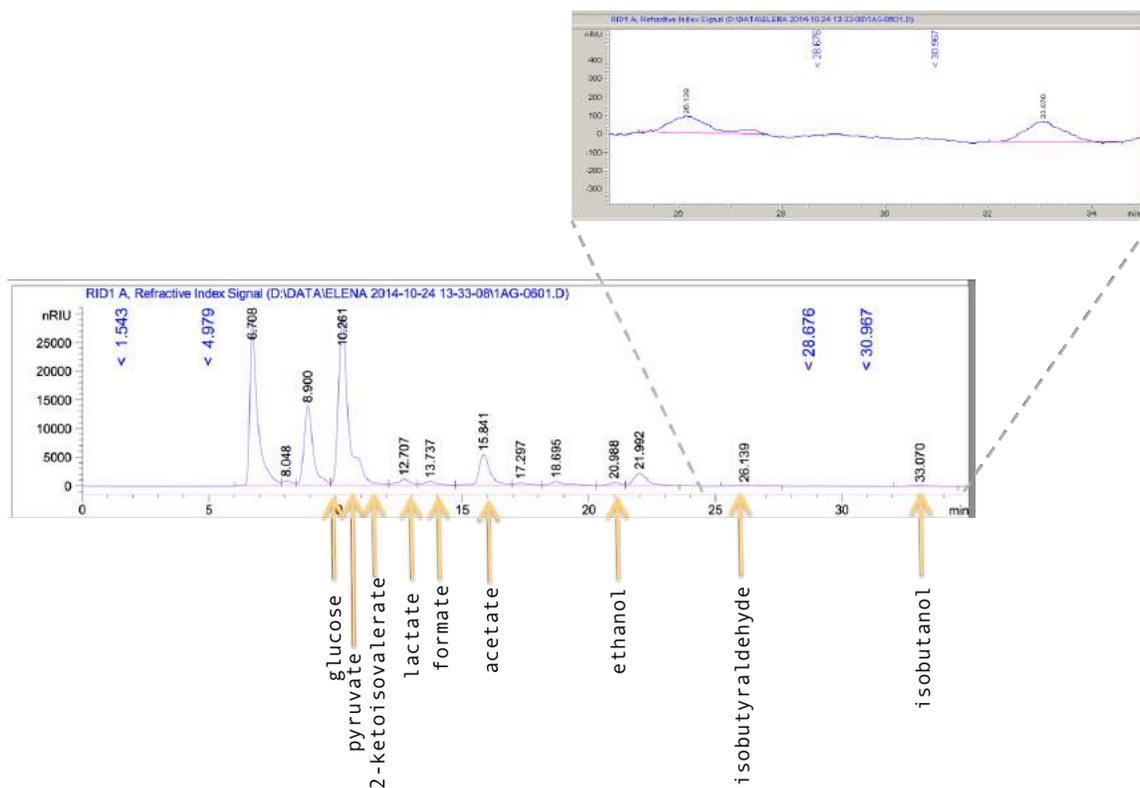


Figure 105. HPLC chromatogram of the supernatant of DL44 p11AK+Geoth_ALDH grown in ASM supplemented with 0.05M ketoisovalerate. The isobutanol peak is seen at 33.004 minutes. Peaks have been annotated according to the known standards. The area of the peak can be found in Table 16.

It is interesting to observe that, even though no band for Geoth_ALDH was observed in the Western blot, a signal for isobutanol is detected by HPLC. This indicates that the plasmid containing the *G. thermoglucosidans* DL33 ALDH recombined into the genome and the DL33 ALDH is being expressed plasmid bound, or it indicates that DL44 possesses the same ALDH as DL33 and the genomic expression of this protein is capable of producing isobutyraldehyde from isobutyryl-CoA.

To inquire into the possibility that the untransformed strain can produce isobutanol, DL44 was grown in ASM medium supplemented with 0.05M 2-ketoisovalerate. The chromatogram obtained by HPLC is shown in **Figure 128**.

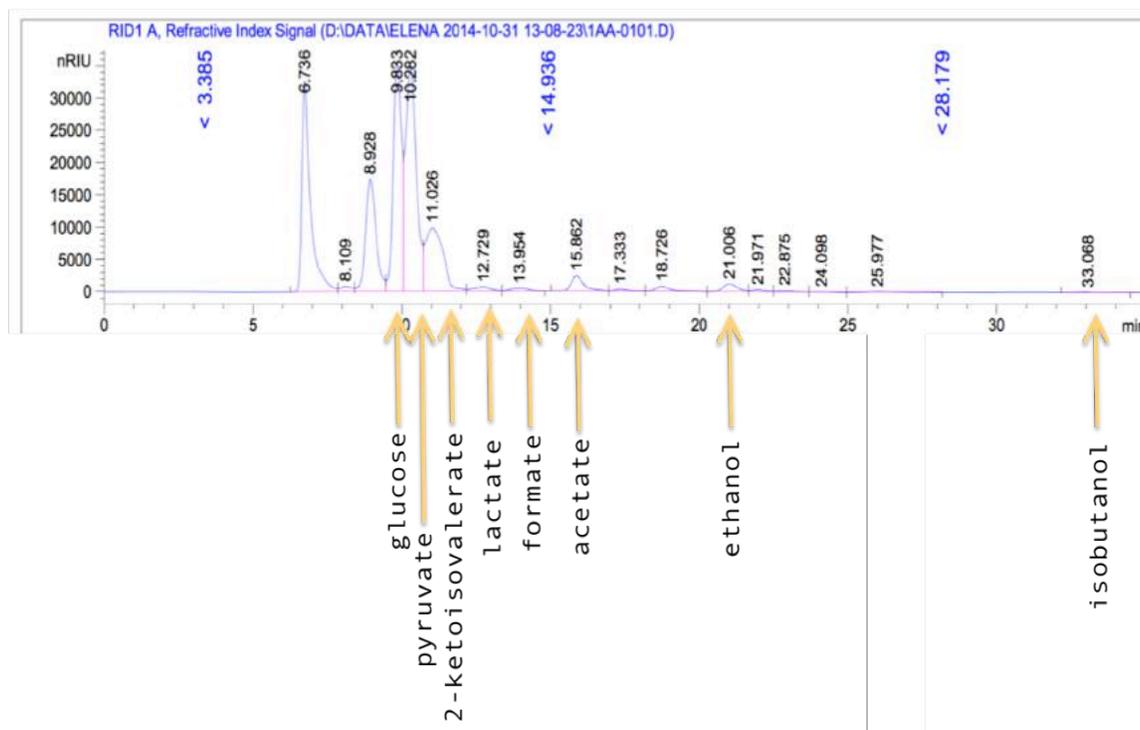


Figure 106. HPLC chromatogram of the supernatant of DL44 grown in ASM supplemented with 0.05M ketoisovalerate. Peaks have been annotated according to the known standards. The isobutanol peak can be attributed to the activity of the intrinsic ALDH of *G. thermoglucosidans* when supplemented with 2-ketoisovalerate given that, when 2-ketoisovalerate is not supplemented, no peak is observed.

When this culture was supplemented with 0.05 M 2-ketoisovalerate, it can be observed that there is a small isobutanol peak at 33.066 minutes. These retention times correspond well to the retention times of the standards shown in **Figures A143, 108** and **111** (2-ketoisovalerate, isobutanol and isobutyraldehyde, respectively).

When DL44 is grown in ASM without supplementing with 2-ketoisovalerate, no peaks for isobutanol are seen, as shown in **Figure 107**.

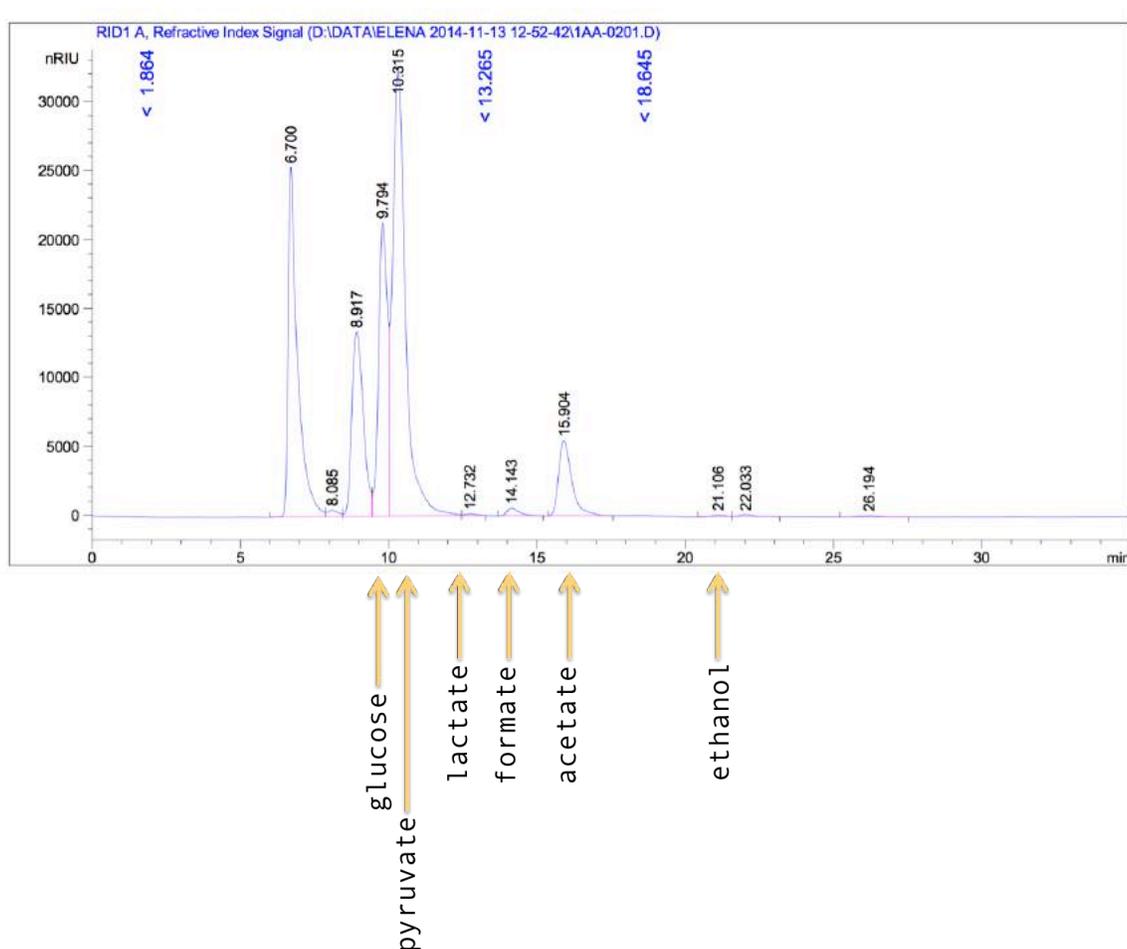
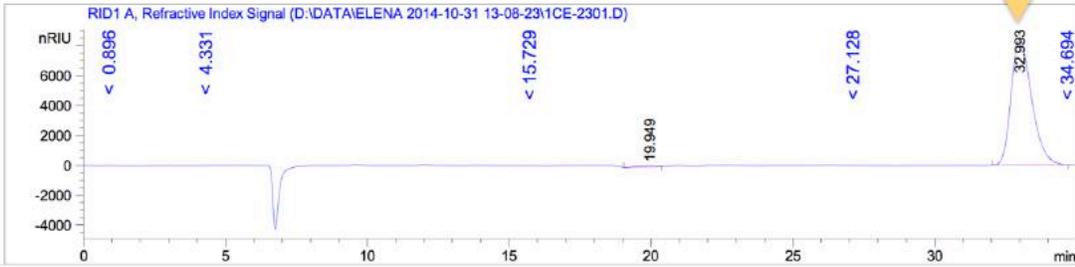


Figure 107. HPLC chromatogram of the supernatant of DL44 grown in ASM not supplemented with ketoisovalerate. Peaks have been annotated according to the known standards. No peak for isobutanol was detected in this sample due to absence of the precursor in the growth medium. The absence of the peak agrees with the hypothesis that isobutanol is obtained only when the precursor is administered to the growth medium to counteract the burden imposed by depletion of this essential metabolite.

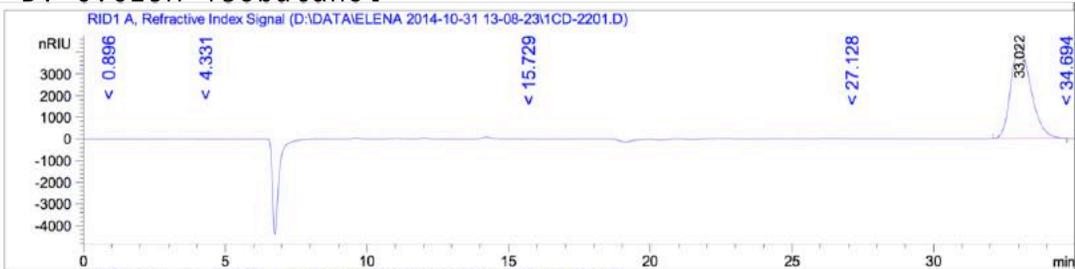
In order to be able to estimate the amount of isobutanol being produced, a calibration curve was constructed. Various decreasing dilutions of known concentration were prepared and they were analyzed on the HPLC. The peaks they produced are shown in **Figure 108**.

Isobutanol peak

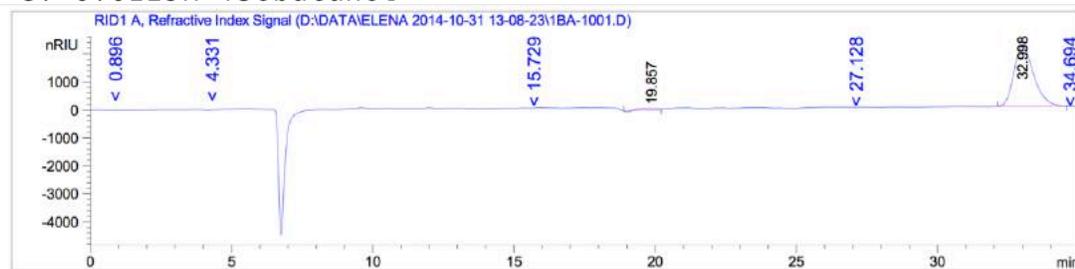
A. 0.05M isobutanol



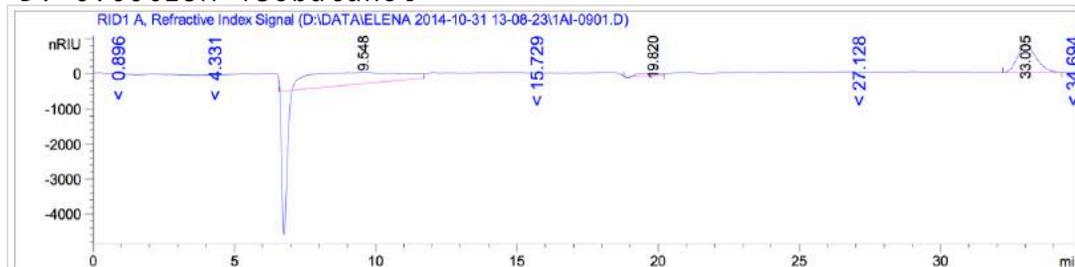
B. 0.025M isobutanol



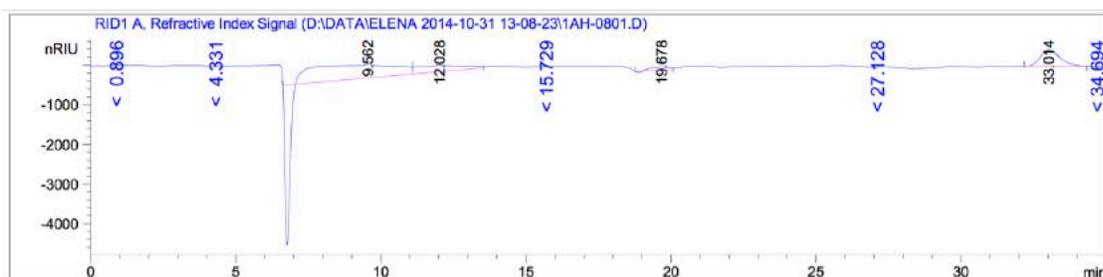
C. 0.0125M isobutanol



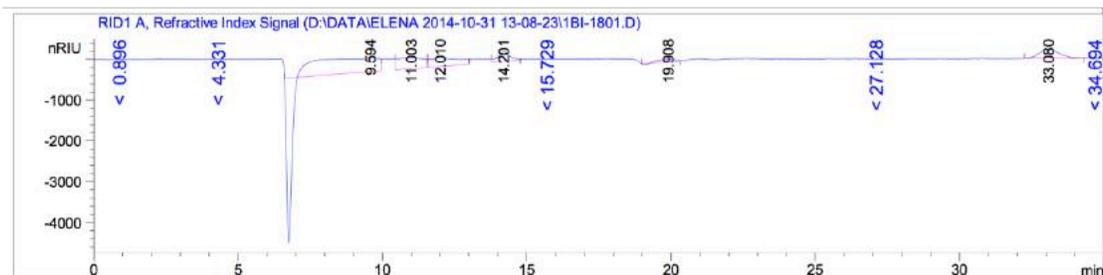
D. 0.00625M isobutanol



E. 0.003125M isobutanol



F. 0.0015625M isobutanol



G. 0.00078125M isobutanol

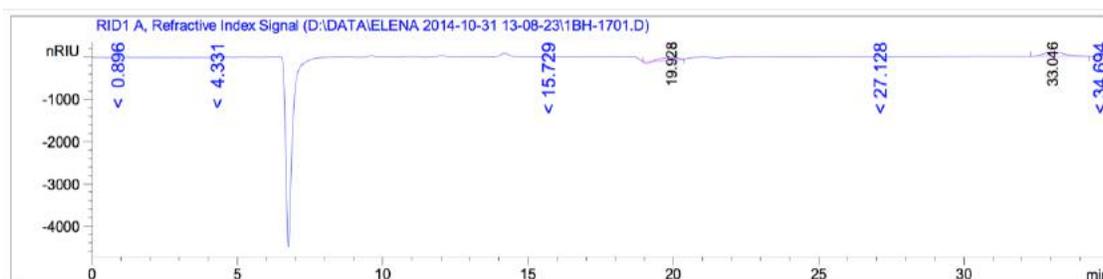


Figure 108. HPLC chromatograms obtained with RI detection from the analysis of decreasing concentrations of isobutanol. The highest concentration was 0.05 M and the lowest concentration was 0.00078125 M. All isobutanol peaks occurred at a retention time between 32.993 and 33.080 minutes. All dilutions were made in water. The initial peak, at around 7 minutes corresponds to water.

To additionally confirm the presence of the isobutanol peak, the same supernatant from cultures of DL44 transformed with p11AK+Tt_ALDH grown in ASM supplemented with 2-ketosivalerate that were analyzed by HPLC were subjected to a toluene extraction and analyzed on a GC-MS.

The peak detected is shown in **Figure 109** and it was confirmed to be isobutanol by the Mass Spectrometry module. To further confirm that it was indeed isobutanol, an

isobutanol standard was also analyzed on the GC-MS and the peak (**Figure 110**) appears at exactly the same retention time.

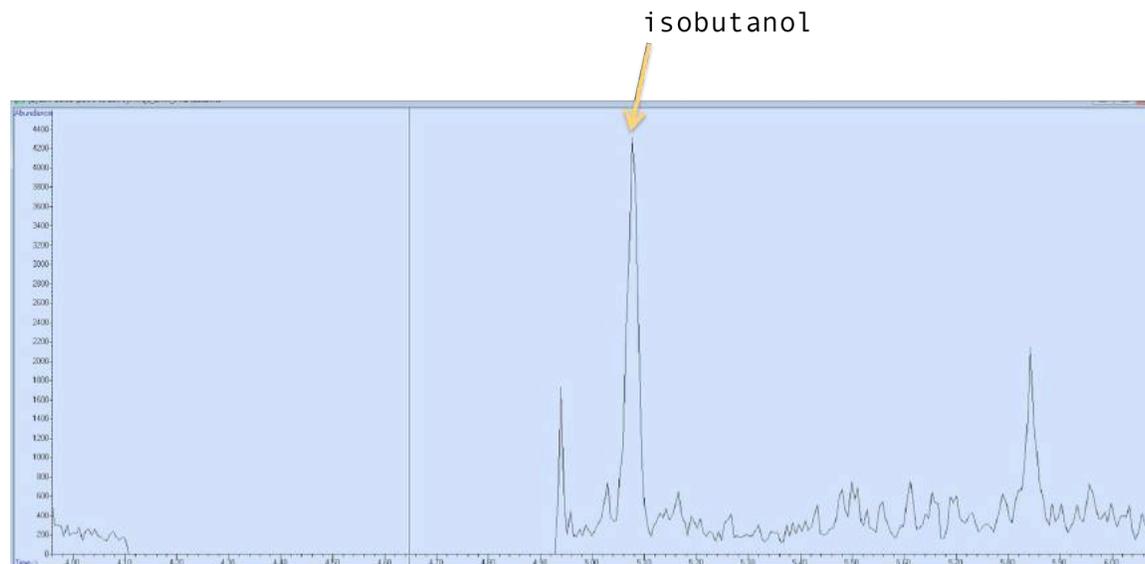


Figure 109. GC-MS chromatogram of the supernatant of DL44 grown in ASM supplemented with 0.05M ketoisovalerate. The peak at 5:00-5:10 was identified as isobutanol by Mass Spectrometry.

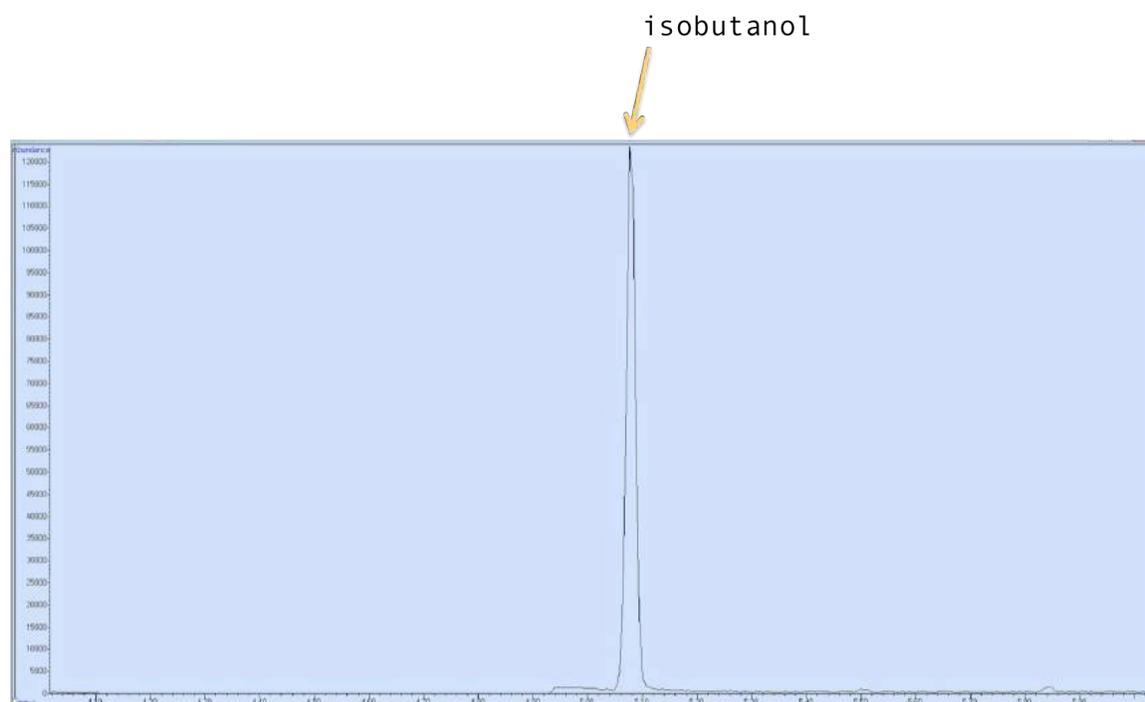


Figure 110. GC-MS chromatogram of an 0.1M isobutanol standard in HPLC grade toluene. The peak at 5:00-5:10 was identified as isobutanol by Mass Spectrometry.

In addition, there is a peak that appears in all HPLC chromatograms at 26.194 minutes. To determine whether the peak could be isobutyraldehyde, an isobutyraldehyde standard curve was also analyzed on the HPLC. However, The unknown peak cannot be isobutyraldehyde because the retention time of isobutyraldehyde has a retention time of 28.446 (see **Figure 111**), so it's more likely that it is a peak from the tryptone added to ASM.

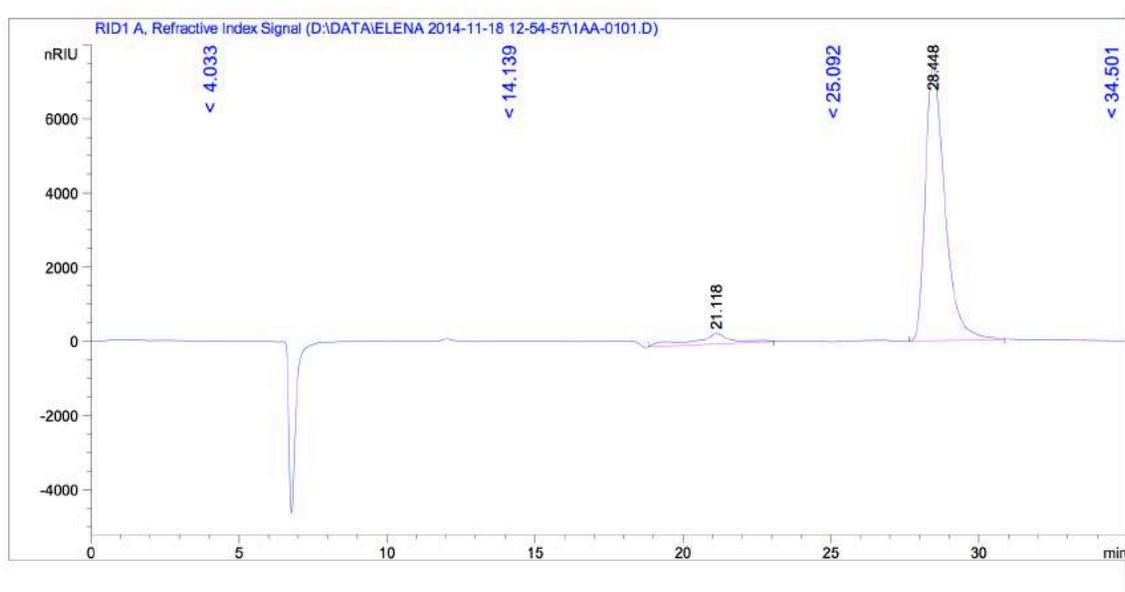


Figure 111. HPLC chromatograms obtained with RI detection of a 0.1M isobutyraldehyde standard in water. The retention time is at 28.448 minutes.

Table 16. Results from the calibration curve obtained after analyzing decreasing concentrations by HPLC. Average area (N=2) (nRIU*s) units per peak are shown per dilution. The area per 0.001 moles of isobutanol was interpolated from this.

| Molar concentration | Area (nRIU*s) | Area (nRIU*s) per 0.001 molar isobutanol |
|---------------------|---------------|--|
| 0.05 | 424124.5 | 8482 |
| 0.025 | 213221.5 | 8529 |
| 0.0125 | 105743 | 8459 |
| 0.00625 | 43324.7 | 6932 |
| 0.003125 | 21852.5 | 6993 |
| 0.0015625 | 11964.5 | 7657 |
| 0.00078125 | 7011 | 8974 |
| | | 8004 = average |
| | | 811 = standard deviation |

The area corresponding to each peak was extracted and it is shown in **Table 16**. Based on these figures, the average concentration of isobutanol per 0.001 M isobutanol was interpolated. This number (8004 nRIU*s per 0.001 M isobutanol), was multiplied with the peak area corresponding to the amount of isobutanol produced per sample (**Table 17**). The results of this are shown in **Table 18**.

Table 17. Peak areas extracted from the HPLC analysis of the supernatant ASM of *G. thermoglucosidans* (nRIU*S).

| DL44 | Geoth_ALDH | Tt_ALDH | Kdc |
|------|------------|---------|--------------|
| 4189 | 5608 | 7582 | Not detected |
| 3763 | 5373 | 3173 | Not detected |
| | 5274 | | |

Table 18. Moles of isobutanol obtained per 1 mole of glucose after 24 hours growth in a 5 mL culture, calculated according to the calibration curve concentrations and the peak areas obtained by HPLC analysis of *G. thermoglucosidans* grown in ASM supplemented with 0.05 M 2-ketoisovalerate.

| DL44 | Geoth_ALDH | Tt_ALDH | Kdc |
|---------------------------|---------------------------|---------------------------|--------------------|
| 0.005233736 | 0.007006634 | 0.009472949 | Not detected |
| 0.004701491 | 0.006713025 | 0.003964346 | Not detected |
| | 0.006589335 | | |
| 0.000496761 ¹⁸ | 0.000676966 ¹⁹ | 0.000671865 ²⁰ | average |
| 0.0000376 | 0.0000214 | 0.0003895 | standard deviation |

The average values from **Table 18** are illustrated in **Figure 112**.

¹⁸ 0.000496761 moles of isobutanol are equivalent to 36.82 mg/L (isobutanol f.w. 74.122 g/mol)

¹⁹ 0.000676966 moles of isobutanol= 50.178 mg/L isobutanol

²⁰ 0.000671865 moles isobutanol = 49.8 mg/L isobutanol

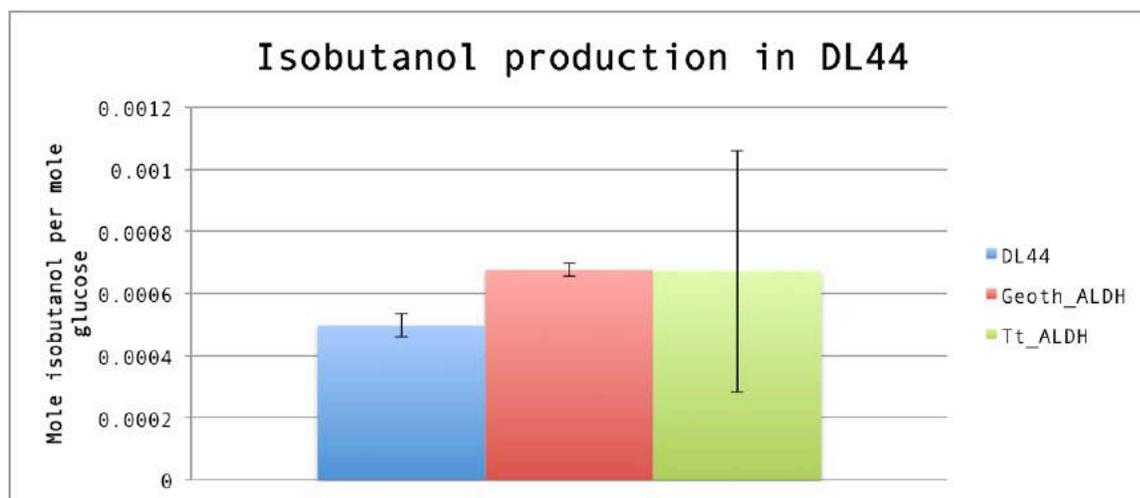


Figure 112. Average production of moles of isobutanol per moles of glucose achieved by strains Geoth_ALDH and Tt_ALDH in comparison to DL44. Error bars represent the standard deviation of at least two samples. Further experiments need to be carried out with Tt_ALDH are needed under strictly defined conditions, such as a bioreactor, to understand the large error bars in the Tt_ALDH sample. It is possible that they are due to differences in oxygenation or cell density at the time when the sample was taken.

6.2.6. OVERPRODUCTION OF 2-KETOISOVALERATE

In order to increase the biosynthesis of 2-ketoisovalerate, the possibility of the plasmid-bound expression of *ilvIHC* and *ilvD* was considered. 2-Ketoisovalerate is an important intermediate **because** it enters alternative pathways.

Since *ilvIHC* and *ilvD* are located in different operons, the creation of an artificial operon would allow the simultaneous expression of both. This would require placing a very strong promoter and immediately downstream from it, *ilvIHC*, followed by *ilvD*. Placing *sfGFP* directly downstream from *ilvD*, would allow to easily confirm the transcription of the entire operon by the presence of green fluorescence. The design is shown in **Figure 113**.

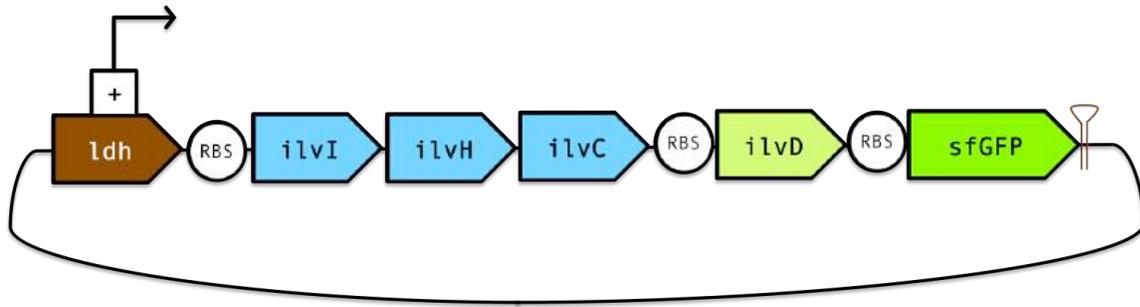


Figure 113. Design of artificial operon to overproduce 2-ketoisovalerate, composed of *ilvIHCD* and *sfGFP*, all downstream from the *ldh* promoter. No synthetic RBS was included between *ilvI* and *ilvH* or between *ilvH* and *ilvC* because, since they are already part of an operon, they have their own native RBS.

In order to construct this artificial operon, *ilvIHC* were PCR-amplified from the *G. thermoglucosidans* DL33 genome using primers PstI-*ilvIHC* and RC-*ilvIHC*-RBS. *ilvD* was amplified using primers RBS-*ilvD* and RC-*ilvD*-RBS. The RBS encoded in primers was the *pheB* RBS used previously in the pUCG18+*ldh*+*sfGFP* construct. These primers had 35 bases of overlap for Gibson assembly. The vector, pUCG18+*ldh*+*sfGFP*, which already contained the *ldh* promoter and *sfGFP* was PCR amplified using primers *ilvD*-RBS-*sfGFP* and RC-*ldh*-RBS. The Gibson assembly reaction was carried out and electrocompetent JM109 was transformed using 1 μ L of desalted Gibson reaction mix. No correct transformant colonies were recovered possibly due to the complexity of the assembly reaction and large size of the final construct. All the transformants had only *ilvD* and *sfGFP*, but *ilvIHC* could not be incorporated.

6.3. DISCUSSION

Although the toxicity of butanols is often ascribed to membrane damage, it is possible that one of *G. thermoglucosidans*' alcohol dehydrogenases oxidizes isobutanol to a by-product that is toxic for the cells. However, no growth was observed, indicating that even in the rare case that isobutanol was oxidized, it is converted to a by-product that the cells cannot use.

Tang et al. [2009] discovered that most of the fatty acids in *G. thermoglucosidans* were saturated, which would also explain why *G. thermoglucosidans* is significantly less

tolerant to isobutanol than *B. subtilis*. Saturated fatty acids lead to decreased membrane fluidity [Daron, 1970], so the membrane damage caused by butanols might be higher than in *B. subtilis*.

Heat allows the membrane of the cell to be more fluid. In the case of butanols, which potentially accumulate in the membrane, this is particularly advantageous because it suggests that strain tolerance to butanols can be improved by artificial selection.

Based on the results obtained with Kdc_CAB34226, and as reported by Lin et al. 2014, it can be confirmed that this protein is neither thermophilic nor thermostable and it is not suitable for producing isobutanol in *G. thermoglucosidans* at industrially relevant temperatures. Keto acid decarboxylases require TPP as a cofactor for correct folding, and it is possible that the high temperatures alter the three dimensional structure of the protein, and the binding of TPP is altered, thus producing a non-functional enzyme. This is in accordance with a paper by Blanchard et al. [2014], which also raises the point that heterologously expressed proteins from mesophiles do not fold properly in thermophiles because they are marginally stable. Small changes in temperature, chemical environment or amino acid sequence can cause misfolding.

As reported by Lin *et al.*, *G. thermoglucosidans* possesses an alcohol dehydrogenase (Geoth_3237), which is capable of reducing isobutyraldehyde into isobutanol. In the Lin *et al.* paper, it was reported that, surprisingly, the isobutanol yield decreased when the Geoth_3237 Adh was overexpressed.

Both Adh and ALDH require NADH as the electron carrier, which in turn affects the redox balance. An Adh that requires NADH instead of NADPH should have a lower impact on the redox imbalance, when growing *Geobacillus* spp. aerobically.

Kdc appears to be toxic for *E. coli*. This would explain why BL21 transformants could not be obtained, but BL21 (DE3) pLysE transformants could be obtained. The T7 lysozyme stopped the T7 RNA polymerase from transcribing *kdc*, and only by allowing them to slowly lose the T7 lysozyme were the cells able to develop tolerance to the toxicity of this protein. Similarly, Kdc appears to be significantly toxic for *G. thermoglucosidans* as well. As demonstrated by the fact that no growth was observed by cells transformed with p11AK+kdc

at 50 °C, when the protein might still have some residual activity, whereas the cells did grow at 55 °C, when the protein is most likely inactive.

6.3.1. ALDH AND CLONING

The publication by Lin *et al.* [2014] about the production of isobutanol via a keto acid decarboxylase resulted in a change of approach for this project. Instead of continuing to pursue the Kdc route, an alternative route had to be sought. This led to the discovery that ALDH enzymes have the capacity to produce isobutyraldehyde. Having had a vector like p11AK1 made the entire process incredibly quick and straightforward. The only thing that was needed was a simple digestion and ligation, given that the plasmid already had a promoter and an RBS. This was a great example to show how useful it is to have the toolbox of shuttle vectors created as part of this project.

The most likely explanation for the absence of a band for Geoth_ALDH in the Western blot is that the plasmid was digested and religated by *G. thermoglucosidans* after they were transformed, seeing as the Geoth_ALDH was identical in sequence to the genomic ALDH of DL44, although this cannot be confirmed because the genomic sequence of DL44 is not available.

Geoth_KDH was indeed present in the plasmids isolated from *E. coli*, as confirmed by a colony-PCR. Isobutanol was seen on the HPLC anyway, even though the cells were not expressing Geoth_ALDH because they were supplemented with 2-ketoisovalerate.

If the p11AK+Geoth_ALDH plasmid indeed recombined with the genomic ALDH, it would explain why the amount of isobutanol observed is higher than that of untransformed DL44, possibly due to two copies of Geoth_ALDH being expressed from the genome. Considering that the entire plasmid becomes integrated into the genome during the first recombination event, it would explain why two copies of the gene remained genome-bound after the first recombination event. But this doesn't explain why there was no band on the Western blot. If the his-tagged copy of the gene had been retained in the genome, then a band should have appeared. It is not **advisable** to express *G. thermoglucosidans* genes in the same strain they come from because recombination might always be an issue.

The construction of the artificial operon to overexpress the *ilvIHCD* pathway failed. It is difficult to understand the reason for this. The *G. stearothermophilus ldh* promoter is not active in *E. coli*, so the whole operon cannot have been expressed in *E. coli*, but it is possible that part of the operon was indeed being expressed: the green fluorescence observed in some of the transformants that contained *ilvD* and sfGFP could be explained if there is a sequence in the 3'-region of *ilvD* that acts as a promoter. It is possible that the same occurs at the 3'-end of either *ilvH* or *ilvC*, in which case the expression of either of these cause too much toxicity and stress, explaining why no transformant colonies would be observed. It is also possible that Gibson assembly was not the ideal cloning strategy, but in an artificial operon, the genes have to be separated by a minimal RBS to ensure transcription of the entire operon as a polycistronic unit so adding restriction sites to separate the parts might have disrupted translational initiation.

A way to overcome the absence of overexpression of *ilvIHCD* was to supplement the growth medium with 2-ketoisovalerate, in order to overcome the deficit of this metabolite. It would be interesting to see if higher isobutanol titres can be achieved by overexpressing *ilvIHCD* rather than by supplementing the growth medium with 2-ketoisovalerate, which in the long term, is far from being a cost-effective solution for the production of isobutanol.

The production of isobutanol was slightly higher with the addition of 2-ketoisovalerate than without, so it's likely that 2-ketoisovalerate is indeed being diffused into the cells from the culture medium, but more experiments are necessary to confirm this.

6.3.2. ISOBUTANOL YIELDS

Compared to the Lin *et al.* paper, who achieved isobutanol production in yields of 0.223 mole of isobutanol per mole of glucose (3.3 g/L of isobutanol with the supplementation of 0.2M glucose) at 50 °C, the isobutanol production achieved as part of this work was a lot lower (0.00068 mole of isobutanol per mole of glucose which is equivalent to about 50 mg/L). The engineered strains (Geoth_ALDH and Tt_ALDH) show a higher production of isobutanol than DL44. The very large error bars seen in **Figure 112**, indicate that more biological replicates are needed to know how repeatable and statistically significantly higher the production is when 2-ketoisovalerate is added to the medium. There is no mistake however, that when 2-ketoisovalerate is not added to the medium, no isobutanol production is seen.

The non-fermentative 1-butanol pathway from *Clostridium acetobutylicum* has been transferred to *E. coli* by Atsumi and Liao [2008], with a yield of 14 mg/L of 1-butanol. *C. acetobutylicum* can convert acetyl-CoA to butanol. The highest yield obtained in this study from producing isobutanol from isobutyl-CoA was 50.178 mg/L, which is very similar. It appears that the non-fermentative pathway for isobutanol production produces lower yields than the fermentative pathway.

If *G. thermoglucosidans* possesses a native ALDH which can produce isobutyraldehyde when supplemented with 2-ketoisovalerate, it raises the question about the method used by Lin *et al.* to achieve isobutanol production in this thermophile. All the strains they report had a Kdc, as shown in **Table 15**. There was no strain that only expressed *ilvIHCD*, without Kdc. If the isobutanol production they achieved was due to the overexpression of *ilvIHCD*, then a strain without Kdc would have shown isobutanol production as well, but such reference strain was not included in their report. It would explain why the highest isobutanol yields were obtained from cells overexpressing *ilvIHCD*. The 2-ketoisovalerate they were producing was feeding ALDH instead of Kdc, and it was the native ALDH producing the isobutanol, not the heterologous Kdc.

6.3.3. ISOBUTANOL AS A BIOFUEL

In theory, the scientific literature praises isobutanol as an ideal replacement for fossil-based based claiming that the octane number and energy content of isobutanol are much higher than ethanol [Atsumi and Liao, 2008], however, engine tests show that it is not recommendable to have more than 20% isobutanol in a isobutanol:gasoline mix [Karabektas and Hosoz, 2009] due to poor performance of the engine when higher concentrations of isobutanol are added. Also, isobutanol is a toxic compound not just for bacterial cells but for organs as well. It is a hazardous substance responsible shown to cause organ toxicity, skin irritation and eye damage (Sigma safety data sheet)²¹. In addition, using *Geobacillus thermoglucosidans* to produce isobutanol defeats the point of using a thermophile for the production of useful compounds because isobutanol is not volatile and therefore, cannot be as easily separated from the reaction mixture as ethanol.

²¹ <http://www.sigmaaldrich.com/catalog/product/sigma/i0273?lang=en®ion=GB>

In the present work, the production of isobutanol was carried out simple as a proof-of-principle demonstration of the usefulness of the toolbox constructed because it was very easy to exchange the reporters in the expression cassettes for ALDH or Kdc using *XbaI/SacI* digestion and ligation. The promoter and the RBS of the modular vectors are ready for the expression of heterologous genes placed downstream and "rho1" will terminate transcription of any gene placed upstream from it.

6.4. FUTURE WORK

In order to overexpress the genes required for 2-ketoisovalerate overproduction, it would be recommendable to place each gene in the control of a single promoter, instead of assembling an artificial operon, because this should allow for a more controlled expression of each of the genes composing the pathway. The expression of each gene could be tuned to the most efficient levels by using promoters of different strengths coupled with synthetic RBSs of appropriate strengths as well.

Seeing as the isobutanol that is being produced endogenously by the cells is not being secreted outside the cells, it is possible that active removal of isobutanol by an effective method, such as gas stripping will help remove most of the isobutanol that remains inside the cell of trapped within the cell walls.

It is necessary to grow the DL44 and ALDH strains in a large scale bioreactor to determine what's the isobutanol production when a constant flow of metabolites such as glucose and 2-ketoisovalerate are added, and because growing at such a large scale would permit comparing how different the yields are between growing in a large scale as opposed to small scale.

HPLC is a very quick and efficient method to detect the production of metabolites and to quantify them. GC-MS requires more sample preparation work but it might be much more sensitive to the detection of isobutanol, but more samples need to be analyzed and compared by these two methods.

Michaelis-Menten kinetics reveal that there are two routes to improve productivity. One is to have more enzyme available and the other is to improve the catalytic properties.

A directed evolution approach requires the screening of multiple mutants simultaneously. If a vector yields low transformation efficiencies, only a few mutants can be obtained at a time, but with the vectors included in the toolbox, the transformation efficiencies have shown that they have the potential to generate thousands of colonies at a time, each of which could contain a possible mutation if error-prone PCR is used.

For example, it would be possible to improve isobutanol yields in *G. thermoglucosidans* by obtaining, through directed evolution, variants of ALDH or ADH that have a higher affinity for isobutyryl-CoA or isobutyraldehyde, respectively, which could then be selected based on how much of the metabolite they produce.

A directed evolution approach could also change cofactor dependencies of enzymes, which could solve cofactor imbalance and have a big impact on theoretical yields. A rapid genome evolution approach could also contribute towards finding the right combination of mutations that produce a more isobutanol-tolerant phenotype.

In the future, if the isobutanol production pathway by *G. thermoglucosidans* is to be improved, the following key points must be considered:

- Competing pathways must be eliminated. The strains that have the *ldh* and *pfl* pathways knocked out already show a significant increase in the production of ethanol compared to the wild-type strain [Cripps *et al.* 2009]. *G. thermoglucosidans*, however, is a natural ethanol producer, so simply knocking-out pathways won't be enough to increase isobutanol production. The precursor pathways for 2-ketoisovalerate and isobutyryl-CoA need to be upregulated.
- Redox imbalance needs to be considered. Redox imbalance might not be such a concern when growing *G. thermoglucosidans* in a large-scale bioreactor because, under anaerobic conditions, glycolysis produces only NADH. Whereas, the only enzyme in the pathway that requires NADPH is IlvC. Both ALDH and ADH in *G. thermoglucosidans* require NADH. The production of 2-ketoisovalerate and valine require NADPH. If cofactor imbalance becomes a deterrent for the production of isobutanol, there are two possible solutions. The first would be to follow the approach of Bastian *et al.* [2011], and overexpress *pntAB*, which encodes a pyridine nucleotide transhydrogenase, which catalyzes the transfer of a hydride from NADH to NADP⁺.

This strongly contributed to solving cofactor imbalance but it is an enormous load for cell metabolism and it is not an ideal solution. The second option is to alter the cofactor dependence of ilvC by directed evolution to NADH. This has been shown to work effectively by Bastian et al. [2011]. Using an NADH-dependent IlvC would result in a pathway that is entirely NADH dependent, and a much higher yield of isobutanol could be achieved.

- The enzyme expression levels also need to be optimized. Ideally, ALDH and ADH should be expressed by a strong constitutive promoter and with a strong RBS, but it needs to be understood how much stress this would cause to the cells and if perhaps a weaker promoter would result in a healthier balance and therefore, higher titres.

CHAPTER 7: GENERAL DISCUSSION

7.1. SYNTHETIC BIOLOGY

Advances in the fields of synthetic biology and metabolic engineering demonstrate that cells can be engineered into factories that produce high-value chemicals and proteins but such an approach requires appropriate genetic tools (see **Figure 114**). According to Holtz and Keasling [2010], the purpose of characterizing bioparts is to confer them a modular character and make them reusable between applications. This statement makes evident the close connection between characterization, standardization and modularization.

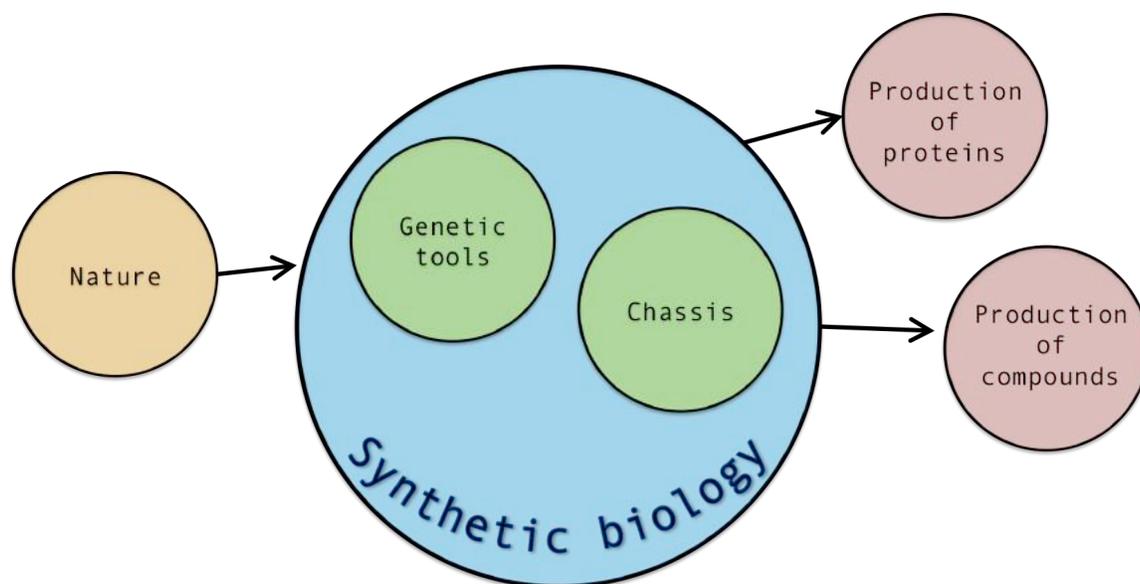


Figure 114. Synthetic biology's input to biotechnology. Synthetic biology provides new tools and approaches to biotechnology and biotechnology aims to discover, improve and industrialize products that can be generated through the use of recombinant DNA technology.

7.2. ABSTRACTION IN SYNTHETIC BIOLOGY: BIOPARTS

Biological functions are complex but they are written in DNA, and DNA is a defined entity and patterns can be found within it that can be linked to begin to understand biological function [Endy 2005]. Such patterns in DNA become bioparts. The same patterns that have the same functions constitute one type of biopart. Eventually, with enough understanding of patterns, more patterns can be deciphered until enough patterns are combined together to generate novel functions and ultimately, build biological systems from scratch.

This work has focused on the abstraction of bioparts to identify functional parts in the unknown sequences of plasmids to find patterns which could be attributed to minimal functional units. Those units were isolated and assembled in new synthetic constructs that resulted in novel functional outputs previously unknown to science. For instance, it is now known which fluorescent reporters are active in *G. thermoglucosidans* under what conditions and it is known that the design of plasmids needs to take into account the engineering side of synthetic biology to make sure that all the required combinations of bioparts achieve a functional outcome. More specifically, antibiotic resistance markers need a transcriptional terminator to be functional.

If a synthetic biology point of view hadn't been applied to solve these questions, the *G. thermoglucosidans* community would continue working with highly inefficient and inappropriate plasmids without understanding the reason for the characteristic low transformation efficiencies.

7.3. SYNTHETIC BIOLOGY AND INDUSTRY

Synthetic biology needs to find application in the real world outside the laboratory, and *G. thermoglucosidans* was a good place to start because it already is being implemented under realistic industrial scale conditions as a production host of valuable compounds (TMO Renewables). Ideally, the robust production of costly compounds should require only minimal low cost inputs, such as a simple growth media and appropriate incubation temperatures and techniques to be a cost effective sustainable process worth investing in, in order to compete with the pre-existing chemical synthesis processes that have already been optimized to be

as cost-effective as possible over time. Industrial scale production requires a host that can withstand the conditions of large-scale fermenters.

Other organisms could have been chosen. For example, Nickel *et al.* [2014] suggest the use of *Pseudomonas* spp. as another chassis for synthetic biology, but the development of this strain is clearly not as advanced as the development of *Geobacillus* spp. There is a lack of tools and *Pseudomonas* appear to be much more difficult to manipulate genetically. And they haven't been tested in an industrial bioreactor.

The principles of synthetic biology of standardization and modularization are important for industry because the industrial scaling of a process that occurs in the lab in small volumes needs to be predictable, controlled and reproducible every time. The characterization of parts under various conditions is of the utmost relevance to be able to predict the behavior of a synthetic system.

7.4. TOOLBOX

The development of tools is necessary to control the expression of genes and metabolic pathways [Keasling, 2012]. The toolbox created in this work offers a starting point for the transcriptional and translational control of plasmid bound protein expression in both *Geobacillus thermoglucosidans*.

The present work has created tools for a substantive and dramatic transition in the capability to genetically manipulate *G. thermoglucosidans* by allowing:

1. **The re-use of components and plasmids** is now possible due to their modular nature. Different combinations of components in diverse plasmid contexts will no longer serve one isolated specific cloning and characterization effort.
2. **To limit the number of time-consuming steps** required for either introducing a new promoter, RBS, or CDS as well as for replacing any of the parts (origin of replication or selection marker) composing the vectors.
3. **The characterization of promoters for *E. coli* or *G. thermoglucosidans*.** The presence of the *Pst*I and *Xba*I restriction enzyme recognition sites allow to easily interchange promoters while leaving the downstream RBS sequence intact. The

promoters may be characterized using the reporters *pheB*, sfGFP, mCherry or BsFbFP, depending on the conditions being studied.

4. **The characterization of RBSs.** The presence of the *XbaI* restriction site upstream from the Shine-Dalgarno sequence permits to excise the CDS together with the RBS whilst leaving the promoter sequence located upstream intact. Unfortunately, a restriction site cannot be placed at the 3'-end of the RBS because it will most likely affect the rate of translation. However, given that the RBS is such a short sequence it can easily be reintroduced with synthetic oligonucleotides.
5. **The characterization of transcriptional terminators.** The effectiveness of transcriptional terminators may be measured by the output of reporter placed before or after the terminator.
6. **The characterization of origins of replication.** The copy numbers of origins of replication may be estimated in various ways: qPCR, comparison of reporter output and growth rates.
7. **The availability of multiple reporters** permits testing the possibility of expressing two plasmids simultaneously, which in turn would permit the plasmid-bound expression of more genes at once.
8. **The heterologous expression of proteins.** The presence of the *XbaI* and *SacI* restriction sites allow to easily exchange the reporter gene currently included in the vectors for a gene encoding a protein that has a desired function.

This work provides evidence that a synthetic biology approach should be implemented for the construction of tools that would permit developing other non-model organisms as chassis:

- As it has been demonstrated in Chapter 5, a careful analysis of DNA sequences is necessary in order to understand what are the functional parts that compose them, and there are freely available online software tools that facilitate this prediction.
- Bioparts can either be obtained from nature or designed *de novo*.
- Origins of replication and selection markers are the main components necessary to construct a shuttle vector, which in turn is the easiest and most common route for gene transfer among hosts.
- Cryptic plasmids are common throughout microorganisms, and host-specific origins of replication could be derived from them if they can be isolated and sequenced.

7.5. HETEROLOGOUS EXPRESSION BY *G. THERMOGLUCOSIDANS*

The toolbox was used to successfully express pheB, sfGFP, mCherry and BsFbFP, as well as Kdc and ALDH. Based on this, it is evident that the toolbox permits the expression of heterologous proteins in *G. thermoglucosidans* at high temperatures. For future efforts that require the expression of protein, it is strongly recommended to use the toolbox constructed in this work.

7.6. IMPROVEMENTS TO THE TOOLBOX

A toolbox can never be fully complete as there is no limit to the number of bioparts that can be included in it, it can only be built-up in complexity. In order to continue the development of *G. thermoglucosidans* as a chassis for synthetic biology, more bioparts have to be added to the toolbox. Fortunately, the modular nature of the toolbox implemented since the design stage and which has been tested for practicality in the course of this investigation is an efficient method to easily swap and “plug-in and play” new bioparts and will encourage the use of the toolbox for the subsequent inclusion of parts and characterization efforts.

If the vectors differ only in one part, for example, the origin of replication, then the results of characterization experiments can be directly comparable and suitable for publication. The modular nature of vectors created in this permits that and it is highly recommendable that they are implemented by the *Geobacillus* research community.

G. thermoglucosidans will be developed as a chassis when new promoters can be added to the toolbox to be characterized under any of the four reporter proteins, depending on the application. New RBSs can be added and they can also be characterized using any of reporters proteins or a combination of them to test for context-dependency.

7.7. THE IMPORTANCE OF PREDICTIVE SOFTWARE FOR SYNTHETIC BIOLOGY

Bioparts need optimizing so that when bioparts interact together, the most optimal, reliable functionality is achieved. The construction and testing of bioparts is time consuming and it is not possible yet to avoid trial-and-error optimization [Kosuri *et al.* 2013]. One way to

mitigate trial-and-error experimentation is by being aided during the design stage of the experiment by predictive software. Such software was used in the course of this investigation, namely:

1. **The RBS Calculator** by Salis *et al.* [2009] allows checking the functionality of RBSs by predicting the secondary structures that arise when three toolbox parts are put together (promoter, RBS, coding sequence). Such predictions were essential for the course of this investigation to seek confirmation that the lack of expression was due to the sequence of the RBS being responsible for unwanted interactions. The Calculator also allowed designing completely synthetic RBS *in silico*, which were then tested *in vivo* with positive results, demonstrating that the RBS Calculator can be used in the future to continue the development of *G. thermoglucosidans* as a chassis.
2. **The IDT OligoAnalyzer** has proved to be extremely valuable for the prediction of potential rho-independent transcriptional terminators. This tool could be used in the future to study rho-independent transcriptional terminators that are intrinsic to thermophiles in order to test them as terminators for devices.

The optimization of the bioparts required to achieve reliable function is time-consuming and it is not possible yet to avoid trial-and-error optimization of the tools that constitute a toolbox because the predictability of how parts behave together is still not 100% certain: idiosyncratic interactions and context effects may arise which may not have been contemplated by the models and need to be tested. It is possible to mitigate trial-and-error optimization by using tools such as the RBS Calculator by Salis *et al.* [2009] by predicting the secondary structures that may arise when three toolbox parts are put together (promoter, RBS, coding sequence).

7.8. NEW SELECTION MARKERS

It is necessary to include in the toolbox selection markers that are not reliant on antibiotics because antibiotic use not viable for industrial scale production due to the cost of the antibiotics and to the capacity cells have to develop resistance. Currently, antibiotic selection is necessary to ensure the retention of the plasmid, but if heterologous proteins are integrated into the genome, then they shouldn't be necessary any longer. It is necessary to pursue an auxotrophic marker for integration in the genome experiments because a second

selection marker is of the utmost necessity to force the second recombination event. It would be particularly advantageous if the marker has both positive and negative selection like *pyrE* and its interaction with 5-FOA.

7.9. LESSONS TO BE LEARNT FROM THIS PROJECT

The approach used in this work to create a toolbox for *G. thermoglucosidans* to empower its development as a chassis for synthetic biology can be implemented across nature. When interesting organisms are found that have desirable intrinsic properties, the creation of a toolbox would permit the genetic modification of such organisms to optimize their natural capacities for human applications.

The synthetic biology approach used in this project resulted in the production of a modular toolbox of genetic components that is efficient and easy to work with, which supports the idea that synthetic biology is a very powerful a field of knowledge to be limited to the traditional *E. coli* or *S. cerevisiae*, which are useful for some applications but not for specific specialized processes. This project was aimed at developing synthetic biology beyond the scope of tradition because if synthetic biology can prove that its ethos can be applied to develop a thermophile chassis, it will be the evidence that shows that the principles of engineering can be applied to the entire biological spectrum of the planet and not only to a few well-known hosts, especially when more specialized hosts are more suitable for certain applications. Synthetic biology could be used to solve some of the problems the world currently faces, such as the energy or the environmental crises.

Creating a toolbox for non-model organisms requires bioparts, which in turn can be obtained from the natural world. For example, origins of replication can be obtained from cryptic plasmids, and they can be paired with known selection markers to create new host specific shuttle vectors. Such origins of replication may have a wide range and work in species that are phylogenetically related, thus contributing to the aggregation of useful bioparts and tools.

7.10. BIOPARTS: *E. COLI* ↔ *G. THERMOGLUCOSIDANS*

The shuttle of parts between *E. coli* and *Geobacillus* does not have to focus on what the effect of the biopart is for *G. thermoglucosidans*. As it was observed in the kanamycin selection transformation efficiency results obtained in this chapter, *E. coli* obtains a benefit as well. For instance, it is clear that the rho-independent transcriptional terminator placed downstream from the kanamycin nucleotidyltransferase gene had extremely beneficial effects in both *E. coli* and *G. thermoglucosidans*.

Thermostability in the proteins of thermophiles has a physical basis. Thermostable proteins are attracting interest because the same physical basis that makes them resist heat degradation is the same basis that makes them more resistant to proteolysis and chemical denaturation, so thermophilic proteins produced by thermophiles are hyperstable and some proteins obtained from thermophiles such as DNA polymerases, xylanases and proteases already have industrial applications [Taylor and Vaisman 2010].

Increased hydrogen bonding and more salt bridges are some of the properties that confer thermostability to thermostable proteins. DNA also forms hydrogen bonds and high temperature in PCR for example, is required to separate both strands. The DNA structures of thermophiles could be extremely suitable to be tested for functionality in mesophiles because they should be more stable and perhaps even more functional than their mesophilic counterparts. This is especially true for bioparts such as transcriptional terminators, which clearly, as demonstrated in this work, are important bioparts and yet very few studies have examined the influence of transcriptional terminators on gene expression. For yeast and *E. coli*, only a handful of terminators are available [Redden 2015]. This should not be the case as terminators are just as important as promoters and RBSs to drive gene expression.

As this work demonstrated, an appropriate rho-independent transcriptional terminator can resolve the host-specificity barrier of antibiotic resistance markers. This work also demonstrated that transcriptional terminators from the sequences of the genomes of thermophiles or from cryptic plasmids of thermophiles can be easily identified by analyzing for hairpins using a free online software (Oligoanalyzer) which can be readily implemented in any research project in any part of the world. This is true especially now that it is known that four reporter proteins could be used to characterize transcriptional terminators from

thermophiles both in *G. thermoglucosidans* and *E. coli* and lead to the construction of a library of rho-independent transcriptional terminators that could become BioBrick standardized parts if they are effective and functional.

The thermostable version of the kanamycin nucleotidyltransferase used in this work [Liao and Kanikula 1991] was developed by expressing the protein in a thermophile and subjecting the thermophile to directed evolution. Small changes in amino acids modified the folding behavior and lead to activity at increased temperatures. If *Geobacillus thermoglucosidans* becomes the thermophilic platform for protein production, the same approach could be applied to obtain more thermostable proteins such as other antibiotic resistance markers or fluorescent reporter proteins which have mutations to confer stability to proteins without compromising the fluorescent output. For this purpose, the high transformation efficiencies obtained with the new modular shuttle vectors are ideal for directed evolution experiments that require screening large libraries of mutants, such as promoters libraries or the evolvability of error-prone PCR variants. Directed evolution could also be applied to improve product yields. Furthermore, as Abatermarco *et al.* [2013] suggest, metabolic pathway engineering can be achieved through evolving parts, *in vivo*, following a directed evolution approach as opposed to traditional gene overexpression or gene deletion by knock-out.

Now more than ever it is necessary to have strong internet based sharing platforms to share bioparts and the knowledge generated from them. Synthetic biology produces standardized parts and vectors are meant to be shared by the scientific community. Such information needs to be available in formats that are easy to understand and to request to permit the prompt advancement of research. BioBricks, BglBrick and pSEVA demonstrate the advantages of standardization: conventional cloning is preferable because it is convenient and has enabled widespread sharing and re-use of parts.

Our current knowledge about the responses of *G. thermoglucosidans* and *E. coli* to synthetic stimuli will grow as long further experiments are carried out with these hosts. The growing availability of genomic information and advances in the fields of synthetic biology and metabolic engineering will contribute to the knowledge aggregation that both research fields need for evolving.

7.11. WHAT THE FUTURE MAY HOLD FOR *G. THERMOGLUCOSIDANS* SYNTHETIC BIOLOGY

An interesting question that should be answered in the future is whether high copy number is actually a necessary property of a plasmid in order to achieve high product production? Or whether it is only necessary to have a low copy number plasmid but with a strong promoter and strong RBS? Would the burden on the cell diminish and result in higher production rates? According to Tyo *et al.* [2009], expression levels after recombination in the genome are similar as those obtained with a high copy number plasmid when you a strong promoter and strong RBS were used. High copy plasmids can magnify the effects of promoter activity levels and produce an imbalance of enzyme levels [Du *et al.* 2012], so lower copy number origins of replications should be sought.

The toolbox could be used in the future to test inducible promoters that may allow the pathway to be switched on after the cells have completed growth stages or the testing of orthogonal elements.

The potential of *Geobacillus thermoglucosidans* should be explored both for the production of valuable chemical compounds, as long as they are volatile so they are easy to separate by gas stripping, or as an alternative, for the production of thermostable functional biocatalysts.

Having a thermophile producing thermophilic enzymes could mean that enzymes won't lose the thermostable properties that confer them greater stability over mesophilic counterparts as long as such enzymes offer an advantage to the host. For example, if *G. thermoglucosidans* is used for the production of xylanases and if the only carbon source available is biomass, then the xylanases obtained will be functional because the survival of *Geobacillus* depends on them.

G. thermoglucosidans is being proposed as a CBP organism (Consolidated BioProcessing), during which hydrolysis and fermentation are all combined to occur within the same environment to reduce costs and increase throughput [Alper and Stephanopoulos,

2009]. The toolbox could be used to express lignocellulosic biomass degrading enzymes that *G. thermoglucosidans* does not currently possess or it could be used to engineer a variety of strains of *G. thermoglucosidans*, each with special properties for a Consortium BioProcessing approach. Or alternatively, *G. thermoglucosidans* could be a participant in a Consortium BioProcess together with other lignocellulose degrading microorganisms like *Clostridium thermocellum*.

7.12. ISOBUTANOL PRODUCTION BY *G. THERMOGLUCOSIDANS*

The production of isobutanol by *G. thermoglucosidans* in this work was a proof-of-concept application that illustrates that the ability to write genetic code to demonstrate the usefulness of engineering concepts in biology to progress the field towards the production of useful compounds.

The advantage of using *G. thermoglucosidans* in particular for the production of ethanol is that ethanol is a volatile product and it can readily be separated from the growth medium by gas stripping. In the future, the production of other volatile compounds should be contemplated, such as the production of volatile isoprenoids which are naturally produced by plants. Isoprenoids can be used as insect attractants and repellents, or as pharmaceuticals, flavors, fragrances, fuels or fuel additives [Vickers *et al.* 2014]. In particular, squalene, isopentenol, bisabolene could be used as fuel replacements. The isobutanol produced as part of this project demonstrates that the toolbox can be readily used for heterologous gene expression and it should be a straightforward process to express a bisabolene synthase for example. Volatile isoprenoids should be as easy to separate from the growth mixture as ethanol.

The unique capabilities *G. thermoglucosidans* has to digest lignocellulosic biomass have the potential to make it even more powerful for biotechnology and biotechnology is powerful to the economy.

8. CONCLUSIONS

The merits of this work are:

1. The creation of a new set of shuttle vectors that follow the synthetic biology principles of modularization and componentization. The new plasmids are more efficient than all previously existing shuttle vectors for *G. thermoglucosidans*.
2. The new plasmids can be used for the overexpression of genes and for the quantification of promoter activities.
3. The development of the toolbox and the demonstration of the use of computer aided design programs to predict synthetic parts should lower development times for research as well as costs in the future.
4. The production of isobutanol as a proof-of-principle application of the toolbox resulted in isobutanol titres of about 50 mg/L.
5. This work demonstrates that organic chemistry can be joined with synthetic biology. The precursor for isobutanol (ketoisovalerate) was added to the growth medium instead of overexpressing the ketoisovalerate pathway and resulted in an increased production of isobutanol.
6. This work confirmed that the ALDH proteins are capable of producing isobutyraldehyde from isobutyryl-CoA, which is knowledge that has only been suggested in literature but had not been experimentally confirmed until this work.
7. A native pathway for isobutanol production was discovered in *G. thermoglucosidans*.
8. This is the second report on the production of isobutanol in a thermophile and the first report of an alternative pathway for the biosynthesis of isobutanol in a thermophile which does not rely on the expression of Kdc.
9. The toolbox has the capability to improve the quality and the speed of the development, to its full potential, of *G. thermoglucosidans* as a chassis for synthetic biology. For instance, the tools created as part of this work have the potential to accelerate undergoing research to implement *G. thermoglucosidans* as the microorganism of choice for a Consolidated Bioprocessing approach to produce ethanol from lignocellulose because minimal vectors offer higher probabilities of introducing large or multiple genes into this thermophile.

9. REFERENCES

1. **Akbar, S., Gaidenko, T. A., Kang, M. C., O'Reilly, M., Devine, K. M. And Price, C. W.** (2001). New family of regulators in the environmental signalling pathway which activates the general stress transcription factor σ_B of *Bacillus subtilis*. *Journal of Bacteriology* 183:4. Pp. 1329-1338.
2. **Alper, H. and Stephanopoulos, G.** (2009). Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential?. *Nature Reviews: Microbiology*. 715-723.
3. **Altun, S., Oner, C., Yasar, F. and Adin, H.** (2011). Effect of n-Butanol Blending with a blend of diesel and biodiesel on performance and exhaust emissions of a diesel engine. *Industrial and Engineering Chemistry Research* 50: 9425-9430
4. **Antoni, D., Zverlov, V. V., Schwarz, W. H.** (2007). Biofuels from microbes. *Applied Microbiology and Biotechnology* 77: 1. Pp. 23.35.
5. **Aravali, R. N. and Garrett, R. A.** (1997). Shuttle vectors for hyperthermophilic archaea. *Extremophiles* 1. Pp. 183-191.
6. **Atsumi, S. and Liao J. C.** (2008). Metabolic engineering for advanced biofuels production from *Escherichia coli*. *Current Opinion in Biotechnology* 19: 414-419
7. **Atsumi, S. And Liao, J. C.** (2008). Directed evolution of Methanococcus jannaschii citramalate synthase for biosynthesis of 1-propanol and 1-butanol by *Escherichia coli*. *Applied Environmental Microbiology* 74. Pp. 7802-7808.
8. **Atsumi, S., Hanai, T. and Liao, J. C.** (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature Letters* 451: 86-90.
9. **Atsumi, S., Li, Z., Liao, J. C.** (2009). Acetolactate synthase from *Bacillus subtilis* serves as a 2-ketoisovalerate decarboxylase for isobutanol biosynthesis in *E. coli*. *Applied and Environmental Microbiology* 75 (19). Pp. 6306-6311.
10. **Atsumi, S., Wu, T.Y., Eckl, E.M., Hawkins, S.D., Buelter, T and Liao, J.** (2010). Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes. *Applied Microbiology and Biotechnology* 85: 651-657.
11. **Baldwin, G., Bayer, T., Dickinson, R., Ellis, T., Freemont, P. S., Kitney, R. I., Polizzi, K. and Stan G.B.** (2012). Synthetic biology: a primer. UK. Imperial College Press. Pp. 73-78.
12. **Balogh, G., Horvath, I., Nagy, E., Hoyk, Z., Benko, S., Benaude, O., Vigh, L.** (2005). The hyperfluidization of mammalian cell membranes acts as a signal to initiate the heat shock protein response. *FEBS Journal* 272. Pp. 6077-6086.
13. **Bartosiak-Jentys, J.** (2009). Metabolic Engineering and Metabolic Flux Analysis of Thermophilic, Ethanologenic *Geobacillus* spp. PhD thesis. Imperial College London.
14. **Bartosiak-Jentys, J. Eley, K. and Leak, D. J.** (2012). Application of pheB as a reporter gene for *Geobacillus* spp., enabling qualitative colony screening and quantitative analysis of promoter strength. *Applied and Environmental Microbiology* 78:16. Pp. 5945-5947.
15. **Bartosiak-Jentys, J., Hussein, A., Lewis, C. J. And Leak, D. J.** (2013). A modular system for assessment of glycosyl hydrolase secretion in *Geobacillus thermoglucosidasius*. *Microbiology* 159:7. Pp. 1267-1275.
16. **Bastian, S., Liu, X., Meyerowitz, J. T., Snow, C. D., Chen, M. M. Y. and Arnold, F. H.** (2011) Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in *Escherichia coli*. *Metabolic Engineering* 13: 345-352

17. **Bern, M. and Goldberg, D.** (2005). Automatic selection of representative proteins for bacterial phylogeny, *BMC Evolutionary Biology* 5:34.
18. **Blanchard, K., Robic, S. and Matsumura I.** (2014). Transformable facultative thermophile *Geobacillus stearothermophilus* NUB3621 as a host strain for metabolic engineering. *Applied Microbiology and Biotechnology*
19. **Blombach, B. and Eikmanns, B. J.** (2011) Current knowledge on isobutanol production with *Escherichia coli*, *Bacillus subtilis* and *Corynebacterium glutamicum*. *Bioengineered Bugs* 2 (6): 346-350
20. **Bruno, T., J., Wolk, A. and Naydich, A.** (2009). Composition-explicit distillation curves for mixtures of gasoline with four-carbon alcohols (butanols). *Energy and Fuels* 23: 2295-2306
21. **Brynildsen, M. P. and Liao, J. C.** (2009). An integrated network approach identifies the isobutanol response network of *Escherichia coli*. *Molecular Systems Biology* 5: 27.
22. **Cabantous, S., Rogers, Y., Terwilliger, T. C. And Waldo, G. S.** (2008). New molecular reporters for rapid protein folding assays. *PLoS ONE* 3:6. e2387
23. **Cannio, R., Contursi, P., Rossi, M. and Bartolucci, S.** (2001). Thermoadaptation of a mesophilic hygromycin B phosphotransferase by directed evolution in hyperthermophilic Archaea: selection of a stable genetic marker for DNA transfer into *Sulfolobus solfataricus*. *Extremophiles* 5:3. Pp. 153-159.
24. **Carere, C. R., Sparling, R., Cicek, N. and Levin, D. B.** (2008). Third generation biofuels via direct cellulose fermentation. *International Journal of Molecular Sciences*. 9:7. Pp. 1342-1360.
25. **Cava, F., de Pedro, M. A., Blas-Galindo, E., Waldo, G. S., Westblade, L. F., Berenguer, J.** (2008). Expression and use of superfolder green fluorescent protein at high temperatures in vivo: a tool to study extreme thermophile biology. *Environmental Microbiology* 10(3): 605-613.
26. **Cava, F., de Pedro, M. A., Blas-Galindo, E., Waldo, G. S., Westblade, L. F., Berenguer, J.** (2008). Expression and use of superfolder green fluorescent protein at high temperatures in vivo: a tool to study extreme thermophile biology. *Environmental Microbiology* 10(3): 605-613.
27. **Chang, D.E., Jung, H.C., Rhee, J. S. and Pan, J. G.** (1999). Homofermentative production of D-or L-lactate in metabolically engineered *Escherichia coli* RR1. *Applied and Environmental Microbiology* 65:4: Pp. 1384-1389.
28. **Chang, Y. C., Lee, W. J., Yang, H. H., Wang, L. C., Lu, J. H., Tsai, Y. I., Cheng, M. T., Young, L. H., Chiang, C. J.** (2014). Reducing emissions of persistent organic pollutants from a diesel engine by fueling with water-containing butanol diesel blends. *Environmental Science and Technology* 48:10. Pp. 6010-6018.
29. **Chen, G. and Inouye, M.** (1990). Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. *Nucleic Acids Research* 18, 1465-1473.
30. **Chen, G. and Inouye, M.** (1994). Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*. *Genes and Development* 8:2641-2652.
31. **Chen, H., Bjerknes, M., Kumar, R. and Jay, E.** (1994). Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Research* 22, 4953-4957.
32. **Chen, X., Nielsen, K. F., Borodina, I., Kielland-Brandt, C. and Karhumaa** (2011). Increased isobutanol production in *Saccharomyces cerevisiae* by overexpression of genes in valine metabolism. *Biotechnology for Biofuels* 4: 21. Pp. 1-12.
33. **Chung, C. T. and Miller, R. H.** (1993). Preparation and storage of competent *Escherichia coli* cells. *Methods in Enzymology* 218: 621-627.

34. **Chung, C. T., Niemela, S. L. and Miller, R. H.** (1989). One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution. *Proceedings of the National Academy of Sciences* 86(7): 2172-2175.
35. **Clomburg, J. and Gonzalez, R.** (2010). Biofuel production in *Escherichia coli*: the role of metabolic engineering and synthetic biology. *Applied Microbiology and Biotechnology* 86:2. Pp. 419-434.
36. **Connor, M. R. and Liao, J. C.** (2009). Microbial production of advanced transportation fuels in non-natural hosts. *Current Opinion in Biotechnology* 20: 307-315
37. **Coorevits, A., Dinsdale, A. E., Halket, G., Lebbe, L., De Vos, P., Van Landschoot, A. and Logan, N. A.** (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 thermantarticus* to the genus as *G. thermantarticus*; proposal of *Caldibacillus debilis* gen. nov., comb. nov., transfer of *G. tepidamans* to *Anoxybacillus* as *A. tepidamans* and proposal of *Anoxybacillus caldiproteolyticus* sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 62. Pp. 1470-1485.
38. **Cripps, R. E., Eley, K., Leak, D. J., Rudd, B., Taylor, M., Todd, M., Boakes, S., Martin, S. and Atkinson, T.** (2009). Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. *Metabolic Engineering* 11: 398-408.
39. **Crowhurst, N.** (2010). Exploiting the anaerobic expression of pyruvate dehydrogenase for the production of biofuels. PhD thesis. Imperial College London.
40. **Daron, H. H.** (1970). Fatty acid composition of lipid extracts of a thermophilic *Bacillus* species. *Journal of Bacteriology* 101:1. Pp. 145-151.
41. **Davidson, Anthony** (2009). Development of a thermostable Pyruvate Decarboxylase. MRes second rotation project. Imperial College London.
42. **Dickinson, R. J., Harrison, S. J. and Hewlins, M. J. E.** (1998). An investigation of the metabolism of valine to isobutyl alcohol in *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry* 273 (40): 25751-25756
43. **Donaldson, G. K., Eliot, A. C., Flint, D., Maggio-Hall, A., Nagaraja, V.** (2010). Fermentative production of four carbon alcohols. United States Patent No. US 7,851,188 B2
44. **Dong, H., Nilsson, L., Kurland, C. G.** (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *Journal of Bacteriology* 177:6. Pp. 1497-1504.
45. **Drepper, T., Eggert, T., Circolone, F., Heck, A., Krauss, U., Guterl, J-K., Wendorff, M., Losi, A., Gartner, W. And Jaeger, K-E.** (2007). Reporter proteins for in vivo fluorescence without oxygen. *Nature Biotechnology* 25(4): 443-445.
46. **Durfee, T., Nelson, R., Baldwin, S. et al.** (2008). The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse. *Journal of Bacteriology* 190:7. Pp. 2597-2606
47. **Ellis, T., Wang, X. and Collins, J. J.** (2009). Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotechnology* 27:5. Pp. 465-471.
48. **Fischer, C, Klein-Marcuschamer and Stephanopoulos, G.** (2008). Selection and optimization of microbial hosts for biofuels production. *Metabolic Engineering* 10:6. Pp. 295-304.
49. **Fong, J. C. N., Svenson, C. J., Nakasugi, K., Leong, C. T. C. , Bowman, J. P., Chen, B., Glenn, D. R., Neilan, B. A. and Rogers, P. L.** (2006). Isolation and

- characterization of two novel ethanol-tolerant facultative-anaerobic thermophilic bacteria strains from waste compost. *Extremophiles* 10. Pp. 363-372.
50. **French, C.** (2009). Synthetic biology and biomass conversion: a match made in heaven? *Journal of the Royal Society Interface* 6:4. Pp. S547-558.
 51. **Gaidenko, T. A., Kim, T. J., Weigel, A. L., Brody, M. S. And Price, C. W.** (2006). The blue-light receptor YtvA acts in the environmental stress signalling pathway of *Bacillus subtilis*. *Journal of Bacteriology* 188:17. Pp. 6387-6395.
 52. **Galanie, S., Siddiqui, M. S. and Smolke, C. D.** (2013). Molecular tools for chemical biotechnology. *Current Opinion in Biotechnology* 24(6). Pp. 1000-1009.
 53. **Georgieva TI, Mikkelsen MJ, Ahring BK** (2008). Ethanol production from wet-exploded wheat straw hydrolysate by thermophilic anaerobic bacterium *Thermoanaerobacter* BG1L1 in a continuous immobilized reactor. *Appl Biochem Biotechnol* 145: 99-110. **Ghetti, F., Checcucci, G. And Lenci, F.** (1992). Photosensitized reactions as primary molecular events in photomovements of microorganisms. *Journal of Photochemical Photobiology B: Biology* 15. Pp. 185-198.
 54. **Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchinson III, C. A. and Smith, H. O.** (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* 6(5): 343-345.
 55. **Gilbert, H. J. and Hazlewood, G. P.** (1993). Bacterial cellulases and xylanases. *Journal of General Microbiology* 139, 187-194.
 56. **Gocke, D., Nguyen, C. L., Pohl, M., Stillger, T., Walter, L. and Müller, M.** (2007). Branched-chain keto acid decarboxylase from *Lactococcus lactis* (KdcA), a valuable thiamine diphosphate-dependent enzyme for asymmetric C-C bond formation. *Advanced Synthetic Catalysis* 349. Pp. 1425-1435.
 57. **Gryczan, T. J., Contente, S. and Dubnau, D.** (1978). Characterization of *Staphylococcus aureus* plasmids introduced by transformation into *Bacillus subtilis*. *Journal of Bacteriology* 134. Pp. 318-329.
 58. **Hanahan, D.** (1983). Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology* 166: 557-580.
 59. **Hanai, T., Atsumi, S. and Liao, J. C.** (2007). Engineered synthetic pathway for isopropanol production in *Escherichia coli*. *Applied Environmental Microbiology* 73: 7814-7818.
 60. **Heinemann, M. and Panke, S.** (2006). Synthetic biology -putting engineering into biology. *Bioinformatics Oxford England* 22:22. Pp. 2790-2799.
 61. **Henstra, A. M. and Stams, A. J. M.** (2011). Deep conversion of carbon monoxide to hydrogen and formation of acetate by the anaerobic thermophile *Carboxydotherrmus hydrogenoformans*. *International Journal of Microbiology* ID 641582.
 62. **Higashide, W., Li, Y., Yang, Y., Liao, J.** (2011). Metabolic engineering of *Clostridium cellulolyticum* for isobutanol production from cellulose. *Applied Environmental Microbiology*. In Press
 63. **Hirata, H., Fukazawa, T., Negoro, S. and Okada, H.** (1986). Structure of a β -galactosidase gene of *Bacillus stearothermophilus*. *Journal of Bacteriology* 166:3. Pp. 722-727.
 64. **Holtz, W. and Keasling, J.** (2010). Engineering static and dynamic control of synthetic pathways. *Cell* 140:1. Pp. 19-23.
 65. **Howard, T. P., Middelaufe, S., Moore, K., Edner, C., Kolak, D. M., Taylor, G. N., Parker, D. A., Lee, R., Smirnoff, N., Aves, S. J. and Love, J.** (2013). Synthesis of customized petroleum-replica fuel molecules by targeted modification of free fatty acid pools in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 110:19. Pp. 7636-7641.
 66. <http://blast.ncbi.nlm.nih.gov/>

67. http://ec.europa.eu/energy/renewables/biofuels/biofuels_en.htm
68. <http://tmcalculator.neb.com/#/>
69. http://www.chathamhouse.org/sites/default/files/public/Research/Energy,%20Environment%20and%20Development/0413pp_biofuels.pdf
70. <http://www.fermentas.com/en/products/all/dna-electrophoresis/generuler-dna-ladders/sm031-generuler-1kb>
71. http://www.genome.jp/kegg-bin/show_pathway?gth00290
72. <http://www.kazusa.or.jp/codon/>
73. http://www.loesemittel.basf.com/portal/load/fid229183/Technical%20Spec%20-%20Isobutanol_BPC.pdf
74. <http://www.rnasoft.ca/cgi-bin/RNASoft/AveRNA/AveRNA.pl>
75. <http://www.utoronto.ca/virology/labwork/gram/p4.gif>
76. <https://salis.psu.edu/software/RBSLibraryCalculatorSearchMode>
77. <http://eu.idtdna.com/analyzer/applications/oligoanalyzer/>
78. **Huala, E., Oeller, P. W., Liscum, E., Han, I. S., Larsen, E. and Briggs, W. R.** (1997). *Arabidopsis* NPH1: a protein kinase with a putative redox-sensing domain. *Science* 278. Pp. 2120-2123.
79. **Inokuma, K., Liao, J. C., Okamoto, M., and Hanai, T.** (2010). Improvement of isopropanol production by metabolically engineered *Escherichia coli* using gas stripping. *Journal of Bioscience and Bioengineering* 110:6. Pp. 696-701.
80. **Inui, M., Suda, M., Kimura, S., Yasuda, K., Suzuki, H., Toda, H., Yamamoto, S., Okino, S., Suzuki, N. And Yukawa, H.** (2008). Expression of *Clostridium acetobutylicum* butanol synthetic genes in *E. coli*. *Applied Genetics and Molecular Biotechnology* 77. Pp. 1305-1316.
81. **Jaenicke, R. and Böhm, G.** (1998). The stability of proteins in extreme environments. *Current Opinion in Structural Biology* 8:6. Pp. 738-748.
82. **Jia, X., Li, S., Xie, S. and Wen, J.** (2011). Engineering a metabolic pathway for isobutanol biosynthesis in *Bacillus subtilis*.
83. **Jin, H., Chen, L., Wang, J. and Zhang, W.** (2014). Engineering biofuel tolerance in non-native producing microorganisms. *Biotechnology Advances* 32:2. Pp. 541-548.
84. **Kaminski, W., Tomczak, E. And Gorak, A.** (2011). Biobutanol: production and purification methods. *Ecological Chemistry and Engineering* 18(1): 31-37
85. **Kananavičiūtė, R. and Čitavičius, D.** (2015). Genetic engineering of *Geobacillus* spp. *Journal of Microbiological Methods* 111. Pp. 31-39.
86. **Karabektas, M. and Hosoz, M.** (2009). Performance and emission characteristics of a diesel engine using isobutanol-diesel blends. *Renewable Energy* 34: 1554-1559
87. **Keasling, J. D.** (2008). Synthetic biology for synthetic chemistry. *ACS Chemical Biology* 3:1- Pp. 64-76.
88. **Keasling, J. D.** (2012). Synthetic biology and the development of tools for metabolic engineering. *Metabolic Engineering* 14:3. Pp. 189-195.
89. **Kelly, J. R., Rubin, A. J., Davis, J. H., Ajo-Franklin, C. M., Cumbers, J., Czar, M. J., de Mora, K., Gliberman, A., Monie, D. D. And Endy, D.** (2009). Measuring the activity of BioBrick promoters using an in vivo reference standard. *Journal of Biological Engineering* 3:4. Pp. 1-13
90. **Kingsford, C. L., Ayanbule, K. and Salzberg, S.** (2007). Rapid, accurate computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology* 8: R22.
91. **Kittleson, J. T., Wu, C. G., Anderson, C. J.** (2012). Successes and failures in modular genetic engineering. *Current Opinion in Chemical Biology* 16:3-4. Pp. 329-336.

92. **Kizer, L., Pitera, D. J., Pfleger, D. F. and Keasling, D. J.** (2008). Application of functional genomics to pathway optimization for increased isoprenoid production. *Applied and Environmental Microbiology* 74:10. Pp. 3229-3241.
93. **Kosuri, S., Goodman, D. B., Camray, G., Mutalik, V., Gao, Y., Arkin, A. P., Endy, D., Church, G. M.** (2013). Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 110:34. Pp. 14024-14029.
94. **Krauss, U.** (2007). Bacterial blue-light photoreceptors of the LOV family. Doctoral dissertation. Heinrich-Heine Universität Düsseldorf.
95. **Krauss, U., Losi, A., Gärtner, W., Jaeger, K. E. And Eggert, T.** (2005). Initial characterization of a blue-light sensing, phototropin-related protein from *Pseudomonas putida*: a paradigm for an extended LOV construct. *Physical Chemistry Chemical Physics* 14. Pp. 2804-2811.
96. **Kuk Lee, S., Cho, H., Ham, T. S., Lee, T. S., Keasling J. D.** (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology* 19:6. Pp. 556-563.
97. **Kung, Y., Rungtaphan, W. And Keasling, J. D.** (2012). From fields to fuels: recent advances in the microbial production of biofuels. *ACS Synthetic Biology* 1:11. Pp. 498-513.
98. **Lamsen, E. And Atsumi, S.** (2012). Recent progress in synthetic biology for microbial production of C3-C10 alcohols. *Frontiers in Microbiology* 3: 196. Pp.
99. **Lawther, R., Wek, R. C., Lopes, J. M., Pereira, R., Taillon, B. E. and Hatfield, G. W.** (1987). The complete nucleotide sequence of the ilvGMEDA operon of *Escherichia coli* K-12. *Nucleic Acids Research* 15 (5): 2137-2155
100. **Leonardo, M. R., Dailly, Y. and Clark, D. P.** (1996). Role of NAD in regulating the adhE gene of *Escherichia coli*. *Journal of Bacteriology* 178:20. Pp. 6013-6018.
101. **Li, S., Wen, J. and Jia, X.** (2011). Engineering *Bacillus subtilis* for isobutanol production by heterologous Ehrlich pathway construction and the biosynthetic 2-ketoisovalerate precursor pathway overexpression. *Applied Microbial Biotechnology* 91: 577-589
102. **Liao, H. H. and Kanukula, A. M.** (1990). Increased efficiency of transformation of *Bacillus stearothermophilus* by a plasmid carrying a thermostable kanamycin resistance marker. *Current Microbiology* 21: 301-306.
103. **Lin L. et al.** (2013). Microevolution from shock to adaptation revealed strategies improving ethanol tolerance and production in *Thermoanaerobacter*. *Biotechnology for biofuels* 6:103.
104. **Lin, L. and Xu, J.** (2013). Dissecting and engineering metabolic and regulatory networks of thermophilic bacteria for biofuel production. *Biotechnology Advances* 31. Pp. 827-837.
105. **Liu, S., Skinner-Nemec, K. A., Leathers, T. D.** (2008). *Lactobacillus buchneri* strain NRRL B-30929 converts a concentrated mixture of xylose and glucose into ethanol and other products. *Journal of Industrial Microbiology and Biotechnology* 35. Pp. 75-81.
106. **Livshits et al.** (2004). United States Patent 6,737,255 B2
107. **Lobo, L. A., Smith, C. J. And Rocha, E. R.** (2011). Flavin mononucleotide (FMN)-based fluorescent protein (FbFP) as reporter for gene expression in the anaerobe *Bacteroides fragilis*. *FEMS Microbiology Letters* 317:1. Pp. 67-74.
108. **Losi, A.** Flavin-based photoreceptors in bacteria. (2006). In *Flavins: photochemistry and photobiology* Royal Society of Chemistry Publishing, UK. Pp. 217-269.

109. **Madigan, M. T., Martinko, J. M., Dunlap, P. V. And Clark, D. P.** (2009). Brock Biology of Microorganisms. USA. Pearson Education. Pp. 446-477.
110. **Maheswar Rao, J. L. U. and Satyanarayana, T.** (2009). Hyperthermostable, Ca²⁺-independent, and high maltose-forming α -amylase production by an extreme thermophile *geobacillus thermoleovorans*: whole cell immobilization. *Applied Biochemistry and Biotechnology* 159: 464-477.
111. **Marchler-Bauer, A.** (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", *Nucleic Acids Research* 39. Pp. D225-D229.
112. **Martínez-García, E., Aparicio, T., Goñi-Moreno, A., Fraile, S. and de Lorenzo, V.** (2014). SEVA 2.0: an update of the Standard European Vector Architecture for de-/re-construction of bacterial functionalities. *Nucleic Acids Research*. Pp. 1-7.
113. **Martinez-Klimova, E.** (2011). Synthetic biology approaches to the metabolic engineering of *G. thermoglucosidasius* for isobutanol production. MPhil transfer report. Imperial College London.
114. **Melcher, Ulrich** (2000). <http://bioinfosu.okstate.edu/MG/MGW1/MG1337.html>
115. **Nazina, T. N., Sokolova, D. S., Grigoryan, A. A., Shestakova, N. M., Mikhailova, E. M., Poltarau, A. B., Tourova, T. P., Lysenko, A. M., Osipov, G. A., Belyaev, S.** (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. Pp. 43-53.
116. **Nazina, T. N., Tourova, T. P., Poltarau, A. B., Novikova, E. V., Grigoryan, A. A., Ivanova, A. E., Lysenko, A. M., Petrunyaka, V. V., Osipov, G. A., Belyaev, S. S., and Ivanov, M. V.** (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 *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. *International Journal of Systematic and Evolutionary Microbiology* 51: 433-446.
117. **Nelson, D. L. and Cox, M. M.** (2005). Lehninger Principles of Biochemistry. W. H. Freeman, USA. Pp. 1082-1111.
118. **Nozzi, N. E., Desai, S. H., Case, A. E. And Atsumi, S.** (2014). Metabolic engineering for higher alcohol production. *Metabolic engineering* 25. Pp. 174-182.
119. **Osburn, O. L., Brown, R. W., Werkman, C. H.** (1937). The butyl alcohol-isopropyl alcohol fermentation. *The Journal of Biological Chemistry* 121. Pp. 685-695.
120. **Ozcelik, I. L., Calik, P., Calik, G., Ozdamar, T. H.** (2004). Metabolic engineering of aromatic group amino acid pathway in *Bacillus subtilis* for L-phenylalanine production. *Chemical Engineering Science* 59:22-23. Pp. 5019-5026.
121. **Park, J. H. and Lee, S. Y.** (2010). Fermentative production of branched chain amino acids: a focus on metabolic engineering. *Applied Microbial Biotechnology* 85: 491-506
122. **Park, J. H., Lee, K. H., Kim, T. Y. and Lee, S. Y.** (2007). Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *PNAS* 104: 19. 7797-7802
123. **Pédelacq, J-D., Cabantous, S., Tran, T., Terwillinger T. C., Waldo, G. S.** (2006). Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology* 24(1): 79-88.

124. **Peralta-Yahya, P. P., Ouellet, M., Chan, R., Mukhopadhyay, A., Keasling, J. D. and Soon Lee, T.** (2001). Identification and microbial production of a terpene-based advanced biofuel. *Nature communications* 2. Pp. 483-
125. **Pfleger, B. F., Pitera, D. J., Smolke, C. D. and Keasling, J. D.** (2006). Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nature Biotechnology* 24:8. Pp. 1027-1032.
126. **Philippot, L., Andersson, S. G. E., Battin, T. J., Prosser, J. I., Schimel, J. P., Whitman, W. B. and Hallin, S.** (2010). The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews. Microbiology* 8:7. Pp. 523-529.
127. **Platko, J. V., Willins, D. A. and Calvo, J. M.** (1990). The *ilvH* operon of *Escherichia coli* is positively regulated. *Journal of Bacteriology* 172 (8): 4563-4570
128. **Platko, J. V., Willins, D. A., Calvo, J. M. and Ryan, C. W.** (1991). Characterization of Lrp, an *Escherichia coli* regulatory protein that mediates a global response to leucine. *The Journal of Biological Chemistry* 266 (17): 10768-10774
129. **Plaza, M., Palencia, P. F., Peláez, C. and Requena, T.** (2004). Biochemical and molecular characterization of α -ketoisovalerate decarboxylase, an enzyme involved in the formation of aldehydes from amino acids by *Lactococcus lactis*. *FEMS Microbiology Letters* 238: 367-374
130. **Porat, I., Vinogradov, M., Vyazmensky, M., Lu, CD., Chipamn, D. M., Abdelal, A. T., Barak, Z.** (2004). Cloning and characterization of acetohydroxyacid synthase from *Bacillus stearothermophilus*. *Journal of Bacteriology* 186 (2): 570-574.
131. Protein calculator v3.4: <http://protcalc.sourceforge.net/>
132. Protein Weight Calculator: <http://protcalc.sourceforge.net/cgi-bin/protcalc>
133. PubChem, NCBI:
<http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=473>
134. **Ramey, D. E.** Butanol: the other alternative fuel. In Eaglesham, A. and Hardy, R. W. F. (2007). *Agricultural biofuels: technology, sustainability and profitability*. Pp. 137-147.
135. **Reeve, B.** (2010). Promoter Regulation in *Geobacillus thermoglucosidasius*. 9-month progress report. Imperial College London. 21 pp.
136. **Reeve, B.** (2013). Promoter Synthesis and Regulation in a Biofuel-Producing Thermophile. Poster presented at the Synthetic Biology 6 conference at Imperial College London.
137. **Reynolds, R., Bermudez-Cruz, R. M., Chamberlin, M. J.** (1992). Parameters affecting transcription termination by *Escherichia coli* RNA polymerase: analysis of 13 rho-independent terminators. *Journal of Molecular Biology* 224:1. Pp. 31-51.
138. **Roelof van der Meer, J. and Belkin, S.** (2010). Where microbiology meets microengineering: design and applications of reporter bacteria. *Nature Reviews Microbiology* 8:7. Pp. 511-522.
139. **Rubin, E. M.** (2008). Genomics of cellulosic biofuels. *Nature* 454:7206. Pp. 841-845.
140. **Salis, H. M.** (2011) The Ribosome Binding Site Calculator. In *Methods in Enzymology* 498, *Synthetic Biology Part B*, edited by Christopher Voigt. Elsevier. USA. Pp. 19-42.
141. Salis, H. M., Mirsky, E. A. And Voigt, C. A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27:10. Pp. 946-950.
142. **Salomon, M. J., Christie, M., Knieb, E., Lempert, U. And Briggs, W. R.** (2000). Photochemical and mutational analysis of the FMN-binding domains of the plant blue-light receptor phototropin. *Biochemistry* 39. Pp. 9401-9410.

143. **Savrasova, E. A., Kivero, A. D., Shakulov, R. S. and Stoyanova N. V.** (2011). Use of the valine biosynthetic pathway to convert glucose into isobutanol. *Journal of Industrial Microbiology and Biotechnology* 38: 1287-1294.
144. **Shaner, N. C., Campbell, R. E., Steinbach, P. A. Giepmans, B. N. G., Palmer, A. E. and Tsien, R. Y.** (2004). Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein.
145. **Shi, D., Wang, C., Wang, K.** (2009). Genome shuffling to improve thermotolerance, ethanol tolerance and ethanol productivity of *Saccharomyces cerevisiae*. *Journal of Industrial Microbiology and Biotechnology* 36. Pp. 139-147.
146. **Shin, J. H., Kim, H. U., Kim, D. I. and Lee, S. Y.** (2013). Production of bulk chemicals via novel metabolic pathways in microorganisms. *Biotechnology Advances* 31:6. Pp. 925-935.
147. **Silva-Rocha, R., Martínez-García, E., Calles, B., Chavarría, M., Arce-Rodriguez, A., de las Heras, A., Páez-Espino, D., Durante-Rodríguez, G., Kim, J., Nikel, P. I., Platero, R. and de Lorenzo, V.** (2013). The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Research* 41. Pp. D666-D675.
148. **Singh, V.** (2014). Recent advances in synthetic biology: current status and challenges. *Gene* 535:1. Pp. 1-11.
149. **Smit, B. A., Vlieg, J. E. T., Engels, W. J. M., Meijer, L., Wouters and G. Smit** (2005). Identification, cloning and characterization of a *Lactococcus lactis* branched-chain α -keto acid decarboxylase involved in flavor formation. *Applied and Environmental Microbiology* 71 (1): 303-311
150. **Smith, K. M., Cho, K-M., Liao, J.** (2010). Engineering *Corynebacterium glutamicum* for isobutanol production. *Applied Microbial Biotechnology* 87: 1045-105.
151. **Solovyev, V. and Salamov A.** (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies* (Ed. R.W. Li), Nova Science Publishers, p. 61-78
http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfin_db
152. **Sonderegger, M., Jeppsson, M., Larsson, C., Gorwa-Grauslund, M.F., Boles, E., Olsson, L., Spencer-Martins, I.** (2004). Fermentation performance of engineered and evolved xylose-fermenting *Saccharomyces cerevisiae* strains. *Biotechnology and Bioengineering* 87:1. Pp. 90-98.
153. **Song, X., Wang, Y., Shu, Z., Hong, J., Li, T., and Yao, L.** (2013). Engineering a more thermostable blue light photoreceptor *Bacillus subtilis* YtvA LOV domain by a computer aided rational design method. *PLOS Computational Biology* 9:7. Pp. e1003129
154. **Steen, E. J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., Cardayre S., and Keasling, J. D.** (2010). Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 463: 7280. Pp. 559-562.
155. **Stephan, B. and Gabriele, K.** (2004). Blue-light perception in Bacteria. *Photosynthesis Research* 79. Pp. 45-57.
156. **Stephanopoulos, G.** (2007). Challenges in engineering microbes for biofuels production. *Science* 315:5813. Pp. 801-804.
157. **Studier, F. W.** (2005). Protein production by auto-induction in high-density shaking cultures. *Protein expression and Purification* 41: 207-234.
158. **Sullivan, M. A., Yasbin, R. E., and Young, F. E.** (1984). New shuttle vectors for *Bacillus subtilis* and *Escherichia coli* which allow rapid detection of inserted fragments. *Gene* 29. Pp. 21-26.

159. **Suzuki, H., Murakami, A. and Yoshida, K.** (2012). Counterselection system for *Geobacillus kaustophilus* HTA426 through disruption of *pyrF* and *pyrR*. *Applied and Environmental Microbiology* 78:20. Pp. 7376-7383.
160. **Suzuki, H., Yoshida, K. and Oshima, T.** (2013). Polysaccharide-degrading thermophiles generated by heterologous gene expression in *Geobacillus kaustophilus* HTA426. *Applied Environmental Microbiology* 79:17. Pp. 5151-5158.
161. **Suzuki, H., Kobayashi, J., Wada, K., Furukawa, M., Doi, K.** (2015). Thermoadaptation-directed enzyme evolution in an error-prone thermophile derived from *Geobacillus kaustophilus* HTA426. *Applied and Environmental Microbiology* 81:1. Pp. 149-158.
162. **Szostkova, K. and Horakova, D.** (1998). The effect of plasmid DNA sizes and other factors on electrotransformation of *Escherichia coli* JM109. *Bioelectrochemistry and Bioenergetics* 47: 319-323
163. **Takami, H., Nishi, S., Lu, J., Shimamura, S. and Takaki, Y.** (2004) Genomic characterization of thermophilic *Geobacillus* species isolated from the deepest sea mud of the Mariana Trench. *Extremophiles* 8: 351-356
164. **Taylor M. P., Esteban, C. D. and Leak, D. J.** (2008). Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. *Plasmid* 60: 45-52.
165. **Taylor, M. P.** (2007). Metabolic engineering of *Geobacillus* species for enhanced ethanol production. PhD Thesis. Imperial College of Science, Technology and Medicine and University of London.
166. **Taylor, M. P., Eley, K. L., Martin, S., Tuffin, M. I., Burton, S. G. and Cowan, D. A.** (2009). Thermophilic ethanologenes: future prospects for second-generation bioethanol production. *Trends in biotechnology* 27 (7): 398-405
167. **Tielker, D., Eichhof, I., Jaeger, K. E. And Ernst, J. F.** (2009). FMN-based fluorescent protein as an oxygen independent reporter in *Candida albicans* and *Saccharomyces cerevisiae*. *Eukaryotic cell* 8:6. Pp. 913-915.
168. **Tracy, B. P., Jones, S. W., Fast, A. G., Indurthi, D. C., Papoutsakis, E. T.** (2012). Clostridia: the importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. *Current Opinion in Biotechnology* 23:3. Pp. 364-381.
169. **Tramier, M., Zahid, M., Mevel, J. C., Masse, M. J., Coppey-Moisan, M.** (2006). Sensitivity of CFP/YFP and GFP/mCherry pairs to donor photobleaching on FRET determination by fluorescence lifetime imaging microscopy in living cells. *Microscopy Research and Technique* 69. Pp. 933-939.
170. **Tummala, S. B., Welker, N. E. And Papoutsakis, E. T.** (1999). Development and characterization of a gene expression reporter system for *Clostridium acetobutylicum* ATCC 824. *Applied and Environmental Microbiology* 65:9. Pp. 3793-3799.
171. **Turner, S. L., Ford, G. C., Mountain, A. and Moir, A.** (1992). Selection of a thermostable variant of chloramphenicol acetyltransferase Cat-86. *Protein Engineering* 5:6. Pp. 535-541.
172. **Tyo, K. E., Alper, H. S. And Stephanopoulos G.** (2007). Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends in Biotechnology* 25:3- Pp. 132-137.
173. **Wachter, R. M.** (2006). The family of GFP-like proteins: structure, function, photophysics and biosensor applications. *Photochemistry and Photobiology* 82. Pp. 339-344.
174. **Wang, P-Z and Doi, R. H.** (1984). Overlapping promoters transcribed by *Bacillus subtilis* σ_{55} and σ_{37} RNA polymerase holoenzymes during growth and stationary phases. *The Journal of Biological Chemistry* 259:13. Pp. 8619-8625.

175. **Ward, C.** (2014). Application of metabolic flux and transcript analyses to understanding the physiology of engineered *Geobacillus thermoglucosidasius*. PhD thesis. Imperial College London.
176. **Wek, R. C. and Hatfield, G. W.** (1988). Transcriptional activation at adjacent operators in the divergent-overlapping *ilvY* and *ilvC* promoters of *Escherichia coli*. *Journal of Molecular Biology* 203: 643-663
177. **Willquist, K., Zeidan, A. A. and van Niel, E. W. J.** (2010). Physiological characteristics of the extreme thermophile *Caldicellulosiruptor saccharolyticus*: an efficient hydrogen cell factory. *Microbial Cell Factories* 9:89.
178. **Wu, T.-Y., Atsumi, S., Machado, I. MP., Huang, W.-C., Chen, P.-Y., Pellegrini, M. and Liao, J.** (2010). Evolution, genomic analysis, and reconstruction of isobutanol tolerance in *Escherichia coli*. *Molecular Systems Biology* 6: 449
179. **Xiao, Z., Wang, X., Huang, Y., Huo, F., Zhu, X., Xi, L. and Lu, J. R.** (2012). Thermophilic fermentation of acetoin and 2,3-butanediol by a novel *Geobacillus* strain. *Biotechnology for Biofuels* 5:88.
180. **Yanisch-Perron, C., J. Vieira, and J. Messing.** (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* 33(1): 103-119.
181. **Yao, S., Gao, X., Fuchsbaue, N., Hillen, W., Vater, J., Wang, J.** (2003). Cloning, sequencing and characterization of the genetic region relevant to biosynthesis of hte lipopeptides Iturin A and Surfactin in *Bacillus subtilis*. *Current Microbiology* 47:4. Pp. 272-277.
182. **Zaki, A. M., Wimalasena, T. T. and Greetham, D.** (2014). Phenotypic characterization of *Saccharomyces* spp. for tolerance to 1-butanol. *Journal of Industrial Microbiology and Biotechnology* 41:11. Pp. 1627-1636.
183. **Zhang, A., Liu, H., Yang, M. M., Gong, Y. S. and Chen H.** (2007). Assay and characterization of a strong promoter element from *B. subtilis*. *Biochemical and Biophysical Research Communications* 354. Pp. 90-95.
184. **Zhou, J. and Rudd, K. E.** (2013). EcoGene 3.0. *Nucleic Acids Research* 41(D1). D613-D624.
- 185.

CHAPTER 10: APPENDIX

Table A19. List of primers used in this study.

| | |
|----------------------------|--|
| emk001 | ggattctagataaggagtgattcgaatgcgtaaaggcgaagagctg |
| emk002 | ggatggatgaactgtacaaatgataatcgcgccccgaaagggggcggtttttgcg |
| emk003 | attacgaattcgagcttcgcaaaaaaacgccc |
| emk004 | atggctagtttcaatcattggg |
| emk005 | ttacataatcgggaagcactttaacg |
| pUC_ori-Pacl-p43 | ccaagcttgcacgctcagttaattaagtgcacgagccggggcatat |
| RC-p43-XbaI-YtvA(pheB-RBS) | agccattcgaatcactccttatctagatataatggtaccgctatcacttta |
| emk006 | ggaaatfttagaagaacgcacgctcttacagggg |
| emk007 | cccctgtaagaagcgtgcttcttctaaaatttcc |
| YtvA(pheB-RBS) | atctagataaggagtgattcgaatggctagtttcaatcattggg |
| p43-XbaI-YtvA(pheB-RBS) | taaagtgatagcggtagcattatctagataaggagtgattcgaatggct |
| RC-YtvA-rho-SacI | cgagctcgcaaaaaaacgcccccttcggggcgcgattacataatcgggaagcactttaacg |
| RC-rho-SacI-pUC_ori | gcgataacaatttcacacaggaaacagctatgaccatgattacgaattcgagctcgcaaaaacg |
| rho-SacI-pUC_ori | attatgtaatcgcgccccgaaagggggcggtttttgagctcgaattcgtaatcatgg |
| pUC_ori-PstI-Idh | gtaaacgacggccagtgccaagcttgcacgctcagggcgagcgggagctgagtgctc |
| RC-Idh-XbaI-(pheB-RBS) | cattcgaatcactccttatctagaatcctccctcaataatgccaacactttcatataaata |
| PpFbFp | cattgtctagaattcacagcccaatctatagagg |
| RC-PpFbFp | aaattgagctcttataatgatgggatgatgatgatg |
| NotI-repBST1 | attaagcggccgctgctgagaattttggaaaatagacct |
| RC-repBST1-Pmel | ccggtgttaaaactcataatcggcgctataatcgcggttta |
| Pmel-kanR | ggcccgttaaacgaagattagatgctataattgttataaaaggattg |
| RC-kanR-Ascl | aatgtggcgccttatcaaaatggatgctggtttgac |
| FseI-AmpR | attaaggccggcctgacagttaccaatgcttaacagtg |
| RC-AmpR-NotI | tttggcgccgcccctcgtagacgctatttttata |
| RC-pUC_ori-NotI | ggaaagcggccgcatggttaactgtcagaccaagtttactc |
| Ascl-SwaI-pUC_ori | ctgaaggcgcgcctgaaatftaaatgtaatcatggtcatagctgtttcctg |
| Pmel-AmpR | gggggttaaacctgtaagggattttggtcatga |
| AmpR-Ascl | tttggcgcgcgacgaaagggcctcgtagata |
| BR068 | ggagtgttctagactagaagcttactgagct |
| BR069 | cagtgaagcttctagctagaacactcctgca |
| XbaI-RBS-sfGFP | cattgtctagataaggagtgattcgaatgcgtaaaggcgaagagct |
| RC-sfGFP-SacI | cgcgagagctcttatcattgtacagttcatccataccatgc |
| Pmel-catE_ATG | tacgggttaaacctttttctcctgccacatg |
| RC-catE-Ascl_ATG | tatttggcgcgcctgtggaacacctacatctgtattaacg |
| Pmel-pMK4_catE_JUN | tacgggttaaacaccaataacttaagggtaactagcctc |
| RC-pMK4_catE_JUN-Ascl | tatttggcgcgccaacgaaaactcacgtaagggattt |
| NotI-repC | attaagcggccgctgttggtaactaatgggtgc |
| RC-repC-Pmel | ccggtgttaaacacaggggttaaaatgtataacgaaagta |
| RC-pUC_ori-NotI | ggaaagcggccgcatggttaactgtcagaccaagtttactc |
| NotI-repB3 | attaagcggccgctaaaataagcgatttaagcgtgat |
| RC-repB-Pmel | ccggtgttaaaccaattatagcatctaatctcaacaaactggc |
| rhs1-kdc | actagcccaacactcttaaggaggttacaatgtatacggcggcgattattgttg |
| RC-T7prom | /5phos/aattcgccctatagtgagtcg |
| RC-His-tag_Kdc | /5phos/ tgaacctcctaaagagtggttg |
| His-tag_Kdc | atgcatcatcaccatcaccactatacggctggcgattattt |
| XbaI-RBS-kdc | cattgtctagataaggagtgattcgaatgtatacggcggcgattatttg |
| RC-kdc-SacI | aaattgagctcttattagtggtgatggtgatgg |
| Tt_ALDH | cattgtctagagtcggtataacacgagaag |
| RC-Tt_ALDH | aaattgagctcttataatgatgatggtgatggtg |

| | |
|--------------------------|---|
| Geoth_ALDH | catatgaaagattcattttctacacagatcagagc |
| RC-Geoth_ALDH | atgtcgtctcacctcctcatttttagaaggtg |
| XbaI-RBS-Geoth_ALDH | attgtctagatccatcttaaaacagacagaaaggattaccatataaaagattcattttct |
| RC-6XHis-Geoth_ALDH-SacI | aaattgagctcttatcagtggtgatgatggatgctcgtctcacctcctcattttt |
| NotI-repB | attaagcggccgcctaaaataagcgatttaaagcgtgat |
| RC-repB-PmeI | ccgggtgttaaaccaattatagcatctaacttcaacaaactggc |
| Ldh6-sfGFP | gtttcaciaaattcatttttgaaaggatgacagacagcgatgcgtaaaggcgaagagct |
| RC-Ldh6-sfGFP | tagggacgacaccagtgaacagctctcgcctttacgcatcgctgtctgtcatccttcc |
| pUC_ori-PstI-Ldh5 | ttgtaaacgacggccagtgccaagcttgcagcctgcagcatcgctcgtccgctatat |
| sfGFP-Ldh7 | gggcatcacgcatggtatggatgaactgtacaaatgataacaagattccgctcttgcta |
| RC-sfGFP-Ldh7 | actgaaaaatctacttagcaagaagcggaatctgttatcattgtacagttcatcca |
| RC-Ldh8-SacI-pUC_ori | ggaaacagctatgaccatgattacgaattcgagctcttattttacatcatcaaaaataacg |
| SEQ-Ldh-KO | atgccgatgatgtacaaca |
| SEQ-RC-Ldh-KO | ctttactcccctcatcctaaa |
| AscI-400 | aattaggcgcgccaagctttatcgtcatcggca |
| RC-400 | agctctcgcctttacgcatttccacttttcttcaatttgttt |
| AvrII-1200 | aatcgtaattaagcgtactagtcttaggcatgaatgtgaacaggggaga |
| RC-1200-Bsu36I | tgataccttaggaaatgaagaaatcgctatttcga |
| emk013 | ctctcgtctattacgccag |
| RC-ilvIHC-ilvD | cgctgcgaagtttcccaatctcaatccctccatgagcttaattttctcactcgtaac |
| RC-ilvD-XhoI-sfGFP | ctttacgcattttgcactctcctctttgttacggctcgagtaaatttcataatcccgc |
| ilvD-XhoI-sfGFP | gattatgaaaatttaaccgtaacaaagaggagagtgcaaaatcgtaaaaggcgaagagc |
| pilvIHC-sfGFP | catagtgatgaaaaatgtgagggaggatggaaaatgcgtaaaaggcgaagagc |
| pilvD-sfGFP | cgagataataaaactgctcatggaggattgagaatgcgtaaaaggcgaagagc |
| RC-pilvD-sfGFP | caccagtgaacagctctcgcctttacgcattctcaatccctccatgagcag |
| RC-ilvIHC-sfGFP | ggacgacaccagtgaacagctctcgcctttacgcattttccatcctccctcacattttc |
| ilvIHC-ilvD | gagtgagaaaaattaataaggagtgattcgattgggaaaactcgcagcg |

CTGCAGcggggacgggagctgagtgctcccgtgtttgccgcggcgtctgtcatgaaatggacaaacaatagtcaacaatc
gccacaatcgcgcatgcattgcggtgccttcgctgtaaaatatttatgaaagtgttcgcattatattgagggaggatTCTAG
AaaggagtgattcgaATGgctattatcggatcggcaaggccgaataagagtcattgctcgaagaatctgtcaagtatta
cacgaatgtgattggcctggaggaagtggaaggaggaaggaagagttatttaaggcatgggatgaattcgatcacatag
cctcattctcaagaagccgattcccctggccttgatcacattgctttaaggtggaacatgaagacgatttagccaagtcagagaa
gaagatcgagcaattcgggtgtacgttaaaacggatttcaaagggaaggccttcagaaggagaagcaattcgcttcgagct
tccaacagggcatcaagtggaattgtaccatgaaatcgtgcgtgtaggcacgaagacaggaaattgaaatccagccccatggcc
ggatggaatgcgcgggattgcaccgcaccgcctagatcacttagcgtgacaggagaagatatcaacacagtgacaagatttt
cacagaagccttgatattcattagcgaaaaaattatgacagtagatggggaagagatggtagggagctttatattgccagaa
acggaaaagcgcacgatgttgcctttattaaaggccagataagaaaatgcatcatgtcgcattctatgtggacaattggatgaa
gtgttaaaggcagcggatatttaccaaaaataatgtccaattcgatgtgacaccgaccgcatgggattacgcgtggacaaac
cacctacttcttgatccttcaggtaatcgcaatgaagcttttgaagcggttacattacgtatcctgattttctaccataacatggaca
gaagacaaaatcggtaagggaattcttcatatagaagagaattgacggagtcattcatcaaggcgtgacaTAAaatcgcgc
cccgaaggggcggttttttgcGAGCTCGAATTC

PstI site

ldh promoter from NCA1503

XbaI site

pheB RBS

pheB

rho-independent transcriptional terminator "rho1"

SacI site

EcoRI site

Figure A114. DNA sequence of *Ldh* promoter and *pheB* as found in "pGR002" located between the XbaI and EcoRI sites of the plasmid. Sequence taken from Bartosiak-Jentys, 2009.

CTGCAGcggggacgggagctgagtgctcccgtgtttgccgcggtctgtcatgaaatggacaaacaatagtcaacaatc
gccacaatcgcgcatgcattgcggtgcgcttccgctgaaatatttatgaaagtgttcgcattatattgagggaggatTCTAG
AaaggagtgattcgaATGcgtaaaggcgaagagctgttcactgggtcgtccctattctggtggaactggatggtgatgtcaac
ggtcataagtttccgtgctggcgaggggtaaggtagcgaactaatggtaaactgacgctgaagttcatctgtactactggtaaa
ctgccggtacctggccgactctggtaacgacgctgactatggtgttcagtgctttgctcgttatccggaccatagaagcagcatg
acttctcaagtccgcatgccggaaggctatgtgcaggaacgcacgattccttaaggatgacggcacgtacaaaacgcgtgc
ggaagtgaaattgaaggcgataccctggtaaaccgcattgagctgaaaggcattgacttaaagaagacggcaatatcctggg
ccataagctggaatacaatttaacagccacaatgtttacatcaccgcccataaacaataatggcattaaagcgaattttaa
attgccacaacgtggaggatggcagcgtgcagctggctgatcactaccagcaaaacactccaatcggatggtctgttctgct
gccagacaatcactatctgagcagcgaagcgttctgtctaaagatccgaacgagaaacgcgatcatatggttctgctggagttc
gtaaccgcagcgggcatcacgcatggtatggtgaactgtacaaaTGATAAaatcgcgccccgaaagggggcgtttttgcg
aagctcGAATTC

PstI site

ldh promoter from NCA1503

XbaI site

pheB RBS

sfGFP

rho-independent transcriptional terminator "rho1"

EcoRI site

Figure A115. DNA sequence of *Ldh* promoter and sfGFP as found in "pUCG18+pLdh+sfGFP" located between the XbaI and EcoRI sites of the plasmid.

Submitted: 02/19/2013 CPU Time: 7:53:19.82 Organism: Escherichia coli str. K-12 substr. DH10B (ACCTCCTTA) [Export](#)

Version: v1.1

mRNA Sequence

CATTATATTGAGGGAGGATTCTAGATAAGGAGTGATTCGAATGCGTAAAGGCGAAGAGCTGTTCACTGGTGCCTCCCTAT

| Start Position | Translation Initiation Rate (au) | ΔG_{total} | $\Delta G_{mRNA-rRNA}$ | $\Delta G_{accuracy}$ | ΔG_{start} | ΔG_{mRNA} | Structure | Accuracy[?] |
|----------------|----------------------------------|--------------------|------------------------|-----------------------|--------------------|-------------------|------------|-------------|
| 7 | 161.97 | 6.08 | 0.92 | 0.01 | 0.0 | -0.04 | Click here | NEQ |
| 31 | 2944.73 | -0.36 | -14.28 | 4.99 | 0.0 | -0.07 | Click here | NEQ |
| 40 | 1105.57 | 1.81 | -9.98 | 0.29 | 0.0 | -1.19 | Click here | NEQ |
| 58 | 129.79 | 6.57 | -3.28 | 0.29 | 0.0 | -0.03 | Click here | CDS NEQ |
| 65 | 455.16 | 3.78 | -4.58 | 0.0 | -1.7 | -0.03 | Click here | CDS NEQ OLS |
| 68 | 276.04 | 4.9 | -4.58 | 1.15 | -1.7 | -0.07 | Click here | CDS NEQ OLS |

All Gibbs free energies (ΔG) are reported in units of kcal/mol.

Figure A116. Highlighted in yellow are the results of the reverse engineering analysis to determine the strength of the pheB RBS placed downstream from the *G. stearothermophilus* Ldh promoter in the "pUCG18+ldh+sfGFP" construct.

Geobacillus stearothermophilus (unspecified strain) genome
Lactate dehydrogenase gene

GenBank: M14788.1

```
gcgggacggggagctgagtgtcccgtgttggccggtctgtcatgaaatggacaaacaatagtcaacaatcgccacaa  
tcgcgcatgcattgcggtgcgcttccgctaaaaatatttatgaaagtgttcgcattatattgagggaggatgaatgcaATGAA  
AAACAACGGTGGAGCCCGAGTAGTGGTCATCGGCGCCGGGTTTGTGGCGCCAGTTAT  
GTGTTTGCCTTAATGAATCAAGGGATTGCCGATGAGATCGTGCTCATCGATGCGAATGA  
AAGCAAGGCCATAGGCGATGCGATGGACTTCAACCATGGGAAAGTATTTGCGCCGAAG  
CCGTTGACATTTGGCACGGCGATTACGATGATTGCCGCGATGCCGATTTGGTTGTCAT  
TTGCGCCGGCGCCAACCAAAAACCGGGCGAGACGCGGCTTGATCTTGTGGACAAAAAC  
ATTGCCATTTTCCGCTCGATCGTTGAGTCGGTCATGGCATCCGGATTTCAAGGACTGTT  
TCTCGTCGCCACCAATCCGGTCGACATTTTAACGTACGCGACGTGGAAATTCAGCGGCC  
TGCCGCATGAGCGGGTGATCGGTTCCGGGACGATTTTAGATACGGCGCGGTTCCGCTT  
TTGTTGGGCGAGTATTTCTCTGTGCTCCGCAAAATGTTTCATGCCTATATTATTGGGGA  
ACACGGCGACACTGAACTCCCGGTCTGGAGCCAGGCTTATATCGGCGTCATGCCGATC  
CGCAAGCTGGTCGAGTCCAAAGGGGAAGAAGCGCAAAAAGATCTCGAGCGCATTTTTG  
TCAATGTGCGCGATGCCGCCTACCAAATTATTGAGAAAAAAGGAGCGACGTACTACGGG  
ATTGCGATGGGGCTTGCCCGCGTGACGCGCGCCATTTTGCATAACGAAAACGCTATTTT  
GACCGTATCAGCCTACCTCGATGGCCTATATGGGGAGCGCGACGTCTACATCGGAGTG  
CCGGCTGTCATTAACCGCAATGGCATCCGCGAGGTGATCGAAATTGAATTGAATGATGA  
CGAAAAAATCGATTCCATCATAGCGCAGCTACATTA AAAAGCGTGCTAGCCCGTGCTT  
TTACGCGATGA
```

pLdh promoter
Ldh CDS

Figure A117. Reverse engineering analysis of the native RBS found downstream from the *G. stearothermophilus* Ldh CDS.

Submitted: 02/13/2013 CPU Time: 2:37:22.36 Organism: Escherichia coli str. K-12 substr. DH10B (ACCTCCTTA) [Export](#)

Version: v1.1

mRNA Sequence

TTCGCATTATATTGAGGGAGGATGAATGCAATGAAAAACAACGGTGGAGCCCGAGTAGTGTCATCGGCGCCGGGTTTGTCCGGCCAGTT

13968.611602

| Start Position | Translation Initiation Rate (au) | ΔG_{total} | $\Delta G_{mRNA-rRNA}$ | $\Delta G_{spacing}$ | $\Delta G_{standby}$ | ΔG_{start} | ΔG_{mRNA} | Structure | Accuracy[?] |
|----------------|----------------------------------|--------------------|------------------------|----------------------|----------------------|--------------------|-------------------|----------------------------|-------------|
| 11 | 141.85 | 6.38 | 1.62 | 0.0 | 0.0 | -0.04 | -4.8 | Click here | NEQ |
| 21 | 4010.19 | -1.05 | -9.08 | 1.52 | 0.0 | -1.19 | -7.7 | Click here | OLS |
| 25 | 38056.16 | -6.05 | -14.08 | 1.52 | 0.0 | -1.19 | -7.7 | Click here | OLS |
| 30 | 13968.61 | -3.82 | -13.58 | 1.15 | 0.0 | -1.19 | -9.8 | Click here | NEQ OLS |
| 43 | 16.68 | 11.13 | -0.98 | 0.29 | 0.0 | -0.07 | -11.9 | Click here | OK |
| 57 | 7.76 | 12.83 | -6.48 | 0.29 | 0.0 | -0.07 | -19.1 | Click here | CDS |
| 76 | 3.09 | 14.88 | -5.28 | 0.01 | -2.2 | -0.04 | -18.0 | Click here | CDS |

All Gibbs free energies (ΔG) are reported in units of kcal/mol.

Figure A118. Highlighted in yellow are the results of the reverse engineering analysis to determine the strength of the native RBS of the *G. stearothermophilus* Ldh promoter.

DNA sequence of the YtvA gene encoding the blue-light GTP-binding receptor from *Bacillus subtilis* subsp. *subtilis* str. 168

GenBank AL009126.3

```
ATCgctagtttcaatcatttgggataaccaggacagctggaagtcataaaaaagcacttgatcacgtgagtcggtgtgtaa
ttacagatcccgcaactgaagataatcctattgtctacgtaaatcaaggcttgtcaaatgaccggctacgagaccgaggaaat
aggaaagaactgtcgcttctacaggggaaacacacagatcctgcagaagtggaacaacatcagaaccgctttacaaaataa
gaaccggtcaccgttcagatccaaaactacaaaaagacggaacgatgttctggaatgaattaaatattgatccaatggaaata
gaggataaaacgtatgttcggaattcagaatgatatccaagcaaaaagaatatgaaaagcttctcgaggattccctcacgg
aaattactgcacttcaactcctattgtcccattcgcaatggcatttcggctctccgctagtcggaaacctgacagaggagcgatt
aattccatcgttgcacattgacgaatatctatcaacatccaaagatgattattgatcattgattatccggattggccaagtgaac
gaacaaacggccgaccaaatttcaagctgagccatttgctgaaattgaccggaactgagtaatactactggcattaagcctga
attggctatgaaaatgaataaactggatgccaattttcgtcgtgaaaacatattcaaatgtaaaggatgccgttaaagtgttccg
attatgTAA
```

Amino acid sequence of YtvA
NCBI Accession number CAB15012

```
MASFQSGIPGQLEVIKALDHVRVGVVITDPALEDNPVYVNQGFVQMTGYETEEILGKNC
RFLQGKHTDPAEVDNIRTALQNKEPVTVQIQNYKKGDMFWNELNIDPMEIEDKTYFVGIQN
DITKQKEYEKLLLEDLSTEITLSTPIVPIRNGISALPLVGNLTEERFNSIVCTLTNILSTSKDDYLI
DLGLAQVNEQTADQIFKLSHLLKLTGTELIITGIKPELAMKMNKLDANFSSLKTYSNVKDAVK
VLPIM*
```

Figure A119. DNA and amino acid sequence of YtvA as found in the genome of *B. subtilis* 168

DNA sequence of BsFbFP as found in "pUCG18+p43+BsfBFP"

```
ATCgctagtttcaatcattgggataccaggacagctggaagtcataaaaaagcacttgatcacgtgagtcggtgtgtaa
ttacagatcccgcactgaagataatcctattgtctacgtaaatcaaggcttgtcaaatgaccggctacgagaccgaggaaat
aggaaagaacgcacgctcttacagggaaacacacagatcctgcagaagtggacaacatcagaaccgcttcaaaaataa
gaaccggtcaccggtcagatccaaaactcaaaaaagacggaacgatggtctggaatgaattaaatgatccaatggaata
gaggataaaacgtattttgctggaattcagaatgatcaccaagcaaaaagaatatgaaaagcttctcaggattccctcacgg
aaattactgcacttcaactcctattgtcccgattcgcaatggcatttcggctctccgctagtcggaaacctgacagaggagcatt
aattccatcgttgacattgacgaataatctatcaacatccaaagatgattttgatcattgattatccggattggccaagtgaac
gaacaaacggccgaccaaattttcaagctgagccattgctgaaattgaccggaactgagttaatcattactggcattaagcctga
attggctatgaaaatgaataaactggatgccaattttcgctgctgaaaacatattcaaatgtaaaggatgccgttaaagtgctccg
attatgTAA
```

Amino acid sequence of BsFbFP

```
MASFQSFGIPGQLEVIKKALDHVRVGVVITDPALEDNPIVYVNQGFVQMTGYETEEILGKNA
RFLQGKHTDPAEVDNIRTALQNKEPVTVQIQNYKKGDMFWNELNIDPMEIEDKTYFVGIQN
DITKQKEYEKLLEDSLTEITALSTPIVPIRNGISALPLVGNLTEERFNSIVCTLTNILSTSKDDYLI
DLSGLAQVNEQTADQIFKLSHLLKLTGTELIITGIKPELAMKMNKLDANFSSLKTYSNVKDAVK
VLPIM*
```

Figure A120. DNA and amino acid sequences of BsFbFP. The codon mutation required to convert the protein from YtvA to the fluorescent BsFbFP is highlighted in green.

CTGCAGTTAATTAAgtgcatgcaggccggggcatatgggaaacagcgcggacggagcgggaattccaattcatgccg
cagccgcctgcgctgttctcattgctggtccttctgtagagctcagcattatgagtggtgattatattcctttgatagggtgatgtttc
gctgaacttttaatacagccattgaacatacgggtgattaataactgacaaacatcacctcttgctaaagcggccaaggacgc
tgccgcccgggctgtttgcgttttaaccgtatttcgtgtatcattggttactatTTTTTgccaagctgtaatggtgaaaattcttacatt
tatttacatttttagaaatgggctgaaaaaaagcgcgcgattatgtaaaatataaagtgatagcggaccattataTCTAGAta
aggagtgattcgaATGgctagtttcaatcatttgggataccaggacagctggaagtcacaaaaagcacttgatcacgtgcg
agtcggtgtggaattacagatcccgcacttgaagataatcctattgtctacgtaaatcaaggctttgtcaaatgaccggctacgag
accgaggaaattttagaaagaacgca cgcttcttacaggggaaacacacagatcctgcagaagtggacaacatcagaaccg
ctttacaaaataaagaaccggtcaccgttcagatccaaaactacaaaaagacggaacgatgttctggaatgaattaaatattga
tccaatggaaatagaggataaaacgtattttgtcggaaattcagaatgatcaccaagcaaaaagaatgaaaagcttctcgag
gattccctcacggaaattactgcactttcaactcctattgtcccgattcgcaatggcatttcggctctccgctagtcggaaacctgac
agaggagcgatttaattccatcgtttgacattgacgaatatctatcaacatccaaagatgattattgatcattgattatccggattg
gcccaagtgaacgaacaaacggccgaccaaatttcaagctgagccatttctgaaattgaccggaactgagttaatcattactg
gcattaagcctgaattggctatgaaaatgaataaactggatgccaaattttcgtcgcgtgaaaacatattcaaatgtaaaggatgccgt
taaagtgcctccgattatgTAAtcgccccgaaagggcgctttttgcGAGCTC

PstI site

p43 promoter from *B. subtilis* 168 (-10 box and -35 box, according to BPROM [XX])

XbaI site

pheB RBS

BsFbFP

rho-independent transcriptional terminator "rho1"

SacI site

Figure A121. Detail of sequence corresponding to the p43 promoter and BsFbFP located between the PstI and SacI sites in the pUCG18+p43+BsFbFP construct.

CTGCAG ggggacgggagctgagtgctcccgtgtttgcccggcgtctgtcatgaaatggacaaaacaatagtcaacaate
gccacaatcgcgcatgcattgcggtgcgctttcgcgtaaaaattttatgaaagtgttcgcattatattgaggaggatTCTAG
AaaggagtgattcgaATGgctagtttcaatcattgggataccaggacagctggaagtcatcaaaaaagcacttgatcacgtg
cgagtcgggtgtgtaattacagatcccgcactgaagataatcctattgtctacgtaaatcaaggcttgtcaaatgaccggctacg
agaccgaggaaatfttaggaagaacgca cgcttctacaggggaaacacacagatcctgcagaagtggacaacacagaa
ccgctttacaaaataaagaaccggtcaccggtcagatccaaaactacaaaaagacggaacgatgttctggaatgaattaata
ttgatccaatggaaatagaggataaaacgtattttgctggaattcagaatgatcaccaagcaaaaagaatatgaaaagctctc
gaggattccctcacggaaattactgcactttcaactcctattgtcccgattcgcaatggcatttcggctctccgctagtcggaaacct
gacagaggagcgaatttaattccatcgtttgcacattgacgaatatctatcaacatccaaagatgattattgatcattgattatccgg
attggcccaagtgaacgaacaaacggccgaccaaattttcaagctgagccatttgctgaaattgaccggaactgagttaatcatta
ctggcattaagcctgaattggctatgaaaatgaataaaactggatgccaattttcgctgctgaaaacatattcaaatgtaaaggatgc
cgtaaagtgtctccgattatgTAAtcgccccgaaaggggctgtttttgcGAGCTC

PstI site

ldh promoter from NCA1503

XbaI site

pheB RBS

BsFbFP

rho-independent transcriptional terminator "rho1"

SacI site

Figure A122. Detail of sequence corresponding to the *Ldh* promoter and BsFbFP located between the PstI and SacI sites in the pUCG18+p43+BsFbFP construct.

AscI - RsrII – Rho2 terminator – PaeI – (SpeI AvrII NheI) - SexAI – XmaI – (Sall XhoI) – PstI/SbfI- RplS – XbaI – OPT_PpFbFP – SacI – Rho1 terminator – Bsu36I– SwaI

GGCGCGCCtctaCGGACCGagtatcaacaacggggccagtttggtgaaatcgTTAATTAAgcgtACTAGTC
CTAGGCTAGCtgatACCAGGTcagtCCCGGGacagGTCGACCTCGAGtacgCCTGCAGGAACA
ATCGTTAAAGCGGACGTTTTTTCGCGCGCCCGGATTTGCTTGAAAACACTACCCGCTGACAG
AAAAGCAAAAACGATGGATCGAAGAGTGGAAAAAGAAAAACAGTAGCTATTGCGCATG
ATACAAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTGCTTAAGCGAGGAAA
ACGATGTTCCGCTGCAATGATGAAAAAGCATTGTCTAGATGTTAATTTATACATTCATCTA
AGGAGGTTCTTCATGATCAACGCCAAATTGTTGCAACTGATGGTTCGAACATAGCAACGA
TGGCATTGTTCGTCGCCGAACAAGAAGGCCAACGAAAGCATTGTTGATTTATGTCAACCCAG
CCTTTGAACGCTTGACGGGCTATTGTGCCGATGATATCTTGTATCAAGATGCTCGCTTTT
TGCAAGGCCAAGATCATGATCAACCGGGAATTGCCATTATTCGCGAAGCCATTTCGCGAA
GGACGCCCATGCTGCCAAGTCTTGCGCAACTATCGCAAAGATGGCAGCTTGTTTTGGAA
CGAATTGAGCATTACACCGGTCCATAATGAAGCCGATCAATTGACGTATTACATTGGCAT
TCAACGCGACGTCACGGCTCAAGTCTTTGCCGAAGAACGCGTCCGCGAATTGGAAGCC
GAAGTCGCCGAATTACGCCGCCAACAAGGCCAAGCCAAACATCATCATCACCATCA
TTAATAAGAGCTCaatttcgcgccccgaaagggcgTTTTTtgcgaagCCTAAGGtatcATTTAAAT

RplS promoter

Synthetic RBS designed for RplS+PpFbFP (predicted strength of 1028279 au in *E. coli* and 572829 au in *Geobacillus* sp. C56-T3)

OPTIMIZED_PpFbFP

Figure A123. Detailed sequence of synthetic MCS containing the RplS promoter, a synthetic RBS and the codon optimized PpFbFP.

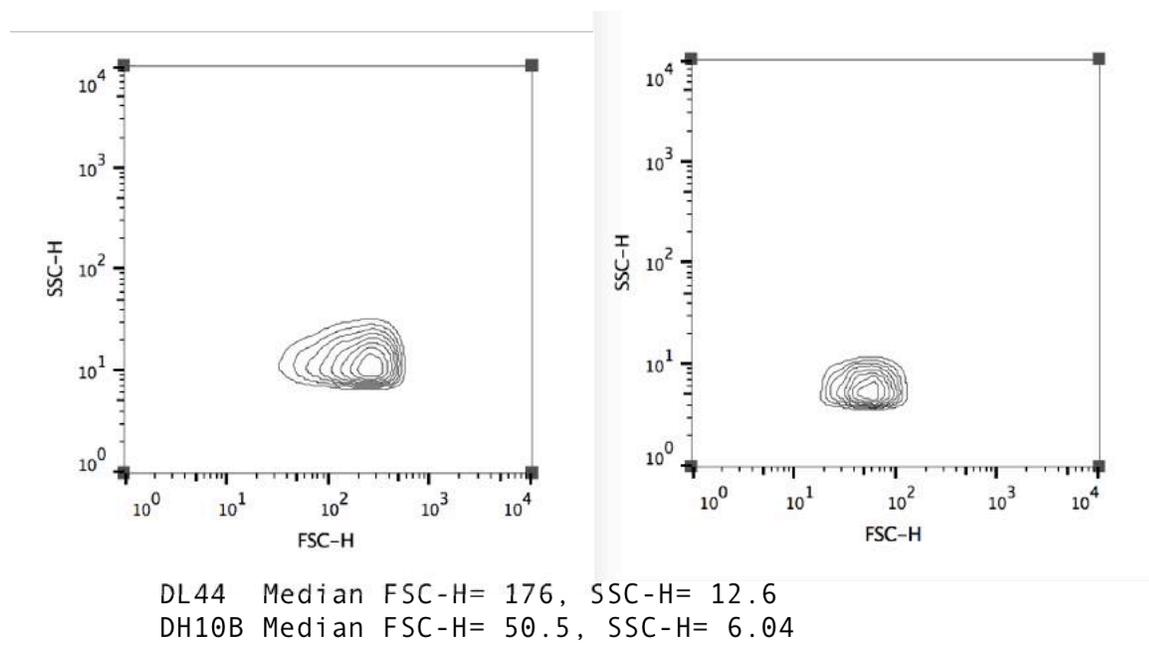


Figure A124. Comparison between the size and agglomeration of *E. coli* and *G. thermoglucosidans* when compared by flow cytometry.

tgaaaaatagaccttgcgactcg**aagattag**tagtacaat**tagaga**caacataacagaacg**ggacaaaat**agaagagcttc
cgaagc**gtgtgta**agtggtgctcccaacaccggttcgccgtactgcacctacggcaaacacctgctcgaaagctcttggttaacat
atztatgccttgccttcaatgcatagatatggtgaatgcaagagcgaaacgcttgaaaatttcgttttctccgatgagagcgtgt
ttgcctgcgtattaatgtacggcaacgcgctctttgttgcctccgaaaaattcgggagaaacgggaggagaacggaa**ATG**taca
faatcactagcgaacaattcacgcaactgaaagaaatcgagagagtcacagatttcgtttataaacgccgtaagattaaaaatga
gctgatggagcgtatcgccgacgcattaggccgcaagttccgcaaagcgcaaaaaacaacgcgcgctcatcgacgaaattgttt
ggctcgcaacagaccggttttctccagggcgcaaacgcttgcaaaaagttaggcgtgccttgccgactgtgatcgtg
ctactcgcgatgctcaaagactcgggcaagtcgctgtgataatccgaatagtaacggacctaaaacgcctgtattc
atctccggagccatgcaaatctcagcgtattgcagccgttttaacttgctgacaacgaagctgataaagtagaaaacgccga
aaagcctacggaatcaagcgaattagcgcgtaaaacggacgctaccattagttcaccagttgaaaaacatgacgatattaataa
attaatagacaatacacatccgttgataaaaattgtaaaatacgtcactattaaagtgaacgaagtcagaagcgtcatccttacg
gcattcgcgatctatccgcatacataaacaacactcgcacgacctaatgcaacgtgcaaaagcgaaggtcaaagccgaaaa
agctcgcgcccacaacgcaacaagccgactcaaccgaccgacccgcttaatttggtacgactcaccgagcagacaca
agcgcagccagccgacaagccagcgaagccagcgtttgacagcgaagacatgtttgcgccgtattgcaggctgttcacg
aattaagggcgttgagggt**TGA**cacactcgttttaaggcgttaacaggcgctagagacgctttaagtcgggggaatattaa
aattccacgcctcctaaaacgcgatatacg

-10 box

-35 box

transcription factor binding sites

transcriptional terminator

Figure A125. Detailed sequence of *repBST1* as found in pUCG18. The promoter region was analyzed by BPRM and the -10 and -35 boxes and transcription factors found are highlighted.

ttcaacaaacgggccagttgtgaagattagatgctataattgtattaaaggattgaaggatgcttaggaagacgagttattaat
agctgaataagaacgggtgctctcctcaaatattcttatttagaaaagcaaactctaaaattatctgaaaaggaATGagaatagta
atggaccaataataatgactagagaagaaagaatgaagattgttcatgaaattaaggaaacgaatattggataaatatggggatg
atgtaaggctattggtgttatggctctcttggtcgtcagactgatgggccctattcggatattgagatgatgtgtcatgtcaacaga
ggaagcagagttcagccatgaatggacaaccggtagtgaagggtgaagtgaatttatagcgaagagattctactagattat
gcatctcaggtggaatcagattggccgcttacacatggcaatcttctctatcttccgatttatgattcaggtggatacttagagaaag
tgtatcaaactgctaaatcggtagaagcccaaaagtccacgatgcgattgtgccctatcgtagaagagctgttgaatatgcag
gcaaatggcgtaaatattcgtgtgcaaggaccgacaacatttctaccatcctgactgtacaggtagcaatggcaggtgccatgtga
ttggtctgcatcatcgcactgttatacgacgagcgttcggcttaactgaagcagttaagcaatcagatcttctcaggttatgacc
atctgtgccagttcgtaatgtctgtgcaacttccgactctgagaaaactctggaatcgctagagaatttctggaatgggattcaggag
tggacagaacgacacggatataatagtgatgtgtcaaaacgcataccattTGCATATGcgggtgtgaaataccgcacaga
tgcgtaaggagaaaaataccgcatcagggcgccattcgccatcaggctgcgcaactgttgggaagggcgatcgggtcggggcctc
tcgctattacgccagctggcgaaaggggatgtgctgcaaggcgattaagttgggtaacgccaggggtttcccagtcacgacgtt
taaacgacggcca

-35 box

-10 box

kanR coding sequence

mutations for thermostability

NdeI

pUC18

lacZ

Figure A126. Detailed sequence of *kanR* as found in pUCG18. The promoter region was analyzed by BROM and the -10 and -35 boxes and transcription factors found are highlighted.

gattatcaaaaaggatcttcacctagatcctfttaaaattaaaaatgaagttttaaatacaatctaaagtatatatgagtaaacttgggtctg
acagTTAccaatgcttaacagtgaggcacctatctcagcgatctgtctatttcgltcatccatagttgcctgactccccgtcgtgtag
ataactacgatacgggagggcttaccatctggccccagtgctgcaatgataccgcgagaccacgctaccggctccagatttat
cagcaataaaccagccagccggaagggccgagcgcagaagtggtcctgcaactttatccgcctccatccagctattaattgttg
ccgggaagctagagtaagtagttcgccagtaaatagttgcgcaacggtgtgcccattgctacaggcatcgtgggtgcacgctcgtc
ttggatggcttcattcagctccggttcccaacgatcaaggcaggttacatgatccccatgttgcaaaaaagcgggttagctcctt
cggctcctccgatcgttgtcagaagtaagtggccgagtggtatcactcatggttatggcagcactgcataattctctactgcatgcc
atccgtaagatgctttctgtgactggtagtactcaaccaagtcattctgagaatagtgatgcgggcgaccgagttgctcttgcggg
cgtcaatacgggataataccgcgccacatagcagaactttaaaagtgtcatcattggaaaacggttctcggggcgaaaactctc
aaggatcttaccgctggtgagatccagttcgatgtaaccactcgtgcacccaactgatcttcagcatctttactttaccagcgtttct
gggtgagcaaaaacaggaaggcaaaatgccgcaaaaaaggggaataagggcgacacggaaatgttgaatactCATactctt
ccttttcaatatttgaagcatttatcagggttattgtctcatgagcggatacatatttgaatgtatttagaaaaataaacaataggg
gttccgcgacatttccccgaaaagtgccacctgacgtctaagaaaccattattatcatgacattaacctaataaaatagggcgtatc
acgaggcccttctcgtc

-10 box

-35 box

transcription factor binding sites

transcriptional terminator

Figure A127. Detailed sequence of *ampR* as found in pUCG18. The promoter region was analyzed by BPROM. The -10 and -35 boxes and transcription factors found are highlighted.

ctaaaataagcgatttaaagcgtgatgtgagtttaactatgaaagcactaaggaattccttaaggaacgtacagacggcttaaaa
gcctttaaaccggttttaaggggttagacaaggtaaaggataaaacagcacaattccaagaaaacacgatttagaaccta
aaaagaacgaattgaaactaactcataaccgagaggtaaaaaaagaacgaagtcgagatcagggaatgagttataaaataa
aaaaagcacctgaaaagggtctttttgatggtttgaactgttcttctatcttgatacatatagaaataacgtcattttatgttg
ctgaaagggtcggtgaagtgttggtatgtatgttttaagattgaaaa**ccttaaa**attg**ttgcac**agaaaaaccccat**tgta**
aagttataagtgactaaacaa**ataactaa**atag**ATC**ggggttctttaaattattatgtgtcctaataagtagcatttattcagatgaaa
atcaagggttttagtggacaagacaaaaagtggaagtgagaccatggagagaaaagaaaatcgctaattgtgattacttga
acttctgcatattctgaatttaaaaaggctgaaagagtaaaagattgtgctgaaatattagagtataaacaataatcgtaaacagg
cgaaagaaaagttgatcagagtggtttgtaaatccaggcttgtccaatgtgcaactggaggagagcaatgaaacatggcattc
agtcacaaaagggtgtgctgaagtattataaacaagccaacagttcggtggtttctcacattaacagttaaaaatgtttatgatg
gcaagaataaataagagttgtcagatatggctcaaggatttcgccgaatgatgcaataaaaaaataataaaaatctgttg
gtttatgctgcaacggaagtgaataaataataaagataattcttataatcagcacatgcatgtattggtatgtgtggaaccaac
ttatttaagaatacagaaaactacgtgaatcaaaaacatggattcaattttgaaaaaggcaatgaaattagactatgatccaa
atgtaaaagtcaaatgattcgaccgaaaaataataataatcggatatacaatcggcaattgacgaaactgaaaatcctgt
aaaggatcggattttatgaccgatgatgaagaaaagaattgaaacgttgtctgattggaggaaggttacaccgtaaaagggt
aatctcctatgggtgtttgtaaaagaaatacataaaaaattaaccttgatgacacagaagaaggcgatttgattcatacagatga
tgacgaaaaagccgatgaagatggatttctattatgcaatgtggaattggaacggaataattttattaaagag**TAG**ttcaa
caaacgggccagttgtgaagattagatgctataattg

-10 box

-35 box

transcription factor binding sites

transcriptional terminator

Figure A128. Detailed sequence of *repB* as found in pUB190. The promoter region was analyzed by BPROM. The -10 and -35 boxes and transcription factors found are highlighted.

GCGGCCGCctgtcgagaatTTTTGGAAAatagaccttgcgcgactcgAAGATTtagtacaatAGAGACAacataacag
aacGGACAAAatagaagagcttccgaagcGTGTGTAAGTGGTGTCCCAACACCgTTCGCGTACTGCCTACGGCAACA
cctgtcgaagctctgtgttaacataTTTatgCCTGTCTCATTtaTatgcatagatatgttgaaTgcaagagcgaaacgcttgaaa
TTTTGTTTTCTCCgatgagagcgtgttgcctgcgtattaatgtacggcaacgcgctctttgtttgTCCCgAAAatcgggagaaac
gggaggagaacggaaATGTacataatcactagcgaacaattcagcaactgaagaaatcgagagagtcacagatttcgTt
ataaacgccgtaagattAAAAatgagctgatggagcgtatcggcgacgcattaggccgcaagtccgcaaacgcaaaaaac
aacgcgcgtcatcgacgaaattgTtggctcgaacagaccgCGTTTTcgttccagggcgcaaacgcttgcaaaaagTta
ggcgtgtccttgcgactgttgatcgtgctactcgcgatgctcaaagactcgggcgaagtgcgtgttgcTatcgtgagaatccgaata
gtaacggacctaaaacgcctgtattcattttccggagccatgcaaatttcgagcgtattgcagccgttttaacttcggtgacaacga
agctgataaagtagaaaacgcccgaagcctacggaatcaagcgtattagcgcgtaaaacggacgctaccattagttcaccag
ttgaaaaacatgacgatataaataatagacaatacacatccgttggataaaattgtaaaatacgtcactattaaagtgaaacg
aagtcgagaagcgtcatccttacggcattcgatctatccgcatacataaacaacactcgcgacctaatgcaacgtgcaa
aagcgaaggtcaaagccgaaaaagctcgcgcccaacaacgcaacaagccgactatcaaccgaccgacggcctaatt
ggTACgacttcacggcggacacaagcgcagccagccgacaagccagcgaagccagcgtttgacagcgaagacatgtttg
cgccgttattgcaggctgttcacgaattaagggcgttgaggTGAcacactcgttttaaggcgttaacagggcgcgtagagac
gtcttfaagtcgggggaatattaaaattccacgcctcctaaaacgcgatatacggcgattatgaGTTTAAACgaagattaga
tgctataattgttattaaaaggattgaaggatgcttaggaagacgagttattaatagctgaataagaacgggtgctcTcaaatattctt
atttagaaaagcaaatcTaaaattatctgaaaagggAATGagaatagtgaaatggaccaataataatgactagagaagaaga
atgaagattgttcatgaaattaaggaacgaatattggataaatatgggatgatgtaaggctattgggtttatggctctcttggctg
cagactgatggccctattcggatattgagatgatgtgtgtcatgtcaacagaggaagcagagttcagccatgaatggacaaccg
gtgagtggaagtggaagtgaattttatagcgaagagattctactagattatgcatctcagggtggaatcagattggccgcttacac
atggTcaattttctctattttgccgatttatgattcagggtgatacttagagaaagtgatcaaaactgctaaatcgggtagaagccaaa
agttccacgatgcgattgtgcccctatcgtagaagagcgtttgaaatgagggcaaatggcgtaatattcgtgtgcaaggaccga
caacattctaccatcctgactgtacaggtagcaatggcagggtccatgttgattggctcgcacatcgcacatcgttatacgcgagc
gctcggcttaactgaagcagttaaagcaatcagatctccttcagggttatgaccatctgtgcccagttcgtaatgctggTcaacttccg
actctgagaaactctggaatcgttagagaatttctggaatgggattcaggagtgacagaacgacacggatataatagtgatgt
gtcaaaacgcataaccattTGA TAAGGCCGCCctcaCGGACCGagattcaacaaacggggcagttgttgaaatc
gTTAATTAAGcgtACTAGTCTTAGGCTAGCTgatACCAGGTcagtCCCGGacagGTCGACCTC
GAGTACgCCTGCAGGAACAATCGTTAAAGCGGACGTTTTTTCGCGCCGCCGGATTTGCTTG
AAAAC TACCCGCTGACAGAAAAGCAAAAACGATGGATCGAAGAGTGGAAAAAAGAAAA
CAGTAGCTATTGCGCATGATACAAGTTTATGCTACTATATTCTTGTGCAACTTAAGCGA
TTTGCTTAAGCGAGGAAAACGATGTTCCGCTGCAATGATGAAAAAGCATTGTCTAGATAag
gagtgattcgaATGGCATCGTTTCAGTCGTTCCGGTATTCCAGGCCAGCTTGAAGTGATCAAAA
AAGCGTTAGATCATGTACGCGTGGGGTGTGATTACTGATCCTGCCTTAGAAGATAAT
CCGATTGTATATGTGAACCAAGGCTTTGTCCAAATGACAGGCTATGAAACAGAAGAAAT
CTTGGGGAAAAACGCGCTTTCTGCAAGGCAACATACGGATCCGGCGGAAGTAGAC
AATATTCGGACTGCTTTACAAAACAAAGAACCGGTCACGGTTCAGATTCCAGAACTACAAA
AAAGACGGAACCATGTTTTGGAACGAATTATACATTGACCCTTTTGAATCGAAGACAAA
ACATACTTCGTGGGAATCCAATACGACATTACAAAACAAAAGAATATGAAAAATTGCTT
GAAGATTCTTTAACTGAAATTACGGCTCTTTCTACACCTATTGTCCCAATTAGAAATGGTA
TCTCGGCCTTGCCGCTTGTGGTAATTTAACCGAAGAACGCTTTAATTCCATTGTGTGCA
CACTTACAAATATCTTGAGCACGAGCAAAGATGATTATCTGATTATCGATTTGTCAGGCT
TAGCGCAAGTCAACGAACAAACCGCAGATCAAATCTTCAAATTGTCCCATTTACTGAAAT
TAACAGGAACGGAACCTTATCATTACGGGAATTAACCCAGAACTGGCTATGAAAATGAATA
AATTAGATGCAAATTTTAGCTCATTGAAAACGTATAGCAATGTAAAGATGCGGTCAAAG
TTTTACCGATTATGtaataaGAGCTCaattcgcgccccgaaagggcgTTTTttgcgaagCCTAAGGtataAT
TTAAATgtaatcatggTcatagctgttctgtgtgaaattgttatccgctcaaatccacacaacatacgcgaccggaagcataa
agtgtaaagcctgggtgcctaagtagtgagctaaactcacattaattgcgttcgctcactgccgcttccagtcgggaaacctgt
cgtgccagctgcattaatgaatcggccaacgcgCGGGGagaggcggtttgcgtattgggcgctctccgcttctcgcctcactgact
cgctgcgctcggctcgttcggctcggcgagcggatcagctcactcaaaggcgttaatacggttatccacagaatcaggggata
acgcaggaagaacatgtgagcaaaagccagcaaaagccaggaaccgtaaaaaaggccgctgtcgtggcttttccata

ggctccgccccctgacgagcatcaaaaaatcgacgctcaagtcagaggtggcgaaacccgacaggactataaagatacc
aggcgtttccccctggaagctccctcgctgctctcctgttccgacctgccccttaccggatacctgtccgcttttcccttcggga
agcgtggcgctttctcatagctcacgctgtaggtatctcagttcgggtgtaggtcgctccaagctgggctgtgtgcacgaacccc
ccgttcagcccgaccgctgctgcttatccggaactatcgctctgagtcacaacccggaagacacgacttatcgccactggcagca
gccactggtaacaggattagcagagcgaggtatgtaggcgggtctacagagttcttgaagtggtggcctaactacggctacacta
gaagacagatatttggtatctgctgctgctgaagccagttacctcgaaaaagagttggtagctcttgatccggcaacaacc
accgctggtagcgggtgtttttgttgaagcagcagattacgctcagaaaaaaggatctcaagaagatccttgatctttctac
gggtctgacgctcagtggaacgaaaactcacgttaagggatttggctcatgagattatcaaaaaggatcttcacctagatccttta
aataaaaaatgaagtttaaatcaatctaaagtataatgagtaaaacttggctgacagttaccaatGGCCGGCCctgacagT
TAccaatgcttaatcagtgaggcacctatctcagcgatctgtctatcttctgctcatccatagttgctgactccccgctgtagataact
acgatacggggagggtaccatctggccccagtgctgcaatgataccgagaccacgctcaccggctccagatttatcagca
ataaaccagccagccggaagggccgagcgcagaagtggctcgaactttatccgctccatccagtctattaattgttccggg
aagctagagtaagtagttccagtaataagttgccaacgttggccattgctacaggcatcgtgggtcacgctcgtctgttgg
atggctcattcagctccggttcccaacgatcaaggcagttacatgatccccatgttgcaaaaaagcgggttagctcctcgg
ctccgatcgttgtagaagtaagttggccgagtgatcactcatggtatggcagcactgcataattcttactgtcatgccatccg
taagatgctttctgtgactggtagtactcaaccaagtcattctgagaatagtgatgcggcgaccgagttgctcttggccggcgtca
atacgggataataccgcccacatagcagaactttaaaagtgctcatcattggaaaacgctcttggggcgaaaactcgaagga
tcttaccgctgttgagatccagttcagtgtaaccactcgtgcaaccaactgatcttcagcatctttactttcaccagcgtttctgggtg
agcaaaaacaggaaggcaaaaatgccgcaaaaaaggggaataagggcgacacggaaatggtgaataactcattactcttcttttc
aatattattgaagcatttatcaggggtattgtctcatgagcggatacatattgaatgtatttagaaaaataaacaataggggtccgcg
gcacatttccccgaaaagtgccacctgacgtctaagaaaccattattatcatgacattaacctataaaaaataggcgtatcacgagg
cc

NotI
repBST1
PmeI
kanR
AclI
MCS
RplS promoter
pheB RBS
hbLOV
MCS
SwaI
M13R
ColE1 (mutation highlighted)
FseI
AmpR

Figure A129. Complete sequence of plasmid p11AK0.

actcttcttttcaatattattgaagcatttatcagggtattgtctcatgagcggatacatattgaatgatGCGGCCGCctgtcg
agaatgttgaaaatagacctgtcgcgactcgaagattagtagtacaatagagacaacataacagaacggacaaaataga
agagcttccgaagcgtgtgtaagtggtgctcccaacaccgttcgccctactcgcctacggcaaacacctgctcgaagctctgt
gttaacatatttatgcctgtcttcatttaaatgcatagatatgttgaaatgcaagagcgaaacgcttgaaatttctgtttctccgatga
gagcgtgttgcctgcgtattaatgtacggcaacgcgctctttgttgcgcaaaaattcgggagaaacgggaggagaacggaa
ATGtacataatcactagcgaacaattcacgcaactgaaagaaatcgagagagtcacagatttcgtttataaacgccgtaagatt
aaaaatgagctgatggagcgtatcgccgacgattaggCccgcaagttccgcaagcgcaaaaaacaacgcgctcatcga
cgaaattgttggctcgaacagaccgctgttttctgtCcccagggcgcgAaaacgcttgcAaaaaagttaggcgtgtcctgc
cgactgttgatcgtgctactcgcgatgctcaagactcggcggaagtcgtctgttctatcgtgagaatccgaatagtaacggacct
aaacgcctgtattcatttccggagccatgcaaatctcagcgtattgcagccgttttaacttgcgtgacaacgaagctgataaagt
agaaaacgccgaaaagcctacggaatcaagcgttagcgcgtaaaacggagcctaccattagttaccagttgaaaaacatg
acgatattaataaattaatagacaatacacatccgttgataaaattgtaaaatcgtcactattaagtgaaacgaagtgacagaag
cgtcatccttaccgacattcgatctatccgcatacataaaacaaactcgcgacctaatacgaacgtcaaaaagcgaaggtca
aagccgaaaaagctcgcgcccacaacgcaacaagccgactcaaccgaccgacccgcttaatttggtagacttcac
ggcggacacaagcgcagccagccgacaagccagcgaagccagcgtttgacagcgaagacatgttgcgcccgttattgcag
gctgttcacgaattaagggcgttgaggTGAcacactcgttttaaggcgttaacagggcgttagagacgtcttaagtcggg
ggaatattaataaaattccacgcctcctaaacgcgatatacgcgattatgaGTTTAAACgaagattagatgctataattgttatt
aaaaggattgaaggatgcttaggaagacgagttattaatagctgaataagaacgggtgcttccaaatattctatttagaaaagca
aatctaaaattatctgaaaaggggATGagaatagtgatggaccaataataatgactagagaagaagaatgaagattgtca
tgaaattaaggaacgaatattggataaataatgggatgatgttaaggctattgggtttatggctctcttggctcagactgatgggc
cctattcggatattgagatgatgtgtcatgtcaacagaggaagcagagttcagccatgaatggacaaccggtagtggaaggt
ggaagtgaaatttatagcgaagagattctactagattatgcatctcagggtggaatcagattggccgcttacacatggcaatttctc
atttgcgatttatgattcagggtgatacttagagaaagtgatcaaaactgctaaatcggtagaagcccaaaagttccacg

Figure A130. Sequence detail of the mutations produced in repBST1 (p11AC1) after artificial selection rounds to obtain chloramphenicol resistance gene with improved function in *G. thermoglucosidans*. Yellow highlight signifies insertion.

tcgcgcttcggtgatgacgCTCGAGgatccaggagaacaaaaacgatttttgaggaaagttataaatttttccgaacgatatg
gcaagcaaaatattgctatgcaacagttcataatgatgagcaaacccctcacatgcatttaggtgtgtgcctatgctgatggaaaact
gcaaggaaaaatgtgttaactgcaagaactgtatggctacaagataaattccccgagcataatgaaaaaacgggttttgagttgaa
gctgtggaacgtggctctgaccgtaaacatattgagacagctaaatttaaaaaacaaacttggaaaaagagattgatttctagaaaa
aaatttagcagttaaaaagatgaatggactgctatagcgtataaagttaaatcagatttagaagtaccagcgaaacgcacacatgaaa
agtgtgaagtgccaacgggtgaaaagtcctattgtttgggaaaagaataatgaaaacagaaaagaaaccaacaaaaatgtt
gttatacggagcgtgattataaaaacttagtactgctgagagataacgataggtaaacagcatgttagaaatctcatgagactg
atatggcgagagaatataaaaaattaagtaagaacatggcaagttaaagaaaaatagtggtctgtagagcgatttaatagaaaat
gtaaatgattataatgagttgctgaagaaaacagctcttaaaagctaaaataagcgatttaaacggtgatgtgagtttaactatgaaag
cactaaggaattcctaaggaacgtacagacggcttaaaagccttaaaaacgcttttaaggggtttgtagacaaggtaaggataaaa
cagcacaattccaagaaaaacagatttagaacctaaaagaacgaattgaactaactcataaccgagaggtaaaaaaagaacg
aagtcgagatcaggggaatgagtttaTAAaataaaaaaagcacctgaaaagggtctttttgatggtttgaactgttcttctatctgat
acatatagaataaacgctattttatttagtctgaaagggtcgttgaaggttggtatgtatgtgttttaagattgaaaacccttaaaatg
gttcacagaaaaaccccatctgttaaaagtataagtgactaaacaaataactaaatagATCggggtttctttaatattatgtctcta
agtagcatttattcagatgaaaaatcaagggttttagtgacaagcaaaaaagtggaagtgagacctggagagaaaagaaaat
cgtaatgttgattactttgaacttctcatattctgaatttaaaaaggctgaaagagttaaagattgtgctgaaatattagagtataacaa
aatcgtgaaacaggcgaaagaaagtgtatcgagtggtttgtaaatccaggctttgtccaatgtgcaactggaggagagcaatgaa
acatggcattcagtcacaaaagggtgtgctgaagttataacaaaaagccaacagttcgttgggtttctcacattaacagttaaaaatgtt
tatgatggcgaagaatataaagagttgtcagatattggctcaaggatttcgccgaatgatgcaatataaaaaaataataaaaaatctgt
tggtttatgctgcaacggaagtgaacataaataaagataattcttataatcagcacatgcatgtattggtatgttgaaccaactta
tttaagaatacagaaaactacgtaacaaaaaatgattcaattttgaaaaaggcaatgaaattagactatgatccaaatgtaa
agttcaaatgattcgaccgaaaaataataatcgatatacaatcgcaattgacgaaactgcaaaatacctgtaaaggatcgg
atthtatgaccgatgatgaagaaaagaattgaaacggttctgattggaggaaggttacaccgtaaaaggtaactctctatggtggtt
gttaaaagaaatacataaaaaaataaaccttgatgacacagaagaaggcgatttgattcatacagatgatgacgaaaaagccgatga
agatggatttctatttgcaatgtggaattgggaacggaaaaaatttttataaagagTAGttcaacaaacgggcccagtttgtgaaga
ttagatgctataattgttataaaaggattgaaggatgcttaggaagacgagttattaatagctgaataagaacgggtgctctccaaatctt
atttagaaaagcaaatctaaaattatctgaaaaggaATGagaatagtgaaatggaccaataataatgactagagaagaaagaatga
agattgtcatgaaataaggaacgaatattggataaataatgggatgatgtaaggctattggtttatggctctgtgctcagactgat
gggcccattcggatattgagatgatgtgtgctatgcaacagaggaagcagagttcagccatgaatggacaaccgggtgagtggaagg
ggaagtgaattttatagcgaagagattctactagattatgcatctcaggtggaatcagattggccgctacacatggcaatcttctat
gccgattatgattcaggtggatacttagagaaagtgtatcaaacgtctaaatcggtagaagccaaaagtccacgatgagatttgcc
ctatcgtagaagagctgttgaatgcaaggcaaatggcgaatattcgtgtgcaaggaccgacaacatttctaccatcctgactgtaca
ggtagcaatggcaggtgccatgttattggtctgcatcatcgcatctgttatacgaagcagcgtcggcttaactgaagcagttaaagcaat
cagatctctcaggttatgaccatctgtccagttcgtaatgtctgtgcaacttccgactctgaaactctggaatcgttagagaattc
tggatgggattcaggagtgacagaaacgacagcggatataatggtatgttcaaacgcataccattTGAcatatgcggtgtaaa
taccgacagatcgtaaggagaaaaataccgcatcaggcgccattcgccatcaggctgcaactgttgggaagggcgatcgggtc
gggctctcgtatcagccagctggcgaagggggatgtgctgcaaggcgattaagtgggtaacgccagggtttccagtcacga
cgtgtgaaaacgacggccagtgccaagctgcatgcCTCGAGcatcgcgctcgcgctatatttgacatggattttgcccgtcgcgctc
gtcggcgtgtatccggcttctattttctgcaaaaaagaatgttacggatagcgttttaactccaatcatggaaatcatcttttgcgatt
gccgtgtttgtggaatgcgggagttaaacgatcgcggagcggggaatataaaagagggcaatctgaaaggaagggaaaatcc
ttcggattgcttttagttttatgggagtgaaatattatagcattacggaaatgataatggcagagttttcattattagactgctga
tgaattgatgatgatacaaaaataatgtgtgtaacaaaatgttaacaaaaagacaaattcattcatagttgataactgataaag
attgtgaaataatgcacaatatatcaatgtatgagcagttcacaattcatttttggaaaggatgacagacagcgATCgtaaggcg
aagagctgtcactggtgctgccctattctgggtggaactggatggtgatgtcaacggctcataagtttccgtgctggtggcaggggtgaagg
gacgcaactaatgtaaacgacgctgaagtcatctgtactactggtaaactgccggtacctggccgactctggtaacgacgctgactt
atggtgtcagtgcttctcgttatccggaccatgaagcagcatgacttctcaagtcgcatgcccgaaggctatgtgcaggaacg
cagatttcttaaggatgacggcagctacaaaacgctgcaaggtgaaattgaaaggcagataccctgttaaaccgattgagctga
aaggcattgacttaagaagacggcaatattcctggccataagctggaatacaatfttaacagccacaatgtttacatcaccgccgata
aacaaaaaatggcattaaagcgaatttaaaatcggcacaacgtggaggatggcagcgtgacagctggctgactactaccagcaaa
acactcaatcggatggtcctgttctgctgcccagacaatcactatctgagcagcaaaagcgttctgtctaaagatccgaacgagaac
ggatcatatggtctgctgagttcgaaccgacgggcatcagcaggtgatggatgaactgtacaaaTGATAAcaagattccgc
tctgtcaagtgaatfttcaagtggtcggaccaatgtacatgctatatttggcgagcatggggatacagagctgctgtttggagc
catcggaattggaagcattccagttgagcaaatattgatgcaaacgataactatagaaaagaggatttagacaataatcttgttaag

ATGtatacagtaggagattacctattagaccgattacacgagttaggaattgaagaaatTTTTGGAGTCCCTGGAGACTATAACTTCAATTTTAGATCAAATTATTTCCACAAGGATATGAAATGGGTCGGAAATGCTAATGAATTAATGCTCATATATGGCTGATGGCTATGCTCGTACTAAAAAGCTGCCGATTCTTACAACCTTTGGAGTAGGTGAATTGAGTGCAGTTAATGGATTAGCAGGAAGTTACGCCGAAAATTTACCAGTAGTAGAAATAGTGGGATCACCTACATCAAAAGTTCAAAATGAAGGAAAATTTGTTTCATCATACGCTGGCTGACGGTGATTTAAACACTTTATGAAAATGCACGAACCTGTACAGCAGCTCGAACTTACTGACAGCAGAAAATGCAACCGTTGAAATGACCGAGTACTTTCTGCACTATTAAGAAAGAAAACCTGTCTATCAACTACCAGTTGATGTTGCTGCTGCAAAAGCAGAGAACCCTCACTCCCTTGAAAAAGGAAAACCTCAACTCAAATACAAGTGACCAAGAAATTTGAACAAAATCAAGAAAGCTGAAAAATGCCAAAAACCAATCGTGATTACAGGACATGAAATAATTAGTTTTGGCTTAGAAAAACAGTCACTCAATTTATTCAAAGACAAAACCTATTACGACATTAACCTTTGGTAAAAGTTCAGTTGATGAAGCCCTCCCTTCAATTTAGGAATCTATAATGGTACTCTCAGAGCCTAATCTTAAAGAATCGTGGAAATCAGCCGACTCATCTTGATGCTTGGAGTTAAACTCACAGACTCTCAACAGGAGCCTTCACTCATCAATTAATGAAAATAAATGATTTCACTGAATATAGATGAAGGAAAAATTTAACGAAAGAAATCCAAAATTTGATTTGAAATCCCTCATCTCTCTCTTAGACCTAAGCGAAATAGAATACAAAGGAAAATATATCGATAAAAAGCAAGAAGACTTTGTTCCATCAAATGCGCTTTTATCACAAGACCGCCTATGGCAAGCAGTTGAAAACCTAACTCAAAGCAATGAAACAATCGTTGCTGAACAAGGGACATCACTTTGGCGCTTCAATTTCTTAAATCAAAGAGTCATTTATTGGTCAACCCCTATGGGGATCAATTGGATACATCCAGCAGCATTAGGAAGCCAAATGCAGATAAAGAAAGCAGACACCTTTATTATTGGTATGGTCACTCAACTACAGTCAA GAATTAGGATTAGCAATCAGAGAAAAAATTAATCCAATTTGCTTTATTATCAATAATGATGGTTACAGTCGAAAGAGAAATCATGACCAAATCAAAGTCAATGATATCCAATGTGGAATTACTCAAATACCAGAATCGTTGGAGCAACAGAAGATCGAGTAGTCTCAAAAATCGTTAGAACTGAAAATGAATTTGTCTGTGATGAAAGAAGCTCAAGCAGATCCAAATAGAATGTACTGGATTGAGTTAATTTGGCAAAGAAGGTGCACCAAAGTACTGAAAAAATGGGCAAACCTATTGCTGAACAAAATAATCA

Figure A132. *Lactococcus lactis* subsp. *lactis* *kivd* gene for alpha-ketoisovalerate decarboxylase, strain IFPL730. GenBank: AJ746364.1

ATGtatacggtcggcgattatftgtggatcgcttgcattggaatgggcatcgaagaaatfttggcgtccgggagattacaacttgc
aatfttggatcaaatcatcagccataaagatatgaaatgggtcggcaacgccaacgaattgaacgccagctatatggccgatgg
ctatgctcgcacgaaaaagccgctgctfttggacgacgttggcgtcggcgaattgagcgtgtcaacggctggctggcagct
atgccgaaaactgcccgtcgtcgaattgctggcagcccagcagcaaagtccaaaacgaaggcaaattgtccatcatacgt
tggccgatggcgacttaaacatttcatgaaaatgcatgaaccggcagcggctgctcgcacgttgtgacggctgaaaacgccac
ggtcgaattgatcgcgtctgagcgccttgtgaaagaacgcaaaccggctatattaactgcccgtggatgttggcgtgcca
agccgaaaaaccgagcttggcgtgaaaaaagaaaaacagcagcagcaacacgagcagatcaagaaatfttgaacaaaatcc
aagaaagcttgaaaaacgcaaaaaaccgattgtcattacgggacatgaaatcattagcttggcgttggaaaaacggtcacg
caattatcagcaaacgaaattgccgattacgacgttgaactfttggcaaaagcagcgtcagatgaagccttggcagctfttggc
attataacggcagcttggcgaaccgaactgaaagaattgtcgaagcgtgactttatcttgaattgttggcgtcaaattgacgg
atagcagcacaggcgttftacgcatcatttgaacgaaacaaaatgatcagcctgaacatcgacgaaggcaaaatctttaa
aacgcatccaaaactftgactcgaagcttgaattgacagcttgttggattgagcgaatcgagtataaaggcaaatatcgcata
aaaaacaagaagatfttgcctgagcaacgccttgttggccaagatcgccttggcaagccgtcgaaaactgacgcaagca
acgaaacgattgtcgcgaacaaggcagcagctfttggagccagcagcattfttggaaaagcaaaagccatttgcgcca
ccgttggggcagcattggctatacgttccggctgccttggcagccaaattgccgataaagaatctcgcatttgttattggc
gacggcagcttgaattgacggcgaagaattggcgttggcattcgcgaaaaatcaaccgatttgccttatacattaacaacgat
ggctatacggtcgaacgcgaaattcatggaccgaaccaagctataacgatattccgatgtggaactatagcaaatgcccggaa
agcttggagccacggaagatcgcgtcgcagcaaaatgtccgcagcaaaacgaattgtcagcgtcatgaaagaagcca
agccgatccgaaccgcatgtattgattgatttggccaaagaaggcgtccgaaagtctgaaaaaatgggcaaatgt
tcgccaacaaaacaaaagcTAA

Figure A133. Codon-optimized for *G. stearothermophilus* version of the *Lactococcus lactis* subsp. *lactis* *kivd* gene for alpha-ketoisovalerate decarboxylase, strain IFPL730. GenBank: AJ746364.1

taatacgactcactatagggcgaattACTAGCCCAACTCTTTAAGGAGGTTACAATG tacaacggtcggcg
attatttggatcgcttgcataaattgggcatcgaagaaatggggtccgggagattacaacttgcaattttgatcaaatcatc
agccataaagatatgaaatgggtcggcaacgccaacgaattgaacgccagctatatggccgatggctatgctcgacgaaaa
agccgctgcctttgacgacgcttggcgtcggcgaattgagcgtgtcaacggctggctggcagctatgccgaaaactgcccgt
cgtcgaattgtcggcagccccgacgagcaaatgcaaaaacgaaggcaaatgtccatcatacgttggccgatggcgacttaa
acatttcatgaaaatgcatgaaccggtcacggctgctcgacgctgttgacggctgaaaacgccaggtcgaattgatcgcgtct
tgagcgccttgttgaagaacgcaaaccggctatattaactgcccgggtgatgttgcgctgccaagccgaaaaaccgagctt
gccgtgaaaaaagaaaacagcacgagcaacacgagcgatcaagaaatgtgaacaaaatccaagaaagctgaaaaacg
caaaaaaccgattgtcattacgggacatgaaatcattagcttggcgttggaaaaacggtcacgcaattatcagcaaacga
aattgccgattacgacgttgaacttggcaaaagcagcgtcgatgaagccttgcgagcttttggcattataacggcagctga
gcgaaccgaactgaaagaattgtcgaagcgtgacttattctgtatgttggcgctcaaattgacggatagcagcacaggcgct
ttacgcatcattgaacgaaaacaaaatgatcagcctgaacatcgacgaaggcaaaatcttaacgaacgcatacaaaacttg
actcgaagccttgattgacgcttggattgagcgaatcagagtataaaggcaaatatcagataaaaaacaagaagattttg
tcccgagcaacgccttgttgcgaagatcgcttggcaagccgctcgaaaactgacgcaaacgaaacgattgtcgcgg
aacaaggcagcagcttttggagccagcagcatttttgaagcaaaagccattttatcggccaaccgttggggcagcattg
gctatacgttccggctgccttgggagccaaatgccgataaagaatctcgccatttggtttattggcgacggcagctgcaattg
acggccaagaattgggctggccattcgcgaaaaaatcaaccgatttcttatacattacaacgatggctatacggtcgaacg
cgaattcatggaccgaaccaaagctataacgatattccgatgtggaactatagcaaatgccggaagcttggagccacgga
agatcgcgtcgtcagcaaaatgtccgcacggaaaacgaattgtcagcgtcatgaaagaagcccaagccgatccgaaccgc
atgtattggattgaattgatttggccaaagaaggcgtccgaaagcttgaaaaaatgggcaaatgttcgcggaacaaaaaca
aaagcCATCACCATCACCACTAAgagctcctggcctcatgggccttcttctactgccgcttccagtcggga
aacctgctgtccagctgcattaacatggcatagctgttctt

T7 promoter

RBS

kdc

6X HIS-TAG

M13R

Figure A134. Sequence detail of the T7 promoter+RBS+*kdc*+his-tag construct used to express Kdc under the control of the T7 promoter in *E. coli*.

ACGGCCAAGTGAGCGCGACGTAATACGACTCACTATAGGGCGAATTACTAGCCCAACACTCTTTAAGGAGGTTACAATGTATACGGTCGGCGATTATTTGT
M13-fwd T7 KDC
template sequence vectorKDC

ACGGCCAAGTGAGCGCGACGTAATACGACTCACTATAGGGCGAATTACTAGCCCAACACTCTTTAAGGAGGTTACAATGTATACGGTCGGCGATTATTTGT
aligned sequence seq1

ACGGCCAAGTGAGCGCGACGTAATACGACTCACTATAGGGCGAATTACTAGCCCAACACTCTTTAAGGAGGTTACAATGTATACGGTCGGCGATTATTTGT
aligned sequence seq2

TGGATCGCTTGCATGAATTGGGCATCGAAGAAATTTTGGCGTCCGGGAGATTACAACCTGCAATTTTGGATCAAATCATCAGCCATAAAGATATGAA
KDC
template sequence vectorKDC

TGGATCGCTTGCATGAATTGGGCATCGAAGAAATTTTGGCGTCCGGGAGATTACAACCTGCAATTTTGGATCAAATCATCAGCCATAAAGATATGAA
aligned sequence seq1

TGGATCGCTTGCATGAATTGGGCATCGAAGAAATTTTGGCGTCCGGGAGATTACAACCTGCAATTTTGGATCAAATCATCAGCCATAAAGATATGAA
aligned sequence seq2

ATGGGTCGGCAACGCCAACGAATTGAACGCCAGCTATATGGCCGATGGCTATGCTCGCACGAAAAAGCCGCTGCCTTTTACGACGTTTGGCGTCGGC
KDC
template sequence vectorKDC

ATGGGTCGGCAACGCCAACGAATTGAACGCCAGCTATATGGCCGATGGCTATGCTCGCACGAAAAAGCCGCTGCCTTTTACGACGTTTGGCGTCGGC
aligned sequence seq1

ATGGGTCGGCAACGCCAACGAATTGAACGCCAGCTATATGGCCGATGGCTATGCTCGCACGAAAAAGCCGCTGCCTTTTACGACGTTTGGCGTCGGC
aligned sequence seq2

CCGAGCTTGCCGTTGAAAAAGAAAAAGCAGCACGAGCAACACGAGCGATCAAGAAATTTTGAACAAAATCCAAGAAAGCTTGAAAAACGCCAAAAACCGA
KDC
template sequence vectorKDC

CCGAGCTTGCCGTTGAAAAAGAAAAAGCAGCACGAGCAACACGAGCGATCAAGAAATTTTGAACAAAATCCAAGAAAGCTTGAAAAACGCCAAAAACCGA
aligned sequence seq1

CCGAGCTTGCCGTTGAAAAAGAAAAAGCAGCACGAGCAACACGAGCGATCAAGAAATTTTGAACAAAATCCAAGAAAGCTTGAAAAACGCCAAAAACCGA
aligned sequence seq2

TTGTCATTACGGGACATGAAATCATTAGCTTCGGCTTGAAAAACGGTCACGCAATTTATCAGCAAAACGAAATGCGGATTACGACGTTGAACTTTGG
KDC
template sequence vectorKDC

TTGTCATTACGGGACATGAAATCATTAGCTTCGGCTTGAAAAACGGTCACGCAATTTATCAGCAAAACGAAATGCGGATTACGACGTTGAACTTTGG
aligned sequence seq1

TTGTCATTACGGGACATGAAATCATTAGCTTCGGCTTGAAAAACGGTCACGCAATTTATCAGCAAAACGAAATGCGGATTACGACGTTGAACTTTGG
aligned sequence seq2

CAAAAGCAGCGTCGATGAAGCCTTGCCGAGCTTTTGGGCATTTATAACGGCACGTTGAGCGAACCGAACTTGAAAGAATTTGTCGAAAGCGCTGACTTT
KDC
template sequence vectorKDC

CAAAAGCAGCGTCGATGAAGCCTTGCCGAGCTTTTGGGCATTTATAACGGCACGTTGAGCGAACCGAACTTGAAAGAATTTGTCGAAAGCGCTGACTTT
aligned sequence seq1

CAAAAGCAGCGTCGATGAAGCCTTGCCGAGCTTTTGGGCATTTATAACGGCACGTTGAGCGAACCGAACTTGAAAGAATTTGTCGAAAGCGCTGACTTT
aligned sequence seq2

ATCTTGATGTTGGCGTCAAATTGACGGATAGCAGCACAGGCGCTTTTACGCATCATTGAACGAAAAACAAAATGATCAGCCTGAACATCGACGAAGGCA
KDC

template sequence vectorKDC

ATCTTGATGTTGGCGTCAAATTGACGGATAGCAGCACAGGCGCTTTTACGCATCATTGAACGAAAAACAAAATGATCAGCCTGAACATCGACGAAGGCA
aligned sequence seq1

ATCTTGATG-----
aligned sequence seq2

AAATCTTTAACGAACGCATCCAAAACCTTGACTTCGAAAGCTTGATTAGCAGCTTGTGGATTGAGCGAAATCGAGTATAAAGGCAAATATATCGATAA
KDC

template sequence vectorKDC

AAATCTTTAACGAACGCATCCAAAACCTTGACTTCGAAAGCTTGATTAGCAGCTTGTGGATTGAGCGAAATCGAGTATAAAGGCAAATATATCGATAA
aligned sequence seq1

aligned sequence seq2

AAAACAAGAAGATTTTGTCCCGAGCAACGCCTTGTGAGCCAAGATCGCTTGTGGCAAGCCGTCGAAAACCTGACGCAAAGCAACGAAACGATTGTCGCC
KDC

template sequence vectorKDC

AAAACAAGAAGATTTTGTCCCGAGCAACGCCTTGTGAGCCAAGATCGCTTGTGGCAAGCCGTCGAAAACCTGACGCAAAGCAACGAAACGATTGTCGCC
aligned sequence seq1

aligned sequence seq2

GAACAAGGCACGAGCTTTTTGGAGCCAGCAGCATTTTTTAAAAAGCAAAAGCCATTTATCGGCCAACCGTTGTGGGCAGCATTGGCTATACGTTTC
KDC

template sequence vectorKDC

GAACAAGGCACGAGCTTTTTGGAGCCAGCAGCATTTTTTAAAAAGCAAAAGCCATTTATCGGCCAACCGTTGTGGGCAGCATTGGCTATACGTTTC
aligned sequence seq1

aligned sequence seq2

CGGCTGCCTTGGGCAGCCAAATTGCCGATAAAGAATCTCGCCATTTGTTGTTTATTGGCGACGGCAGCTTGCAATTGACGGTCCAAGAATTGGGCTTGGC
KDC

template sequence vectorKDC

CGGCTGCCTTGGGCAGCCAAATTGCCGATAAAGAATCTCGCCATTTGTTGTTTATTGGCGACGGCAGCTTGCAATTGACGGTCCAAGAATTGGGCTTGGC
aligned sequence seq1

aligned sequence seq2

CATTCGCGAAAAATCAACCCGATTGCTTTATCATTAAACAGATGGCTATACGGTCGAACGCGAAATTCATGGACCGAACCAAAGCTATAACGATATT
KDC

template sequence vectorKDC

CATTCGCGAAAAATCAACCCGATTGCTTTATCATTAAACAGATGGCTATACGGTCGAACGCGAAATTCATGGACCGAACCAAAGCTATAACGATATT
aligned sequence seq1

aligned sequence seq2

```

CCGATGTGGA ACTATAGCAAATTGCCGGAAAGCTTTGGAGCCACGGAAGATCGCGTCGTCAGCAAAATTGTCCGCACGGAAAACGAATTTGTCAGCGTCA
> template sequence vectorKDC
CCGATGTGGA ACTATAGCAAATTGCCGGAAAGCTTTGGAGCCACGGAAGATCGCGTCGTCAGCAAAATTGTCCGCACGGAAAACGAATTTGTCAGCGTCA
aligned sequence seq1
-----
aligned sequence seq2

TGAAAGAAGCCCAAGCCGATCCGAACCGCATGTATTGGATTGAATTGATTTGGCCAAAGAAGGCGCTCCGAAAGTCTTGAAAAAATGGGCAAATTGTT
> template sequence vectorKDC
TGAAAGAAGCCCAAGCCGATCCGAACCGCATGTATTGGATTGAATTGATTTGGCCAAAGAAGGCGCTCCGAAAGTCTTGAAAAAATGGGCAAATTGTT
aligned sequence seq1
-----
aligned sequence seq2

CGCGGAACAAAACAAAAGCCATCACCATCATCACCCTAAGAGCTCCTGGGCCTCATGGGCCTTCCTTTCACTGCCCGCTTCCAGTCGGGAAACCTGTC
> template sequence vectorKDC
CGCGGAACAAAACAAAAGCCATCACCATCATCACCCTAAGAGCTCCTGGGCCTCATGGGCCTTCCTTTCACTGCCCGCTTCCAGTCGGGAAACCTGTC
aligned sequence seq1
-----
aligned sequence seq2

```

Figure A135. Sequencing of the *kdc* gene as in BL21 cells after successive subculturing that resulted in the cells loss of pLysE. Sequencing results of two independent sequencing reactions. seq1 was done using forward primer emk013. seq2 was done using primer M13R.

acetaldehyde dehydrogenase [Thermus thermophilus HB8]
NCBI Reference Sequence: YP_145486.1

MSERVKVAILGSGNIGTDLMYKLLKNPGHMELVAVVGIDPKSEGLARARALGLEASHEGIAYI
LERPEIKIVFDATSAKAHVRHAKLLREAGKIAIDLTPAARGPYVPPVNLKEHLDKDNVNLITC
GGQATIPLVYAVHRVAPVLYAEMVSTVASRSAGPGTRQNIDEFTFTTARGLEAIGGAKKGKA
IIILNPAEPPILMTNTVRCIPEDEGFDREAVVASVRAMEREVQAYVPGYRLKADPVFERLPTP
WGERTVVSMLLEVEGAGDYLPKYAGNLDIMTASARRVGEVFAQHLLGKPVVEEVA

NC_006462.1

ATGTCCGAAAGGGTTAAGGTAGCCATCCTGGGCTCCGGCAACATCGGGACGGACCTGA
TGTACAAGCTCCTGAAGAACCCGGGCCACATGGAGCTTGTGGCGGTGGTGGGGATAGA
CCCCAAGTCCGAGGGCCTGGCCCGGGCGCGGGCCTTAGGGTTAGAGGCGAGCCACG
AAGGGATCGCCTACATCCTGGAGAGGCCGGAGATCAAGATCGTCTTTGACGCCACCAG
CGCCAAGGCCACGTGCGCCACGCCAAGCTCCTGAGGGAGGCGGGGAAGATCGCCAT
AGACCTCACGCCGGCGGCCCGGGCCCTTACGTGGTGCCCCCGGTGAACCTGAAGGA
ACACCTGGACAAGGACAACGTGAACCTCATCACCTGCGGGGGGCAGGCCACCATCCCC
CTGGTCTACGCGGTGCACCGGGTGGCCCCCGTGCTCTACGCGGAGATGGTCTCCACG
GTGGCCTCCCGCTCCGCGGGCCCCGGCACCCGGCAGAACATCGACGAGTTCACCTTC
ACCACCGCCCCGGGGCCTGGAGGCCATCGGGGGGGCCAAGAAGGGGAAGGCCATCAT
CATCCTGAACCCGGCGGAACCCCCATCCTCATGACCAACACCGTGCGCTGCATCCCC
GAGGACGAGGGCTTTGACCGGGAGGCCGTGGTGGCGAGCGTCCGGGCCATGGAGCG
GGAGGTCCAGGCCTACGTGCCCGGCTACCGCCTGAAGGCGGACCCGGTGTGGAGAG
GCTTCCCACCCCCTGGGGGGAGCGCACCGTGGTCTCCATGCTCCTGGAGGTGGAGGG
GGCGGGGGACTATTTGCCCAAATACGCCGGCAACCTGGACATCATGACGGCTTCTGCC
CGGAGGGTGGGGGAGGTCTTCGCCCAGCACCTCCTGGGGAAGCCCGTGGAGGAGGT
GGTGGCGTGA

Figure A136. Sequence detail of the ALDH gene from *Thermus thermophilus* as obtained from NCBI.

acetaldehyde dehydrogenase [*Geobacillus* sp. Y4.1MC1]

NCBI Reference Sequence: YP_003989256.1

MKDSFSTQIRARVSVLLNVKVRLLKGNFMSKVKVGILGSGNIGTDLMYKILNNDGHMELSIIA
GIVPESEGLMRARSMGIRASHEGIAAILEDPDVRIVFDATSAKAHRQNAEALRAAGKIVIDLTP
AAIGPFVPPINLNEHINAWNVMISCGGQATIPLVYAVSRIVSVQYAEVITTTASRSIGPGTR
QNMDEFTYTTAQGLQSIGGAKIARAIPVINSADPTIIMTNTVYAAVEEDFDEKAILSIDMVDE
VSKYVPGYRLKATPFIDYRDPWGRLPVVILNEVEGSEDFLPAYAGNLDIMTASARRVGES
YAQHLLKMKEVRR

ATGAAAGATTTCATTTTCTACACAGATCAGAGCCAGAGTATCGGTCCTGTTGAACGTGAAA
GTCAGACTGAAAGGGGACAATTTTCATGAGCAAAGTGAAAGTTGGAATCTTAGGTTCTGG
GAATATTGGAACGGATTTGATGTACAAAATCCTTAATAATGACGGCCATATGGAATTATC
TATAATAGCTGGAATCGTACCGGAATCAGAAGGTCTTATGAGAGCGCGAAGCATGGGCA
TACGAGCCAGCCATGAAGGAATAGCGGCGATTTTGGAGGATCCCGATGTACGGATTGT
CTTTGACGCTACGAGTGCAAAGGCACATCGCCAAAATGCGGAAGCGTTACGTGCAGCA
GGAAAGATTGTGATAGATTTGACGCCTGCCGCAATTGGCCCTTTCGTGGTGCCTCCCAT
CAATTTGAATGAGCATATAAATGCTTGAATGTAAACATGATTTTCGTGCGGAGGACAGG
CGACAATTCCTCTCGTTTATGCAGTAAGTCGCATTGTTTCTGTCCAATATGCGGAAGTCA
TCACAACCTACCGCCAGCCGCTCCATAGGCCCCGGAACAAGACAAAATATGGATGAGTTT
ACCTATACTACCGCACAGGGGCTGCAAAGCATCGGTGGTGCGAAGATAGCCAGGGCAA
TTCCGGTGATTAATTCAGCAGATCCCAACAATCATTATGACAAATACCGTTTATGCCGCAG
TGGAAGAGGATTTTGTGAGAAAGCGATCATTGTTCCATTAAGGACATGGTTCGATGAA
GTGTCCAAATATGTTCCCTGGATATCGCCTCAAAGCGACTCCATTTATCGATTATCGTGAC
ACACCATGGGGAAGATTGCCTGTCGTTGTTATCTTAAATGAAGTCGAAGGCTCGGAAGA
TTTTCTTCCAGCATATGCTGGTAATTTAGACATTATGACAGCTTCGGCACGGCGTGTGG
CGAGTCATATGCCAGCACCTTCTAAAATGAAGGAGGTGAGACGATGA

Figure A137. Sequence detail of the ALDH gene from *Geobacillus* sp. Y4.1MC1 as obtained from NCBI.

AscI – RsrII – Rho2 terminator – PacI – (SpeI AvrII NheI) – SexAI – XmaI – (Sall XhoI) – PstI/SbfI – RplS Promoter – XbaI – RBS – OPT_Tt_KDH – SacI – Rho1 – Bsu36I – SwaI

GGCGCGCCtctaCGGACCGagtaattcaacaaacgggcccagtttggtgaaatcgTTAATTAAgcgtACTAGTC
CTAGGCTAGCtgatACCAGGTcagtCCCGGGacagGTCGACCTCGAGtacgCCTGCAGGAACA
ATCGTTAAAGCGGACGTTTTTTCGCGCCGCCCGGATTTGCTTGAAAACCTACCCGCTGACAG
AAAAGCAAAAACGATGGATCGAAGAGTGGAAAAAAGAAAAACAGTAGCTATTGCGCATG
ATACAAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTGCTTAAGCGAGGAAA
ACGATGTTCCGCTGCAATGATGAAAAAGCATTGTCTAGAgtcgggtataacacgagaagg
AGTAGAAGTACATGAGCGAACGCGTCAAAGTCGCCATTTTGGGCAGCGGCAACATTGG
CACGGATTTGATGTATAAATTGTTGAAAAATCCGGGACATATGGAATTGGTCGCCGTGG
TCGGCATTGATCCGAAAAGCGAAGGCTTGGCTCGCGCTCGCGCTTTGGGCTTGGAAAGC
CAGCCATGAAGGCATTGCCTATATTTTGGAAACGCCAGAAATCAAATTTGCTTTGATGC
CACGAGCGCCAAAGCCCATGTCCGCCATGCCAAATTGTTGCGCGAAGCCGGAAAAATT
GCCATTGATTTGACACCGGCTGCTCGGGGACCGTATGTCGTTCCACCGGTCAACTTGAA
AGAACATTTGGATAAAGACAACGTCAACTTGATTACGTGCGGAGGCCAAGCCACGATTC
CGTTGGTCTATGCCGTCCATCGGGTCGCTCCGGTCTTGTATGCCGAAATGGTCAGCAC
GGTCGCTTCGCGCAGCGCTGGACCGGGAACACGCCAAAACATTGATGAATTTACGTTTA
CGACGGCTCGCGGATTGGAAGCCATTGGCGGAGCCAAAAAAGGCAAAGCCATTATTAT
CTTGAATCCGGCTGAACCACCGATTTTGTGACGAACACGGTGCCTGCATTCCGGAA
GATGAAGGCTTTGATCGCGAAGCCGTCGTCGCCAGCGTTCGCGCTATGGAACGCGAAG
TCCAAGCCTATGTTCCGGGATATCGCTTGAAAGCCGATCCGGTCTTTGAACGCTTGCCG
ACGCCGTGGGGAGAACGCACGGTCGTCAGCATGTTGTTGGAAGTCGAAGGCGCTGGC
GATTATTTGCCGAAATATGCCGAAACTTGGATATTATGACGGCCAGCGCACGCCGCGT
CGGCGAAGTCTTTGCCCAACATTTGCTGGGCAAACCGGTGCAAGAAGTCGTCGCCAC
CATCACCATCATCATTAAATAAGAGCTCaatttcgccccgaaagggcgTTTTTtgcgaagCCTAAGGtat
cATTTAAAT

Figure A138. Sequence detail of the ALDH gene from *Thermus thermophilus* as obtained from NCBI after codon optimization by GeneART, and placed in the MCS of vector p11AK1, under the control of the RplS promoter and a synthetic RBS.

CTAGAGTCGGTATAACACGAGAAGGAGTAGAAGTACATGAGCGAACCGCTCAAAGTCGCCATTTGGGCAGCGGCAACATTGGCACGGATTGATGTATA
Tt_ALDH

template sequence P6 Tt_KDH

-----TTTTGGGCAGCGGCAACATTGGCACGGATTGATGTATA
aligned sequence seq1

AATTGTTGAAAAATCCGGGACATATGGAATTGGTCGCCGTGGTCGGCATTGATCCGAAAAGCGAAGGCTTGGCTCGCGCTCGCGCTTTGGGCTTGAAGC
Tt_ALDH

template sequence P6 Tt_KDH

AATTGTTGAAAAATCCGGGACATATGGAATTGGTCGCCGTGGTCGGCATTGATCCGAAAAGCGAAGGCTTGGCTCGCGCTCGCGCTTTGGGCTTGAAGC
aligned sequence seq1

CAGCCATGAAGCATTGCCTATATTTGGAACGCCAGAAATCAAAATTGCTTTGATGCCACGAGCGCCAAAGCCCATGTCGCCATGCCAAATTTGTTG
Tt_ALDH

template sequence P6 Tt_KDH

CAGCCATGAAGCATTGCCTATATTTGGAACGCCAGAAATCAAAATTGCTTTGATGCCACGAGCGCCAAAGCCCATGTCGCCATGCCAAATTTGTTG
aligned sequence seq1

CGGAAGCCGAAAAATGCCATTGATTTGACACCGCTGCTCGGGACCGTATGTCGTTCCACCGGTCAACTTGAAAGAACATTTGGATAAAGACAACG
Tt_ALDH

template sequence P6 Tt_KDH

CGGAAGCCGAAAAATGCCATTGATTTGACACCGCTGCTCGGGACCGTATGTCGTTCCACCGGTCAACTTGAAAGAACATTTGGATAAAGACAACG
aligned sequence seq1

TCAACTTGATTACGTGCGGAGGCCAAGCCACGATTCGGTTGGTCTATGCCGTCATCGGGTCGCTCCGGTCTTGATGCCGAAATGGTCAGCACGGTCGC
Tt_ALDH

template sequence P6 Tt_KDH

TCAACTTGATTACGTGCGGAGGCCAAGCCACGATTCGGTTGGTCTATGCCGTCATCGGGTCGCTCCGGTCTTGATGCCGAAATGGTCAGCACGGTCGC
aligned sequence seq1

TTCCGCGACGCTGGACCGGGAACACGCCAAAACATTGATGAATTTACGTTTACGACGGCTCGCGGATTGGAAGCCATTGGCGGAGCCAAAAAGGCAAA
Tt_ALDH

template sequence P6 Tt_KDH

TTCCGCGACGCTGGACCGGGAACACGCCAAAACATTGATGAATTTACGTTTACGACGGCTCGCGGATTGGAAGCCATTGGCGGAGCCAAAAAGGCAAA
aligned sequence seq1

GCCATTATTATCTTGAATCCGGCTGAACCACCGATTTTGTGACGAACACGGTGCCTGCATTCCGGAAGATGAAGGCTTTGATCGCGAAGCCGTCGTCG
Tt_ALDH

template sequence P6 Tt_KDH

GCCATTATTATCTTGAATCCGGCTGAACCACCGATTTTGTGACGAACACGGTGCCTGCATTCCGGAAGATGAAGGCTTTGATCGCGAAGCCGTCGTCG
aligned sequence seq1

CCAGCGTTCGCGCTATGGAACGCGAAGTCCAAGCCTATGTTCCGGGATATCGCTTGAAGCCGATCCGGTCTTTGAACGCTTGCCGACGCCGTGGGAGA
Tt_ALDH

template sequence P6 Tt_KDH

CCAGCGTTCGCGCTATGGAACGCGAAGTCCAAGCCTATGTTCCGGGATATCGCTTGAAGCCGATCCGGTCTTTGAACGCTTGCCGACGCCGTGGGAGA
aligned sequence seq1



Figure A139. Sequencing results of the *T. thermophilus* ALDH gene after it was incorporated to plasmid p11AK+Tt_ALDH.

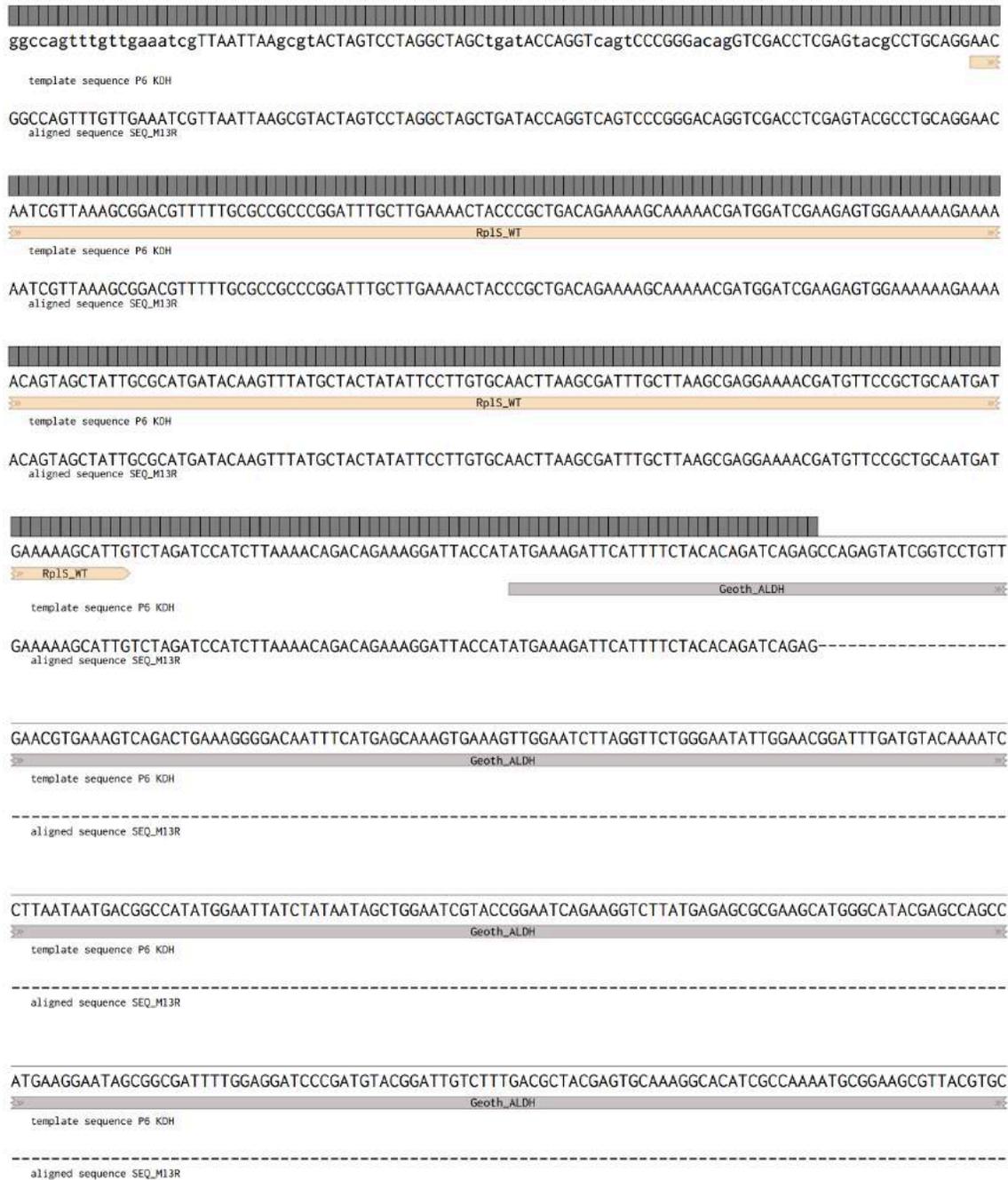


Figure A140. Sequencing results of the *G. thermoglucosidans* ALDH gene after it was incorporated to plasmid p11AK+Tt_ALDH. Primer used was M13R.

Glucose standard 0.05M in water

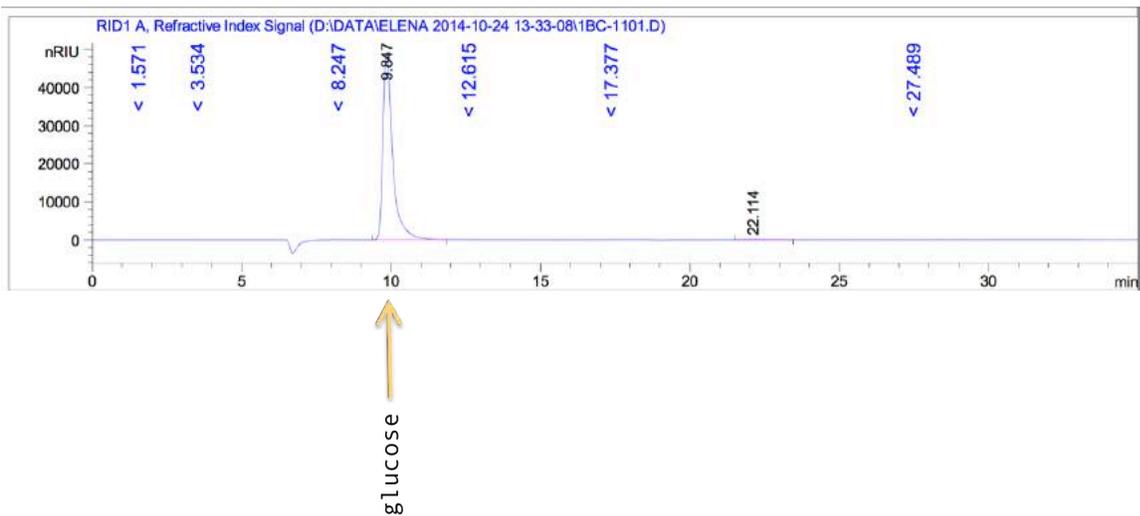


Figure A141. HPLC chromatogram of glucose standard 0.05 M. The peak is observed at a retention time of 9.847 minutes.

2-ketoisovalerate standard 0.05M in water

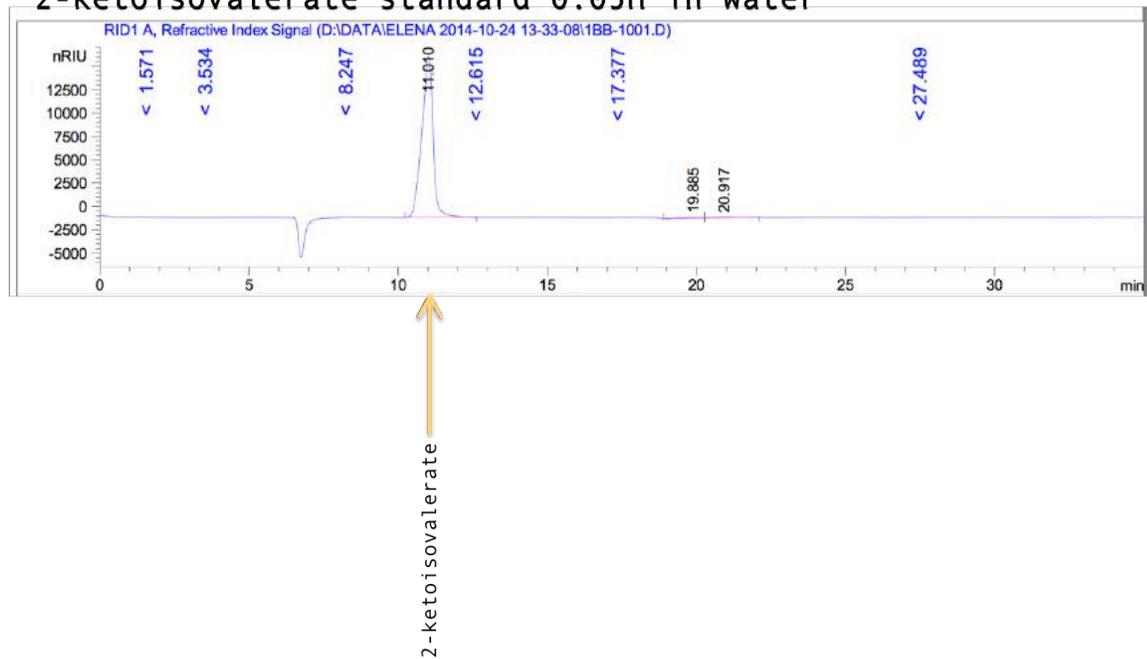


Figure A142. HPLC chromatogram of 2-ketoisovalerate standard 0.05 M. The peak is observed at a retention time of 11.010 minutes.

Components of sterile ASM medium without additional glucose or 2-ketoisovalerate

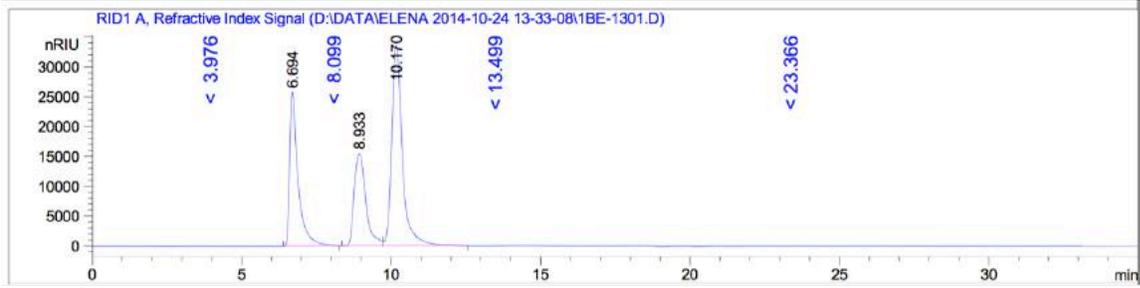


Figure A143. HPLC chromatogram of sterile ASM without additional glucose or 2-ketoisovalerate.

Components of sterile ASM medium supplemented with 0.05M ketoisovalerate and 0.1M glucose

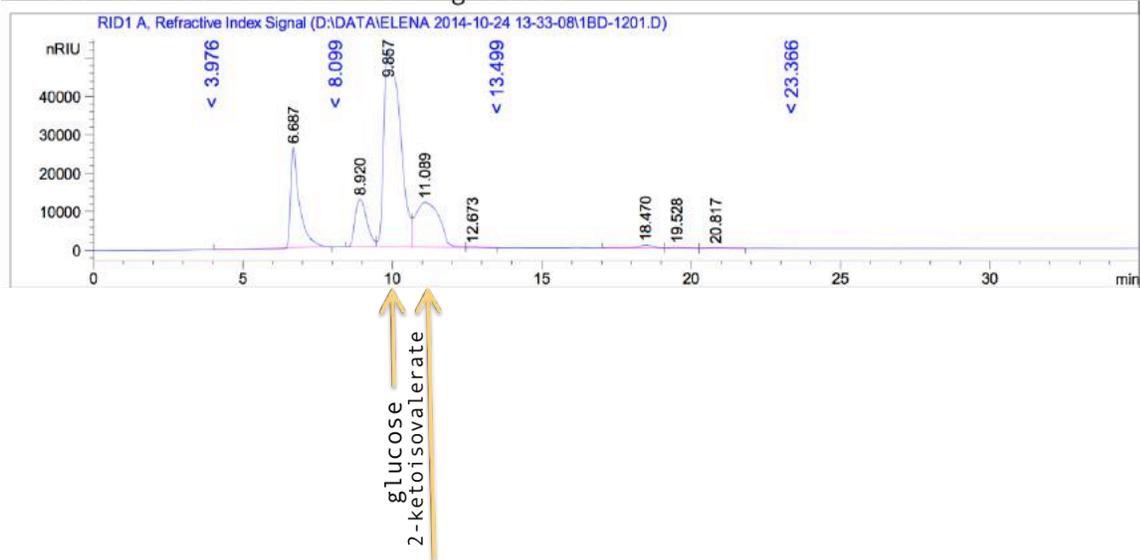


Figure A144. HPLC chromatogram of sterile ASM that has been supplemented with 0.05M ketoisovalerate and 0.1M glucose.

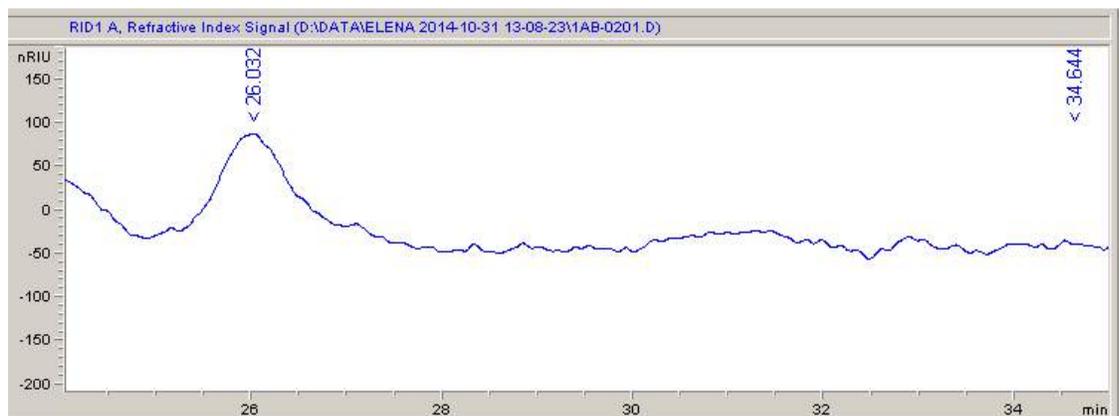


Figure A155. Close up of the isobutyraldehyde peak (26.032 min) and the isobutanol peak (34.644 min) observed from cultures of DL44 grown in the presence of 0.05M 2-ketoisovalerate.



Figure A156. Colony-PCR screening of *E. coli* DH10B transformant colonies. Lane 1, molecular size marker (1 kb GeneRuler, Fermentas). Lanes 2-9, seven colonies transformed with p11AK+Geoth_ALDH were used as template.

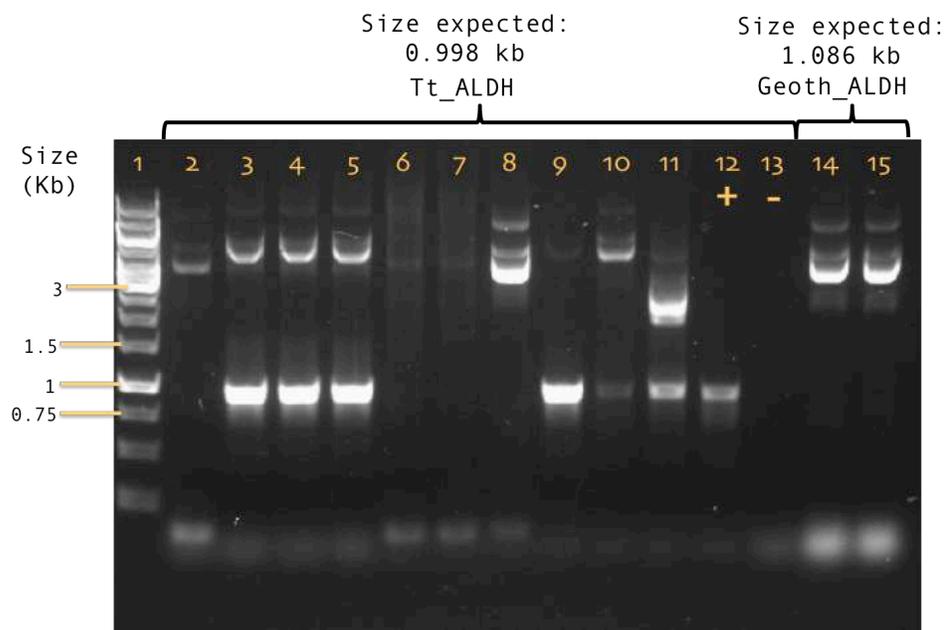


Figure A157. Colony-PCR screening of *G. thermoglucosidans* DL44 transformant colonies. Lane 1, molecular size marker (1 kb GeneRuler, Fermentas). Lanes 2-10, eight colonies transformed with p11AK+Tt_ALDH. Lane 12, positive control for Tt_ALDH. Lane 13, negative control for Tt_ALDH. The primers used were Tt_ALDH and RC-Tt_ALDH. Lanes 14-15, two DL44 colonies transformed with plasmid p11AK+Geoth_ALDH (continues in Figure AXX).

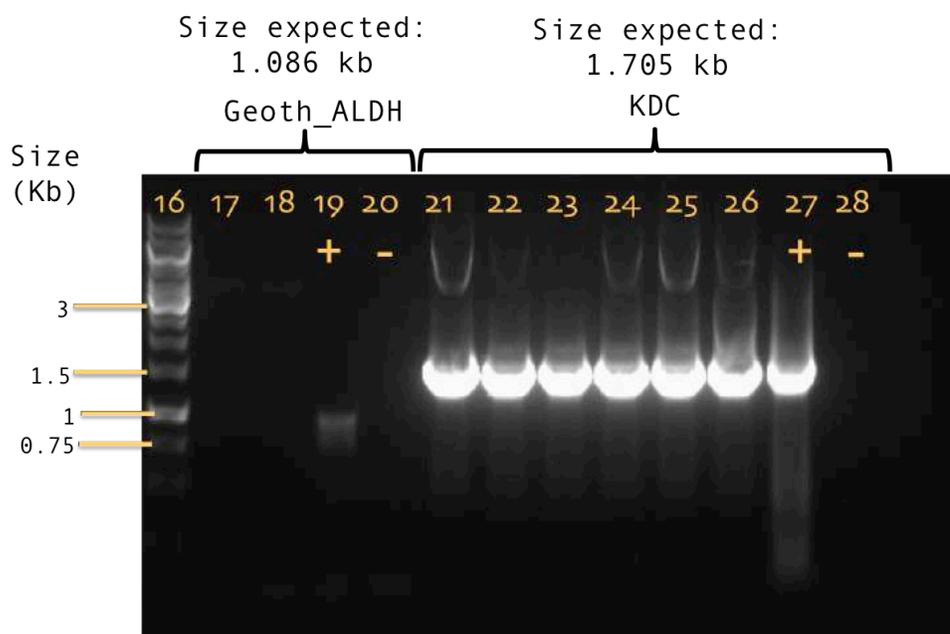


Figure A158. Continuation of Figure A157. Lane 16, molecular size marker (1 kb GeneRuler, Fermentas). Lanes 17-18, two more DL44 colonies transformed with plasmid p11AK+Geoth_ALDH. Lane 19, positive control for Geoth_ALDH. Lane 20, negative control for Geoth_ALDH. Primers used were Geoth_ALDH and RC-Geoth_ALDH. Lanes 21-26, six colonies of DL44 transformed with p11AK+kdc. Lane 27, positive control for kdc. Lane 28, negative control for kdc. Primers used were: XbaI-RBS-kdc and RC-SacI-kdc.

