DEVELOPMENTS IN GENETIC ENGINEERING OF NOVEL ACETOGENS

Robert Mansfield



JANUARY 1, 2018

UNIVERSITY OF NOTTINGHAM

Declaration

Unless otherwise acknowledged, the work presented in this thesis is my own. No part has been submitted for another degree in the University of Nottingham or any other institute of learning.

Robert Patrick William Mansfield

2017

Acknowledgements

Many thanks to Professor Nigel Minton for his support and guidance during this research project. My thanks to the University of Nottingham and the BBSRC for supporting and funding for this research project, without which this PhD would not have been possible.

I would like to express my eternal appreciation for my fellow PhD students and colleagues in the B60 lab and C02 office, for their valuable assistance and fantastic company over the past four years. Beyond the academic domain, I would like to thank my family and friends for their unwavering support and encouragement.

Heroes, without exception.

Abstract

The development of processes for sustainable energy and chemical production is of great importance for the health of our planet. Utilising suitable feedstocks such as renewable resources and existing waste streams is central to making such processes a reality. Microorganisms can be employed in the processing of a diverse range of such feedstocks, offering unique routes of production for useful and valuable chemical products. Acetogenic organisms, capable of fermenting single carbon (C1) feedstocks are especially interesting from the perspective of industrial application. Their natural metabolism and biochemistry enables fixation of low energy C1 compounds, under conditions which would typically be unfeasible with traditional chemical catalysts. Genetic research and development of new acetogenic species opens the door to industrially feasible, and economically attractive, sustainable bioprocessing.

This study outlines the genetic development of the methanol and syngas fermenting acetogen *E. limosum*. This includes establishing gene-transfer and genetic engineering tools in this organism for first time. Additionally, we demonstrate the genetic engineering of synthetic metabolic pathways in this strain to enable production of valuable chemicals, acetone and isopropanol.

Gene-transfer is a cornerstone of modern genetic research in microorganisms, and so effective methods of establishing it are of significant value. We present the development of an improved methodology for enabling and enhancing gene transfer in recalcitrant microorganisms which contain active restriction modification (RM) systems. The method harnesses lambda-red recombineering to support the rapid creation of tailored methylation donor (MD) strains for the preparation and protection of transforming plasmids. The process is uniquely designed in a manner which enables compatibility with both electroporation and conjugation methods of gene transfer.

Abbreviations

%	Percentage
^{o}C	Degrees Celsius
µg	Microgram
ACE	Allele-coupled Exchange
AE	Allelic Exchange
C1	Single carbon containing molecule
Cas	CRISPR-associated
CDS	Coding Sequence
Cla	Clarithromycin
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
CFU	Colony Forming Unit
Cm	Chloramphenicol
CO	Carbon Monoxide
CO_2	Carbon Dioxide
CoA	Co-enzyme A
CRISPR	Clustered Regularly Interspersed Palindromic Repeats
DAM	DNA adenine methyltransferase
DCM	DNA cysteine methyltransferase
dH_2O	Distilled water
DMSO	Dimethyl sulfide
DNA	Deoxyribonucleic acid

DSB	Double Strand Break
Erm	Erythromycin
EtOH	Ethanol
g	Grams
g	Gravity
h	Hour
H2	Hydrogen
HA	Homology Arm
HDR	Homology Directed Repair
Kb	Kilobase Pairs
L	Litre
LB	Luria-Bertani medium
LHA	Left Homology Arm
M	Molar
mg	Milligram
MCS	Multiple Cloning Site
NAD(P)+	Oxidised Nicotinamide Adenine Dinucleotide (Phosphate)
NAD(P)H	Reduced Nicotinamide Adenine Dinucleotide (Phosphate)
NHEJ	Non-Homologous End Joining
nm	Nano Meters
OD_{xxx}	Optical Density (wavelength in nm)
ORF	Open Reading Frame
PAM	Protospacer Adjacent Motif
PBS	Phosphate Buffered Saline

PCR	Polymerase Chain Reaction
PYG	Peptone Yeast Glucose media
RBS	Ribosome Binding Site
RHA	Right Homology Arm
RNA	Ribonucleic Acid
SBRC	Synthetic Biology Research Centre
SNP	Single Nucleotide Polymorphism
SOE	Splicing by Overlap Extension
Tm	Thiamphenicol
V	Volts
v/v	Volume per Volume
w/v	Weight per Volume
WT	Wild Type
Ω	Resistance (ohms)

Table of Figures

Figure 1.1.1 – Global energy consumption by sector ---------- 25

Figure 1.1.2 – Annual worldwide petroleum consumption statistics by region between 1980 and 2013 ---------- 26

Figure 1.3.1 – Phylogenetic tree based on pan-genome analysis of 16 species from the clostridiales Order ---------- 32

Figure 1.4.1 – Schematic of the Wood-Ljungdahl pathway ---------- 36

Figure 1.5.1 – A roadmap for gene system development in Clostridium ---------- 40

Figure 1.5.2 – Schematic of allele-coupled exchange (ACE) to restore *pyrE* in Eubacterium limosum 8486 ---------- 52

Figure 1.5.3 – CRISPR/Cas9 complex interacting with DNA. ---------- 53

Figure 1.5.4 – Illustration of Non-homologous end joining (NHEJ) and Homology-directed repair (HDR) mechanisms following a DNA damage ---------- 55

Figure 1.6.1 – Synthetic Isopropanol and Acetone pathway ---------- 57

Figure 2.14.1 – Vector map for pRM03. ---------- 94

Figure 2.14.2 - Vector map for pRM04 ---------- 95

Figure 2.14.3 – Vector map for pSBRCas9_*Pbt*_1 ---------- 96

Figure 2.14.4 – Vector map for pRM01 ---------- 97

Figure 2.14.5 – Vector map for pMTL8J15-Isopropanol (P_{TcdB}) ---------- 98

Figure 2.14.6 – pJ204-adh ---------- 99

Figure 2.14.7 – Vector map for pMTL8J15-Isopropanol-Dockerin-Scaffoldin. --------- 100

Figure 2.14.8 – Vector map for pMTL8J25x_TT_Cbe_thl_CACthlRBS_catP. ---------- 101

Figure 2.14.9 – Vector map for pRBV1. ---------- 102

Figure 2.14.10 – Vector map for pKD46cm ---------- 103

Figure 3.2.1 – Genome atlas for *E. limosum* ATCC 8486 ---------- 119

Figure 3.2.2 – Genome atlas for *E. limosum* ATCC 8486 with CRISPR loci and *cas* operon marked.---------- 122
Figure 3.2.3 – Schematic of the Wood-Ljungdahl pathway and associated downstream pathways---------- 125
Figure 3.2.4 – Graphical alignment of three complete *E. limosum* genomes ---------- 128
Figure 3.2.5 – Alignment schematic for NCBI and SBRC genomes for *E. limosum* ATCC 8486---------- 129
Figure 3.2.6 - Qualitative analysis of growth of *E. limosum* in PYG broth culture with three different antibiotics---------- 133
Figure 3.2.7 – Voltage optimization of electroporation of *E. limosum* ATCC 8486. -- 136
Figure 3.2.8 – Replicon stability assay data for replicons pCB102 and p19* in *E. limosum* ATCC 8486. ---------- 137
Figure 3.2.9 – Screening PCR confirmation of *pyrE* repair and Lac-TcdR knock-in by Allele-Coupled Exchange in *E. limosum* ATCC 8486---------- 139
Figure 3.2.10 – Vector map for pSBRCas9_*Pbt*_1 ---------- 140
Figure 3.2.11 – Screening PCR for confirmation of *pbt*1 knock-out by CRISPR in *E. limosum* ATCC 8486 ---------- 140
Figure 4.1 – Example dockerin scaffoldin schematic ---------- 164
Figure 4.2 – Synthetic metabolic pathway for isopropanol production in *E. limosum* ---------- 167
Figure 4.3 – Schematics of Acetone and Isopropanol production operon cassettes- 168
Figure 4.4 – Vector map of plasmid pMTL8J15-iso-G1---------- 170
Figure 4.5 – Comprehensive growth curve and product profile data for five transformed variants of the *E. limosum* ATCC 8486 mutant i1B ---------- 174
Figure 4.6 – Endpoint data for Generation I and Generation I-Scaffolded, Acetone and Isopropanol cassettes in five transformed variants of the *E. limosum* ATCC 8486 mutant i1B ---------- 176
Figure 4.7 – Example of raw GC chromatagraph output showing ethanol-isopropanol peak overlap ---------- 178

Figure 4.8 – Promoter strength library for mid-exponential growth phase *E. limosum* ATCC 8486 ---------- 180
Figure 4.9 – Colony PCR to confirm the integrity of plasmids used for testing of the Generation I and Generation I-Scaffolded constructs for acetone and isopropanol production in *E. limosum* ATCC 8486-i1B (i1B)---------- 182
Figure 4.10 – Endpoint data for five transformed *E. limosum* ATCC 8486 variants, including the four isopropanol cassettes developed in this study---------- 184
Figure 4.11 – Endpoint carbon distribution for five transformed *E. limosum* ATCC 8486 variants ---------- 185
Figure 4.12 – The molecular structure and formula for the analysed products. ------ 190
Figure 4.13 – Comparative analysis of scaffolded and unscaffolded acetone/isopropanol production cassettes in *C. autoethanogenum* CH20 grown on PETC media with 40mM fructose ---------- 191
Figure 4.14 – Comparative analysis of scaffolded and unscaffolded acetone/isopropanol production cassettes in *C. autoethanogenum* CH20---------- 193
Figure 4.15 - Schematics for the assembly of G1S2, generation-I scaffolded acetone and isopropanol cassettes ---------- 194
Figure 4.16 – A basic illustration of the enzyme scaffolding principle for a hypothetical metabolic pathway containing two enzymatically catalysed reactions.---------- 204
Figure 5.2.1 – Schematic of representative genomic loci for each RM system Type I-IV. ---------- 213
Figure 5.2.2 –The effects of methylation compatibility on transformation efficacy - 219
Figure 5.2.3 – Six-part workflow for the generation of methylation donor strains of *E. coli*---------- 220
Figure 5.2.4 –PCR screening of OrfD specificity subunit rearrangements ---------- 227
Figure 5.2.5 – *C. carboxidivorans* RM system genome schematic ---------- 229
Figure 5.2.6 – pRBV1 holding vector map. ---------- 234
Figure 5.2.7 – Cassette example schematic (RBV-pAN2) ---------- 234
Figure 5.2.8 – PCR screening of recombineered *E. coli* re-InterStellar ---------- 237

Figure 5.2.9 – Engineered methylation donor to *C. acetobutylicum* ATCC 824 electroporation efficiency comparison ---------- 238
Figure 5.2.10 – RFP promoter assay for recombinant araC-P_{BAD} cassette ---------- 239
Figure 5.2.11 – Engineered methylation donor to *C. difficile* DH1916 trans-conjugation efficiency comparison---------- 241
Figure 5.2.12 - Dot blot for methyltransferase activity in recombinant *E. coli* InterStellar-HsdSM ---------- 243
Figure 5.2.13 – Methylation cassette for *C. botulinum* ATCC 19397---------- 244
Figure 5.2.14 – Engineered methylation donor to *C. botulinum* ATCC 19397 trans-conjugation efficiency comparison ---------- 245
Figure 5.2.15 – Engineered methylation donor to *E. limosum* ATCC 8486 electroporation efficiency comparison ---------- 246
Figure 5.2.16 – Gel analysis of miniprep variability from methylation donors expressing OrfA and OrfD---------- 248
Figure 5.2.17 – Dot blot for methyltransferase by recombinant *E. coli* InterStellar-HsdSM ---------- 249
Figure 5.2.18 – Analysis of pMTL83151 with *C. carboxidivorans* RM motif sequences ---------- 250
Figure 7.1.1 – Optimization of genomic DNA (gDNA) extraction in *E. limosum* ------- 270
Figure 7.1.2 – Protein extraction enhancement with salt wash pre-treatment ------- 271
Figure 7.1.3 – Pan genome for *E. limosum* strains ---------- 274
Figure 7.2.1 – HPLC-UV/IR analysis of *E. limosum* ATCC 8486 fermentation---------- 275
Figure 7.3.1 – *P. mirabilis* pKJ439039 (R702) genomic island ---------- 277
Figure 7.3.2 – Arabinose induction test---------- 277

Table of Tables

Table 2.4.1 – Strain details for organisms targeted for research throughout this project 69

Table 2.5.1 – AWM media additives..... 72

Table 2.5.2 – YTF media additives 72

Table 2.5.3 – Working antibiotic concentrations for *E. coli* and Gram-positive species. 73

Table 2.5.4 – Working media supplement concentrations for *E. coli* and Gram-positive species..... 73

Table 2.10.1 – Q5 polymerase standard thermocycler protocol..... 83

Table 2.10.2 – DreamTaq polymerase standard thermocycler instructions. 84

Table 2.13.1 – Table of strains used in the development of recombinant methyltransferase expressing methylation donor strains of *E. coli* (Chapter 5). 91

Table 2.14.1 – Origin of scaffoldin and dockerin genes used in the construction of scaffolded isopropanol cassettes. 100

Table 2.15.1 – Primers for construction of *pyrE* repair vector pRM03 for use in *E. limosum* ATCC 8486 Δ*pyrE* 104

Table 2.15.2 – Primers for construction of pSBRCas9_*Pbt*_1..... 104

Table 2.15.3 – TIIS assembly of Isopropanol operon into pRM01 backbone 105

Table 2.15.4 – Primers for construction of Cel48A and CohA2 BB2 parts..... 105

Table 2.15.5 – Primers for HiFi construction of Adh scaffoldin ORF and scaffold extension to include CohA2 105

Table 2.15.6: Primers for insertion of *thl* gene downstream of *ctfB* in scaffolded and unscaffolded isopropanol cassettes by HiFi 106

Table 2.15.7 – HiFi assembly primers for construction of pRBV1 107

Table 2.15.8 – Primers for construction of Left and Right Homology Modules (LHM and RHM) pre pRBV1 construction 107
Table 2.15.9 – HiFi assembly primers for gene integration into pRBV1 vectors 108
Table 2.15.10 – Recombineering Primers 109
Table 2.15.11 – OrfD screening Primers..... 109
Table 2.15.12 – pMTL8315y HiFi assembly 109
Table 2.16.1 – List of plasmids used during the establishment of genetic tools in *E. limosum* ATCC 8486 110
Table 2.16.2 – List of plasmids used within this study for the establishment of acetone and isopropanol production in *E. limosum* and *C. autoethanogenum* 111
Table 2.16.3 – Promoter library plasmid collection used within this study (Chapter 4). 112
Table 2.16.4 – List of plasmid constructs generated for generation of recombinant *E. coli* strains (Chapter 5) 113
Table 3.2.1 – General genome statistic for the closed and annotated *E. limosum* ATCC 8486 120
Table 3.2.2 – General COG functional category association of annotated genes from *E. limosum* ATCC 8486 120
Table 3.2.3 – CRISPR locus locations and details..... 121
Table 3.2.4 – Pacbio modified base recognition motifs with putative RM system assignment 123
Table 3.2.5 – Summary of prophages identified by PHAST in three *E. limosum* genomes 124
Table 3.2.6 – Information table for Figure 21 126
Table 3.2.7 – Overview of variation between NCBI and SBRC genomes for *E. limosum* ATCC 8486..... 130
Table 3.2.8 – Breakdown of major regions of genomic variation between NCBI and SBRC genomes..... 130
Table 3.2.9 – Details of ‘other’ variations from Table 33 131

Table 3.2.10 – Initial transformation attempts of *E. limosum* strains *via*electroporation 134

Table 3.2.11 – Replicon testing in *E. limosum* ATCC 8486..... 135

Table 4.1 – Endpoint dataset for the testing of the first-generation Acetone and Isopropanol cassettes in five transformed variants of the *E. limosum* ATCC 8486 mutant i1B 175

Table 4.2 – Calculated LoD values for GC analysis..... 178

Table 4.3 – End point analysis (120 hours) of all tested isopropanol cassettes and the pMTL8J151 plasmid in *E. limosum* ATCC 8486 183

Table 4.4 – *C. autoethanogenum* strain list for scaffold testing..... 187

Table 4.5 – Product yields (mM) for PETC Fructose *C. autoethanogenum* transformants 188

Table 4.6 – Product molar ratio for PETC Fructose grown *C. autoethanogenum* transformants 188

Table 4.7 – Product carbon ratio for *C. autoethanogenum* transformants..... 189

Table 4.8 – Product yields (mM) for *C. autoethanogenum* transformants grown on PETC media with CO..... 192

Table 5.2.1 – RM system details from ReBase for four target organisms 223

Table 5.2.2 - Methylated motifs detected in *E. limosum* ATCC 8486 with best candidate RM systems identified for each..... 225

Table 5.2.3 – RM systems in the *E. limosum* ATCC 8486 genome..... 226

Table 5.2.4 – BLASTp alignment of Elim8486ORFD S subunit against ReBase Type I S subunit database..... 228

Table 5.2.5 – Methylated motifs detected in *C. carboxidivorans* P7 with best candidate RM systems identified for each..... 229

Table 5.2.6 – RM systems in the *C. carboxidivorans* P7 genome 230

Table 5.2.7 – Relevant alignment scores for the assignment of RM systems to modified DNA motifs observed in the *C. carboxidivorans* P7 genome. 232

Table 5.2.8 – Summary of recombineering vectors 235

Table 5.2.9 – Fluorescence intensity values for South-Western Dot Blot of extracted plasmids 243

Table 5.2.10 – *E. limosum* RM targeting motif alignment to pMTL83151 and pMTL86151 247

Table 5.2.11 – *C. carboxidivorans* RM targeting motif alignment to pMTL83151 and pMTL86151 251

Table 7.1.1 – DNA concentrations measurements from Nanodrop and Qubit for four methods of DNA extraction from *E. limosum* ATCC 8486. 271

Table 7.1.2 – pBLAST of Ga0111251_114071 alignment against Swiss-Prot database 272

Table 7.1.3 - Full alignment gap summary for NCBI vs SBRC genome sequences 273

Table 7.2.1 – Potential binding sites for TcdR sigma factor in *E. limosum* ATCC 8486. 276

Table 7.2.2 – pBLAST alignment of TcdR amino acid sequence against genomes of *E. limosum*, *C. acetobutylicum*, and *C. autoethanogenum*. 276

Table of Contents

Declaration 1

Acknowledgements 2

Abstract 3

Abbreviations 4

Table of Figures 7

Table of Tables 11

Table of Contents 15

1. Introduction 25

1.1. Background 25

1.2. Single carbon (C1) waste 28

1.2.1. Syngas 29

1.2.2. Methanol 29

1.3. Acetogens 30

1.3.1. Eubacterium limosum 32

1.3.2. Clostridium carboxidivorans 33

1.4. Wood-Ljungdahl metabolism 33

1.4.1. CO_2 + H_2 34

1.4.2. Methanol 37

1.5. Strain development roadmap 39

1.5.1. The genome 41

1.5.2. Gene transfer 44
1.5.3. Genome editing tools 47
1.5.4. Genome modelling 56
1.6. Aims of this Research 57
2. Materials and Methods 59
2.1. General cloning 59
2.1.1. Cloning strains and culture conditions 59
2.1.2. Preparation of electrocompetent *E. coli* cells 60
2.1.3. Transformation of Electrocompetent *E. coli* cells 60
2.1.4. Preparation of chemically competent *E. coli* cells 61
2.1.5. Transformation of chemically competent *E. coli* cells 61
2.1.6. Plasmid Extraction and purification from *E. coli* 62
2.1.7. Restriction Enzyme Digestion 62
2.1.8. DNA fragment ligations 63
2.1.9. De-phosphorylation of linear DNA 63
2.1.10. Gel Electrophoresis 63
2.1.11. DNA Gel Extraction 64
2.1.12. Standard DNA sequencing 64
2.2. Assembly methods 64
2.2.1. TIIS assembly 64
2.2.2. HiFi assembly 65
2.3. Vector construction 66
2.3.1. ACE vector construction 66
2.3.2. CRISPR vector construction 67

2.3.3. pRBV1 construction 68
2.3.4. pRBV+gene construction 68
2.3.5. Expression cassette vector construction 68
2.4. Research Strains 69
2.5. Culturing and fermentation 70
2.5.1. *E. limosum* culture medium 70
2.5.2. *C. acetobutylicum* medium 71
2.5.3. *C. difficile* culture medium 71
2.5.4. *C. carboxidivorans* culture medium 71
2.5.5. Antibiotic concentrations 73
2.5.6. Media supplementation concentrations 73
2.5.7. Sample analysis 73
2.6. Gene-transfer in target strains 74
2.6.1. Transformation of *E. limosum* 74
2.6.2. Transformation of *C. acetobutylicum* 75
2.6.3. Conjugation of *C. difficile* and *C. botulinum* 76
2.7. Chloramphenicol Acetyltransferase Assay 77
2.1.1. CAT assay sample preparation 77
2.1.2. CAT assay analysis 78
2.1.3. Protein quantification by BCA assay 79
2.8. RFP reporter Assay 79
2.9. Genomic DNA extraction 80
2.9.1. Quick method 80
2.9.2. Standard Gram-positive gDNA extraction 80

2.9.3. Standard Gram-negative gDNA extraction 81
2.9.4. Post extraction gDNA purification 81
2.9.5. High purity gDNA extraction method 82
2.10. Polymerase Chain Reaction (PCR) 83
2.10.1. DNA Polymerases 83
2.10.2. PCR methodologies 84
2.11. DNA quantification and purity assessment 86
2.11.1. Nanodrop 86
2.11.2. Qubit 86
2.12. Genome 87
2.12.1. Sequencing 87
2.12.2. Assembly 88
2.12.3. Annotation 89
2.12.4. BLAST bioinformatics 89
2.12.5. Genome Alignment 89
2.12.6. Pan-Genome assembly 90
2.13. Recombineering 91
2.13.1. Strains 91
2.13.2. Recombineering method 92
2.13.3. Transfer of recombinant pR702 92
2.14. Core Plasmid Maps 94
2.14.1. Chapter 3 94
2.14.2. Chapter 4 97
2.14.3. Chapter 5 102

2.15. Primers 104
2.15.1. Chapter 3 primers 104
2.15.2. Chapter 4 primers 105
2.15.3. Chapter 5 primer 107
2.16. Plasmid list 110
2.16.1. Chapter 3 plasmids 110
2.16.2. Chapter 4 plasmids 111
2.16.3. Chapter 5 plasmids 113
3. Development of *Eubacterium limosum* as a chassis for research 115
3.1. Introduction 115
3.1.1. Research background 115
3.1.2. *E. limosum* 115
3.1.3. Genome 116
3.1.4. Gene transfer 116
3.1.5. Establishing Genome editing tools 116
3.1.6. Project Aims 117
3.2. Results 118
3.2.1. Genome 118
3.2.1. Genome properties 118
3.2.2. Barriers to Gene transfer 121
3.2.3. Prophage analysis 123
3.2.4. C1 metabolism 125
3.2.5. *E. limosum* strain comparison 128
3.2.6. Type strain genome comparison 129

3.2.7. Gene transfer 132
3.2.8. Allele-coupled exchange (Lac-TcdR) 138
3.2.9. CRISPR/Cas9 proof of concept 139
3.3. Discussion 141
3.3.1. Target chassis 141
3.3.2. Implementing the roadmap 142
3.3.3. Closing the Genome 143
3.3.4. Genome structure and features 145
3.3.5. Genome Comparison 147
3.3.6. Establishing gene transfer in *E. limosum* 150
3.3.7. Genome Engineering Proof of Application 155
3.4. Conclusion 159
4. Extending the product spectrum of *E. limosum* 161
4.1. Introduction 161
4.1.1. Isopropanol and Acetone 162
4.1.2. Pathway scaffolding 163
4.1.3. Aims of this Study 165
4.2. Results 166
4.2.1. Development of acetone and isopropanol production cassettes 166
4.2.2. Construct features 169
4.2.3. Generation I & IS 169
4.2.4. GC method re-evaluation 177
4.3. Generation 2 & 3 178
4.3.1. Testing the effects of pathway scaffolding 186

4.4. Discussion 195
4.4.1. General Discussion 195
4.4.2. Methodology limitations 196
4.4.3. Isopropanol Operon Developments 199
4.4.4. Scaffolded Pathways 202
4.4.5. Future work and improvements 207
4.5. Conclusion 209
5. Chapter 5 – Building methylation donor strains to overcome transformation barriers 211
5.1. Abstract 211
5.2. Project background 211
5.3. Introduction 212
5.3.1. RM systems 212
5.3.2. Overcoming gene transfer barriers 215
5.3.3. Methods of Gene Transfer 217
5.3.4. Aims of this study 218
5.4. Results 219
5.4.1. Transformation Workflow 219
5.4.2. Genome RM System Identification 220
5.4.3. Application of PacBio Data 224
5.4.4. Cassette and Plasmid Design 232
5.4.5. Plasmid engineering 233
5.4.6. Cassette design 235
5.4.7. Selection and Screening for Recombinant *E. coli* 236

5.4.8. Proof of concept with *C. acetobutylicum* ATCC 824 237
5.4.9. RFP promoter assay 239
5.4.10. Transforming *C. difficile* DH1916 240
5.4.11. Evidence for methyltransferase activity 242
5.4.12. Transforming *C. botulinum* 19397 244
5.4.13. Improving *E. limosum* ATCC 8486 transformation efficiency 246
5.4.14. Groundwork for *C. carboxidivorans* transformation 248
5.4.15. RM system evasion 250
5.5. Discussion 252
5.5.1. Overview 252
5.5.2. Concept 253
5.5.3. Bioinformatics 254
5.5.4. System design 256
5.5.5. Background genotypes for donors 257
5.5.6. Recombineering backgrounds 258
5.5.7. Proof of concept 259
5.5.8. Applications 260
5.6. A future for *C. carboxidivorans* 262
5.7. Conclusion 263
5.8. Chapter acknowledgements: 264
6. Original contributions and recommendations 266
6.1. Project aims 266
6.2. Original contributions 266
6.2.1. Genome assembly 266

6.2.2. Gene transfer toolkit 266
6.2.3. Strain engineering 267
6.2.4. Gene transfer pipeline 267
6.3. Impacts 268
6.4. Future Studies 268
7. Appendices 270
7.1. Chapter 3 supplementary 270
7.1.1. gDNA extraction protocol development 270
7.1.2. Recommended protein and gDNA extraction method (*E. limosum*) 272
7.2. Chapter 4 supplementary 275
7.2.1. HPLC-UV/IR 275
7.2.1. TcdR off site target analysis 276
7.3. Chapter 5 supplementary 277
7.3.1. Recombineering target locus 277
7.3.2. Effect of L-arabinose concentration on gene-transfer 277
8. Bibliography 279

1. Introduction

1.1. Background

Global energy consumption is at an all-time high and on an increasing trajectory (Figure 1.1.1). Non-renewables such as coal, natural gas and oil account for over three quarters of total energy consumption. The remaining quarter is predominantly made up of energy from nuclear, hydroelectric, and renewable sources. Fossil fuel derived petrochemicals are also the raw feedstock for the production of most of the synthetic and semi-synthetic polymers central to the manufacture of plastics[1].

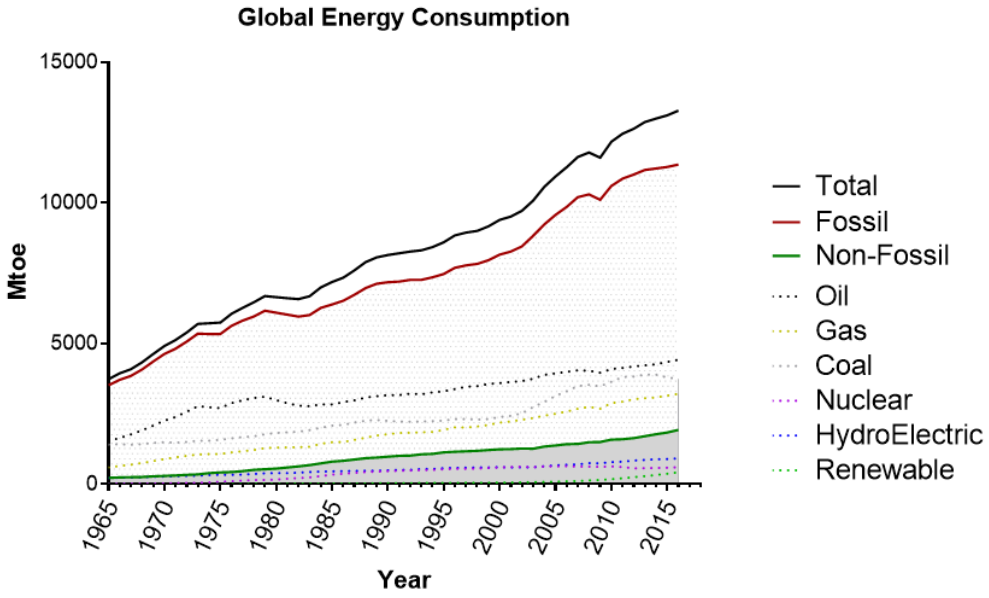


Figure 1.1.1 – Global energy consumption by sector

Data collated from the BP Statistical Review of World Energy June (2017)[2]. A breakdown and summary of energy consumption between 1965 and 2017 measured in Million tonnes of oil equivalent (Mtoe). Solid lines represent the sum totals of the energy categories: Cumulative energy total from all sources (black); Total of oil, gas, coal (red); Total of nuclear, hydro, renewable. Dotted lines represent each energy sector individually: Oil (black dot); Coal (grey dot); Gas (yellow dot); Nuclear (purple dot); Hydroelectric (blue dot); Renewable (green dot).

In spite of a fall in petroleum consumption in Europe and North America in recent years, global consumption continues to increase (Figure 1.1.2). This trend is substantially

driven in large part by demand from developing nations, particularly Asia and Oceanic regions[3].

With ever-growing demand and finite reserves, alternatives to fossil fuels will in the longer term be required to secure future energy and materials markets. On a much shorter timescale, the need to ameliorate the environmental impacts of fossil fuel usage is a key driver in the search for suitable alternatives.

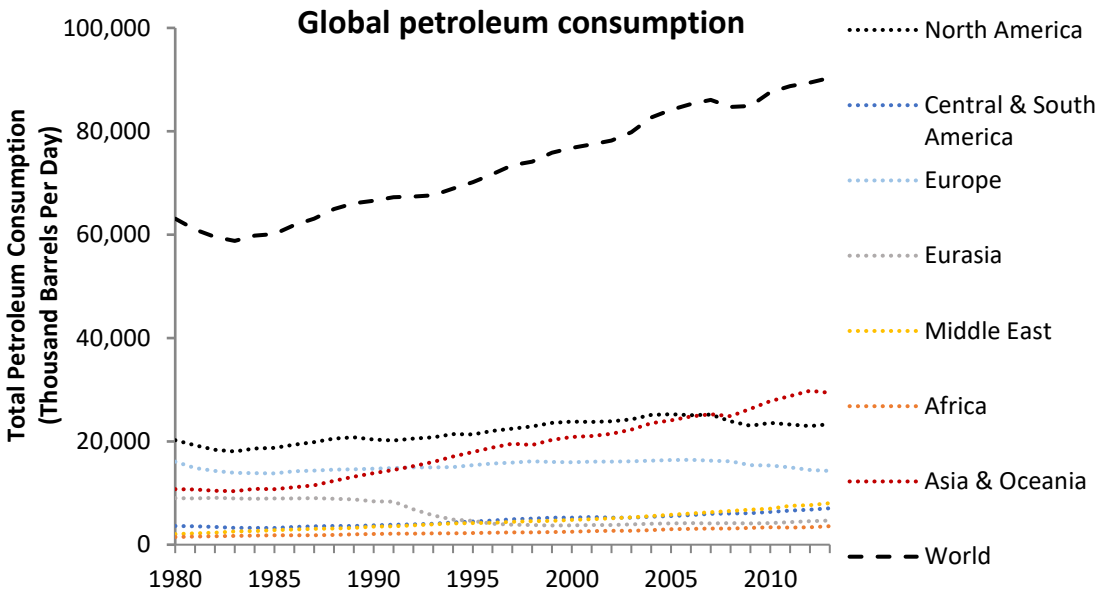


Figure 1.1.2 – Annual worldwide petroleum consumption statistics by region between 1980 and 2013

Consumption values represented as an average number of barrels consumed day throughout the year. Data were provided by the US Energy Information Administration (EIA)[3].

The link between fossil fuel usage, specifically the concomitant emission of greenhouse gases (GHGs), and rising global temperatures has been well documented[4]. As of October 2017, 195 members of the United Nations Framework Convention on Climate Change (UNFCCC) have signed the Paris Climate Change Agreement. This Agreement is a commitment to mitigate GHG emissions in order to limit average global temperatures rises to below 2 °C above pre-industrial levels, in a manner which does not negatively impact food security[5]. The 2 °C temperature rise is predicted to bring with it a myriad of negative environmental changes including rises in sea level, increased extremes of

temperature and the loss of arable land. Collectively these issues are a primary driver for continued investment in research and development of sustainable alternatives to fossil fuels. Feedstocks for these sustainable alternative processes can be derived from one of two broad and frequently overlapping categories. First the leveraging of renewable energy sources such as biomass, solar, geothermal, or hydroelectric. Second the recycling of waste materials produced by existing human activities.

New technologies for generating renewable energy, in particular electrical energy, and for improving energy storage are constantly improving[6]. To this end, the long-term future for energy looks likely to be primarily an electric one. However, without significant technological advancements, it is unlikely that electrical power alone will be able to replace fossil fuels as a platform for global chemical needs. Instead, advancements in bioprocessing seem likely to play a significant role in the solutions for meeting chemical demands in the longer-term. Additionally, bioprocessing is well positioned to support the medium-term transition away from the fossil derived fuels which are currently required to meet energy needs.

Identification of suitable feed material for bioprocessing is a central factor in determining the economic viability of a process. The suitability and sustainability of eacg feed material and bioprocesse must also be evaluated for its broader ethicality. Two key ethical considerations are the competition of feedstocks with food supplies, and any environmental impacts associated with their acquisition. Failure to address this issue remains one of the major points of contention for first generation biofuels. Specifically, where biofuel focused crop plantations compete against food crops for arable land, or are implicated in large scale deforestation in order to create the land required for cultivation, at the expense of the natural carbon cycle[7]. Second generation biofuels aim instead to leverage cellulosic waste materials from agricultural and forestry industries such as wheat straw and other waste biomass. Although the precise definition for what constitutes a 'waste material' is in itself a subject of ambiguity amongst researchers, it is generally accepted that waste utilization should in principle

facilitate economic and environmental cost reductions, when compared to producing renewables from dedicated biofuel crops[8].

Bioprocesses can also operate at a higher level of the carbon-cycle than is achieved with solid biomass feedstocks thereby avoiding direct reliance on pre-existing complex carbon sources. Biological autotrophies can be incorporated into the processes to first asymilate simple carbon sources, for direct or indirect generation of useful fuels, chemicals, and biomass. Such applications are already performed at industrial scale, for example the microbial fermentation of steel mill off-gas and methane in a bioprocess implemented by the industrial biotechnology firms Lanzatech and Calysta respecitvely[9].

Ultimately, biofuels and biomaterials derived from existing waste materials are central to our recycling efforts in the migration towards a circular economy, where resources are regenerated from waste materials and energy[10,11]. Closing such carbon and energy loops reduces the requirement for new fossil fuels to enter the global cycle, a vital goal in minimising GHG levels and associated temperature rises. At a fundamental level, minimising material and energy losses and by improving the efficiencies of existing process and recycling remaining waste streams are central to minimising the levels and effects of human waste.

1.2. Single carbon (C1) waste

The effective recycling of solid organic wastes, such as lignocellulosic biomass and municipal waste, has proved challenging due to the vast physical diversity and molecular complexity of these materials. In order to address this complexity, pre-treatment processes are usually required. These can add significant financial and time costs to the overall process. One possible means of overcoming this challenge is *via*the conversion of complex organic solids into single carbon (C1) intermediates, such as CO, CO_2, and methanol *via*thermochemical decomposition under oxygen limited conditions[12].

1.2.1. Syngas

The primary product of such pyrolysis is a gaseous mixture containing carbon monoxide (CO), carbon dioxide (CO_2) and hydrogen (H_2), termed synthesis gas (syngas). Syngas can be further converted into the liquid C1 compound methanol, *via*a relatively straightforward catalytic conversion[13]. Syngas can be generated from the gasification of an extensive range of carbon-containing feedstocks. These include natural gas, oil, coal, biomass, and is a typical waste product of industrial processes such as steel milling. Syngas contains only ~50% the energy density of natural gas so is not a viable substitute for direct combustion[14]. It is, however, still valuable as a rich fermentation substrate for anaerobic microorganisms. Syngas mixtures can be used as the primary energy and carbon source for catalytic conversion or microbial fermentation processes, for the production of fuels and chemicals. The biological fermentation of syngas has a number of advantages over equivalent chemical processes, namely the lower temperatures and pressures required, and improved tolerance to minor contaminants in the gas feeds such as sulphur compounds[15]. Additionally, biological systems have increased flexibility with regard to the precise ratio of the gases in syngas mixtures.

1.2.2. Methanol

Methanol is the simplest and smallest alcohol, with the chemical formula CH_3OH. Originally named wood-alcohol due to its historic association as a product of destructive distillation of wood, methanol is a highly volatile liquid at ambient temperature and pressure, and is soluble in water. It can be used both as a fuel and as a chemical building block for larger hydrocarbon molecules. Currently, the most prevalent methods for industrial production of methanol employ a catalytic conversion of syngas[16]. Methanol is also a major component of a number of industrial waste streams. The most notable of these waste-streams are processes employing Fischer-Tropsch synthesis, as well as a selection of well-established chemical reactions for the conversion of synthesis gases into liquid hydrocarbons fuels. Another interesting source of methanol is *via*the

reductive conversion of atmospheric CO_2 with H_2, in processes where hydrogen can be produced through the electrolysis of water using energy from renewable sources[16]. In biology, methanol is a natural product of the anaerobic metabolism of some microorganisms and conversely is a substrate for metabolism in others *via*the Wood-Ljungdahl (WL) pathway[17,18].

As with syngas, methanol can be used in microbial fermentations for the production of fuels and chemicals. It represents an attractive alternative to syngas as a feedstock for biological C1 fermentations as it avoids gas-liquid transfer complications in bioreactors. The mass-transfer limitation for gas molecules dissolving into the aqueous phases of fermentations is a major obstacle to syngas fermentations processes. This limitation adds significant restrictions to reactor design, where novel approaches such as hollow fibre membrane technology or microbubble systems are required to allow sufficient substrate-biocatalyst interactions[19]. Without such work-around's undesirable growth rate limitations are imposed on microbial biocatalysts, and the overall rates of conversion between substrate and product are reduced[20].

1.3. Acetogens

Acetogens are a specialised class of prokaryotes with the capacity for reductive synthesis of acetyl-CoA through the fixation of C1 compounds such as methanol, formate, CO, and CO_2 with H_2[21]. A highly conserved metabolic system known as the Wood-Ljungdahl pathway enables these microorganisms to grow autotrophically on C1 compounds. The pathway enables conservation of energy through ATP generation and fixation of carbon into biomass and chemical products[22]. Early emergence of the WL pathway during microbial evolution is often proposed as a critical step in the development of life on earth[23,24], and may explain why acetogens occur with near ubiquity across global ecological niches[25]. One upshot of the early emergence of the pathway and the frequency with which it can still be found, is that a significant degree of variation exists between acetogenic species. This reflects the broad range of

phenotypic adaptions required for survival across a spectrum of environmental conditions.

The first reported isolation of an acetogen occurred in 1936 with the isolation of *Clostridium aceticum* by Klass Wieringa[26]. Since then, a wide range of acetogenic species have been identified and in many instances have undergone in depth characterisation at the genomic and phenotypic levels[15]. Significant research progress has been made into understanding the underlying metabolism of acetogenic species[27]. Such research endeavours have been assisted in recent years by the reduced cost and increased accessibility of whole genome sequencing. By definition, all acetogens are capable of acetate production as a product of WL metabolism, however a number of interesting acetogenic species have demonstrated a capacity to synthesise more complex metabolic end-products. Products of particular commercial interest include, ethanol, butyrate, butanol, lactate, and 2,3-butandiol. The production of ethanol by the acetogen *Clostridium autoethanogenum* in particular been the subject of significant commercial interest and research progress. The successful scale-up of *C. autoethanogenum* catalysed bioprocesses to a 300 Mta pilot plants has been demonstrated by Lanzatech Co. at two Chinese locations, in collaboration with steel manufacturers Beijung Shougang Ltd. and Bao Steel Ltd[28].

In spite of a significant body of research, genetic manipulation of acetogens in general is still in its relative infancy and broadly restricted to a select few model species. Restricting research to a select few species may stifle overall progress of the field, as the extensive diversity of phenotypes, in particular the range of different substrate utilisations and product profiles, suggest that other species might represent a more valuable basis for an industrial chassis. Two good examples of this are *Clostridium carboxidivorans* and *E. limosum*, both of which are capable of synthesising more complex C4 compounds in addition to acetate and ethanol.

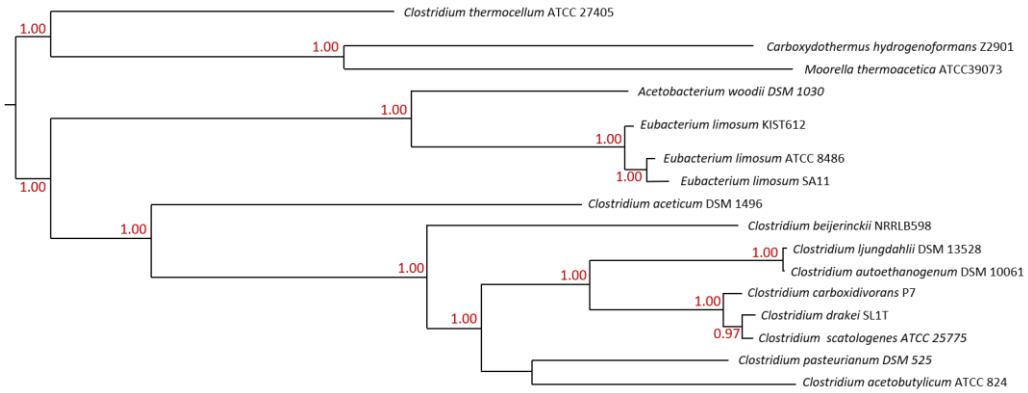


Figure 1.3.1 – Phylogenetic tree based on pan-genome analysis of 16 species from the clostridiales Order

Tree includes a selection of acetogenic species and species with capacity to naturally produce C4 products of metabolism. The tree was generated in Kbase[29], using FastTree 2 – approximately maximum-likelihood algorithm[30], where relatedness was determined by alignment of a subset of Clusters of Orthologous Group (COG) domains between species. Confidence scores for branch points (0-1) are marked in red at the relevant junctions. Species list: *Clostridium thermocellum* ATCC 27405, *Carboxydothermus hydrogenoformans* Z2901, *Moorella thermoacetica* ATCC39073, *Acetobacterium woodii* DSM 1030, *Eubacterium limosum* KIST612, *Eubacterium limosum* ATCC 8486, *Eubacterium limosum* SA11, *Clostridium aceticum* DSM 1496, *Clostridium beijerinckii* NRRLB598, *Clostridium ljungdahlii* DSM 13528, *Clostridium autoethanogenum* DSM 10061, *Clostridium carboxidivorans* P7, *Clostridium drakei* SL1T, *Clostridium scatologenes* ATCC 25775, *Clostridium pasteurianum* DSM 525, *Clostridium acetobutylicum* ATCC 824.

1.3.1. Eubacterium limosum

E. limosum is a Gram-positive, rod-shaped Firmicute, of the Clostridiale class and Eubacteriaceae family. *E. limosum* strains have been isolated from a variety of locations, including mammalian gastrointestinal tracts, mud, and sewage[31]. The reported strains have demonstrated a flexible substrate metabolism with capacity to natively produce acetate, ethanol and butyrate as products of syngas and methanol fermentations[32]. At the time of writing, complete genomes were available for *E. limosum* strains KIST612 and SA11; and a partial sequence available for the type strain ATCC 8486. The most notable close relative of *E. limosum*, by pan-genome alignment and phylogenetic 16S-RNA alignment(Figure 1.3.1) against other well researched acetogens, is *Acetobacterium woodii*[33]. Although there has been a number of papers published on the genomics and metabolism of *E. limosum*[31,34,35], prior to this study there were no reported methods for gene transfer into these strains.

1.3.2. *Clostridium carboxidivorans*

C. carboxidivorans is another Gram-positive, rod-shaped Firmicute of the Clostridiale class, but belongs to the Clostridiale family. Only one isolate of the species has been reported, *C. carboxidivorans* P7 (ATCC BAA-624), isolated from an agricultural settling lagoon[36]. *C. carboxidivorans* is also capable of metabolism of a wide range of growth substrates, including the C1 compounds CO and CO_2/H_2[37]. Unlike *E. limosum*, neither methanol or formate substrate utilisation are supported by *C. carboxidivorans*[38]. This observed substrate incompatibility is reflected in the genome by the absence of a homologue for a MtaABC complex, which would be expected for assimilation of methanol into the WL metabolism. *C. carboxidivorans* is of particular interest commercially, as it demonstrates a natural propensity for production of a number of complex molecules from C1 metabolism, including acetate, ethanol, butyrate, butanol, hexanoate, and hexanol[39]. This extensive native product profile has made *C. carboxidivorans* an attractive target for research as a potential chassis for industrial fermentation. However, in alignment with *E. limosum* research, no successful attempts at genome engineering have been reported. Among acetogens *C. carboxidivorans* is closely related to *C. scatologenes* and *C. drakei* (Figure 1.3.1). *C. carboxidivorans* possesses one of the largest acetogenic genomes (5.7 Mb)[40], which includes a native cryptic plasmid of 19 kb[41].

1.4. Wood-Ljungdahl metabolism

The Wood-Ljungdahl (WL) pathway, otherwise known as the acetyl-CoA pathway, is a highly conserved non-cyclic collection of enzymatically catalysed reactions, which enable autotrophy in acetogenic organisms. Specifically, the WL pathway enables the production of complex organic molecules from C1 compounds such as methanol, CO, CO_2+H_2 and formate, as well as from the methyl groups from a number of methoxylated aromatic compounds[42]. WL metabolism is relatively unique in its capacity to operate as both a bioenergetic system without net ATP requirements and as a system for carbon-

fixation without carbon loss. Details of the pathway were first published in 1969 by Wood and Ljungdahl. It has since been postulated as one of the earliest biochemical pathways on earth and used by some of the first autotrophs in the anoxic atmosphere approximately 3.8 billion years ago[43,44].

Figure 1.4.1 illustrates a generically formatted Wood-Ljungdahl pathway schematic for the fixation of methanol, CO and CO_2. Models for this pathway generally begin as two "reductive branches", the methyl branch (starting CO_2 to formate) and the carbonyl branch (beginning CO_2 to CO), often referred to colloquially as the Eastern and Western branches, respectively. The reducing equivalents represented by hydrogen [H] can be replaced by electrons produced through the oxidation of sugars and other organic substrates. In this manner, CO_2 can act as an electron sink, so there is no requirement for the pathway to be linked to energy production during heterotrophic growth. Instead it is proposed that under heterotrophic conditions acetogens produce energy through glycolysis while the WL pathway maintains intracellular redox balance[27].

1.4.1. $CO_2 + H_2$

Autotrophic growth with CO_2 as the primary carbon source requires the presence of H_2 as a source reducing equivalents, in the absence of alternate reducing molecules such as NADH and reduced ferredoxin. During growth on CO_2 and H_2 the overall reaction for WL pathway is as follows[22]:

$$2CO_2 + 4H_2 + n\text{ADP} + nP_i \longrightarrow CH_3COOH + 2H_2O + n\text{ATP}$$

Equation 1 – Equation for the fixation of CO_2 to acetate

An empirical reaction formula for autotrophic growth on CO_2 and hydrogen.

Although it seems reasonable that this reaction would be committed to ATP production in order to enable autotrophic growth, the net ATP yield for the full WL pathway is

neutral for basic substrate-level phosphorylation. The single ATP generated during this acetate kinase reaction is balanced by one consumed during formyl-H_4F synthase activity. Instead the pathway is coupled to chemiosmotic ATP generation mechanisms, whereby by ion gradient drive ADP phosphorylation[45]. The underlying reactions which comprise the pathway are highly conserved, however the redox coupling to these reactions possesses a high degree of variation between different acetogenic species. This diversity is intrinsically linked to the different methods of coupling the pathway to energy-conservation and ultimately net ATP generation. This coupling is achieved primarily through Rnf or Ech complexes, or by harnessing Na^+ or H^+ chemiosmotic gradients. Consistency is however observed between all known acetogens, in the use of the iron-sulphur clusters of ferredoxin proteins as energy equivalents employed to energise ion-translocating membrane complexes[27].

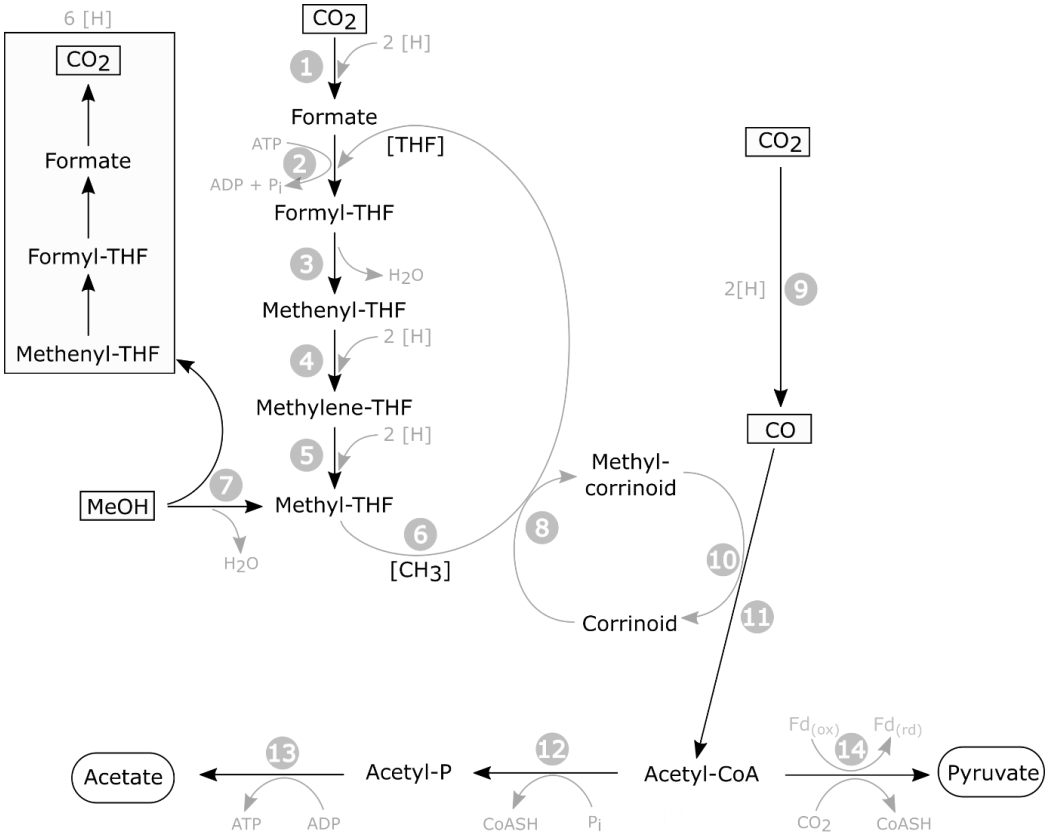


Figure 1.4.1 – Schematic of the Wood-Ljungdahl pathway

Fixation of carbon dioxide (CO_2), carbon monoxide (CO), and methanol (MeOH) to produce the complex organic products acetate and pyruvate. Reactions are not stoichiometrically balanced. Utilisation of methanol proceeds in two directions at a ratio of 3:1, the first as a methyl donor to the THF complex, and second along the reverse methyl-branch (depicted in the box in the top left) to generate CO_2 and 6 reducing equivalents ([H]). 1) Formate dehydrogenase, 2) formate-THF synthase, 3) formyl-THF cyclohydrolase, 4) methylene-THF dehydrogenase, 5) methylene-THF reductase, 6) methyltransferase-1, 7) MtaABC complex, 8) methyltransferase-2, 9) CO dehydrogenase/Acetyl CoA synthase, 10) Correnoid/iron-sulfur protein, 11) CO dehydrogenase/Acetyl CoA synthase, 12) Phosphate acetyltransferase, 13) Acetate kinase, 14) Pyruvate oxidoreductase.

The metabolism of CO_2/H_2 provides the basis for autotrophy of other usable C1 substrates, most of which enter the pathway midway through methyl or carbonyl branches (e.g., CO), or undergo initial reaction steps to themselves generate CO_2 and appropriate reducing equivalents (MeOH). In the presence of [H] ions or other reducing equivalents, both branches are capable of the reduction of CO_2 towards the reactive intermediates required for acetyl-CoA formation. These intermediates, CO and a methylated corrinoid iron-sulphur protein (Co-FeS-P)[46], are bought together during a reaction catalysed by the bifunctional carbon monoxide dehydrogenase/acetyl-CoA

synthase complex (CODH/ACS). For catalysis, CO and the Co-FeS-P bound methyl group are condensed together with coenzyme-A by ACS. Co-FeS-P is recycled to repeat this process and the finalised acetyl-CoA molecule proceeds into further metabolic processing such as acid fermentation, solvent production or for incorporation into cellular biomass. Acetyl-CoA represents the major fork for carbon flux in the pathway, a divergence which is very much species dependent in the potential end products. The rate of each enzymatic step in the pathway is determined by both intra- and extra-cellular conditions at the point of reaction, including pH, temperature, substrate availability and importantly, availability of the enzyme for catalysis[47]. Said conditions will subsequently affect the equilibrium balance pertaining to reactions beyond acetyl-coA towards end products.

1.4.2. Methanol

Methanol metabolism *via*the WL pathway has been best characterised in the model acetogen *Moorella thermoacetica*[48]. The initial stages for metabolism proceed through a well conserved 3-subunit complex, methanol: coenzyme M methyltransferase (MtaABC). This initial processing of methanol by MtaABC to feed into the WL pathway involves an interesting early divergence in the carbon flux. For every three molecules of methanol that proceed to enter the methyl-branch through methylation of THF, a single methanol molecule is processed in the reverse direction of the methyl branch for the production of CO_2 and six reducing equivalents [H]. These reducing equivalents are subsequently used to drive conversion of CO_2 into CO in the carboxyl branch of the pathway. Therefore, in the presence of excess CO_2, flux along both methyl and carboxyl branches of the pathway will result in three molecules of acetate (or equivalent) for four molecules of MeOH following the overall reaction:

$$4MeOH + 2CO_2 + n\text{ADP} + nP_i \rightarrow 3CH_3COOH + 2H_2O + n\text{ATP}$$

Equation 2 – Equation for the fixation of methanol (MeOH) and CO_2 to acetate

An empirical reaction formula for autotrophic growth on MeOH and CO_2.

As such, the CO_2 which was produced in the generation of reducing equivalents through the reverse methyl pathway is assimilated back into the pathway[48]. The excess CO_2 which is assimilated into the pathway increases to the total carbon which can be fixed into biomass and product yields.

It is worth Noting that methylotrophy *via*the WL pathway is just one of a number of routes to biologically process methanol. There are two well characterised aerobic pathways for microbial methanol metabolism, the serine cycle and ribulose monophosphate pathway (RuMP)[18,49]. Both aerobic pathways suffer from biochemical limitations which preclude complete carbon fixation from methanol into either biomass or chemical products. This results in the loss of some carbon as CO_2 which for the most part is not natively recycled by organisms. Although some attempts have been made to engineer pathways for CO_2 recapture through synthetic application of carboxylation reactions[50]. As a result of the limitations of the other methylotrophy pathways, the WL pathway represents a more attractive basis for methanol bioprocessing from an industrial perspective, expecting that it should provide the better overall product yields.

For most acetogens the primary destination for the WL pathway to take from acetyl-CoA to product is towards acetate. In organisms with the appropriate downstream metabolisms extension of acetate metabolism can enable ethanol production. This route contains just two other complete molecular intermediates, acetaldehyde and acetyl-phosphate. Researchers interested in the production of bioethanol from syngas are primarily focused on maximising this conversion of acetyl-CoA. Organisms with naturally elevated levels of ethanol production, such as *C. autoethanogenum* and *C. ljungdahlii*, are typically the focus of such research. Acetate is also a commercially valuable commodity. Wood and Ljungdahl originally proposed that the acetyl-CoA pathway in acetogen *Moorella thermoacetica* be utilised for acetate production from CO_2[51].

Where acetyl-CoA interacts with pyruvate:ferredoxin oxidoreductase in the presence of sufficient CO_2 and hydrogen ions, it is converted into pyruvate *via*pyruvate ferredoxin

oxidoreductase (*pfor*). Through pyruvate, the pathway is coupled with amino acid synthesis and other cellular biomass production. Additionally, lactate *via*lactate dehydrogenase, and 2,3-butanediol *via* acetolactate and acetonin intermediates, can be synthesised *via*pyruvate in a number of acetogenic species[52,53]. Downstream metabolism of pyruvate is an attractive route for strain engineering which could be applied to acetogenic species. Higashide and colleagues at the University of California were able to introduce iso-butanol synthesis to *Clostridium cellulolyticum*, achieved through an engineered valine biosynthesis pathway and resulting in a reported efficiency of 660 mg litre^{-1} on crystalline cellulose substrate[54].

1.5. Strain development roadmap

The variation which has been driven by billions of years of evolution has honed the capacity of different species and indeed different strains, to perform specialised functions to maximise species survival. In the context of harnessing microbes as biological catalysts, the capacities of particular interest are those which enable an organism to convert a given feedstock into a desirable product. As such, organisms which naturally demonstrate favourable characteristics represent a more attractive chassis for development of industrial processes than others. Genetic engineering is a valuable tool in the development of selected strains, both to assist in basic research as well as to improve the commercial applicability of a target strain. To this end, the ability to rapidly open up new and exciting microbes for ground up investigation has significant value for research and potential commercial application.

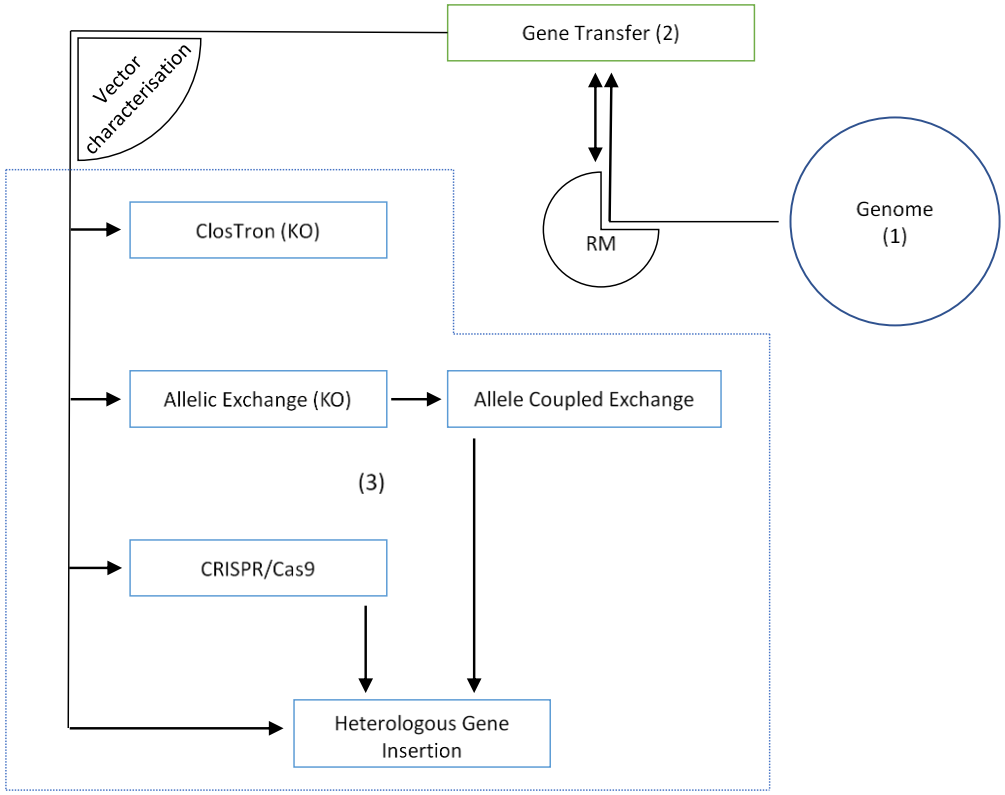


Figure 1.5.1 – A roadmap for gene system development in Clostridium

Adapted from Minton *et al*. 2016[55]. CRISPR/Cas9 has been added as an additional step to the originally proposed roadmap. Three primary stages of development: (1) genome sequencing and annotation (purple), (2) Establishing and optimising gene transfer (green) and where necessary overcoming restriction modification systems (RM) as potential barriers to gene transfer, (3) tools for genome editing and strain engineering (blue).

A workflow proposal for tackling new strains from a research perspective was published in 2016 by the SBRC Nottingham[55]. This 'roadmap' provides a flexible framework for progressing from the initial identification of a strain of interest through to advanced genetic research and strain development. The roadmap can be divided into three distinct sections. First the acquisition of a complete and annotated genome sequence. Second, establishing and optimising robust gene transfer methodologies. Third, the implementation of tools for advanced genome editing. Chapter 3 broadly follows this process, applied to the research and development of an acetogenic bacterial strain of interest, *E. limosum* ATCC 8486.

1.5.1. The genome

In the 22 years since the first complete bacterial genome sequence was published in 1995[56], advancements in the underlying technologies for sequencing and the ability to analyse the results has been immense. Arguably the most significant consequence of these advancements is the overall cost reductions which have been made, making the process widely affordable across research, in a paradigm shift described as the democratisation of sequencing[57].

A complete and annotated genome sequence is a foundation for genetic development. The sequence itself provides a comprehensive dataset for the inner workings of the cell, which can assist in early and late stage research and development processes. Applications range from the basic identification of gene targets for manipulation, to the development of comprehensive in-silico metabolic models for the organism[58]. Collectively the applications of the genome provide valuable insight into the underlying metabolism of organisms and combine with phenotypic datasets, such as fermentation profiles, to become a powerful investigatory tool for strain development.

1.5.1.1. Genome completion

Access to a complete genome offers numerous benefits for basic research, and in the context of biotechnological applications, it is also valuable for the development of strains for industrial application. The generation of a complete genome is achieved in three distinct phases: sequencing, assembly, and annotation.

There are extensive methodology options available for DNA sequencing, however, it was not until the development of high-throughput technologies that complete genome sequencing became a cost effective and practical reality. Next generation sequencing (NGS) is a catchall term used to describe a collection of high-throughput sequencing methods which enable parallelised processing of thousands to millions of nucleotide sequences. The high-throughput nature of these technologies enables significant cost

reductions over traditional PCR-based dye-terminator methods such as Sanger sequencing[59,60]. Assembly is the construction of a longer sequence through the alignment of a number of smaller reads, a process which is generally performed using specialised software such as CLC Genomics Workbench or SPAdes[61]. This can be achieved through the basic alignment of short sequencing reads onto a pre-existing reference sequence, or through a *de-novo* assembly where short reads are aligned to one another and a longer sequence is generated based on the overlapping segments. Following the assembly of a closed genome, annotation of the sequence is performed to identify the genomic features. Automated software is generally used to achieve this goal, where coding sequences are first identified and gene function elucidated based on sequence homology against existing genomes and bioinformatic databases such as TIGR-Fam, Pfam, PRIAM, KEGG and In-terPro[62].

1.5.1.2. Sequencing technologies

The original NGS technologies such as Illumina MiSeq and 454 FLX, employ a sequencing-by-synthesis methodology, and generate average sequencing read lengths of between ~50 and 700 bp[63]. This is not sufficient in many cases to cover the stretches of repeat DNA commonly found in prokaryotic species, often in excess of 5000bp[64]. As a result, the assembly of a complete genome from these data is not possible without an appropriate template on which to align them. Longer reads can be generated by more recently developed sequencing methods such as the PacBio Single Molecule Sequencing Technology (SMRT) which can generate read lengths in excess of 15000bp[65], albeit at significantly greater cost. A drawback of the longer reads generated by PacBio SMRT is a significantly reduced read accuracy of ~87% per single read, in contrast to >99.9% offered by the sequencing by synthesis methods. This generally means that PacBio data is used to produce a *de-novo* 'reference assembly' of the closed genome, which is then error corrected by subsequent mapping of the more accurate shorter reads from another method.

1.5.1.3. Assembly

The extent of sequencing required to assemble a closed genome is largely determined by what is available already with regard to the target organism. This is due to the nature of raw sequencing read data and the limitations of software/algorithms used for automated alignment. For novel or poorly researched organisms this means a *de-novo* sequencing and assembly approach must be taken. This involves aligning overlapping sequencing reads to assess their ordering in the context of the wider closed genome[66]. In contrast, where sequencing an organism for which a sufficiently similar completed genome already exists, raw read data can be mapped directly to this sequence as a template. Template-based assemblies significantly reduce the necessary computing power required to position a given read against the ultimate closed genome. The key hindrance to *de-novo* sequencing and thereby genome closure is the presence of long stretches of repeated DNA sequences which are commonly found in microbial genomes. There are a number of paid and freely available software packages available for the assembly of NGS sequencing reads[67]. These programmes can employ a selection of application specific algorithms, determined by the type of sequencing data (e.g., raw PacBio or Illumina reads), and the availability of appropriate reference sequences.

1.5.1.4. Annotation

Once genome assembly has been performed, the resulting sequence(s) can be annotated. This is generally achieved through the use of automated annotation software, for which there are a number of freely available software packages and online facilities such as RAST or IMG/R from the Joint Genome Institute (JGI)[68,69]. Automated annotation processes for gene identification works primarily on the basis of best-match protein alignment, which although effective in many cases, is prone to a number of recognised inaccuracies[70]. The nature of annotating based on homology to existing genomes means that any errors in the original genome are carried forward. This effect is self-propagating as each new genome is published, increasing the exposure of such

errors for future annotations. Additionally, where erroneous annotations become sufficiently widespread they are liable to pervade secondary bioinformatic databases such as uniprot[71]. Manual annotation curation remains the most robust means of counteracting such algorithm driven error. However, even at the relatively small scale of bacterial genomes manually reviewing each gene would be too laborious to perform as part of standard practice.

1.5.2. Gene transfer

Not all strains are naturally genetically malleable, and this severely hinders the application of the genetic tools necessary for implementing targeted modifications. Even in strains where gene transfer methods have been developed, improving efficiencies is still valuable where the existing efficiencies would be considered low ($<10^4$ CFU μg^{-1} of transformed DNA). High efficiency gene-transfer is central to many genome editing techniques, including high-throughput manipulation based techniques such as by Transposon-Directed Insertion-Site Sequencing (TraDIS)[72]. The rapid advancements in robotics in research is likely to add to the research value which higher transformation efficiency can generate, where high efficiencies can be harnessed with automation. The importance of effective gene transfer extends beyond a tool for basic research, transformed microorganisms are already actively used as chassis for a number of comercial fermentation processes[73,74].

Successful and efficient transformation of microorganisms is principally dependent on two factors. Firstly, it is essential that exogenous DNA is able to physically traverse the microbial cell walls and cytoplasmic membranes, to reach the cytoplasm of the cell. A range of different methods have been employed to drive the internalisation of non-native DNA in bacteria, these include electroporation, sonoporation, chemotransformation and bacterial conjugation[75]. These techniques rely on transient disruption of cell wall and membranes integrity, or active penetration, allowing exogenous DNA to enter the cell body. Secondly, once the transforming DNA enters the

cell, a sustained genetic alteration must occur. The way in which this change is implemented is dependent on the nature of the transforming DNA. Generally, this DNA is either incorporated into the host's genome or has access to suitable cellular machinery to self-replicate. Direct genomic incorporation generally requires the assistance of intracellular machinery. Examples include native DNA repair mechanisms and viral derived systems, as applied to the λ-red and RecET recombineering approaches[76,77]. More typically, transformations are performed using self-replicating plasmid DNA, physically independent from the genome. Plasmid based transformations are typically less stable than genomic integrations and can require maintained selective pressure to ensure plasmid maintenance[78]. Genome engineering methodologies, such as allelic exchange and transposon mutagenesis methods[79,80], often leverage the relative ease and transience of plasmid transformations to generate permeant genomic changes. This is achieved by coupling an initial plasmid-based transformation to a subsequent genomic integration which results in the insertion of a section of the plasmid sequence into the host genome.

Gene transfer has been well established in a number of model miroorganisms[81], but the methodologies used vary widely and are not universally applicable across all strains and species. The most obvious variation exists in the method used to drive the internalisation of exogenous DNA (see above), but further variation exists even within these methods. For example the pre-treatments by which 'competence' for DNA uptake is established and the physical parameters which are employed during electroporation has significant baring on the efficiency of gene transfer[82]. This diversity in methodologies reflects the necessity of a tailored, strain-specific, transformation approachs.

1.5.2.1. Barriers to Gene Transfer

There are a number of specific factors which affect the genetic tractability of microbes, these include passive physical barriers, as well as naturally occurring active systems

which protect cells from foreign DNA. Generally speaking, transformation efficiencies in Gram-positive species is poor, particularly by comparison to Gram-negative species[82]. It is anticipated that this is in part due to the generally increased physical impediment of extensive peptidoglycan networks in the cell walls of Gram-positive species. This hampers the entry of exogenous DNA into the cytoplasm where passive uptake methods such as electroporation and sonoporation are employed[83]. Physical cellular barriers can in some cases be overcome by appropriate pre-treatments to encourage competence for DNA uptake. This includes the use of cell wall weakening agents such as penicillin and lysozyme[75]. Arguably, the more challenging barriers to transformation are the active barriers, which actively seek and destroy non-native DNA. The two most well characterised active defence-mechanisms are restriction modification (RM) systems and CRISPR associated systems. Both systems have evolved as means to recognising and destroying foreign DNA which enters the cell from sources such as bacteriophage and potentially harmful environmental DNA[84].

CRISPR based defence systems operate through the specific recognition of a targeted nucleotide sequences. This is achieved primarily through the direct base-pairing interaction between a guiding RNA molecule and target DNA, which directs the activity of an exonuclease protein or complex to cleave matching DNA molecules. The DNA sequences which are recognised and targeted by CRISPR systems are stored as templates within regions of microbial genomes called CRISPR loci, which essentially act as adaptive immune systems for microbes. These loci are filled with DNA fragments which have been attained during a previous exposure to a given sequence, for example during bacteriophage attack. This targeting mechanism is highly specific due to the high degree of fidelity, which multiple base pairing interactions generate across the length of a target sequence. In the case of the most well characterised system of its kind, CRISPR/Cas9 has a target sequence of 20 bases in addition to 3 bases of protein-based DNA targeting[85].

In contrast to the highly sequence specific targeting of CRISPR systems, RM systems target DNA based on relatively short recognition sequences and determine cleavage according to nucleotide modification patterns at that loci[86]. In prokaryotes these modifications are the methylation of adenine and cytosine bases. There are four types of RM system, Type I-IV. Types I-III cleave DNA based on the absence of methylation at a defined recognition site, whereas type IV systems actively cleave inappropriately methylated DNA[87]. Type I systems are formed of a multi-subunit complex of restriction, specificity and methylation enzymes. Generally, these systems cleave DNA arbitrarily at sites outside their recognition sequence. Conversely, type II systems cleave DNA adjacent to or within their recognition sequence. These systems are the most common but also the most diverse category in terms of overall protein homology. Type III systems are formed from a complex of restriction and methylation proteins, though in most cases the 47ethyltransferase is functional without necessity of complex formation. Type IV systems are unlike the other three types in that they do not possess methylation functionality and actively target methylated (rather than unmethylated) DNA for cleavage.

1.5.3. Genome editing tools

A number of genome editing tools exist which can be applied in target organisms once an effective transformation methodology has been designed. Historically these tools were limited to random UV/chemical mutagenesis, or targeting a small number of predefined genomic loci, which while valuable, were limited in their scope for directed application[88,89]. Advancement in genome editing tools, in combination with increased accessibility to full genome sequencing, has greatly improved the scope for targeted research and modification of biological systems[90]. The original roadmap paper describes two key tools in the researcher's arsenal for targeted genome editing in clostridial species, although applicable in theory to most bacteria, ClosTron and Allele Couple Exchange (ACE)[55]. In addition to these two methods, more recently developed

CRISPR/Cas based genome editing has been successfully applied in many microorganisms. CRISPR/Cas systems are currently driving a new wave of genome engineering developments across life sciences as a result of its effectiveness and relative ease of use[91,92].

1.5.3.1. ClosTron

Traditionally, group II intron-based technologies have proved a popular and simple means of generating stable mutants in bacterial species. The group II introns are self-catalytic mobile ribozymes which target specific DNA sequences for integration of RNA based intermediates. This is achieved by forward and reverse splicing reactions, followed by reverse transcriptase activity to convert the integrated RNA back into DNA[93]. Most of these approaches, including ClosTron[94] and TargeTron[95], are based on the manipulation the broad host range LI.LtrB intron of *Lactococcus lactis*. ClosTron in particular has been widely used and exemplified in most well researched clostridial species[96].

ClosTron combines the use of group II introns with conditional antibiotic selection to overcome low intron insertion frequencies. The initial splicing reaction serves to generate and activate a previously silent *ermB* gene, conferring erythromycin resistance through the removal of an interrupting type I intron in the initial ClosTron plasmids. The transient interruption, termed a retrotransposition-activated marker (RAM), ensures that the resistance gene is not expressed until the type II intron has successfully inserted into the genome. This enables quick and effective screening of initial transformants for the desired knockout mutants.

There are a number of drawbacks to the ClosTron system, and indeed all systems relying on screening by genomic insertion. Namely the polar effects which can arise as a result of inserting active genes into operonic regions, where the transcription of downstream genes may be inadvertently affected[97]. Furthermore, in some cases the

laxity of ClosTron targeting specificity results in multiple genomic insertions outside of the desired target. This means that laborious Southern-blot analysis of mutants is required to confirm their appropriate integrity. Additionally, the nature of the antibiotic selection applied to ClosTron mutagenesis is principally irreversible. This means that iterative rounds of knockouts with this technique relies on multiple appropriate antibiotics for selection as well as complementary resistance genes, both of which are limited in terms of availability and inconvenient to have to individually tailor to new strains. So, whilst the improved screening capacity afforded by the ClosTron technology is central to the system's success, it also represents one of its major weaknesses. These issues have driven development of other genome modification tools with capacity for iterative rounds of mutant generation *via*recyclable selection and screening, as well as capacity for cleaner 'in-frame' knockouts.

1.5.3.2. Allelic exchange and Allele-coupled Exchange

Homologous recombination-based techniques, Allelic Exchange (AE) and Allele-coupled Exchange (ACE), provide an alternate approach to genome engineering. These techniques, first demonstrated in clostridial species in 2012[79], harness the native DNA repair machinery of target organisms to drive homologous recombination events between the genome and plasmid borne template sequences. This allows for the generation of precise in-frame knock-outs and even knock-ins, based on the replacement of a genomic locus with a template DNA sequence. The most basic template sequences consist of two homology arms, generally of 300 and 1200 bp[79,98–100], matching the sequence flanking the target genomic locus. Where knock-in is desired, the two homology arms flank the heterologous sequence which is to be inserted.

In organisms with established and high transformation efficiencies, homology directed genome editing can be achieved through the transformation of non-replicating plasmids or linear DNA[79,101]. Gram-positive bacteria however, are generally associated

with relatively low transformation efficiencies, particularly prior to time-intensive optimisation processes[82]. This, combined with relatively low rates of natural homologous recombination events, means that self-replicating plasmids are generally required to improve the longevity of inserted homologous DNA sequences within target cells. This allows more time for initial recombination events to occur, but has more significant screening burdens than other gene engineering methods and necessitates the often time-consuming process of plasmid loss prior to downstream work. Screening for ultimately desired double 'cross-over' events can be particularly troublesome where selective pressures work against the desired homologous recombination events, such as in the context of knocking out genes with high cellular fitness contributions.

AE is essentially a two-step process which relies initially on antibiotic selection to remove wildtype cells from the population[79]. This antibiotic selection is coupled to a first recombination event, where an antibiotic resistance marker is integrated into the chromosome. Where self-replicating plasmid vectors are employed, the selection of this first cross-over event is performed based on the slower growth phenotype of plasmid harbouring cells. This requires prior identification of autonomously replicating but segregationally unstable plasmid replicons in the target host, so called pseudo-suicide replicons, with which this growth phenotype can be generated[94]. Once the faster growing single crossover mutants have been identified the recombination event must be selected for, which can be readily achieved in many cases through counter-selectable chromosomal gene targets. Orotate phosphoribosyl transferase (*pyrE*) and orotidine 5' phosphate decarboxylase (*pyrF*) are examples of such targets which are available in many clostridiales[98]. Both are involved in the pyrimidine biosynthesis pathway for uracil, so their deletion generates a uracil auxotroph. Conversely, the complete and functional uracil biosynthesis pathway can be selected against using the fluorinated nucleotide 5-Fluoroorotic Acid (5'FOA), which results in lethal incorporation of fluorinated nucleotides into DNA or RNA[102]. 5'FOA selection allows for isolation of double crossover mutants where the *pyrE* or *pyr*F gene loci are the target of AE. This

results in the creation of a strain for which future knock-ins and knock-outs can be selected for by restoration of uracil autotrophy, which can itself be reversed for sequential mutant generation cycles by ACE. There have been a number of identified counter-selectable markers and loci e.g., *codA*, *galK*, and *mazF*[103–105].

ACE operates much in the same way as AE but is applied in the AE generated auxotrophic strains, where the reversible auxotrophy is leveraged to select for crossing over events. As for AE, ACE requires two distinct homology mediated cross-over events to occur. The first, usually the longer homology arm, integrates the entire plasmid into the genome. The second event simultaneously excises the plasmid backbone along with the original genomic locus, leaving the desired insertion or deletion between the homology regions on the genome. Homologous recombination events in ACE are coupled to the restoration of the background strains auxotrophy. In the case of *pyrE* deletion strains, this is achieved through either introduction of a heterologous *pyrE* gene or through restoration of the truncated *pyrE*.

Figure 1.5.2 illustrates the theoretical application of ACE to repair the *pyrE* locus and insert 'gene X' in an *E. limosum* ATCC 8486 mutant with a non-functional truncated *pyrE* gene. The regions of homology in this example are the initially plasmid bourne '*pyrE* gene (orange dashed) and a 1200 bp region from downstream of the native *pyrE* locus (grey dashed). The two crossover events are shown in Figure 1.5.2 a and b, first the long RHA, followed by the short LHA, resulting in a repaired *pyrE* locus with insertion of an insert sequence (geneX).

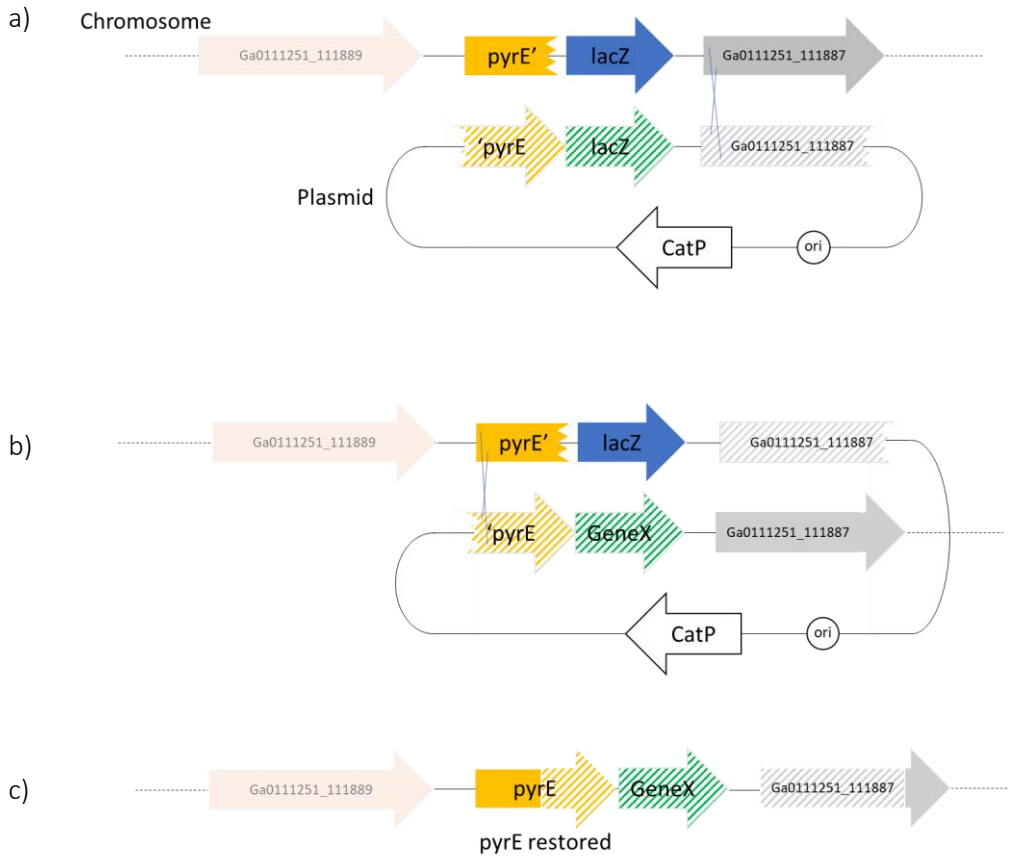


Figure 1.5.2 – Schematic of allele-coupled exchange (ACE) to restore *pyrE* in Eubacterium limosum 8486

Schematic adapted from work by Heap *et al.* (2012)[79]. a) Initial crossover event: the long homology arm (~1.2 kb) crosses into the genome resulting in integration of the entire plasmid. b) Secondary crossover event: the short homology arm (0.3 kb) crosses, restoring a fully functional *pyrE* and releasing the plasmid backbone outside of the homology arms. c) The resultant genomic locus following successful double crossover. The *lacZ* gene can be replaced with a payload of choice for gene insertion or removed entirely for clean *pyrE* restoration.

1.5.3.3. CRISPR/Cas9 mediated genome editing

CRISPR/Cas9 based genome editing is the latest tool for targeted in-*vivo* manipulation of target genomes[92]. These tools combine targeted DNA cleavage, with naturally occurring DNA repair machinery to achieve desired genomic manipulations[91]. The fundamental operation the CRISPR/Cas9 for genome editing centres around a two-component complex. A DNA endonuclease, typically the *cas9* from *Streptococcus pyogenes*, for cleavage activity. In addition to a guiding RNA molecule (gRNA) for targeting the complex to a specific DNA sequence. Engagement of the DNA repair machinery occurs automatically in response to DNA damage in most organisms, and is therefore coupled directly to CRISPR/Cas9 mediated cleavage events. Through relatively trivial manipulation of gRNA targeting and understanding of available repair mechanisms, researchers have been able to harness CRISPR/Cas systems to introduce targeted knock-in, knock-out, and base modification at genomic loci.

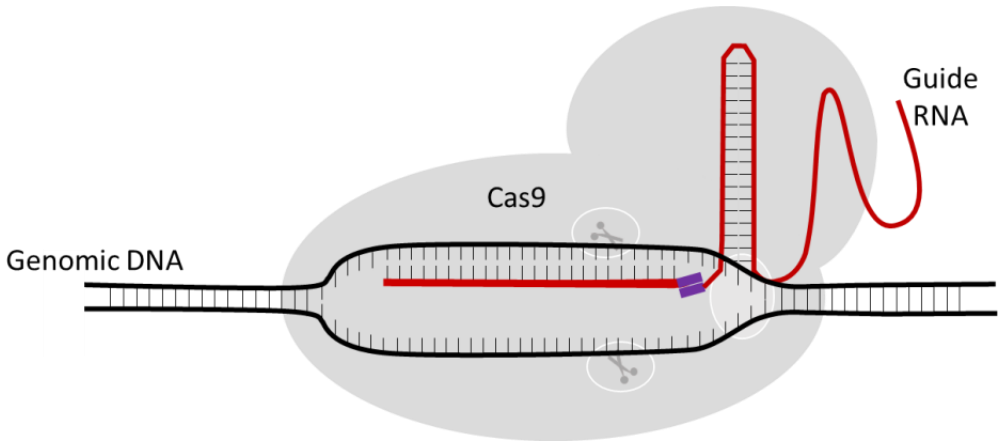


Figure 1.5.3 – CRISPR/Cas9 complex interacting with DNA.

Schematic adapted from DiCarlo *et al.* (2013)[106]. Genomic DNA (black) unzipped by the Cas9 protein (grey) and probed for sequence consistency against the crRNA region of the guide RNA duplex (red) by RNA-DNA base pairing interactions. Two endonuclease domains of Cas9 (scissors) are engaged if base pairing is successful and if a PAM (purple) is also present in the target DNA (black).

In its native context, CRISPR/Cas systems operate as an adaptive immune system for prokaryotes[107,108]. Foreign DNA from viruses (bacteriophage), plasmids, or environmental debris can be integrated into a unique region of prokaryotic genome called the CRISPR locus. This locus is characterised by Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) sequences, which separate each of the integrated fragments in regions termed spacers. This locus can then be transcribed as a single CRISPR RNA (crRNA) molecule before being processed into smaller fragments each containing just a single spacer[109]. The processed crRNA molecules are then able, usually through a secondary interaction with other specific RNA molecules known as maturation, to associate with the Cas9 endonuclease to form a mature CRISPR/Cas9 complex.

Mature CRISPR/Cas complexes are guided to a target sequence of DNA (Figure 1.5.3) through strict base pairing interaction between a single strand of DNA and the 20 bp crRNA sequence. Sequences matching this 20 bp region are termed protospacers. Additionally, the Cas protein confers a small amount of its own specificity through direct protein-DNA interactions of 3-5bp in length outside of the crRNA sequence, a sequence termed the Protospacer Adjacent Motif (PAM). The initial spacers which are transcribed from the CRISPR locus do not contain these PAMs to avoid self-targeting by the systems, as such the combined crRNA-PAM specificity should only appear in foreign DNA which enters the cell. Where the crRNA-PAM specificity has been met, the Cas9 endonuclease is engaged resulting in two cleavage events, one on either strand of the target DNA molecule, termed a double strand break (DSB)[107]. DNA cleavage generally results in irreparable damage and thereby quickly renders any targeted DNA inactive. In the context of harmful DNA such as that from bacteriophage, this system means that harmful foreign DNA molecules can be disarmed in their attempt to hijack the host cell.

Employing CRISPR/Cas9 to introduce targeted DSBs in the genome of a target cell is a death sentence to cells in which the damage cannot be quickly and effectively repaired.

There two classes of DNA repair mechanism, homology directed repair (HDR) and non-homologous end joining (NHEJ).

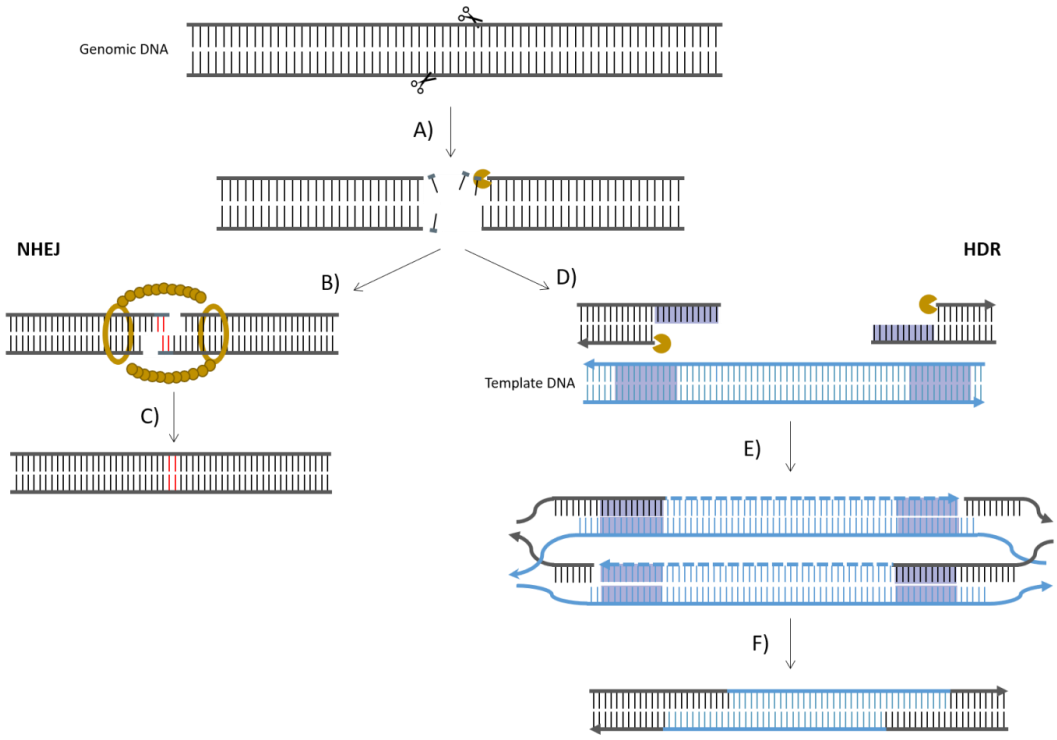


Figure 1.5.4 – Illustration of Non-homologous end joining (NHEJ) and Homology-directed repair (HDR) mechanisms following a DNA damage

A) Double strand break (DSB) event as a result of DNA cleavage by dual restriction endonucleases (e.g., Cas9 or an RM system), the exposed DNA is prone to degradation by exonucleases; B) Broken DNA is recognised and NHEJ repair complexes are recruited and assembled at the broken ends; C) NHEJ machinery then process the ends of the DNA fragment by removal of damaged or mismatched nucleotides, or adds nucleotides *via*DNA polymerases resynthesis, before ligating the two fragments. D) HDR can occur in the presence of an appropriate DNA template (blue) containing two regions of homology (purple) either side of the DSB, HDR complex assembles on both sides of the break and begins to performs resection of the 5' ends of DNA with exonuclease activity, exposing a single stranded overhang; E) The single strand overhang can then invade and bind to appropriate homologous DNA sequences with the assistance of further HDR proteins, and subsequently polymerase activity fills in the joining region using the homologous DNA strand as a template. F) The final gaps are closed by ligase activity and the intertwined DNA helices are separated by a process called resolution, yielding two separate DNA strands each containing an identical sequence between the regions of homology regions.

HDR mechanisms repair DNA damage only when an appropriate template DNA strand for the repair is available. An appropriate repair template requires regions of sequence homology to the native sequence, adjacent to both sides of a damaged or cleaved locus. The basic mechanism for this process is described in Figure 1.5.4. Ultimately, employing

HDR mechanisms enables the accurate replacement of sequences between the two regions of homology, for sequence knock-out, knock-in, or replacement. HDR is a highly conserved process which pervades the three primary domains of life[110]. As such it can be effectively applied in most living cells where effective introduction of CRISPR/Cas9 mediated cleavage is achievable, and a suitable repair template is available. NHEJ on the other hand is a more efficient but less flexible repair mechanism, which repairs DSBs through direct ligation of the exposed ends of DNA following a cleavage event (Figure 1.5.4). These ligation events are generally coupled with base insertion, deletion, or replacement, which ultimately results in SNPs in the repaired sequence. This can be effective in scenarios where gene inactivation is desirable through the introduction frame shifts. However, the unpredictability of the exact nature of mutations which are introduced can necessitate a significant screening burden to ensure anticipated outcomes[111]. NHEJ machinery is less widely conserved than HDR and is not found in most prokaryotic species[112]. For this reason, it is not immediately applicable to most research or engineering in bacterial biocatalysts.

1.5.4. Genome modelling

Applied synthetic biology has the capacity to extend metabolic processes beyond that which can be achieved through directed evolution alone, principally through the introduction of heterologous enzymatic pathways. It is clear however, that any genetic modifications need to be accounted for and balanced at the level of global cellular metabolism. Metabolic modelling based on complete genome data allows quantitative analysis of existing metabolic pathway operations and can be used for predicting the most appropriate avenues for strain engineering. In silico genome modelling has existed for nearly two decades, since the first genome-scale metabolic model was produced for the pathogenic bacterium *Haemophilus influenza* GSMM[113]. Numerous approaches have since been developed for analysing metabolic networks at the levels of basic structure, metabolic flux, metabolic control, and pathway kinetics[114]. These insights can

be valuable for both genetic modification by rational design as well as directed evolution engineering approaches. A combination of targeted genetic modification and directed evolution, based on robust understandings generated by the in-*silico* modelling of biological systems, is likely to yield the most effective results in the production of new chassis for industrial fermentation processes.

1.6. Aims of this Research

The primary aim of this research project was to establish and apply genetic engineering tools in the C4-producing, acetogenic bacterium, *Eubacterium limosum*. Specifically, this required basic genetic characterisation, establishment of a gene-transfer methodology, and demonstration of gene manipulation tools ACE and CRISPR/Cas9 (Chapter 3). The study then aimed to showcase this progress through the proof-of-concept engineering of *E. limosum* metabolism for production of acetone and isopropanol using the pathway presented in Figure 1.6.1 below (Chapter 4). The final aim of the project was to develop an improved methodology for establishing gene-transfer in recalcitrant microorganisms (Chapter 5).

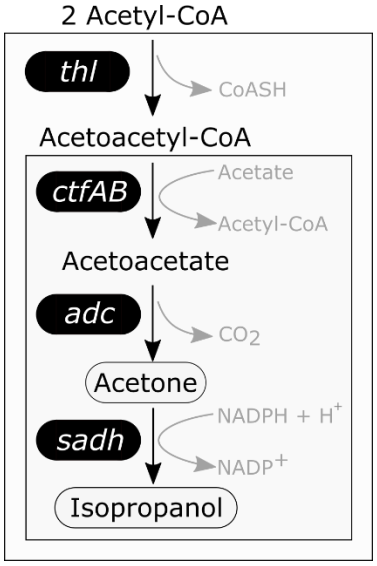


Figure 1.6.1 – Synthetic Isopropanol and Acetone pathway

Schematic based on publication by Hanai *et al*. (2007)[115]. The enzymes responsible for catalysing four distinct enzymatically catalysed reactions for conversion of acetyl-coA to isopropanol are marked (black bubbles). Acetyl-CoA acetyltransferase *(thl)*, E.C. 2.3.1.9; CoA transferase *ctfAB*, E.C. 2.8.3.9; Acetoacetate decarboxylase (*adc)*, E.C. 4.1.1.4; Secondary alcohol dehydrogenase (*sadh*), E.C. 1.1.1.80.

2. Materials and Methods

2.1. General cloning

2.1.1. Cloning strains and culture conditions

E. coli strains were cultured with standard LB medium supplemented where appropriate with appropriate antibiotics[116]. LB medium contained 5 g/L NaCl; 5 g/L yeast extract; 10 g/L tryptone; 15 g/L agar (solid media only). Cultures were generally grown at 37 °C with 225 rpm shaking for broth growth. For strains harbouring larger plasmids or where the products on constructs were considered to be potentially toxic, incubations were performed at 30 °C. Antibiotic concentrations for *E. coli* selection were as follows: apramycin 50 µg/ml; kanamycin 50 µg/ml; erythromycin 500 µg/ml; chloramphenicol 12.5 µg/ml (25 µg/ml on solid media); spectinomycin 200 µg/ml.

Table 1 –*Escherichia coli* (*E. coli*) strains used for cloning in this study

Strain	Function and properties	Genotype	Source
E. coli TOP10	Plasmid storage and cloning strain	F- *mcrA, Δ(mrr-hsdRMS-mcrBC), φ80lacZΔM15 ΔlacX74, nupG, recA1 araD139, Δ(ara-leu)7697 galE15, galK16, rpsL(Str^R) endA1, λ$^-$*	NEB
E. coli CA434	*E. coli* HB101 containing cryptic plasmid pR702 - Conjugative transfer strain	F$^+$ *mcrB mrr hsdS20(r$_B^-$ m$_B^-$) recA13 leuB6 ara-14 proA2 lacY1 galK2 xyl-5 mtl-1 rpsL20(Sm^R) glnV44 λ$^-$* - *R702-TcR, SmR, SuR, HgR Tra$^+$, Mob*	M. Young, UCW, Aberystwyth, UK (Purdy *et al*, 2002)
E. coli DH5α	Basic cloning applications	F$^–$*, endA1, glnV44, thi-1, recA1, relA1, gyrA96, deoR, nupG, purB20, φ80dlacZΔM15, Δ(lacZYA-argF)U169, hsdR17(r$_K^–$m$_K^+$), λ$^–$*	NEB
E. coli Stellar	Cloning and strain development in a methylation free background (DCM$^–$, DAM$^–$)	*F–, ara,Δ(lac-proAB)* [*Φ80d lacZΔM15*], *rpsL(str), thi, Δ(mrr-hsdRMS-mcrBC), ΔmcrA, dam, dcm*	CloneTech™
E. coli Express	Strain development in a DCM$^–$, DAM$^+$ background	*fhuA2 [lon], ompT, gal, sulA11, R(mcr-73::miniTn10--TetS)2, dcm R(zgb-210::Tn10--TetS), endA1, Δ(mcrC-mrr)114::IS10*	NEB
E. coli HB101	Strain development in a DCM$^+$, DAM$^+$ background	*F$^-$, mcrB, mrr hsdS20(r$_B^-$ m$_B^-$), recA13 leuB6, ara-14, proA2, lacY1, galK2, xyl-5, mtl-1, rpsL20(Sm^R), glnV44, λ$^-$*	Promega*

*Promega genotype of HB101 does not list F-, *mcrB*, or *mrr* genotypes.

2.1.2. Preparation of electrocompetent *E. coli* cells

Overnight *E. coli* starter cultures were set up in 5 ml of LB broth with appropriate antibiotic selection where appropriate, and incubated for ~18 hours at 37 °C with 220 rpm orbital shaking agitation. The starter culture was subsequently inoculated into 500 ml conical flask containing 250 ml fresh LB broth with antibiotic selection maintained where required, for a starting OD_{600} of 0.05. The 250 ml overday culture was incubated under the same conditions as the overnight starter until an OD_{600} of 0.4-0.8 was achieved (~3 hours). The culture was then decanted into six pre-chilled 50 ml falcon tubes for cell washing and harvesting steps. All remaining steps of the preparation were performed on ice or at 4 °C, including use of prechilled reagents/buffers. Cells were harvested by centrifugation at 4000 x g for 10 minutes, and the resulting supernatant was removed. Cell pellets were then washed twice by process of resuspension in 25 ml dH_2O before centrifugation under the same conditions as previously, followed by supernatant removal. Cell pellets were collated into a single 25 ml solution during the second wash step prior to centrifugation. Cells were harvested by resuspending the final washed pellet in 1.5 ml of 10% glycerol (v/v in dH_2O), before separation into 50 µl aliquots for storage at -80 °C.

2.1.3. Transformation of Electrocompetent *E. coli* cells

Single aliquots of pre-prepared electrocompetent *E. coli* were removed from storage at -80 °C and thawed on ice. Once thawed, 1 µl of plasmid DNA or 5-10 µl of dialysed ligation or fragment assembly mixture was added to the cells. The mixture was transferred to a pre-chilled 2 mm gap electroporation cuvette (Biorad) and allowed to incubate on ice for 10 minutes. The cuvette was then pulsed using the Biorad Micropulsar™ electroporation apparatus, using the pulse pre-set condition EC2. Immediately following the pulsation, 850 µl of room temperature SOC medium (ThermoFisher) was added to the cuvette and sample gently resuspended. Samples were then transferred to 1.5 ml microfuge tubes and allowed to recover at 37 °C with

agitation for 45-60 minutes. Recovered samples were then spread onto LB agar plates supplemented with an appropriate antibiotic for selection. Selection plates were incubated overnight at 37 ^{o}C.

In the instances where temperature sensitivity was relevant (e.g., in strains carrying pKD46-cm) all incubation steps were carried out at 30 ^{o}C rather than the standard 37 ^{o}C, unless plasmid loss was desired.

2.1.4. Preparation of chemically competent *E. coli* cells

E. coli chemically competent cells were prepared using an adaption. A standard incubation temperature of 37 °C was employed for all culturing except where temperature sensitive plasmids were employed, in which case temperatures were reduced to 30 °C. Fresh 5 ml starter culture of the desired *E. coli* strain in LB medium with appropriate selection was incubated for 12-18 hours. The starter culture was then sub-cultured into a 250 ml volume of fresh LB with selection maintained and grown to an OD_{600} of 0.3-0.6. The 250 ml culture was decanted into pre-chilled falcon tubes and incubated for 15 minutes on ice. Samples were kept on ice or at 4 °C for the remainder of the preparation procedure. Harvesting of the cells was performed by centrifugation at 3,000 g for 10 minutes and washing of the resultant pellet by resuspension in a chilled 10 ml 100 mM $CaCl_2$, 20% v/v glycerol solution. At this stage samples were incubated on ice for up to two hours before a final pelleting and resuspension in 10 ml fresh $CaCl_2$ solution. Aliquots of 100 µl were divided into 1.5 ml Eppendorf tubes and stored at -80 °C until use.

2.1.5. Transformation of chemically competent *E. coli* cells

Single aliquots of pre-prepared chemically competent *E. coli* were removed from storage at -80^{o}C and thawed on ice. Once thawed, 1 µl of plasmid DNA or 10 µl of dialysed ligation or fragment assembly mixture was gently added to the cells followed

by a 30 minute incubation on ice. The cells were then heat shocked at 42 °C for 45 seconds before returning to ice for a further 5 minutes. Next, 850 µl of room temperature SOC medium (ThermoFisher) was added to the cells followed by incubation at 37 °C with constant agitation for recovery. Recovered samples were then spread onto LB agar plates supplemented with an appropriate antibiotic for selection. Selection plates were incubated overnight at 37 °C. In the instances where temperature sensitivity was relevant (e.g., in strains carrying pKD46-cm) all incubation steps were carried out at 30 °C (rather than 37 °C).

2.1.6. Plasmid Extraction and purification from *E. coli*

Plasmids extractions were performed from 1-4 ml of overnight *E. coli* cultures incubated at 37 °C with appropriate selective pressure. Extraction was performed using the Monarch® Plasmid Miniprep Kit (New England Biolabs), following the manufacturer's instructions.

2.1.7. Restriction Enzyme Digestion

Restriction digest reactions were performed using enzymes and buffers from New England Biolabs (NEB). Standard digest reactions were performed in a reaction mixture with the following components: ~1 µg of DNA (purified linear or circular DNA), 0.5 µl of each restriction enzyme, 2 µl of appropriate 10x reaction buffer (defined manufacturer recommendation for each restriction enzyme), made up to a total volume of 20 µl with nuclease free H_2O. Standard reactions conditions were 1 hour incubation of the reaction mixture at 37 °C. Digestions were followed by gel electrophoresis of samples for screening and size selection (Method 2.1.10), followed by extraction and purification of the digested fragments (Method 2.1.11).

2.1.8. DNA fragment ligations

Ligation reactions were performed in order to attach two or more DNA fragments together, by ligation of compatible 5' and 3' ends. DNA T4 ligase (NEB) was used for the ligation of purified linear DNA fragments into plasmid backbones. Typically, a molar ratio of 3:1, insert:backbone was used, where inserts were linearized DNA fragments and backbones were linearized plasmids containing replicons and selection markers. The ligation mixture consisted of ~50 ng of backbone DNA, 37.5 ng of insert DNA, 2 µl T4 ligase buffer (10x), 1 µl of DNA T4 ligase, made to a total volume of 20 µl with nuclease free dH_2O.

2.1.9. De-phosphorylation of linear DNA

De-phosphorylation of 5' ends of linear DNA fragments was sometimes required to prevent unwanted re-circularisation plasmid backbones and other self-ligation events. De-phosphorylation was performed using Antarctic Phosphatase (NEB) and Antarctic Phosphatase Buffer (NEB) following the manufacturers recommended reaction conditions. DNA 1-2 µg (for backbones of 3-6 kb in size), Antarctic Phosphatase reaction buffer (10x) 2 µl, Antarctic Phosphatase 1 µl. Samples were incubated for 30 minutes at 37°C, followed by heat inactivation of the Antarctic Phosphatase enzyme with a 2 minute incubation at 80°C.

2.1.10. Gel Electrophoresis

Gel electrophoresis was used to separate and visualise nucleic acids based on sequence length. Gels for electrophoresis were produced at 0.7-1% w/v agarose (ThermoFisher) in TAE buffer. TAE buffer was made by dilution of a 50x stock solution to give a 1x working solution containing 40mM Tris, 20mM glacial acetic acid, and 1mM EDTA. Agarose gels were microwaved until molten and poured into a gel tray once temperature had reached ~55 °C. While gels were molten, appropriate volumes of SYBR

Safe DNA Gel Stain (1:10000 dilution) were added and mixed thoroughly. Gels were allowed to set at room temperature with a comb added to create sample wells. Once set, the tray end buffers were carefully removed and the gel was submerged in a TAE containing Gel Electrophoresis Tank (Cleaver Scientific), and the comb removed. Samples mixed with 6x Gel Loading Dye (NEB) were then loaded into each well, with DNA ladders flanking each sample set, O'generuler 1kb plus (ThermoFisher) or 2-log ladder (NEB). Electrophoresis was standardly run at 100 V for 45 minutes in 1x TAE buffer.

2.1.11. DNA Gel Extraction

Agarose Gels were visualised using a blue light box or by LiCor Odyssey DNA analyser before carful extraction of desired bands using a scalpel. DNA containing gel slices were then purified using the Wizard® SV Gel and PCR Clean-Up System (Promega) following the supplier's instructions.

2.1.12. Standard DNA sequencing

Purified DNA fragments, PCR product or plasmid DNA, were sequenced by Eurofins Scientific or Source Biociences. Samples were sent along with appropriate primers for sequencing using Sanger sequencing (chain termination method) based methodologies[117].

2.2. Assembly methods

2.2.1. TIIS assembly

The type IIS restriction enzyme BsaI was utilised to assembly of multiple fragments in a 'one pot' reaction. PCR primers with BsaI compatible overhangs were designed to contain appropriate sticky ends following digestion, as described by Engler 2008[118].

Insert fragments were generated by PCR (Method 2.10) and gel purified alongside a BsaI digested pRM01 backbone (Method 2.14.2). Inserts and backbone were combined in a ratio of 2:1 in a total volume of 20 µl, containing 1 µl BsaI (NEB), 2 µl CutSmart buffer (NEB), 1 µl T4 ligase (NEB) and 2 µl 10mM ATP (Sigma). Sample mixtures were transferred to a thermocycler and run on the following protocol:

Time (minutes)	Temperature (°C)	Cycles
10	40	
10	16	5x
10	50	
20	80	
-	15	

Samples were then transformed directly into chemically competent *E. coli*, or dialysed before transformation into electrocompetent *E. coli*.

2.2.2. HiFi assembly

NEBuilder® HiFi Assembly master mix (NEB) was used for the assembly of multiple DNA fragments with terminal overlapping regions of homology. Fragments were generated by either PCR, or digestion and linearization of an existing plasmid. Primer design and assembly were performed following the manufacturer recommended guidelines. Fragments were designed to contain overlapping regions of 15-25 base pairs and an overlap Tm of >50 °C. Up to five fragments were combined in 1:1 ratios for PCR products, or 2:1 (insert:backbone) where linearized vector was used. The reaction mixtures of 20 µl total volumes consisted of 10 µl 2x HiFi Master Mix, and fragments as appropriate diluted in dH_2O. Reaction mixtures were then incubated for 1 hour at 50 °C before immediate transformation into chemically competent *E. coli*.

2.3. Vector construction

A number of different plasmid-based constructs were generated throughout the course of the project. Below is an overview of the protocols used to produce plasmids from four vectors or vector categories used during this project: ACE vectors, CRISPR vectors, Recombineering vectors, and gene expression cassette vectors.

2.3.1. ACE vector construction

Construction of allele-coupled exchange (ACE) vectors were in this study were based on the pMTL80000 format backbone, pMTL83151[119]. Underlying plasmid design is based on the original format for ACE vector format as described by Heap *et al*. (2012)[79]. The target strain for ACE was the *pyrE* knockout strain, *E. limosum*-Δ*pyrE* (ATCC 8486), developed by James Millard (Millard, J. SBRC Nottingham. Thesis 2017) using the allele-exchange *pyrE*-KO-vector, pMTL-JM101. This vector consisted of a long right homology arm (RHA) of ~1000 bp, and a short left homology arm (LHA) of ~300bp, each flanking a spacer region containing the *lacZ* multiple cloning site (MCS) of the pMTL80000 vector series. The *E. limosum*-Δ*pyrE* strain contains a truncated *pyrE* gene, with the *lacZ* gene inserted at the site of truncation. The ACE correction (pMTL-RM03, Figure 2.14.1) and complementation (pMTL-RM04, Figure 2.14.2) vectors used in this study were comparable to the pMTL-JH14 used to repair an equivalent *pyrE* mutation in *C. acetobutylicum* ATCC 824[79]. Both pMTL-RM03 and pMTL-RM04 contain the LHA from pMTL-JM101, with an extended LHA containing the full native *pyrE* gene. pMTL-RM03 contains the *lacZ* MCS downstream of the full native *pyrE* gene, resulting in its insertion following repair of the locus. The pMTL-RM04 vector was generated by replacing the *lacZ* MCS with an insert fragment containing a lactose inducible TcdR expression module, from the plasmid construct pMTL-C24 (Wang, H. SBRC, Nottingham. Unpublished). This was achieved using traditional restriction-digestion cloning methodology, utilising NotI and NheI restriction sites on the backbone and insert.

Additional information regarding the construction and structure of the ACE vectors is available in the supplementary information.

2.3.2. CRISPR vector construction

The construction of CRISPR vectors was based on the pSBRCas9 plasmid backbone, developed by Peter Rowe Nottingham SBRC (Rowe, P – Thesis 2017). This vector is backwards compatible in most regards with the pMTL80000 vector series, in that it is separated into the standard four quadrant format. Specifically, the standard pSBRCas9 vector is akin to a pMTL8315z vector, where the 'z' quadrant harbours a unique MCS which is further subdivided into three functional regions. First, a Cas9 exonuclease (*Streptococcus pyogenes*) containing region. Second, guide RNA (sgRNA) module with a sequence adjustable protospacer. Third, an adjustable homology directed repair module. The spacer sgRNA region was modified to include desired protospacer using HiFi assembly method (Method 2.2.2) following SalI digestion and Antarctic phosphatase treatment (method 2.1.9) of a pSBRCas9 backbone. sgRNA sequences were identified as 20 base pair regions adjacent to appropriate PAM sequence NGG, using the algorithms of Donch et al. (2014) and Hsu et al. (2013), in the online software platform, Benchling. The HiFi insert was a single stranded oligo of 70 nucleotides (Table 2.15.2), which contained overlap homology to both sides of the SalI digestion site, between the P_{araE} promoter and the sgRNA. Following insertion of the desired guide, homology arm (HA) module was generated by SOEing PCR of two ~750 bp fragments (Table 2.15.2), from up- and down-stream of the target knockout region. SOEing products were designed to contain terminal AscI and AsiSI restriction overhangs. HA modules were inserted into the pSBRCas vector backbones *via*digenstion-ligation cloning with AscI and AsiSI sites.

2.3.3. pRBV1 construction

The recombineering base vector (RBV) pRBV1 was used in the construction and storage of gene cassette modules prior to recombineering in *E. coli* strains containing the pR702 conjugation assistance plasmid (Chapter 5). pRBV1 is based on pMTL8314x with a custom 'x' module containing two homology arms flanking an arabinose inducible promoter module (AraC-P_{BAD}), a multiple cloning site (MCS), and a selection marker (Apramycin). The vector was built from four PCR amplified fragments (Table 2.15.7) using the NEB HiFi assembly method (Method 2.2.2).

2.3.4. pRBV+gene construction

A number of vectors for recombineering were constructed based on the pRBV1 backbone. Target genes (Table 2.15.9) were PCR amplified and inserted by HiFi (Method 2.65) at the pRBV1 spacer region following BsaI digestion and linearization of the backbone.

2.3.5. Expression cassette vector construction

Vectors carrying gene expression cassettes (e.g., Figure 2.14.5) were used to instigate acetone and isopropanol production from transformed *E. limosum* and *C. autoethanogenum* strains. Initial vectors were created using Type IIS assembly method (Method 2.2.1), and subsequently modified by basic restriction cloning. Promoters were interchangeable with a library of NotI-NdeI digestion cloning compatible promoter modules (Method 2.16.2). Removal and addition of genes to the basic isopropanol operon was possible through SpeI, NheI, and XbaI sites. The Gram-positive replicon was exchangeable with any pMTL80000 compatible equivalent *via*FseI-AscI digestion cloning.

2.4. Research Strains

Table 2.4.1 – Strain details for organisms targeted for research throughout this project

Strain	Applications	Genotype	Source
E. limosum ATCC 8486	Multiple applications (Chapter 3,4,5)	WT	DSMZ – Eggerth 1935
E. limosum ATCC 8486 Δ*pyrE*	Proof of ACE target strain (Chapter 3)	Δ*pyrE*	Millard, J. SBRC Nottingham
E. limosum ATCC 8486 i1B	Proof of ACE target strain (Chapter 3)	Lac-TcdR$^+$ (at repaired *pyrE* locus)	This study
E. limosum DSM 2593	Early genome engineering efforts (Chapter 3)	WT	DSMZ – Eggerth 1935
E. limosum DSM 2594	Early genome engineering efforts (Chapter 3)	WT	DSMZ – Eggerth 1935
E. limosum ATCC 8486 i1B	Proof of ACE result (Chapter 3), Overexpression chassis for acetone and isopropanol cassettes (Chapter 4)	Lac-TcdR$^+$	This study
E. limosum ATCC 8486 Δ*pbt*1	Proof of CRISPR result (Chapter 3)	Δ*pbt*1	This study
C. autoethanogenum DSM 10061	Overexpression chassis for acetone and isopropanol cassettes (Chapter 4)		SBRC Nottingham culture collection
C. carboxidivorans P7	Early genome engineering efforts and target for recombineering project (Chapter 5)	WT	DSMZ – Liou *et al*. 2005
C. acetobutylicum ATCC 824	Recombinant methylation donor proof of concept (Chapter 5)	WT	SBRC Nottingham culture collection
C. difficile DH1916	Recombinant methylation donor testing (Chapter 5)	WT	*C. difficile* DH1916 (Anaerobic Reference Laboratory: Cardiff)
C. botulinum A ATCC 19397	Recombinant methylation donor testing (Chapter 5)	WT	SBRC Nottingham culture collection

2.5. Culturing and fermentation

All Gram-positive strains (i.e., clostridia and eubacteria) were standardly cultured at 37 ^{o}C in an anaerobic cabinet, MG1000 Mark II Anaerobic Work Station (Don Whitley Scientific, UK). Where appropriate each media was supplemented with antibiotic prior to inoculation (Table 2.5.3). Solid media was produced with the addition of 15 g/L agar to relevant broth media before autoclaving. Broth media was placed into cabinets at least 24 hours before use, and plates at least 6 hours before use.

2.5.1. E. limosum culture medium

E. limosum strains were cultured on AWM1 or AWM2 media. Fermentations were performed in 30 ml media volumes in 120 ml serum bottles at 37 ^{o}C with constant shaking at 220 RPM.

AWM1 per litre: K2HPO4, 0.45 g; KH2PO4, 0.33 g; NH4Cl, 1 g; NaCl, 20 mg; MgSO4.7H2O, 0.16 g; CaCl2, 20 mg; Yeast, 2 g; NaHCO3, 10.0 g; SL9 solution, 1 ml; Selenite/tungstate solution 1 ml; Vitamin solution 20 ml (Table 2.5.1); pH adjusted to 7.2 with NaOH. Application dependent carbon sources glucose, fructose, or methanol were sterilely added after filter sterilisation and degassing (24 hour in anaerobic cabinet) of the bulk media at working concentrations of 5 g/L, 5 g/L, 62mM respectively. Selection for uracil autotrophy was performed on AWM1 with yeast extract substituted for an equal amount of casamino acids, and plates supplemented with 0.05% w/v L-cystine.

AWM2 per litre: KH_2PO_4, 1.76 g; K_2HPO_4, 8.44 g; NH_4Cl, 1 g; cysteine hydrochloride, 0.5 g; NaCl, 2.9 g; yeast extract, 2 g; $KHCO_3$, 6 g; Resazurin 1% (w/v), 0.1 ml; SL9 solution, 1.0 ml; selenite/tungstate solution, 1 ml; vitamin solution, 2.0 ml (Table 2.5.1); pH adjusted to 7.2 with NaOH. Application dependent carbon sources glucose, fructose, or methanol were sterilely added after autoclave sterilisation and degassing (24 hour in anaerobic cabinet) of the bulk media at working concentrations of 5 g/L, 5 g/L, 62mM

respectively. Additionally, 1.5mM $MgSO_4$ was added once media had been made anaerobic.

2.5.2. *C. acetobutylicum* medium

C. acetobutylicum strains were cultured on 2xYTG (Yeast, Tryptone, Glucose). 2xYTG per litre: tryptone, 16 g; yeast extract, 10 g; NaCl, 5 g; glucose 5 g. Medium was pH adjusted to 5.2 and sterilised by autoclave.

2.5.3. *C. difficile* culture medium

C. difficile was cultured in brain heart infusion (BHIS) medium (Oxoid), supplemented with 0.1% w/v L-cysteine (Calbiochem) and 5 mg/ml yeast extract (Oxoid).

2.5.4. *C. carboxidivorans* culture medium

C. carboxidivorans was grown on YTF medium. YTF medium per litre: tryptone, 16 g; yeast extract, 10 g; fructose, 10 g; NaCl, 0.2 g; 1000x acid trace element solution, 1 ml; 1000x basic trace element solution, 1 ml; vitamin stock solution 1 ml (Table 2.5.2).

Table 2.5.1 – AWM media additives

Vitamin solution – Source: DSMZ 141	per Liter
Biotin	2.0 mg
Folic acid	2.0 mg
Pyridoxine-HCl	10.0 mg
Thiamine-HCl x 2 H_2O	5.0 mg
Riboflavin	5.0 mg
Nicotinic acid	5.0 mg
D-Ca-pantothenate	5.0 mg
Vitamin B12	0.1 mg
p-Aminobenzoic acid	5.0 mg
Lipoic acid	5.0 mg

Selenite-tungstate solution	Per Liter
NaOH	0.5 g
Na_2SeO_3 x 5 H_2O	3.0 mg
Na_2WO_4 x 2 H_2O	4.0 mg

Selenite-tungstate solution	Per Liter
Nitrilotriacetic acid	12.8 g
$FeCl_2$ x 4 H_2O	2.0 g
$ZnCl_2$	0.070 g
$MnCl_2$ x 4 H_2O	0.1 g
H_3BO_3	0.006 g
$CoCl_2$ x 6 H_2O	0.19 g
$CuCl_2$ x 2 H_2O	0.002 g
$NiCl_2$ x 6 H_2O	0.024 g
Na_2MoO_4 x 2 H_2O	0.036 g
Nitrilotriacetic acid	12.8 g

Table 2.5.2 – YTF media additives

1000x Acid trace element solution	Per Liter
HCl	50 mM
H_3BO_3	100 mg
$MnCl_2.4H_2O$	230 mg
$FeCl_2.4H_2O$	780 mg
$CoCl_2.6H_2O$	103 mg
$NiCl_2.6H_2O$	602 mg
$ZnCl_2$	78 mg
$CuSO_4.5H_2O$	50 mg
$AlK(SO_4)_2.12H_2O$	50 mg

1000x Basic trace element solution	Per Liter
NaOH	10 mM
Na_2SeO_3	58 mg
Na_2WO_4	53 mg
$Na_2MbO_4.2H2O$	52 mg

B-vitamin solution	Per Liter
p-aminobenzoate	114 mg
Riboflavin	104 mg
Thiamine	200 mg
Nicotinate	206 mg
Pyridoxin	510 mg
Pantothenate (calcium)	104 mg
Cyanocobalamin	78 mg
d-biotin	22 mg
Folate	48 mg
Lipoate/thioctic acid	50 mg

2.5.5. Antibiotic concentrations

Table 2.5.3 – Working antibiotic concentrations for *E. coli* and Gram-positive species.

Antibiotic	Diluent	Working concentration (μg/ml)	
		E. coli	Gram-positive organisms
Ampicillin	dH_2O	*100*	
Apramycin	dH_2O	*50*	
Chloramphenicol	EtOH (100%)	*25*	
Clarithromycin	DMSO (100%)		6
Erythromycin	EtOH (100%)	*500*	
Kanamycin	dH_2O	*50*	
Spectinomycin	dH_2O		250
Thiamphenicol	DMSO (100%)		15

2.5.6. Media supplementation concentrations

Table 2.5.4 – Working media supplement concentrations for *E. coli* and Gram-positive species.

Antibiotic	Diluent	Working concentration (μg/ml)	
		E. coli	Gram-positive organisms
Lactose	dH_2O		20
Uracil	dH_2O		20
L-arabinose	dH_2O	10-100	

2.5.7. Sample analysis

Analysis of metabolites was performed on culture supernatant of fermentation reactions using gas-chromatography with a Focus GC (Thermo Scientific), using flame ionisation detection (FID) for analytes detection and quantification. Analytes measured were: acetate, acetone, butyrate, butanol, ethanol, iso-propanol, and internal-standard valerate. Standards were quantified against a standard curve produced by dilution series of 0, 0.2, 0.5, 1, 2.5, 5, 10, 50, 100 mM of each standard. Samples were prepared by centrifugation of 1-2 ml fermentation samples for 5 minutes at maximum speed, in order to pellet cell biomass for supernatant extraction. Supernatant was stored at -20 ^{o}C until subsequent preparation steps. Once all sample supernatants had been

collected, 5 µl of 10M H_2SO_4 was added to 500 µl aliquots of sample, before addition of 500 µl of propyl-propionate containing 50mM valerate. Sample mixtures were then vortexed vigorously for 10 seconds and then centrifuged at maximum speed for 1 minute. The top phase of the centrifuged mixtures was extracted and added to a GC sampling vial ready for analysis.

Samples were run on a Thermo Scientific TR-FFAP 30m x 0.25mm x 0.25 µm column, using a hydrogen flow rate of 0.8 ml/min, and priming conditions of: 50 °C hold, 1 min; 210 °C ramp at 40 °C/min; 210 °C hold for 1 minute; inlet temperature of 240 °C with split ratio 50; 1 µl injection. Detection was performed with FID at 270 °C. Total column runtime was 6 minutes, with ~6 addition minutes of cool-down and pre-injection washes. Chromatogram analysis was performed using ChromQuest, with automated peak calling and integration.

2.6. Gene-transfer in target strains

2.6.1. Transformation of *E. limosum*

2.6.1.1. *E. limosum* competent cell preparation

Electrocompetent cells were prepared in 300 ml volumes of AWM1 media, inoculated from an appropriately grown exponential phase starter culture. This culture was grown from a starting OD_{660} of <0.1 until mid-stationary phase, in the OD_{660} range 0.4-0.8 (>5 hours). Cells were then transferred to into chilled 50 ml falcon tubes, and maintained at 4 °C or on ice for the remainder of the preparation process. The chilled cell culture was then centrifuged at 7000 x *g* and 4 °C for 10 minutes and the resulting supernatant was discarded. Cell pellets were then resuspended in chilled SMP buffer (270 mM sucrose, 7mM sodium phosphate, 1mM $MgCl_2$, pH 7.2) and collated into a total volume of ~25 ml, before centrifugation and supernatant removal steps were repeated. This SMP wash step was repeated once more before a final resuspension in 1-2 ml of SMP with DMSO 10% v/v. The resuspended cell volumes were subsequently split into 100 µl

aliquots for immediate use or storage at -80 °C. All centrifugation and handling steps were performed in an anaerobic tent (Coy Laboratories).

2.6.1.2. *E. limosum* electroporation

All *E. limosum* transformations were performed using a GenePulsar Xcell™ electroporator, with both CE and PC modules (BioRad,USA). Competent cell aliquots (100 µl) were added to ≥500 ng of plasmid DNA in 2 mm gap Gene Pulser® electroporation cuvette (BioRad), and incubated on ice for 10 minutes before electroporation. Electroporation parameters were as follows: exponential wave, voltage 1.5 kV, resistance 200 Ω, capacitance 50 µF. These settings resulted in a time constant of between 8 and 10 ms. Fresh AWM1 media was added to a total volume of 1 ml directly to the plasmid-cell mix following cessation of the electrical pulse. Shocked cultures were then transferred to a pre-prepared recovery tube containing a further 1 ml AWM1, and allowed to recover for 3½ hours. Following recovery cells were resuspended by light vortex and inversion before application on AWM + antibiotic selection plates. Recovery volumes of 15 µl, 150 µl, and 1.5 ml (pelleted first by centrifugation to a spreadable volume ~100 µl), were generally spread across three selection plates. Growth of antibiotic-resistant colonies were typically observed after 72 hours where Gram-positive replicon pCB102 was employed, and 48 hours with the replicon p19.

2.6.2. Transformation of *C. acetobutylicum*

2.6.2.1. *C. acetobutylicum* competent cell preparation

Exponentially growing (OD_{600} 0.2-1) *C. acetobutylicum* was used to inoculate 75 ml of 2xYTG medium. Cells were cultured to an OD_{600} 0.4 to 0.8 and harvested by centrifugation at 4 °C for 10 minutes at 5000 x *g*. Supernatant was subsequently removed and pellets resuspended in EPB buffer (274 mM sucrose, 5 mM sodium phosphate). Samples were centrifuged as before, and resultant pellets resuspended in

5 ml fresh EPB. Centrifugation was repeated once more before final resuspension in 1 ml of EPB. Samples were either used immediately or stored at -80 following the addition of 10% DMSO.

2.6.2.2. *C. acetobutylicum* electroporation

All *C. acetobutylicum* transformations were performed using a GenePulsar Xcell^{TM} electroporator, with both CE and PC modules (BioRad,USA). Competent cell aliquots (570 µl) were added to 500 ng of plasmid DNA, and allowed to incubate for 2 minutes on ice. Samples were then electroporated using the following settings: exponential wave, voltage 2.0 kV, resistance ∞ Ω, capacitance 25 µF. Pulsed samples were transferred to 10 ml fresh 2xYTG medium and recovered for 4 hours. Samples were then centrifuged at 5000 x *g* and plated onto selection plates with appropriate antibiotic selection

2.6.3. Conjugation of *C. difficile* and *C. botulinum*

Strains were restored from glycerol or DMSO -80 ^{o}C freezer stocks by plating onto respective solid growth media. Colonies were picked from these recovery plates for the inoculation of broth media. After 24-48 hours colonies were passaged into fresh media as overnight starter cultures. Donor strains of *E. coli* containing the transfer plasmid were set up in parallel overnight cultures. L-arabinose was added to donors where required for expression of methylation cassettes (Chapter 5). 1 ml of donor strain overnight cultures were centrifuged at 3000 x *g* for 2 minutes and supernatant removed, before washing with 500 µl PBS. Donor pellets were then transferred to the anaerobic cabinet. The pellets were resuspended in 200 µl of target-host cell culture and spread onto mating plates. Mating plates were L-arabinose (100 mM) containing agar medium for the respective target-host, without antibiotic. Mating plates were left for 24 hours before growth was harvested by flooding plates with PBS (~1 ml), and transferring the resultant slurry to appropriate selection plates. Dilution series of

transformants was produced by serial dilution of slurry in PBS prior to transfer to selection plates. D-cycloserine was used to prevent growth of *E. coli* donor strains for all subsequent plating stages following mating.

2.7. Chloramphenicol Acetyltransferase Assay

A protocol derived from Shaw (1975) to measure chloramphenicol acetyltransferase (CAT) activity. Samples were inoculated from an exponentially growing culture and grown until mid-exponential phase, OD_{600} 0.45 and 1.2 (discussion – most relevant to continuous fermentation conditions). Where required, appropriate inducers were added to their relevant concentrations to media prior to inoculation.

2.1.1. CAT assay sample preparation

Samples were balanced to 1 OD_{600} unit of cells, which was harvested by pelleting at 17,000 x *g* before removing the supernatant and storing sample tubes at -20°C until further processing. Once all samples had been collected, cell pellets were washed to remove unwanted extracellular polysaccharide. This was achieved by thawing samples on ice and resuspending them in 1 ml 1.5M NaCl solution. Resuspended samples were pelleted at 21,000 x *g* at 4°C for 10 minutes, and supernatants were removed by careful pipetting. The Salne wash process was repeated one additional time. After the second wash step, samples were resuspended in 250 µl of a lysis master mix, comprised of 1x Bugbuster® protein extraction reagent (EMD Millipore™) supplemented 1:100 with Protease Inhibitor Cocktail Set VII (Calbiochem). Samples were then incubated at 37°C with constant agitation by 220 rpm shaking for 45 minutes, before being subjected to 10 minutes of sonication. Sonications were performed in 1.5 ml Eppendorf tubes in a Bioruptor® (Diagenode) at ~4°C with 5 cycles of high intensity sonication with 30 seconds on, 30 seconds off.

2.1.2. CAT assay analysis

Samples were analysed on a Tecan Infinite® M1000PRO (TECAN) plate reader, maintained at 25°C, and using a 96-well plate configuration with automated injection. An assay master-mix was prepared for automated injection to the following specifications: 0.005% w/v chloramphenicol (Sigma Aldrich); 0.167mM acetyl-CoA sodium salt (Sigma Aldrich); 0.0833mM 5,5'-dithio-bis-[2-nitrobenzoic acid] (DTNB)(ThermoFisher Scientific); 100mM Tris buffer pH7.8 (Fisher Scientific). To each of the wells in the 96 well plate which were to be analysed, 90 µl of 100mM Tris buffer was added, followed by 10 µl of sample. Samples and master-mix were prewarmed to room temperature before loading into the plate reader. The Tecan was prepared for injection by washing and flushing of the system, with 3 full strokes (1 ml) of dH_2O and master-mix respectively. The assay was then run on selected wells, utilising the automatic injection to add master-mix, using the following program settings: Wait for temperature min 24.5 °C, max 25.5 °C; Injection 100 µl at a rate of 100 µl/s; Orbital shaking for 4 seconds at a magnitude of 1.5 mm; 10 kinetic cycles of 412nm absorbance analysis (average of 25 flashes) were performed per sample, set at 9 second intervals. Units (U) of CAT activity in each sample was calculated based on the maximum linear 412 nm absorbance rate increase, using Equation 3. Sample rates were subsequently adjusted for protein concentration (using Method 2.1.3) to give a final activity value for each given promoter construct in U/mg protein.

$$\text{U / ml sample} = \frac{(\text{rate})(0.1)(\text{df})}{(0.0136)(\text{v})}$$

0.1 = Assay volume; df = Dilution factor; 0.0136 = Micromolar extinction coefficient for TNB at 412 nm; v = Volume of protein extraction (ml).

Equation 3 – Formula for CAT activity calculation in tested samples.

2.1.3. Protein quantification by BCA assay

The protein concentration of samples was measured by BCA assay, using the Pierce™ BCA Protein Assay Kit, 96 well microplate format. A dilution series of diluted albumin (BSA) standards was produced, containing nine samples with a range of 0-2000 µg/ml. A working reagent was generated at a 50:1 ratio of BCA reagents, A:B. Samples and standards were added in 25 µl volumes to a 96 well microplate, followed by a 200 µl volume of working reagent. Once all samples had been added, the microplate was covered and thoroughly shaken for 15 seconds to ensure mixing, before a half hour incubation at 37 °C. The plate was then moved to room temperature for 5 minutes before analysis on a Tecan Infinite® M1000PRO (TECAN) plate reader, by measuring absorbance at 562 nm. Concentration values were measured in at least duplicate for all samples and standards. Average absorbance readings for each sample were mapped against the linear standard curve plotted by the BSA dilution series.

2.8. RFP reporter Assay

The intensity of RFP expression was measured using a protocol derived from Bi *et al* 2013[120]. An RFP gene derived from pKTrfp (AddGene), was cloned into a pRBV backbone (Method 2.3.4) and subsequently recombineered into pR702 in *E. coli re-InterStellar* (Method 2.13.2), to create strain rIS-RFP. An overnight culture of rIS-RFP was set up in LB supplemented with kanamycin and apramycin. 5 µl of overnight culture was used to inoculate a flat bottomed Greiner 96 well plate containing 195 µl of LB media supplemented with kanamycin, apramycin, and varying levels of L-arabinose. The tested L-arabinose working concentrations were 0, 0.1, 1, 10, and 100mM. Samples were incubated at 30 degrees for 48 hours in a CLARIOstar microplate reader (BMG Labtech) with constant agitation, 200 rpm orbital shaking. Fluorescence of the samples was measured every 10 minutes with 40 flashes per well of excitation at 584 nm wavelength and an emission detection at 621 nm. OD_{600} was also measured for each

timepoint. Fluorescence intensities were calculated, following adjustment for appropriate sample blanks, as a product of fluorescence intensity over OD_{600}.

2.9. Genomic DNA extraction

Where DNA was required for the purpose of general laboratory purposes (PCR amplifications etc.), and for *E. coli*, genomic DNA extractions were performed using the GenElute Bacterial Genomic DNA kit (Sigma Aldrich), following the relevant manufacturer guidelines (Quick method - 2.9.1). Where high purity genomic DNA was required (Next generation sequencing etc.), DNA was extracted using a phenol chloroform-based extraction methodology (High purity method - 2.9.5).

2.9.1. Quick method

Quick genomic extractions were performed using the GenElute Bacterial Genomic DNA kit (Sigma Aldrich) following the manufacturers recommended protocol. Two distinct extraction protocols were used, one specific to Gram-positive (2.9.2), another to Gram-negative (2.9.3) species. A universal DNA purification protocol (2.9.4) was then applied to the resulting extraction lysate of either extraction method.

2.9.2. Standard Gram-positive gDNA extraction

Cell pellets were harvested by 16000 x *g* centrifugation of 1-4 ml of overnight cultures of Gram-positive bacteria grown in appropriate semi-defined media. Supernatant was discarded and cell pellet resuspended in 200 µl of a lysozyme solution, containing 10 mg/µl of lysozyme from chicken egg white. The resuspended solutions were then incubated at 37 °C for half an hour, followed by a two minute, room temperature, incubation after adding 20 µl of RNase A solution. Subsequently, 20 µl of proteinase K

solution and 200 µl Lysis-solution were added to each sample, vortexed thoroughly, and incubated at 55 °C for 10 minutes.

2.9.3. Standard Gram-negative gDNA extraction

Cells were harvested by 12000 x *g* centrifugation of 1.5 ml of *E. coli* overnight culture, grown in LB medium supplemented with appropriate antibiotics where required. Supernatant was discarded and the remaining cell pellet resuspended in 180 µl of Lysis Solution T and 20 µl of RNase A solution before 2 minute incubation at room temperature. 20 µl of proteinase K solution was then added to the sample and mixed thoroughly before further incubation at 55 °C for half an hour. A second lysis step was performed with the addition of 200 µl of Lysis Solution C, before 10 minutes further incubation at 55 °C.

2.9.4. Post extraction gDNA purification

Purification methodologies was consistent for extraction processes of both Gram-positive and Gram-negative species. GenElute Miniprep Binding Columns, installed in 2 ml collection tubes, were prepared for DNA binding with the addition of 500 µl of column preparation solution directly to the column. Column preparation solution was eluted by 1 minute of centrifugation at 12000 x *g* and discarded. Sample lysates were prepared for column binding with the addition of 200 µl of neat ethanol mixed by gentle inversion before loading onto a prepared binding column. Samples were centrifuged for 1 minute at 10000 x *g* and the columns transferred to fresh 2 ml collection tubes. Samples were washed with the addition of 500 µl Wash Solution 1 and centrifugation at 10000 x g. A second wash was performed after moving the columns to a fresh 2 ml collection tube with the addition of 500 µl Wash solution and centrifugation at max speed for 3 minutes. Columns were then transferred to a final 1.5 ml microfuge tube for elution. An elution step consisting of addition of 25 µl 55 °C dH_2O, 5 minute

incubation at 55 °C, and centrifugation at 12000 x *g* for 1 minute, was performed twice with each sample. The resultant 50 µl of purified gDNA was stored at -20 °C until use.

2.9.5. High purity gDNA extraction method

Where high purity DNA was required from Gram-positive species in this study (*E. limosum* and *C. carboxidivorans*), phenol-chloroform extractions based on the following methodology were employed. Volumes of 1-10 ml of early-mid exponential phase cultures were harvested by centrifugation at 7000 x *g*, 4 °C for 10 minutes. Supernatant was removed and the resultant cell pellet was resuspended in 180 µl lysis buffer (10 mg/ml lysozyme) for 45 minutes at 37 °C. Next, sample were mixed with a 220 µl solution, containing 25 µl of 20 mg/ml protease K solution (Qiagen), 85 µl dH_2O and 110 µl 10% w/v SDS (Sigma-Aldrich). The mixed sample was then subjected to a further 30 minute incubation step at 65 °C with regular mixing by inversion. Equal volumes of prepared sample and Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v/v) were added to a 5PRIME Phase Lock Gel tubes (Quanta Bio) and centrifuged for 3 minutes at 17,000 x *g*. The top phase was removed and added to a fresh phase-lock tube containing Phenol:Chloroform:Isoamyl Alcohol and was centrifuged as previously. This step was repeated a further two times per sample. The top phase was then transferred into a prechilled mixture of 40 µL sodium acetate and 800 µL and incubated at -80 °C for 1 hour. The sample was then centrifuged at 15 minutes at 12,000 x g at 4 °C and the supernatant discarded. The DNA pellet was washed in 70% ethanol and centrifugation repeated. Supernatant was again discarded and the remaining liquid allowed to air dry. The resulting pellet was resuspended in 50 µl dH_2O. Samples were stored at -20 °C for up to six months.

2.10. Polymerase Chain Reaction (PCR)

PCR was used to amplify DNA sequences from a DNA template as defined by complementary pairs of flanking primers. The PCR primers used throughout this research project are available at Method 2.14.2.5. Tm values were calculated using online Tm calculators employing the modified Breslauer annealing temperature calculation algorithms[121].

2.10.1. DNA Polymerases

2.10.1.1. Q5 High-Fidelity

Q5 High-Fidelity Polymerase (NEB) was used for applied genetic work (e.g., cloning) and confirmatory sequencing in this study, where the fidelity of amplified DNA sequences was integral to downstream processes, following the manufacturers recommended instructions for 25 µl PCR reaction volumes. Each reaction mix contained: 12.5 µl Q5 High-Fidelity 2x Master Mix, 2.5 µl Forward primer, 2.5 µl Reverse primer, ~50 ng Template DNA, up to a total volume of 25 µl with nuclease free dH_2O.

Table 2.10.1 – Q5 polymerase standard thermocycler protocol

Cycle Step	Temperature (°C)	Duration (Seconds)
Initial Denaturation	98	30
25-35 cycles	98	10
	50-72	20
	72	30 (/kb)
Final extension	72	120
Hold	10	-

2.10.1.2. DreamTaq

DreamTaq (ThermoFisher) was used for basic size-based screening PCR reactions, following the manufacturer's instructions for 25 µl PCR reaction volumes. Each reaction

mix contained: 12.5 µl DreamTaq Green PCR Master Mix (2X), 1 µl forward primer, 1 µl reverse primer, ~50 ng Template DNA, made up to a total volume of 25 µl with nuclease free dH_2O.

Table 2.10.2 – DreamTaq polymerase standard thermocycler instructions.

Cycle Step	Temperature (°C)	Duration (Seconds)
Initial Denaturation	95	60-180
25-35 cycles	95 Tm –5 72	30 30 60 (/kb)*
Final extension	72	300-900
Hold	10	-

*60 seconds recommended for 2 kb, with an additional 60 seconds added for each additional kb.

2.10.2. PCR methodologies

2.10.2.1. Colony PCR

Colony PCRs were slight variations on the protocols described above, which enabled PCR to be performed directly from cells without requiring prior DNA extraction or purification. The template for colony PCRs were produced by resuspension of a small colony from plate, or dilution of a 2 µl volume of densely grown broth culture, in 20 µl of dH_2O. A 1 µl volume of this solution was used in each 25 µl PCR reaction in place of the standard gDNA or plasmid template. An initial denaturation step of 10 minutes at 98 °C was added to the start each PCR program to denature cells and expose DNA to the PCR reagents. Aside from these changes, the PCR reactions were identical to those using standard DNA templates.

2.10.2.2. Splicing by Overlap Extension PCR

Splicing by overlap extension (SOE) PCR was employed in order to join together homology arm (HA) pairs (LHA & RHA) prior to ligation into CRISPR (Method 2.3.2) and ACE (Method 2.3.1) vectors. Q5 High-Fidelity polymerase was used for all SOE reactions. Four primers were prepared for each SOE PCR reaction. Two standard primers at distal ends of the two HAs to be spliced together, [LHA_Fw] and [RHA_Rv]. Two SOE primers situated at the ends of the HAs to be spliced together, [LHA_SOE_Rv] and [RHA_SOE_Fw]. As well as the standard primer annealing sequence, both SOE primers also contained a 5' overhang which conferred homology to the sequence of the other SOE primer. HAs were first amplified individually, LHA with [LHA_Fw][LHA_SOE_Rv], RHA with [RHA_SOE_Fw][RHA_Rv]. Once amplified, both PCR products were extracted and purified to be used as a template in a second PCR reaction, using total of ~50 ng purified HA at a ratio of approximately 1:1. The second PCR reaction used just the external primers, [LHA_Fw][RHA_Rv], and relied on homology of both HA templates, generated by the SOE primer overhangs, to prime splicing of the two-HA into a single final product.

2.10.2.3. Touch-Down PCR

Where standard PCR protocols (2.10.1.1, 2.10.1.2) failed, a modified thermocycling protocol was employed in order to improve PCR reaction conditions for given primer pairs through progressive annealing-temperature (Tm) adjustments. Tm values were decreased by 0.5-1 ^{o}C per cycle, starting 5 ^{o}C above the predicted Tm value (Method 2.10), until 5 ^{o}C below the predicted Tm value. The PCR protocols were then set to continue at this final Tm for a further 25 cycles before continuing on to the final extension step.

2.11. DNA quantification and purity assessment

DNA was quantified for various purposes to ensure optimal conditions were achieved for molecular biological methods. For general lab quantifications, such as quantifying ligation fragments or pre-sequencing evaluations, a Nanodrop microvolume spectrophotometer was used. Where more accurate quantifications were required, for example prior to next generation sequencing, a Qubit fluorometer was used. Purity and quality of DNA was examined by 260/280 and 260/230 ratios by nanodrop and by visual examination of a gel run sample following electrophoresis. Ratios values of less than 1.8 for 260/280 and 260/230 indicated contamination by protein/phenol and carbohydrates respectively. A smearing effect on gel images suggested denaturation of DNA, while light smearing at less than 5000 bp was used as an indicator of RNA contamination.

2.11.1. Nanodrop

The Nanodrop Lite spectrophotometer (ThermoFisher) was used to quantify purified DNA samples, following the manufacturer's instructions. Following appropriate blanking measurements with the relevant elution buffer or dH_2O, sample volumes of 1-2 µl were loaded directly onto the lower pedestal for quantification. Nucleic acid concentrations were calculated automatically based on an extinction coefficient of one for the absorption of 50 ng/µl of dsDNA at 260 nm.

2.11.2. Qubit

The Qubit 3.0 Fluorometer (ThermoFisher) was used to quantify purified genomic DNA samples following the manufacturer's instructions for broad range dsDNA quantification. A Qubit Working Solution was made by diluting the Qubit reagent 1:200 in Qubit buffer. Next, assay tubes were made up for each sample and both standard controls, in clear thin-wall 500 µl PCR tubes. Volumes of between 1-20 µl of each

sample, and 10 µl of each standard, were added to working assay tubes and made up to a total volume of 200 µl with Qubit Working Solution. Assay tubes were then vortexed for 2-3 seconds before incubation for 2 minutes at room temperature. Finally assay tubes were loaded into the Qubit 3.0 Fluorometer to take readings. Concentrations of dsDNA were calculated based on the reported fluorescence of each samples in relation to an automatically generated standard curve generated from the output control standards pair.

2.12. Genome

2.12.1. Sequencing

2.12.1.1. Sample preparation for NGS and SMRT

DNA samples for NGS and SMRT sequencing were prepared using phenol chloroform extraction (Method 2.9), and sent to the service providers on dry ice following successful completion of internal quality assessment measures. Quality assessment was performed as follows: purity by nanodrop (Method 2.11.1); concentration by Qubit (Method 2.11.2); and integrity be gel electrophoresis (Method 2.1.10).

2.12.1.2. PacBio (SMRT) sequencing

PacBio sequencing of *E. limosum* ATCC 8486 was performed using the PacBio RSII system (PacBio, USA) at the Genome Quebec Innovation Centre (McGill University, Canada). Sequencing employed sheared large insert library and a single SMRT cell. Initial bioinformatic assembly of the genome was provided by Genome Quebec, yielding 3 primary contig files.

2.12.1.3. Illumina (NGS) sequencing

Next-generation sequencing was performed by DeepSeq (University of Nottingham, UK) and MicrobesNG (University of Birmingham, UK) using the Illumina MISeq platform (Illumina, USA). Genomic DNA samples, extracted by phenol chloroform, was sent to the service providers on dry ice following successful completion of internal quality assessment measures. Quality assessment was performed as follows: purity by nanodrop (method 2.11.1); concentration by Qubit (method 2.11.2); and integrity be gel electrophoresis (method 2.1.10).

Genomic DNA libraries at both sequencing facilities were prepared using TruSeq PCR Free Library Preparation Kits (Illumina, USA). DNA was fragmented by sonication using Carvars S2 sonicator before purification and end-repair with AMPure-XP beads (Beckman Coulter). Y-adaptors were then added and purified before quantification of the resultant DNA libraries by qPCR (Kapa Biosystems). Libraries were sequenced using the Illumina MiSeq 500 bp V2. Sequencing data was returned as trimmed paired end read-sets for processing as appropriate.

2.12.2. Assembly

De-novo assembly was performed to close the genome of *E. limosum* ATCC 8486. The overlapping regions of three PacBio-generated contigs were determined manually using alignment functions of the online cloning software suite, Benchling. Following assembly of a closed genome, CLC genomics workbench 8.0 was used to map Illumina-generated paired-end reads to the sequence. A consensus of the Illumina read mappings was then generated to provide a final complete genome sequence.

2.12.2.1. Re-sequencing/Reference Mapping

Reseqeuencing was performed by mapping Illumina read data to a reference sequence generated within this project (*E. limosum* ATCC 8486) or available on NCBI. Mapping was performed using the CLC genomics workbench 8.0.

2.12.3. Annotation

Annotation of assembled genomes was performed using the IMG/R pipeline provided by the Joint Genome Institute[69]. The IMG/R annotation pipeline employs standard bioinformatic tools BLAST[122], MCL[123], and CLUSTALW[124] for genome analysis. Algorithmically determined coding sequences (CDSs) are interrogated against numerous external data repositories (RefSeq, EBI Genome Reviews, GenBank/EMBL, JGI microbial genome data, LIGAND, GO, COG, Enzyme, InterPro, KEGG, KO, Pfam, TIGRFam, CDD databases, and GOLD).

2.12.4. BLAST bioinformatics

The Basic local alignment search tool BLAST was used to compare and analyse DNA and protein sequences[122]. Alignments were typically performed against the NCBI nucleotide collection, using default alignment and search parameters unless stated otherwise. BLAST alignment confidence values (E values) presented in this study have been produced by unrestricted BLAST alignment unless stated otherwise.

2.12.5. Genome Alignment

Genome alignments were performed using the freely available Mauve software package. Genome FASTA files were aligned using progressiveMauve and resulting data were analysed and curated manually. ProgressiveMauve parameters were as follows: default seed weight, seed families not applied, Locally Collinear Blocks (LCB) determined, full alignment, iterative realignment, sum-of-pairs LCB scoring, default

minimum LCB weight. Alignment scoring was calculated using a HOXD pairwise substitution matrix[125].

2.12.6. Pan-Genome assembly

Pan-genome analysis was generated through the online Department of Energy Systems Biology Knowledgebase (Kbase) platform[29]. The pan-genome was determined through grouping of orthologous protein sequences performed using the OrthoMCL genome-scale algorithm on the complete genomes of three *E. limosum* strains, ATCC 8486 (type), SA11, and KIST612. Genomes for SA11 (NZ_CP011914) and KIST612 (CP002273.2) were downloaded from NCBI, ATCC 8486 was produced as in the course of this project (Chapter 3).

2.13. Recombineering

2.13.1. Strains

Table 2.13.1 – Table of strains used in the development of recombinant methyltransferase expressing methylation donor strains of *E. coli* (Chapter 5).

Strain	Function and properties	Genotype	Source
E. coli InterStellar (IS)	*E. coli* Stellar containing cryptic plasmid pR702	*ara,Δ(lac-proAB) [Φ80d lacZΔM15], rpsL(str), thi, Δ(mrr-hsdRMS-mcrBC), ΔmcrA, dam, dcm* - *R702-TcR, SmR, SuR, HgR Tra$^+$, Mob*	Millard, J. SBRC Nottingham
E. coli rInterStellar (rIS)	*E. coli* InterStellar containing pKD46-cm (Figure 2.14.10)	*ara,Δ(lac-proAB) [Φ80d lacZΔM15], rpsL(str), thi, Δ(mrr-hsdRMS-mcrBC), ΔmcrA, dam, dcm* - *R702-TcR, SmR, SuR, HgR Tra$^+$, Mob* - *Eco, gam, beta*	This project
Recombineered *E. coli* InterStellar (rIS-x*)	*E. coli* InterStellar containing a recombinant pR702	*ara,Δ(lac-proAB) [Φ80d lacZΔM15], rpsL(str), thi, Δ(mrr-hsdRMS-mcrBC), ΔmcrA, dam, dcm* - *R702-TcR, SmR, SuR, HgR Tra$^+$, Mob, RE***	
E. coli SExpress-reco	*E. coli* NEB Express transformed to contain a recombinant pR702	*fhuA2 [lon], ompT, gal, sulA11, R(mcr-73::miniTn10--TetS)2, dcm R(zgb-210::Tn10--TetS), endA1, Δ(mcrC-mrr)114::IS10* - *R702-TcR, SmR, SuR, HgR Tra$^+$, Mob, RE***	This project

*: 'x' is replaced by the gene name/designation for the inserted gene(s), as for pRBV1 assemblies (Table 2.16.4)

**: RE refers to a number of possible genotypes, determined by the recombinant insertionmade at the target locus, AadA1 of pR702.

2.13.2. Recombineering method

Linear DNA fragments were generated for recombineering using Q5 polymerase-based PCR of target RBV vectors using primer pair [200, 219] (Table 2.15.10). PCR products were then transformed by electroporation (Method 2.1.3) into *E. coli* rInterstellar competent cells primed for recombineering. Priming was achieved through arabinose induction at 100mM, applied on entry to exponential growth phase (i.e., added at OD_{600} 0.15-0.2) during competent cell preparation (Method 2.1.2). Cells were recovered in SOC media overnight at 30 °C. Samples were then spread onto selective plates containing 50 µg/ml apramycin and incubated at 37 °C until colonies were observed. Colonies were screened to confirm recombineering at the target locus by colony PCR and subsequent sanger sequencing (Source Biosciences, UK) using locus-external primer pair [338, 339] (Table 2.15.10). Loss of pKD46cm was screened using negative selection re-streaking on chloramphenicol plates.

2.13.3. Transfer of recombinant pR702

Transfer of pR702 between *E. coli* strains was mediated direct conjugal transfer and selection by auxotrophy. InterStellar strains with different recombinant modifications (e.g., rIS-RFP) were inoculated into overnight starter cultures in LB supplemented with kanamycin and apramycin. In parallel overnight cultures were set up for *E. coli* Express in LB with no selection. The following morning both strains were inoculated into fresh media and incubated separately for 3 hours at 37 °C with 220 rpm orbital shaking. At the end of the incubation step, the OD_{600} of each culture was measured, and the two strains were mixed to an equivalent OD_{600} ratio of 1:1. Mixed cultures were statically incubated at 37 °C for one hour and then plated onto selection plates. Proline deficient minimal media (Method 2.13.3.1) supplemented with apramycin was used to select for trans-conjugants *via*the antibiotic pressure, and against InterStellar based on auxotrophy (Method 2.13.1).

2.13.3.1. Minimal media for auxotrophic selection

M9 Minimal E. coli growth media consisted of the following, per litre: M9 Salts, 100 ml ($Na_2HPO_4.7H_2O$, 33.7mM; KH_2PO_4, 22mM; NaCl, 8.55mM, NH_4Cl, 9.35mM); CaCl2, 0.3mM; MgSO4, 1mM; biotin 1 µg, thiamine, 1 µg; glucose, 200 g; and amino acids 1g each for:

Included		Excluded
Glycine	Arginine	Glutamine
Alanine	Methionine	Isoleucine
Serine	Phenylalanine	Leucine
Threonine	Aspartic Acid	Proline
Cysteine	Glutamic Acid	
Valine	Asparagine	
Histidine	Lysine	

2.14. Core Plasmid Maps

Visualised of plasmid maps and short DNA sequences was performed on the browser-based platform Benchling.

2.14.1. Chapter 3

2.14.1.1. pRM03

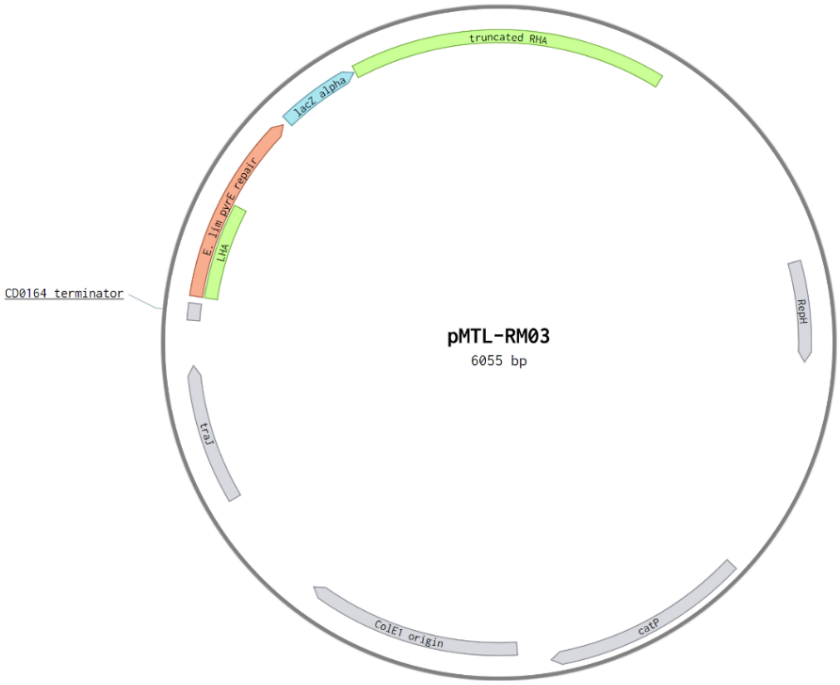


Figure 2.14.1 – Vector map for pRM03.

ACE repair vector for *pyrE* in *E. limosum* ATCC 8486 Δ*pyrE*. Vector is based on pMTL8315x backbone, where module x contains a partial *pyrE* gene (truncated at the N-terminus) derived from WT *E. limosum* ATCC 8486, a *lacZα* (multiple cloning site) from pMTL80000 vector series, and a truncated RHA derived from the sequence downstream of *pyrE* in WT *E. limosum* ATCC 8486.

2.14.1.2. pRM04

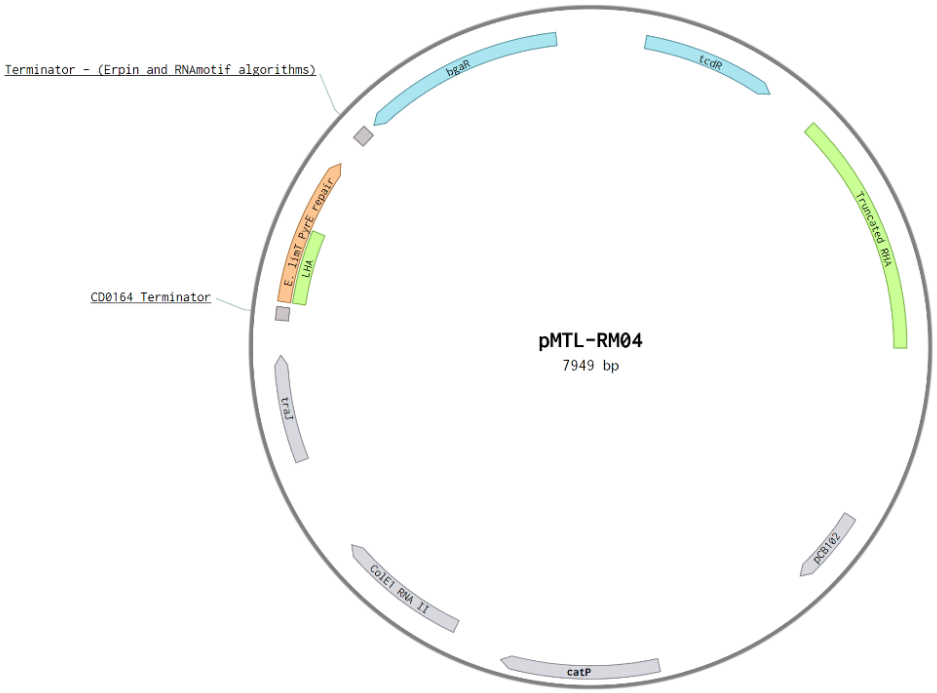


Figure 2.14.2 - Vector map for pRM04

Vector for ACE knock-in of Lactose inducible TcdR expression module at the *pyrE* of *E. limosum* ATCC 8486 Δ*pyrE* Vector is based on a pMTL8315x backbone, where module x contains a partial *pyrE* gene (truncated at the N-terminus) derived from WT *E. limosum* ATCC 8486, a lactose inducible promoter system *bgaR*-P_{bgaL} from *C. perfringens*, sigma factor TcdR from *C. difficile*, and a truncated RHA derived from the sequence downstream of *pyrE* in WT *E. limosum* ATCC 8486. Terminators determined by Erpin and RNAmotif algorithms.

2.14.1.3. pSBRCas9_*Pbt*_1

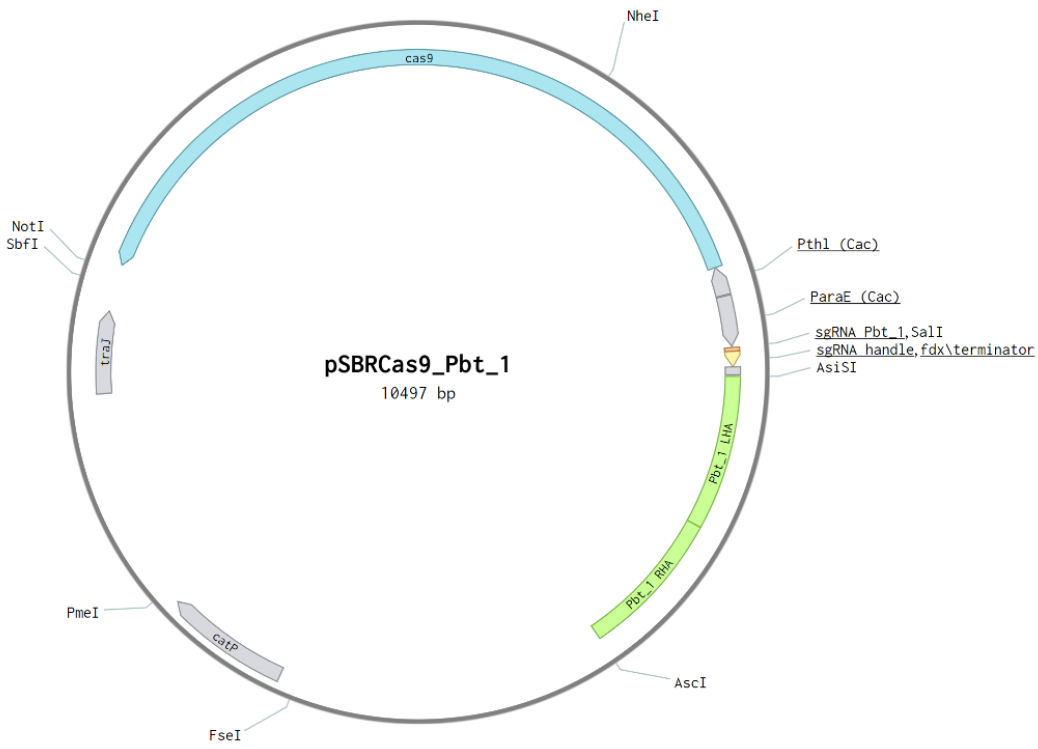


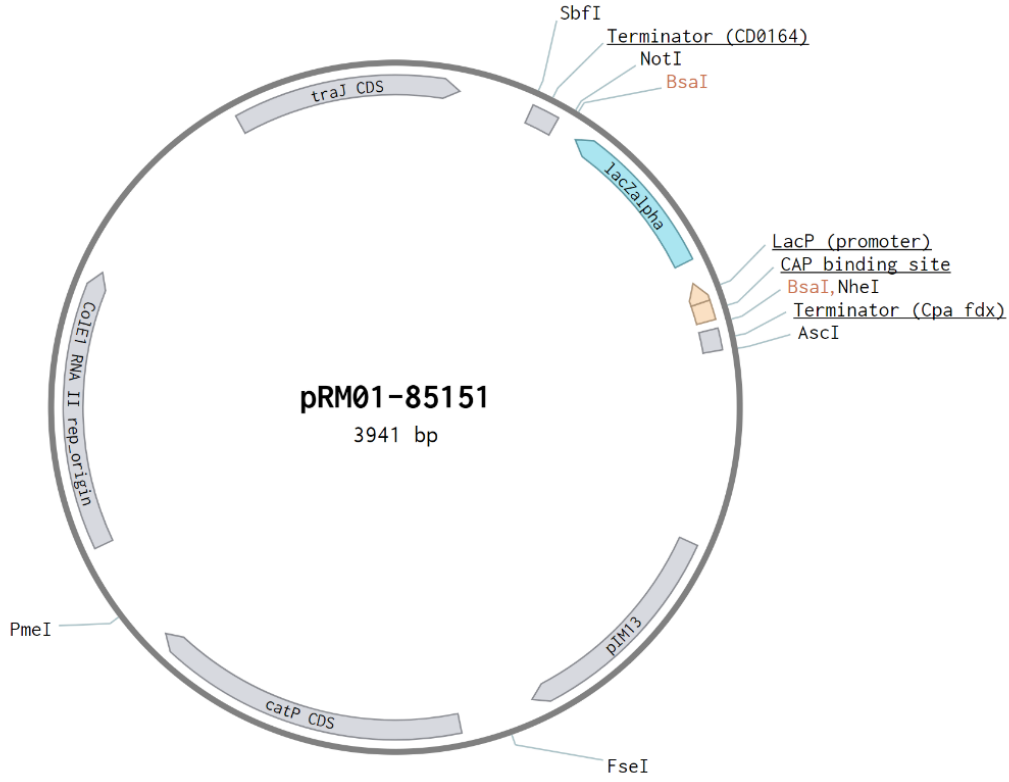
Figure 2.14.3 – Vector map for pSBRCas9_*Pbt*_1

A vector for targeted knockout of phosphate butyryltransferase 1 (Ga0111251_111764). Vector based on the pSBRCas9 backbone. Homology arms LHA and RHA represent 750 bp up and downstream of the *pbt*1 gene in *E. limosum* ATCC 8486. The sgRNA is a 20 bp sequence within the *pbt*1 gene for targeting Cas9 cleavage activity. The *cas9* gene is derived from *Streptococcus pyogenes*. P_{thl} and P_{araE} promoters derived from *C. acetobutylicum* ATCC 824 drive constitutive expression of *cas9* and the sgRNA respectively.

2.14.2. Chapter 4

2.14.2.1. pRM01

A)



B)

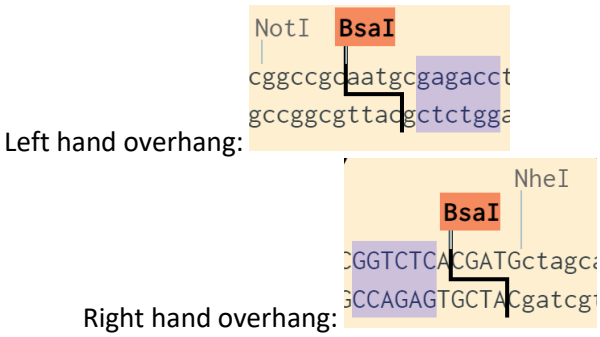


Figure 2.14.4 – Vector map for pRM01

A) pRM01. Type IIS assembly compatible pMTL8000x vector. A modified pMTL85151 vector with the standard MCS replaced with a *LacZ*α flanked by two BsaI sites for compatibility with TIIS assembly protocols. Insertions replacing *LacZ*α allow for blue white screening of integration events in appropriate *E. coli* hosts. B) Overhangs for insert compatibility: left – 5' CATT 3'; right – 5' CGAT 3'.

2.14.2.2. pMTL8J15-Isopropanol (P_{TcdB})

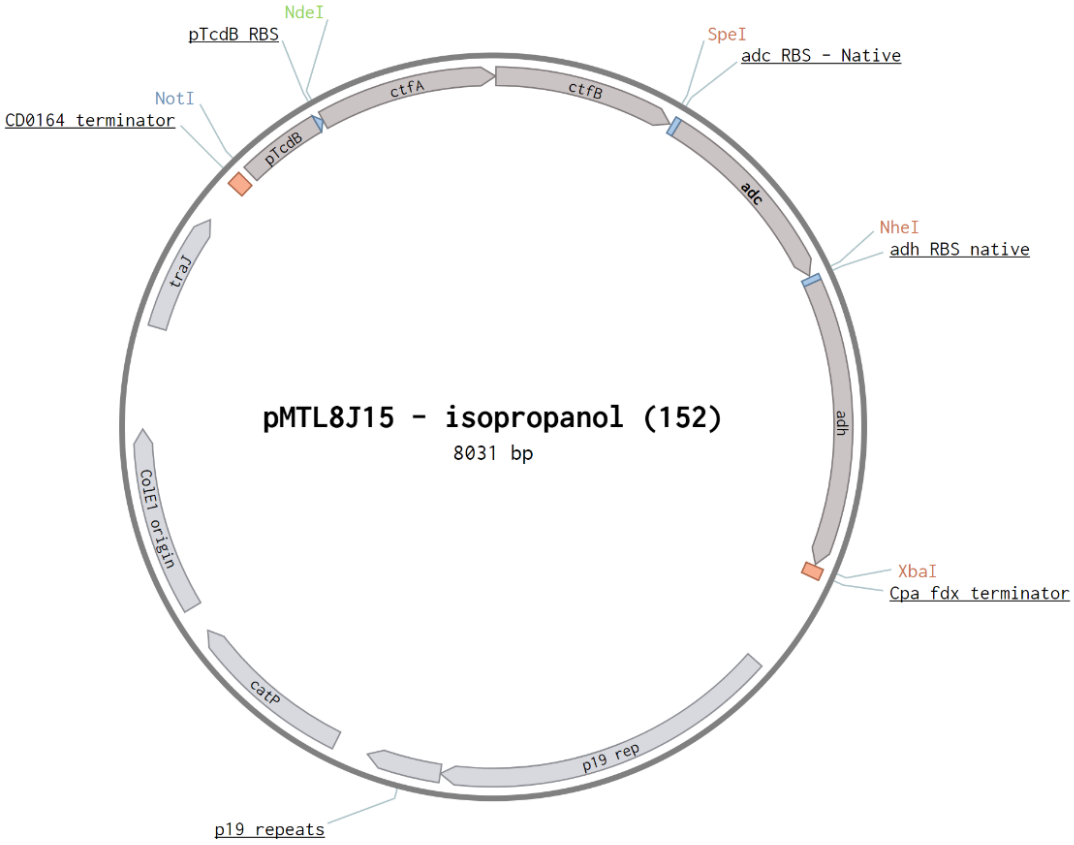


Figure 2.14.5 – Vector map for pMTL8J15-Isopropanol (P_{TcdB})

Example isopropanol cassette expression vector. The cassette (between terminators marked orange) was built originally in pRM01-8515 vector backbone using TIIS assembly methods (Table 2.15.3).

2.14.2.3. pJ204-Adh

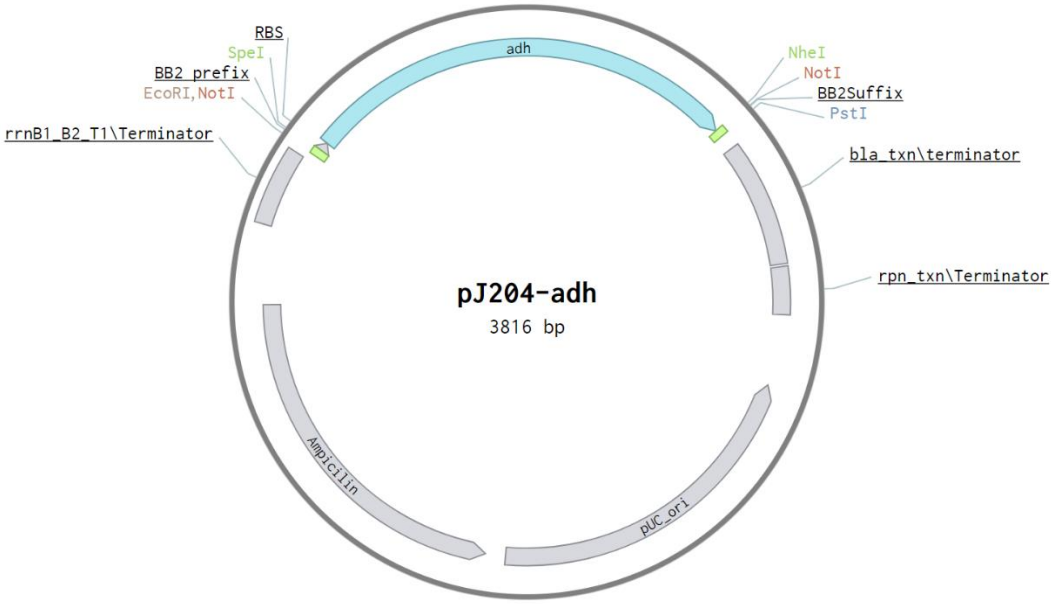


Figure 2.14.6 – pJ204-adh

An example holding vector for BioBrick2 parts used for assembly of dockerin/scaffoldin isopropanol gene expression cassettes.

2.14.2.4. pMTL8J15-Isopropanol-Dockerin-Scaffoldin

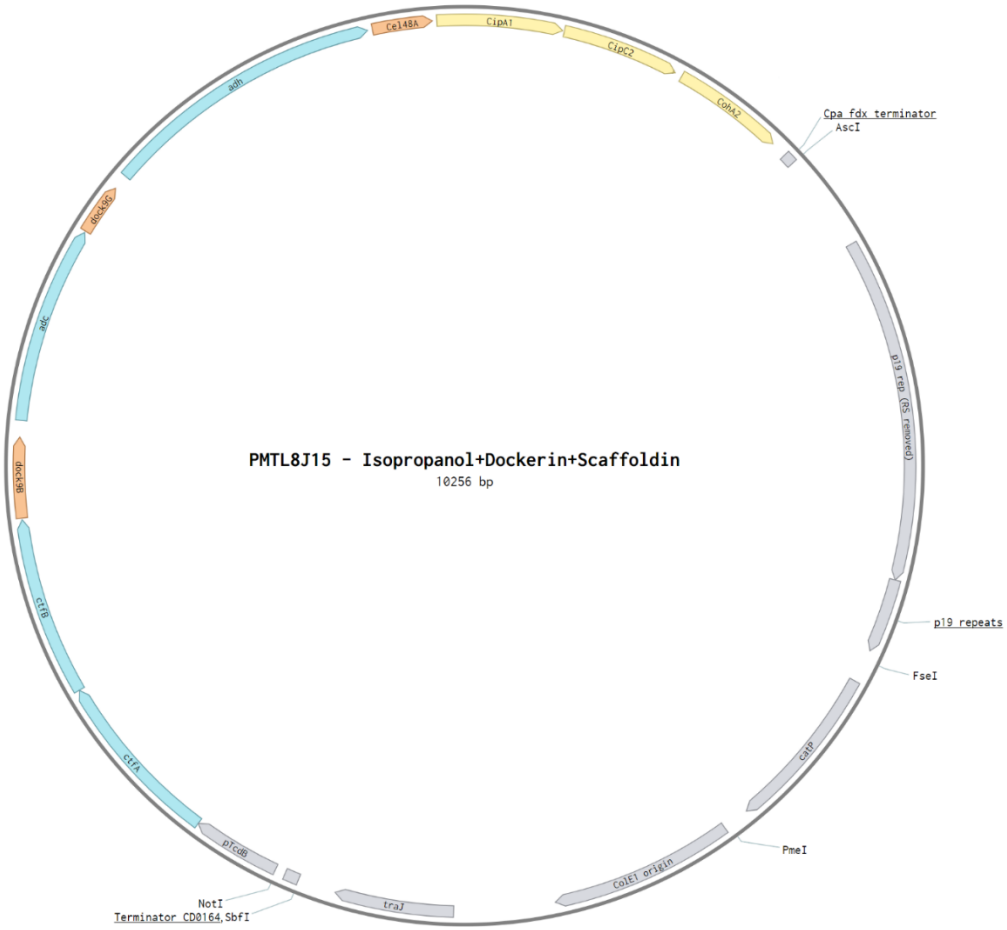


Figure 2.14.7 – Vector map for pMTL8J15-Isopropanol-Dockerin-Scaffoldin.

Isopropanol cassette expression vector with *ctfAB*, *adc*, and adh genes (blue) extended to contain N-terminal dockerin domains (orange) dcok9B, dock9G, and cel48A respectively. Each dockerin corresponds to a domain on a co-expressed synthetic scaffold protein (yellow) – dock9B:cipA2; dock9G:cipC2; cel48A:cohA2. The cassette was built based on the unscaffolded pMTL8J15-isopropanol vector (Figure 2.14.5), through a combination of BioBrick2 (BB2) assembly of premade compatible 'parts' from the SBRC Nottingham culture collection (dock9B, dock9G, CipA1_CipC2), and HiFi assembly of the remaining parts (Cel48A, cohA2). Table 2.14.1 below contains details for scaffoldin and dockerin parts.

Table 2.14.1 – Origin of scaffoldin and dockerin genes used in the construction of scaffolded isopropanol cassettes.

Target for scaffolding	Gene	Type	Origin (project source)	Reference
CtfAB	*cel9B*	Scaffoldin	*C. theromcellum* (BB2 compatible vector in SBRC Nottingham Culture collection)	Kovacs and Willson 2013[100]
	cipA1	Dockerin		
Adc	*cel9G*	Scaffoldin	*C. cellulolyticum* (BB2 compatible vector in SBRC Nottingham Culture collection)	Kovacs and Willson 2013[100]
	cipC2	Dockerin		
Adh	*cel48A*	Scaffoldin	*C. acetobutylicum* ATCC 824 (Table 2.15.4)	Sabathe 2003[126]
	cohA2	Dockerin		

2.14.2.5. pMTL8J25x_TT_Cbe_thl_CACthlRBS_catP

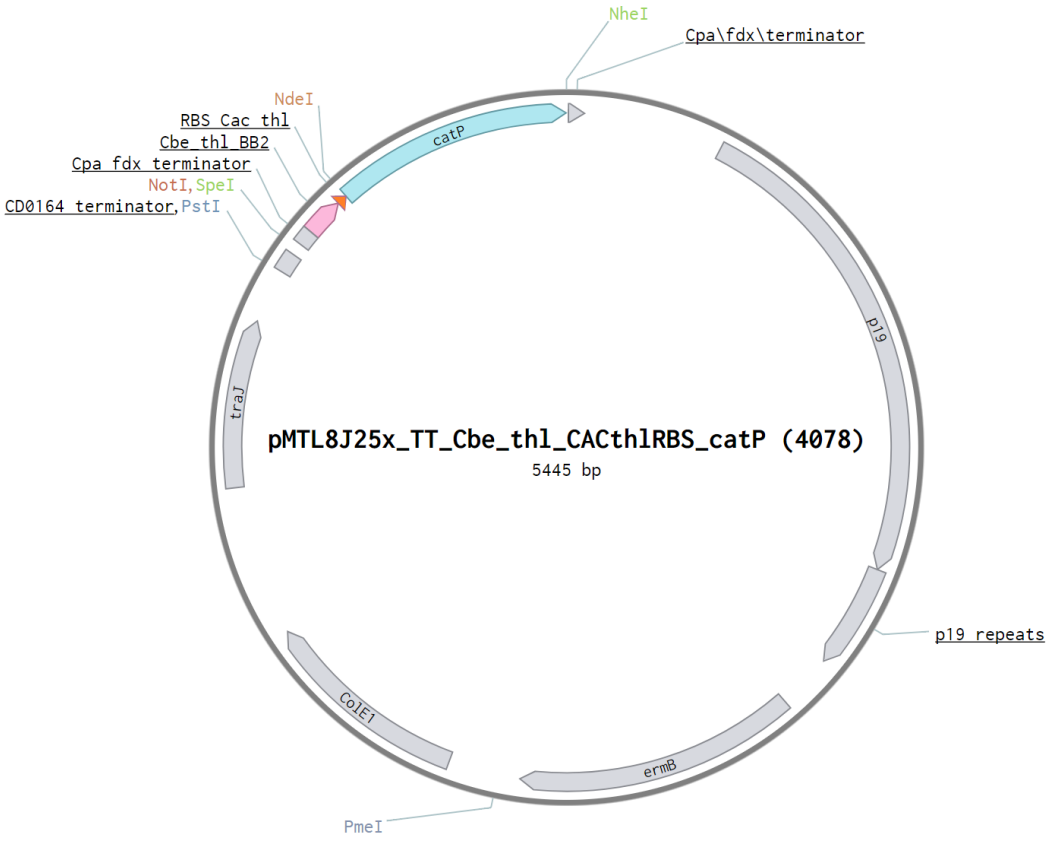


Figure 2.14.8 – Vector map for pMTL8J25x_TT_Cbe_thl_CACthlRBS_catP.

An example vector for *E. limosum catP* based promoter assay. Vector expressing *catP* (blue) under the control of a promoter (pink) derived from the thiolase gene of *C. beijerinckii* with an RBS derived from the thiolase gene *C. acetobutylicum* (orange). The promoter regions of 13 different test vectors was varied, while RBS regions and the remaining vector were consistent.

2.14.3. Chapter 5

2.14.3.1. pRBV1

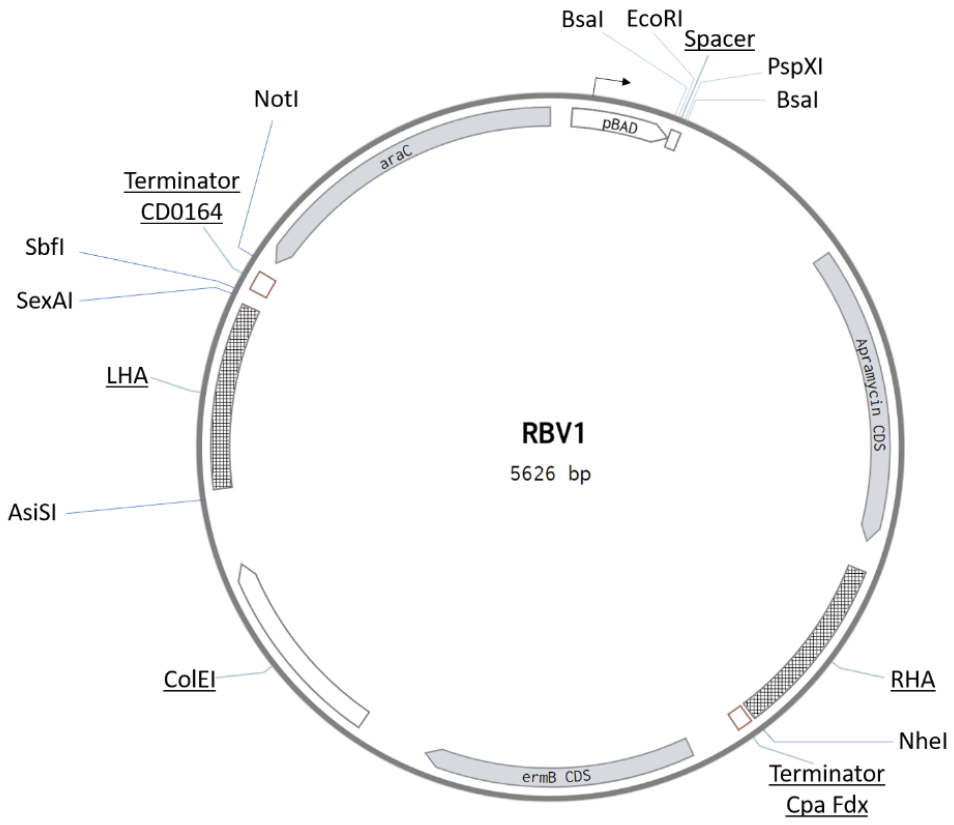


Figure 2.14.9 – Vector map for pRBV1.

A vector for recombineering cassette assembly and storage, pRBV1. Vector assembled by HiFi from four fragments: 1: Left homology arm sequences matches the 500 bp upstream of AadA1 in the pR702 conjugation assistance plasmid. 2. AraC-P_{BAD} promoter system derived from pBMTBX-4 plasmid (AddGene) for arabinose induced expression of genes placed within a synthetic 'spacer' region by HiFi assembly, Type IIS assembly, or traditional restriction-digest cloning methods. 3. A constitutively expressed apramycin resistance gene for cassette selection with a RHA derived from the 500 bp downstream of AadA1 of pR702 plasmid. 4. Backbone consisting of resistance marker *ermB*, and Gram-negative replicon ColEI from the pMTL80000 vector series.

2.14.3.2. pKD46cm

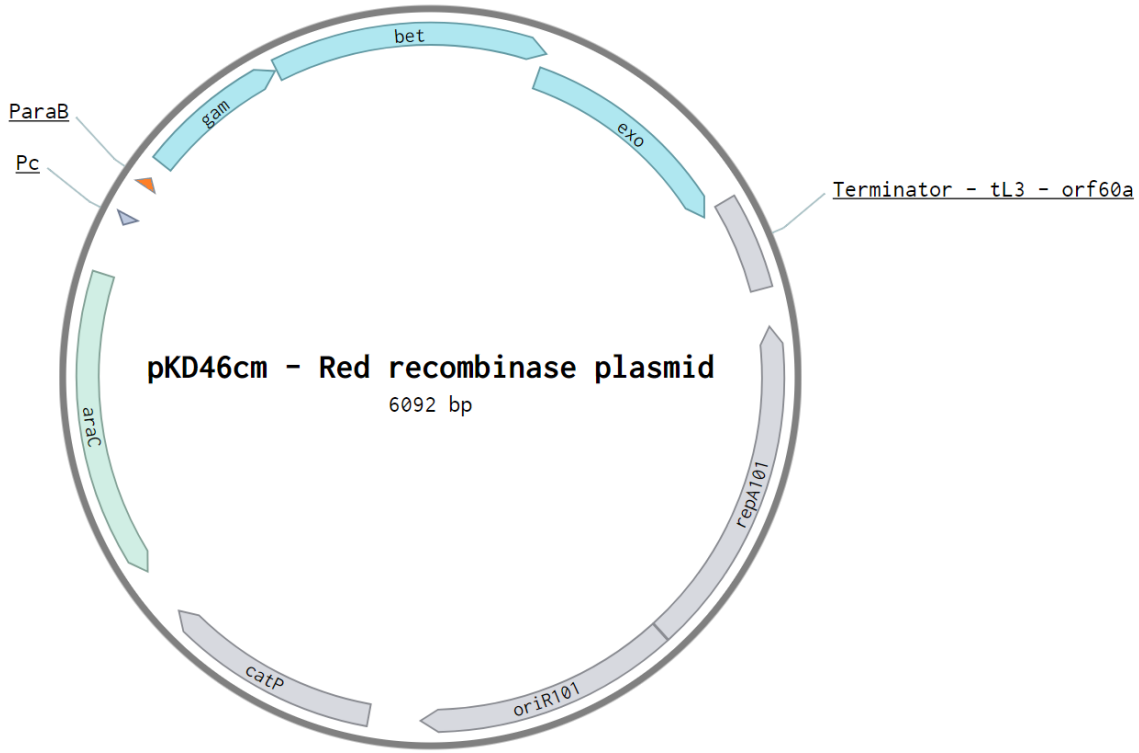


Figure 2.14.10 – Vector map for pKD46cm

A lambda red recombineering assistance plasmid. Expression of three lambda-red genes responsible for recombineering, *gam*, *beta*, and *exo* (blue) are controlled by the arabinose inducible promoter system *araC* (teal), P_{araB} (orange). The plasmid retention in *E. coli* is determined by a temperature sensitive replicon (oriR101-repA101), stable at 30 °C and unable to effectively replicate at ≥37 °C.

2.15. Primers

Primers were typically designed against plasmid maps using the browser-based software platform, Benchling.

2.15.1. Chapter 3 primers

Table 2.15.1 – Primers for construction of *pyrE* repair vector pRM03 for use in *E. limosum* ATCC 8486 Δ*pyrE*

Name	#	Target	Sequence
ELIM-SHA-F (SacI)	J01	*E. limosum* ATCC 8486	atatatCCGCGGgacctttggcgatttcgtg
ELIM-SHA-R (NotI)	J02	*E. limosum* ATCC 8486	atatatGCGGCCGCttaaataccgtttttatcaagatgatc
*pyrE*_scr_F	01	*E. limosum* ATCC 8486	CTGAATATTTCAAGATCAACGG
*pyrE*_scr_R	02	*E. limosum* ATCC 8486	GGAATTCTGTTGCATCAGC
tcdR_int_F	03	*E. limosum* ATCC 8486	AAGAGAGGATGATCATATGC
*bgaR*_int_R2	04	*E. limosum* ATCC 8486	AGGTAGAAATGCAAATATTGTGG

Short homology arm (SHA) PCR fragment (primers [J01, J02]) was inserted into the backbone of pJM_Elm_*pyrE*_KO (Millard, J. SBRC Nottingham. Thesis 2017) following SacI-NotI double digestion of both plasmid and PCR fragment. This resulted in replacement of a 'truncated' short homology arm (316 bp) with an 'untruncated' version (616 bp) to enable functional *pyrE* expression on genomic integration. Screening primers 03-06 were used to confirm the integration status of putative mutants.

Table 2.15.2 – Primers for construction of pSBRCas9_*Pbt*_1

Name	#	Target	Sequence
*Pbt*1_sgRNA	05	*E. limosum* ATCC 8486	atcttaaggaggagttttcgtcgacCCTTCAGACCAAAACCTATTgttttagagc tagaaatagcaagtt
*Pbt*1_LHA_F (AsiSI)	06	*E. limosum* ATCC 8486	ATATATATGGCGCGCCCTCTGCCAGAGTTCTGCC
*Pbt*1_LHA_R (SOE)	07	*E. limosum* ATCC 8486	ACGCGCCGTTTAATCGTACATTGTAAAACCTCGCTTTC
*Pbt*1_RHA_F (SOE)	08	*E. limosum* ATCC 8486	GGTTTTACAATGTACGATTAAACGGCGCGTC
*Pbt*1_RHA_R (AscI)	09	*E. limosum* ATCC 8486	ATATATATGGCGCGCCCGTGGAGCATCAGAATATTGTTGCACTCC
*Pbt*1_scr_F	10	*E. limosum* ATCC 8486	TATTTAAACCCGAGCTCCGGC
*Pbt*1_scr_R	11	*E. limosum* ATCC 8486	ACGTCAATGGGCAGATAATCG

Protospacer region marked in blue for sgRNA oligonucleotide. LHA_R and RHA_F contain compatible overhangs for SOEing PCR. Primers 06-09 are for assembly of the homology arms. Primers 10 and 11 are external primers for screening the targeted *pbt*1 knockout region.

2.15.2. Chapter 4 primers

Table 2.15.3 – TIIS assembly of Isopropanol operon into pRM01 backbone

Name	#	Target (origin)	Sequence
GG_TcdB_F	51	*C. difficile* DH1916	ACAGGACGGTCTCCaatgGCGGCCGCTTAATGAATTTAAAG
GG_TcdB_R	52	*C. difficile* DH1916	ACAGGACGGTCTCCttCATATGATTTTCTCCTTTACTATAATATTTTTATTGA ATATTTTTACATCTAAATGCTAAAAC
GG_*ctfAB*_F	53	*C. acetobutylicum* ATCC 824	ACAGGACGGTCTCCTGAACTCTAAAATAATTAGATTTGAAAATTTAAGGTC
GG_*ctfAB*_R	54	*C. acetobutylicum* ATCC 824	ACAGGACGGTCTCCGTCTAAACAGCCATGGGTCTAAG
GG_adc_F	55	*C. acetobutylicum* ATCC 824	ACAGGACGGTCTCCAGACTAGTTAAATTTAGGAAGGTGACTTTTATGTTAA AGGATGAAGTAATTAAAC
GG_adc_R	56	*C. acetobutylicum* ATCC 824	ACAGGACGGTCTCCAAGCTAGCTTACTTAAGATAATCATATATAACTTCAG CTCTAGG
GG_adh_F	57	*C. beijerinckii* NRRL B592	ACAGGACGGTCTCCGCTTTTTAAGGAGGAACATATCATATGAAAGGTTTT GCAATG
GG_adh_R	58	*C. beijerinckii* NRRL B592	ACAGGACGGTCTCCATCGTCTAGATTATAATATAACTACTGCTTTAATTAA GTCTTTTGGC

Table 2.15.4 – Primers for construction of Cel48A and CohA2 BB2 parts

Name	#	Target (origin)	Sequence
Cel48A-BB2 F	BD1	*C. acetobutylicum* ATCC 824	TATATGAATTCGCGGCCGCACTAGTACTTTATTCGCTAATACAGCTACACC
Cel48A-BB2 R	BD2	*C. acetobutylicum* ATCC 824	CGCGCTAGCCTATTTTGCAATTAATTTAGTAAGTTCC
CohA2 BB2 F	BS1	*C. acetobutylicum* ATCC 824	TATATGAATTCGCGGCCGCACTAGTAATGTTGTAAAGAATGAAACTGTAA AATTGTCAG
CohA2 BB2 R	BS2	*C. acetobutylicum* ATCC 824	CGCGCTAGCGCTTTGAACAACTGGTGCTGAACC

Table 2.15.5 – Primers for HiFi construction of Adh scaffoldin ORF and scaffold extension to include CohA2

Name	#	Target (origin)	Sequence
HF - Adh + Cel48A - Fw	HD1	*pJ204-Adh-Cel48A (this project)*	GACGATAAGTAATGAGCTAGTTTTTAAGGAGGAACATATC
HF - Adh + Cel48A (Scaff) - Rv	HD2	*pJ204-Adh-Cel48A (this project)*	gcataTGAACTAACCTCCTAttattatttgtcatcgtcat
HF-Scaf(3 part) F	HS1	*pJ204-CipA1-CipC2-CohA2_Flag (this project)*	atgacgatgacaaataataaTAGGAGGTTAGTTCAtatgc
HF-Scaf(3 part) R	HS2	*pJ204-CipA1-CipC2-CohA2_Flag (this project)*	gcttcttatttttatgctagTTACTTATCGTCGTCATCCT

A intermediate construct containing dockerin variants of CtfAB and Adc, pMTL8J15-Acetone-*ctfAB*_Cel9B-*adc*_Cel9G, was cleaved at XbaI site at the end of the expression cassette. This linearized vector was used as the backbone in the HiFi reaction. Resulting vector contains a modified Adh ORF coupled to Cel48 by a short linker region.

Table 2.15.6: Primers for insertion of *thl* gene downstream of *ctfB* in scaffolded and unscaffolded isopropanol cassettes by HiFi

Name	#	Target (origin)	Sequence
A1-ctfB R HiFi	TA1	pMTL8J15-Isoprop-Scaff-Dock (this project)	AATTTTGATACactagtCTAaacagccatgggtc
A2-adcRBS F HiFi	TA2	pMTL8J15-Isoprop-Scaff-Dock (this project)	actagttaaatttaggaaggtgacttttatg
B1-thl F HiFi	TB1	*C. acetobutylicum* ATCC 824	gacccatggctgttTAGactagtGTATCAAAATTTAGGAGGTTAGTTAGA
B2-thl R HiFi	TB2	*C. acetobutylicum* ATCC 824	ccttcctaaatttaactagtCTAGCACTTTTCTAGCAATATTG
C1-Dock9B R HiFi	TC1	pMTL8J15-Isoprop-Scaff-Dock (this project)	TAATCACTAGCCTGTTCAGCC
C2-adcRBS F HiFi	TC2	pMTL8J15-Isoprop-Scaff-Dock (this project)	TATTGCTAGAAAAGTGCTAGGCTAGTTAAATTTAGGAAGGTGAC
D1-thlRBS-FLG F HiFi	TD1	*C. acetobutylicum* ATCC 824	CGGCTGAACAGGCTAGTGATTACAAGGATGACGACGATAAGTAAGTATCAAAATTTAGGAGGTTAGTTAG
D2-thl-adcRBS R HiFi	TD2	*C. acetobutylicum* ATCC 824	CCTTCCTAAATTTAACTAGCCTAGCACTTTTCTAGCAATATTGC

2.15.3. Chapter 5 primer

Table 2.15.7 – HiFi assembly primers for construction of pRBV1

Name	#	Target (origin)	Sequence
RecoM1-F	R1	LHM-RecoV2 (this thesis)	GCCCAATACGCAGCGATCGCTTCTACGGCAC
RecoM1-R	R2	LHM-RecoV2 (this thesis)	TAGCCGTCAAGTTGTCATAACACAGGCTACTACACGCGG
AraC-pBAD-F	R3	pBMTBX-4 (Addgene)	GCGTGTAGTAGCCTGTGTTATGACAACTTGACGGCTACATC
AraC-pBAD-R	R4	pBMTBX-4 (Addgene)	ATCGTGAGACCGAATTCGGTCTCGCATTCCCAAAAAAACGGGTATGGAG
RecoM2-F	R5	RHM(ap)-RecoV2 (this thesis)	AATGCGAGACCGAATTCGGTCTCACGATACTCGAGTCATATGATTCCGG
RecoM2-R	R6	RHM(ap)-RecoV2 (this thesis)	CACACTCTTAAGTTTGCTTCTATAGAAAAGAAGTGAATGCGGCG
85241-RBV-F	R7	pMTL85241 (Nottingham SBRC Culture Collection)	GCCGCATTCACTTCTTTTCTATAGAAGCAAACTTAAGAGTGTGTTGATAGTGC
85241-RBV-R	R8	pMTL85241 (Nottingham SBRC Culture Collection)	CGTAGAAGCGATCGCTGCGTATTGGGCGCTCTTC

Table 2.15.8 – Primers for construction of Left and Right Homology Modules (LHM and RHM) pre pRBV1 construction

Name	#	Target (origin)	Sequence
LHM-LHA-F (BsaI)	12a	pR702 (*E. coli* CA434)	CTGGTCTCGAATGTTCTACGGCACGTTTGAAGG
LHM-LHA-R (SexAI)	12b	pR702 (*E. coli* CA434)	ACCTGGTCTGGCTTCAGGAGATCGG
LHM-Prom-F (SexAI)	13a	pMTL83153rs (Canadas, I. SBRC Nottingham)*	ACCAGGTGACGAGCAAGGCAAGACC
LHM-Prom-R (BsaI)	13b	pMTL83153rs (Canadas, I. SBRC Nottingham)*	CAGGTCTCGTGTTGTTGCCTCCTTAGCAG
RBM-Apra-F (PspXI)	16a	pIJ786 (Nottingham SBRC Culture Collection)	ACTCGAGTCATATGATTCCGGGGATC
RBM-Apra-R (SOE)	16b	pIJ786 (Nottingham SBRC Culture Collection)	ATTGTTAGACACATATGTGTAGGCTGGAGCTG
RBM-RHA-F (SOE)	17a	pR702 (*E. coli* CA434)	CCTACACATATGTGTCTAACAATTCGTTCAAGC
RBM-RHA-R (BsaI)	17b	pR702 (*E. coli* CA434)	CTGGTCTCGATCGGCAGCGACTTCCACGATG

*pMTL83153rs is a modified pMTL83153 containing riboswitch in *fdx* promoter region (pMTL8xxx3) – riboswitch not relevant to future pRBV1 construction beyond left homology module (LHM) assembly to join LHA to terminator

Table 2.15.9 – HiFi assembly primers for gene integration into pRBV1 vectors

Name	#	Target (origin)	Sequence
RFP RBV-F	20a	pKTrfp (AddGene)	GTTTCTCCATACCCGTTTTTTTGGGtttaagaaggagatatacatatgg
RFP RBV-R	20b	pKTrfp (AddGene)	CCGGAATCATATGACTCGAGTtttggatccttaagcacc
hsdSM-HiFi-F	21a	*C. difficile* DH1916 (Anaerobic Reference Laboratory: Cardiff)	GTTTCTCCATACCCGTTTTTTTGGGatgAACACTGGAGGGAAAGTC
hsdSM-HiFi-R	21b	*C. difficile* DH1916 (Anaerobic Reference Laboratory: Cardiff)	CCGGAATCATATGACTCGAGTgtaTTTCAGTTGAACAATTCTCC
T3I-Hifi-F	22a	*E. coli* Top10 - pAN2 (Nottingham SBRC Culture Collection)	GTTTCTCCATACCCGTTTTTTTGGGaaaTATATGGAGGTTGTTTGTTGAG
T3I-HiFi-R	22b	*E. coli* Top10 - pAN2 (Nottingham SBRC Culture Collection)	CCGGAATCATATGACTCGAGattTTATTCAGATTCTTTATTAACATATG
Cbo-HiFi-F1	23a	*C. botulinum 19397* (Nottingham SBRC Culture Collection)	gtttctccatacccgttttttgggAatgAATAAGAGAGGTATTTATAAATG
Cbo-HiFi-R1	23b	*C. botulinum 19397* (Nottingham SBRC Culture Collection)	CATTATGTTTCCTCCATCATTAGTTATACCCCCTTATAGATTAACATTAC
Cbo-HiFi-F2	24a	*C. botulinum 19397* (Nottingham SBRC Culture Collection)	GTAATGTTAATCTATAAGGGGGTATAACTAATGATGGAGGAAACATAATG
Cbo-HiFi-R2	24b	*C. botulinum 19397* (Nottingham SBRC Culture Collection)	CCGGAATCATATGACTCGAGTATCGCTATTTAATTAAACAACTCTC
ORFA-Elim-F	25a	*E. limosum* ATCC 8486 (DSMZ)	GTTTCTCCATACCCGTTTTTTTgggCTGCTTTAATGGGAGATATATAGATG
ORFA-Elim-R	25b	*E. limosum* ATCC 8486 (DSMZ)	CCGGAATCATATGACTCgagGTTTAGCTGTCCAAGTTTGAGAC
ORFD-Elim-F	26a	*E. limosum* ATCC 8486 (DSMZ)	GTTTCTCCATACCCGTTTTTTTgggGATGAAAGAAAGGAAAAAATCATG
ORFD-Elim-R	26b	*E. limosum* ATCC 8486 (DSMZ)	CCGGAATCATATGACTCgagGACTATGGAGTGGCAGTCC
catP RBV-F	29a	*pMTL83254*	GTTTCTCCATACCCGTTTTTTTGGGAAATTTAGGAGGTTAGTTCATATGG
catP RBV-R	29b	*pMTL83254*	CCGGAATCATATGACTCGAGCTAGCTTAACTATTTATCAATTCCTG
30 F	30	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGcatatatttaaagcttgaaatatgaaagg
30 R	31	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTttaattaaatagttctccttttgaaagc
32 F	32	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGgaaaagattaaaggagatagaaatgg
32 R	33	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTcactaccaaaaatctaaatttttacacc
34 F	34	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGtgaagataaggagattagctaaagatg
34 R	35	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTctatatatctttactcaatatttccctc
36 F	36	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGgaaaattatacgaagcatcgg
36 R	37	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTctaaacaaatttaagctgccttg
38 F	38	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGtataaaattaagaaggtttaatatgtataaagacg
38 R	39	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTttacagcagtttatctctatacttatcac
40 F	40	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGtagcataaagaaaggtaatatttatgg
40 R	41	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTgccaagattaaattcaaggc
42 F	42	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGagtatataggagggattactttatg
42 R	43	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTttacacattctcaataaacttttc
44 F	44	*C. carboxidivorans* P7 (DSMZ)	gtttctccatacccgttttttttGGGtgtaaatgttggaggtgaaattatgg
44 R	45	*C. carboxidivorans* P7 (DSMZ)	CCGGAATCATATGACTCGAGTctaccatgactttagtgcctcctc

Table 2.15.10 – Recombineering Primers

Name	#	Target (origin)	Sequence
LHA F	200	pRBV vectors (this project)	TTCTACGGCACGTTTGAAGG
RHA R	219	pRBV vectors (this project)	GCAGCGACTTCCACGATG
LHA ext F	339	Recombinant pR702 (this project)	GAATGTCGTAACCGCTGC
LHA ext R	338	Recombinant pR702 (this project)	ATTCAGAATGCCGAACACC
Apra-scr R	234	pRBV vectors (this project)	ACTGCAGGTCGACGGATC
pBAD-scr F	244	pRBV vectors (this project)	GTCCATATTGCATCAGACATTGC

Table 2.15.11 – OrfD screening Primers

Name	#	Target (origin)	Sequence
OrfD F	332	*E. limosum* ATCC 8486	GATGAAAGAAAGGAAAAAATCATG
OrfD R1	333	*E. limosum* ATCC 8486	TTAAATGTCGATGTCTGAAACATC
OrfD R2	334	*E. limosum* ATCC 8486	GACTATGGAGTGGCAGTCC

Table 2.15.12 – pMTL8315y HiFi assembly

Name	#	Target (origin)	Sequence
5y-A-F	y1	pMTL83151	TCAGCCGGGCAGGATAGGTCAAGTAGGCCCACCCGCGAGC
5y-A-R	y2	pMTL83151	GTATTTTTTTTTGGCGCGCCGTCGACTCTAGAGGATCCCCGG
5y-B-F	y3	pMTL83151	TCTAGAGTCGACGGCGCGCCAAAAAAAAATACTTGTTATGTAT
5y-B-R	y4	pMTL83151	TTTATATCTGTTAAGATAAACTCATAAATTGAATATATATAGTTAAT
5y-C-F	y5	pMTL83151	TATATATTCAATTTATGAGTTTATCTTAACAGATATAAATGTAAAT
5y-C-R	y6	pMTL83151	TTTGGTCATGAGGTTTAAACCACCGTATTTCTACGATGTT
5y-D-F	y7	pMTL83151	AGAAATACGGTGGTTTAAACCTCATGACCAAAATCCCTTAACG
5y-D-R	y8	pMTL83151	CTCGCGGGTGGGCCTACTTGACCTATCCTGCCCGGCTGAC

2.16. Plasmid list

2.16.1. Chapter 3 plasmids

Table 2.16.1 – List of plasmids used during the establishment of genetic tools in *E. limosum* ATCC 8486

Plasmid	Description	Reference
pMTL81151	pAMβ1, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL82151	pBP1, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL83151	pCB102, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL84151	pCD6, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL85151	pIM13, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL86151	pIP404, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL87151	p19r, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL88151	pCAc, ColE1+*traJ*, *catP* (Cm^R/Tm^R) containing modular plasmid	SBRC Nottingham Culture Collection
pMTL8J151	Modified pMTL87151 with silent point mutations to remove native restriction sequences from within Gram-positive replicon (module 7)	This study
pSBRCas-*pbt*1	Cas9 expressing vector targeted for genomic cleavage within the *pbt*1 gene (Ga0111251_111764) of *E. limosum* ATCC 8486.	This study
pRM03	ACE vector for *pyrE* repair in *E. limosum* ATCC 8486 Δ*pyrE*	This study
pRM04	Vector for simultaneous *pyrE* repair and LacTcdR knock-in in *E. limosum* ATCC 8486 Δ*pyrE*	This study

2.16.2. Chapter 4 plasmids

Table 2.16.2 – List of plasmids used within this study for the establishment of acetone and isopropanol production in *E. limosum* and *C. autoethanogenum*

Acetone & Isopropanol vectors	Description	Reference
pRM-ISO G1 (pMTL8J15_iso_152)	pMTL8J151 with isopropanol production cassette consisting of c*tfAB*, a*dc*, a*dh* under control of the P_{TcdB} promoter	This study
pRM-ACET G1 (pMTL8J15_ace)	pMTL8J151 with isopropanol production cassette consisting of c*tfAB*, a*dc* under control of the P_{TcdB} promoter	This study
pRM-ISO-152-thl G2.1 (pMTL8J15_iso_thl)	pRM-ISO with the addition of *thl* in the isopropanol production cassette	This study
pRM-ACET-thl G2.1 (pMTL8J15_ace_thl)	pRM-ACE with the addition of *thl* in the acetone production cassette	This study
pRM-ISO-152-77 G2 (pMTL8J15_iso_77)	pRM-ISO with P_{tcdB} replaced with promoter P_{4077}	This study
pRM-ACE-77 G2 (pMTL8J15_ace_77)	pRM-ACE with P_{tcdB} replaced with promoter P_{4077}	This study
pRM-ISO-152-77-thl G3 (pMTL8J15_iso_77_thl)	pRM-ISO-77 with the addition of *thl* in the expression cassette	This study
pRM-ACE-77-thl G3 (pMTL8J15_ace_77_thl)	pRM-ACE-77 with the addition of *thl* in the expression cassette	This study
pRM-ISO-DSF G1S (pMTL8J15_iso_DSF)	pRM-ISO with the C-terminal dockerins added to *ctfB*, *adc*, and *adh* genes, in addition to a synthetic scaffoldin gene downstream of adh	This study
pRM-ACE-DSF G1S (pMTL8J15_ace_DSF)	pRM-ACE with the C-terminal dockerins added to *ctfB*, and *adc* genes, in addition to a synthetic scaffoldin gene downstream of adh	This study
pRM-ISO-DSF-Thl G1S.2 (pMTL8J15_iso_DSF_thl)	pRM-ISO-DSF with *thl* added to the isopropanol production cassette	This study
pRM-ACE-DSF-Thl G1S.2 (pMTL8J15_ace_DSF_thl)	pRM-ACE-DSF with *thl* added to the acetone production cassette	This study

Table 2.16.3 – Promoter library plasmid collection used within this study (Chapter 4).

Promoter testing vectors	Description	Reference
pMTL8J15x-4073	*C.beijerinckii* P_{fdx}	This study
pMTL8J15x-4076	*C.kluyveri* P_{fdx}	This study
pMTL8J15x-4077	*C.perfringens* P_{fdx}	This study
pMTL8J15x-4078	*C.beijerinckii* P_{thl}	This study
pMTL8J15x-4098	*C.cellulovorans* P_{thl}	This study
pMTL8J15x-4099	*C.butyricum* P_{thl}	This study
pMTL8J15x-999	*C.sporogenes* P_{fdx}	This study
pMTL8J15x-1673	*C.accetobutylicum* P_{thl}	This study
pMTL8J15x-Lac	*bgaR*-P_{bgaL}: Lactose inducible system from *C.perfringens* - based on the divergently oriented promoter and AraC family transcriptional regulator P_{bgaL} and *bgaR* respectively	This study
pMTL8J15x-IPTG	P_{fac}: Synthetic lac operator inserted into the promoter of the *C. pasteurianum* ferredoxin gene	This study
pMTL8J15x-AraE	P_{araE}: Promoter from C. acetobutylicum *araE* gene	This study
pMTL8J15x-TcdB	P_{tcdB}: Promoter from *C. difficile*	This study
pMTL8J15x-NoPromoter	No promoter control vector	This study

Thirteen *catP* expression vectors driven by a selection of different promoters, derived from a variety of bacterial sources. Promoter sequences and template plasmid vectors were sourced from Ben Wilson (Wilson, B. SBRC Nottingham (formerly CRG), 2013 Thesis). Original sources are noted in the Description column. Each vector differs only in the promoter region upstream of *catP* and an RBS derived from the thiolase gene (*thl*) of *C. acetobutylicum* ATCC 824.

2.16.3. Chapter 5 plasmids

Table 2.16.4 – List of plasmid constructs generated for generation of recombinant *E. coli* strains (Chapter 5)

Recombineering vectors	Description	Reference
pRBV1	Basic empty recombineering vector, no payload added at the RBV spacer (RBV1 control)	This study
pRBV-RFP	RFP gene inserted for gene expression assessment at the RBV spacer	This study
pRBV-pAN2	RBV spacer carrying pAN2 derived methyltransferase gene, originally from Φ3TI of *Bacillus subtilis* for improved transformation of *C. acetobutylicum* ATCC 824	This study
pRBV-Cdif-HsdSM1	RBV spacer carrying methyltransferase and specificity subunits HsdSM1 derived from *C. difficile* DH1916	This study
pRBV-Cbot-Orf717-Orf2811	RBV spacer carrying methyltransferases Orf717-Orf2811 derived from *C. botulinum* A ATCC 19397	This study (Groothuis Daphne and Raquel Rodrigues)
pRBV-Eli-OrfA	RBV spacer carrying methyltransferase OrfA from *E. limosum* ATCC 8486	This study
pRBV-Eli-OrfD	RBV spacer carrying methyltransferase OrfD from *E. limosum* ATCC 8486	This study
pRBV-Ccar-30	RBV spacer carrying methyltransferase 30 from *C. carboxidivorans* P7	This study
pRBV-Ccar-32	RBV spacer carrying methyltransferase 32 from *C. carboxidivorans* P7	This study
pRBV-Ccar-34	RBV spacer carrying methyltransferase 34 from *C. carboxidivorans* P7	This study
pRBV-Ccar-36	RBV spacer carrying methyltransferase 36 from *C. carboxidivorans* P7	This study
pRBV-Ccar-38	RBV spacer carrying methyltransferase 38 from *C. carboxidivorans* P7	This study
pRBV-Ccar-40	RBV spacer carrying methyltransferase 40 from *C. carboxidivorans* P7	This study
pRBV-Ccar-42	RBV spacer carrying methyltransferase 42 from *C. carboxidivorans* P7	This study
pRBV-Ccar-44	RBV spacer carrying methyltransferase 44 from *C. carboxidivorans* P7	This study
pMTL8315y	pMTL83151 without 12/16 RM motif sequences (*C. carboxidivorans*)	This study

3. Development of *Eubacterium limosum* as a chassis for research

3.1. Introduction

3.1.1. Research background

The 'Strain Development Roadmap' published by Minton *et al*. (2016)[55] provided a framework for the research and development of advanced genetics in previously unmodified microorganisms. A modified version of this roadmap, updated to reflect the introduction of improved genome editing tools as described in Chapter 1, was used to employed in the genetic development of *E. limosum* in this project. Starting first with the production of a complete genome, for the target strain, *E. limosum* ATCC 8486.

3.1.2. *E. limosum*

E. limosum is an attractive starting point for development of an industrially valuable organism. It is capable of metabolising a wide range of organic compounds including, a number of C1 compounds, which is of particular commercial interest[52]. These compounds include syngas (CO, CO_2 and H_2), methanol (CH_3OH), and formate (HCO_2H)[127,128]. In addition to its extensive substrate utilisation, *E. limosum* is noteworthy for its capacity to produce C4 products from these C1 feedstocks from the acetyl-coA generated *via*the Wood-Ljungdahl pathway.

Until this study there were no published records of genetic manipulation of *E. limosum*, in spite of numerous publications pertaining to the metabolism and fermentative properties exhibited by the species. The prior lack of applied genetics is unlikely to be a result of a lack of desire to implement them. Targeted mutant generation, in addition to the plethora of other applications of genome engineering tools, are invaluable avenues to improving the understanding of the inner workings of microbes. The lack of

robust and effective methods for gene transfer are likely the primary barrier, which has precluded active research progress in this area.

3.1.3. Genome

The increased accessibility of full genomic data, supported by an ever-improving understanding of intracellular processes, has bought about significant advancement in the fields of synthetic- and systems-biology[90]. In turn these advancements have opened the door toward effective rational design attempts in microbial systems[129,130]. Access to a complete and closed genome sequence allows researchers to generate robust understandings of the internal metabolic workings of a given microorganism with a reasonable degree of accuracy. Crucially, this is possible without the necessity for extensive prior laboratory experimentation. The genome provides a robust starting point from which to direct future strain research, and in the context of microbial biocatalysts, a platform for strain development.

3.1.4. Gene transfer

Implementation of rationally designed modifications is completely reliant on access to effective tools for performing them through genetic modification. In turn these tools generally rely on effective transformation protocols in order to function, and thus establishment of robust gene transfer methodologies is vital to genetic research and engineering projects.

3.1.5. Establishing Genome editing tools

Appropriate tools for genetic manipulation are important tools in both microbial research. Targeted genome editing systems such as allele-coupled exchange (ACE) and CRISPR/Cas9 are central to the modern microbiologist's toolbox. The controlled genetic modifications which these tools enable are invaluable for strain characterisation, and

can further be used in the context of more intricate strain development. Targeted genome engineering is beneficial for basic research as well as for potential commercial development of microorganisms in both health and biotechnology fields, as it permits a greater level of rational design to be applied to research, than that which can be achieved by more traditional microbial engineering techniques such as random mutagenesis or directed evolution[89,131].

3.1.6. Project Aims

This chapter discusses the development of a robust method of transformation protocol for *E. limosum* ATCC 8486. This includes characterisation of elements and parameters which are essential to applying advanced genetic manipulation techniques. Subsequently the developed methodology is employed to establish a selection of tools for the genetic manipulation in this organism. Specifically demonstrating proof of concept application of two prominent genome editing techniques, Allele-Coupled Exchange (ACE) and CRISPR/Cas9.

3.2. Results

3.2.1. Genome

At the project outset there was no complete and closed genome sequence available for *E. limosum* ATCC 8486. Access to an accurate and complete genome sequence to work from was considered to be an essential point of reference and analysis for effectively planning strain development going forward. The complete genome sequence was generated from an initial *de novo* assembly of three contigs generated from PacBio RS II sequencing, achieved by manual alignment of overlapping contig ends, to produce a closed reference assembly. Subsequently, an Illumina paired-end read library was mapped to the reference assembly for the purpose of error correction in order to establish a final consensus sequence. Annotation of the sequence was performed using the Joint Genome Institute's (JGI) Integrated Microbial Genomes (IMG) system. The resultant genome atlas for *E. limosum* ATCC 8486 is presented in Figure 3.2.1.

3.2.1. Genome properties

General properties and statistics for the assembled genome of *E. limosum* ATCC 8486 are presented in Table 3.2.1 and Table 3.2.2. In summary, the *E. limosum* ATCC 8486 genome consists of a single circular chromosome of 4,421,303 bp and a G/C content of 47.2%. Of 4219 predicted genes, 4129 were predicted to be protein encoding. This now closed genome represents the third complete genome produced of *E. limosum* species, adding to the published genomes of *E. limosum* strains SA11 and KIST612[31,34].

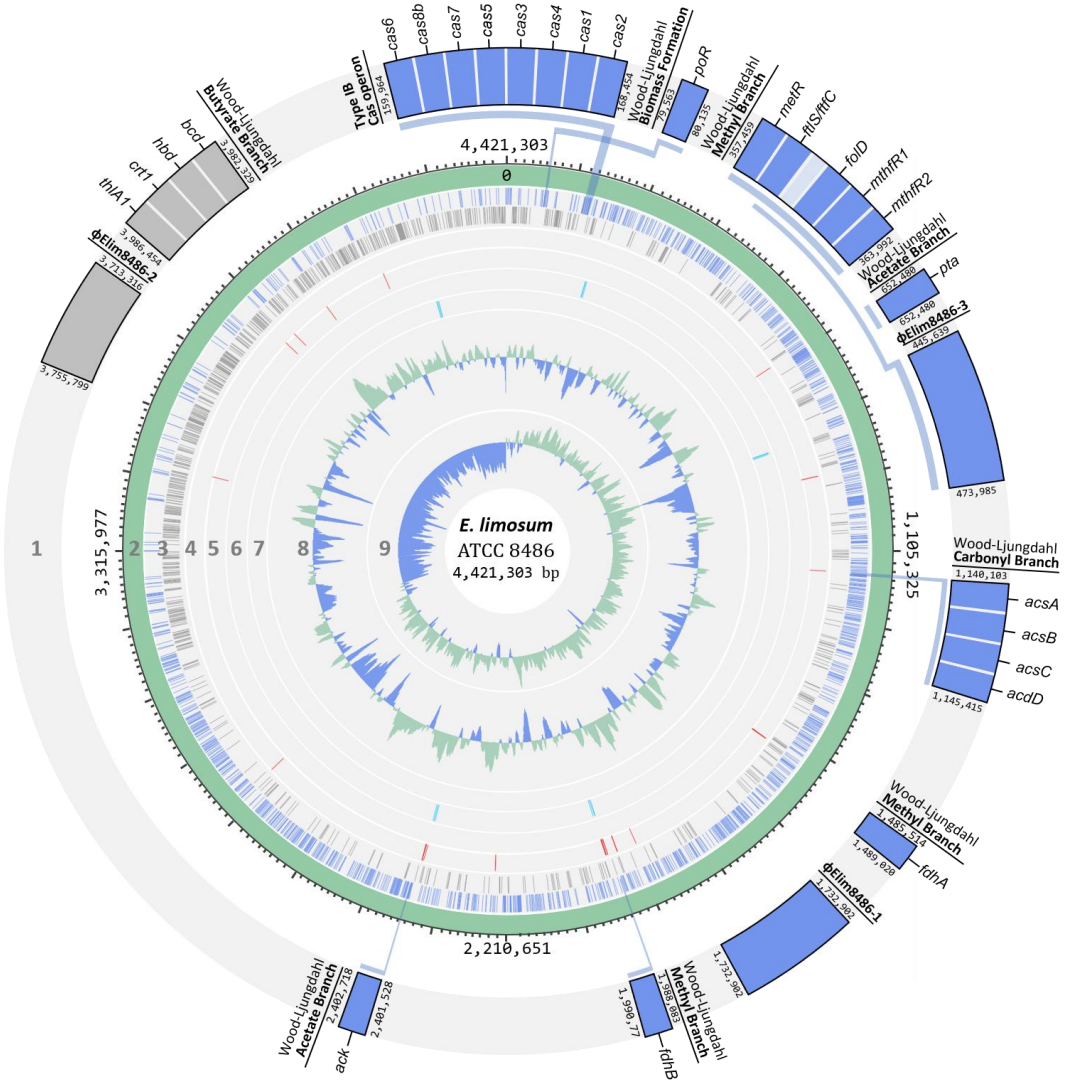


Figure 3.2.1 – Genome atlas for *E. limosum* ATCC 8486

The *E. limosum* ATCC 8486 genome is 4,421,303 base pairs in length, comprised of 4219 predicted genes and a GC content of 47.2%. Of the 4219 predicted genes, 97.87% are predicted to represent protein coding sequences while the remaining 90 genes are expected encode RNAs. Track 1: Key genome features; Core Wood-Ljungdahl metabolism and acidogenic genes, and three prophage loci; Track 2: source; Track 3: Positive (blue) and negative (grey) strand CDSs; Track 4: positive strand tRNA CDS; Track 5: tRNA CDS; Track 6: rRNA CDS; Track 7: rRNA CDS; Track 8: sequence GC content; Track 9: GC skew (Tracks 8&9: blue=GC above average, green=GC below average).

Table 3.2.1 – General genome statistic for the closed and annotated *E. limosum* ATCC 8486

Attribute	Value	% of total
Genome size (bp)	4421303	100.00%
DNA Scaffolds	1	
DNA coding (bp)	3954494	89.44%
DNA GC (bp)	2086531	47.19%
Total Genes	4219	100.00%
Protein Coding Genes	4129	97.87%
RNA genes	90	2.13%
Genes with function prediction	3207	76.01%
Transmembrane Count	1096	25.98%
Signal Peptide Count	275	6.52%
Genes without function prediction	922	21.85%
Genes assigned to COGs	2646	62.72%
CRISPR loci	2	

Table 3.2.2 – General COG functional category association of annotated genes from *E. limosum* ATCC 8486

Code	value	% of total protein coding	Description
All	*4129*	*100.00%*	*All genes*
A	0	0.00%	RNA processing and modification
B	0	0.00%	Chromatin Structure and dynamics
C	201	4.87%	Energy production and conversion
D	31	0.75%	Cell cycle control and mitosis
E	279	6.76%	Amino Acid metabolism and transport
F	90	2.18%	Nucleotide metabolism and transport
G	221	5.35%	Carbohydrate metabolism and transport
H	220	5.33%	Coenzyme metabolism
I	75	1.82%	Lipid metabolism
J	198	4.80%	Translation
K	320	7.75%	Transcription
L	110	2.66%	Replication and repair
M	173	4.19%	Cell wall/membrane/envelop biogenesis
N	29	0.70%	Cell motility
O	92	2.23%	Post-translational modification, protein turnover, chaperone functions
P	152	3.68%	Inorganic ion transport and metabolism
Q	33	0.80%	Secondary Structure
R	219	5.30%	General Functional Prediction only
S	167	4.04%	Function Unknown
T	233	5.64%	Signal Transduction
U	26	0.63%	Intracellular trafficking and secretion
V	88	2.13%	Defence mechanisms
W	12	0.29%	Extracellular structures
X	18	0.44%	Mobilome, prophages, transposon
Y	0	0.00%	Nuclear structure
Z	1	0.02%	Cytoskeleton
Total *:	2647	64.11%	

*Adjusted total accounts for 341 genes with multiple COG categories, reducing the total from 2988 COG matches.

3.2.2. Barriers to Gene transfer

In order to establish tools for advanced genetic manipulation in *E. limosum*, a robust gene transfer methodology was required. Using the newly closed and complete genome it was possible to identify a number of possible barriers to transformation in advance of practical experimentation. Specifically, these were the CRISPR and restriction modification (RM) systems native to *E. limosum* ATCC 8486.

3.2.2.1. CRISPR systems

Two distinct CRISPR loci were identified using the CRISPRfinder webtool[132]. Both loci possessed an identical direct repeat (DR) sequence (Table 3.2.3). Eight Cas genes (Cas1-8) encoding for a full type-I CRISPR system have been identified, all localised to CRISPR locus 1. No sequence identity was detected between the spacer regions at the CRISPR loci and the pMTL80000 vector series which were to be used for transformation testing. The presence of such homology may have led to reduced transfer efficiency of transforming vectors as a result of targeted cleavage by the native CRISPR-Cas9 systems. This indicated that the CRISPR system would not pose a direct problem for transformation. The genome loci for the CRISPR loci and Cas operon are depicted in the genome atlas in Figure 3.2.2.

Table 3.2.3 – CRISPR locus locations and details

Locus ID #	CRISPR location	Number of spacers	Direct Repeat Sequence
1	168635-170680	31	GTTGAAGATTAACATGAGATGTATTTAAAT
2	509602-510614	15	GTTGAAGATTAACATGAGATGTATTTAAAT

Location of two CRISPR loci in the genome sequence for *E. limosum* ATCC 8486 developed for this study.

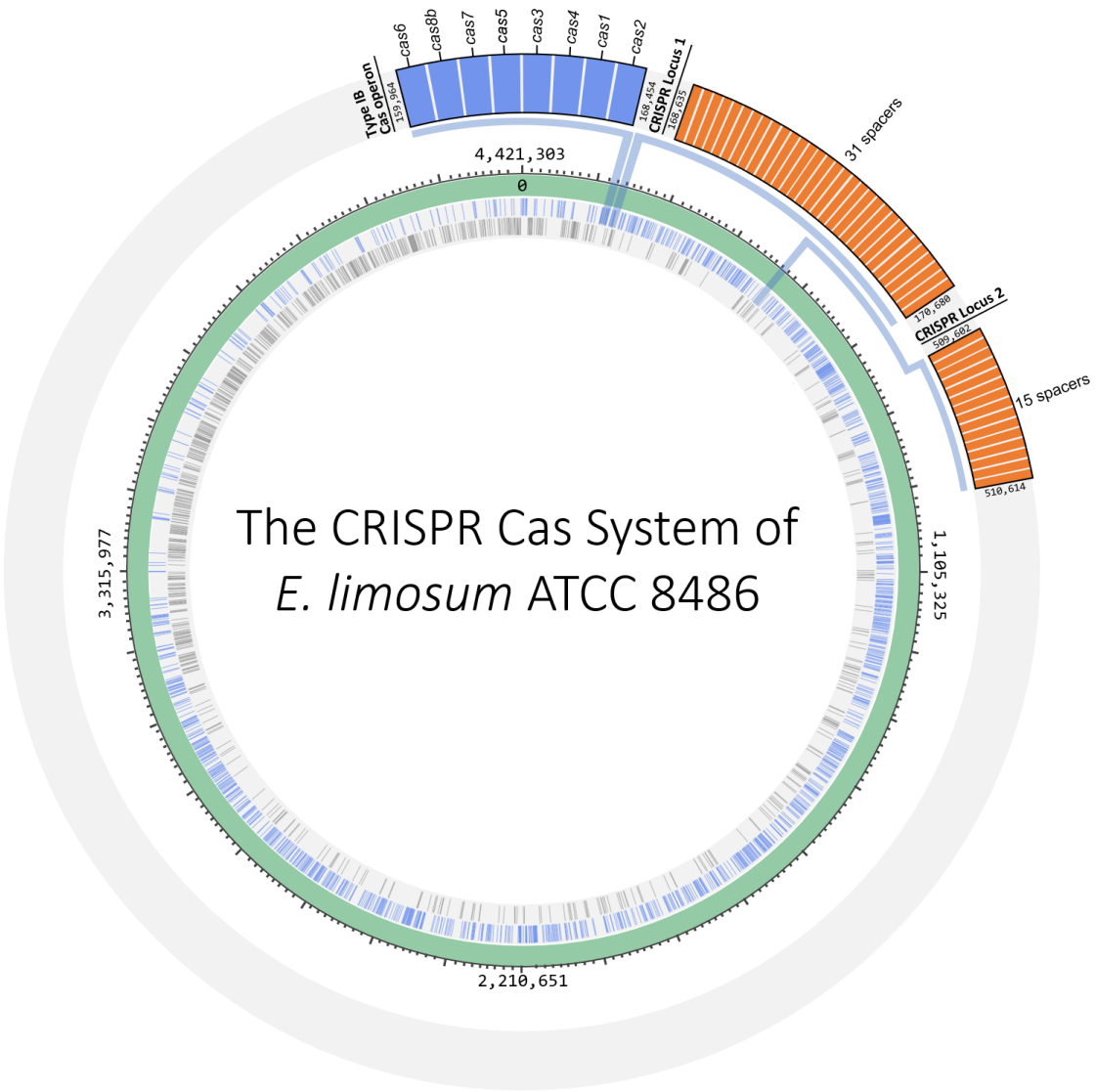


Figure 3.2.2 – Genome atlas for *E. limosum* ATCC 8486 with CRISPR loci and *cas* operon marked.

Genome containing a Type IB *cas* operon (blue): *cas* genes *cas*1-8. Two CRISPR loci (orange): locus 1 containing 31 spacers; locus 2 containing 15 spacers.

3.2.2.2. Restriction Modification Systems

In addition to basic genomic sequence analysis, the single-molecule real time (SMRT) methodology employed in the PacBio sequencing methodology allowed for the detection of modified DNA bases. This PacBio dataset was used to identify repeated methylation patterns on the genomic DNA (Table 3.2.4). This information was then attributed to a number of RM systems identified within the genome *via* the ReBase™

RM analysis pipeline[133]. In total our analysis identified five putative RM systems, two of which appeared to be expressed on prophages and were not expected to be functional. The remaining three (OrfA, OrfB, OrfD) were attributed to a given recognition motif based on homology against enzymes of known motif association. Each of the three recognition sequences reported in the PacBio dataset had multiple matches to sequence regions within the pMTL80000 vectors, indicating that the respective RM systems may pose barriers to successful gene transfer. No type IV restriction systems were identified within the *E. limosum* ATCC 8486 genome. As such, deoxyadenosine methylase (DAM) and deoxycytosine methylase (DCM) containing *E. coli* cloning strains were not expected to be of issue. Detailed analysis of the RM system of *E. limosum* 8486 is available in chapter 5.

Table 3.2.4 – Pacbio modified base recognition motifs with putative RM system assignment

Recognition sequence Motifs (+ strand)	Locus Tag	Nomenclature
CAAAAAR	Ga0111251_111421	Eli8486OrfB
CACNNNNRTANNG	Ga0111251_112370	Eli8486OrfD
GCGRAG	Ga0111251_11256	Eli8486OrfA

Assignments were generated based on the homology of each RM system to previously characterised systems. Systems with known recognition motifs were identified in the Restriction Enzyme Database, ReBase™, using the 'Gold Standard' reference database[133].

3.2.3. Prophage analysis

The PHAST webtool was used to identify prophages within the *E. limosum* ATCC 8486 genome (Table 3.2.5). Three prophage containing regions were identified, including two intact phage genomes and one of questionable completeness. Prophages were also observed in the genomes of both other *E. limosum* strains which were examined, with one in the SA11 strain and three in the KIST612 strain. Overlap between the predicted prophages which were identified by PHAST exist between the type strain and SA11, but not with the KIST612 strain. The predicted lactobacillus derived phage, ple3 is reported for both SA11 and ATCC 8486.

Table 3.2.5 – Summary of prophages identified by PHAST in three *E. limosum* genomes

Strain	Region length	Completeness	#CDS	Region position	Best match phage	GC percentage
8486	28.3 Kb	questionable	31	445639-473985	Phage_lactob_ple2_ nc_031036	47.39%
8486	48.3 Kb	intact	58	1684545-1732902	Phage_lactob_ple3_ nc_031125	47.27%
8486	42.4 Kb	intact	71	3713316-3755799	Phage_propio_pfr2_ nc_031108	46.64%
SA11	58.8 Kb	intact	150	1889524-1948364	Phage_lactob_ple3_ nc_031125	40.86%
KIST612	16.3 Kb	intact	22	1247413-1263743	Phage_lactob_c5_ nc_019449	46.38%
KIST612	13.1 Kb	intact	16	2937932-2951066	Phage_lactob_c5_ nc_019449	47.86%
KIST612	45.7 Kb	Intact	68	2963400-3009191	Phage_clostr_phiCD111_ nc_028905	43.37%

Analysis of three *E. limosum* genomes for prophage content. Details of phage loci in each genome identified with overviews of features.

3.2.4. C1 metabolism

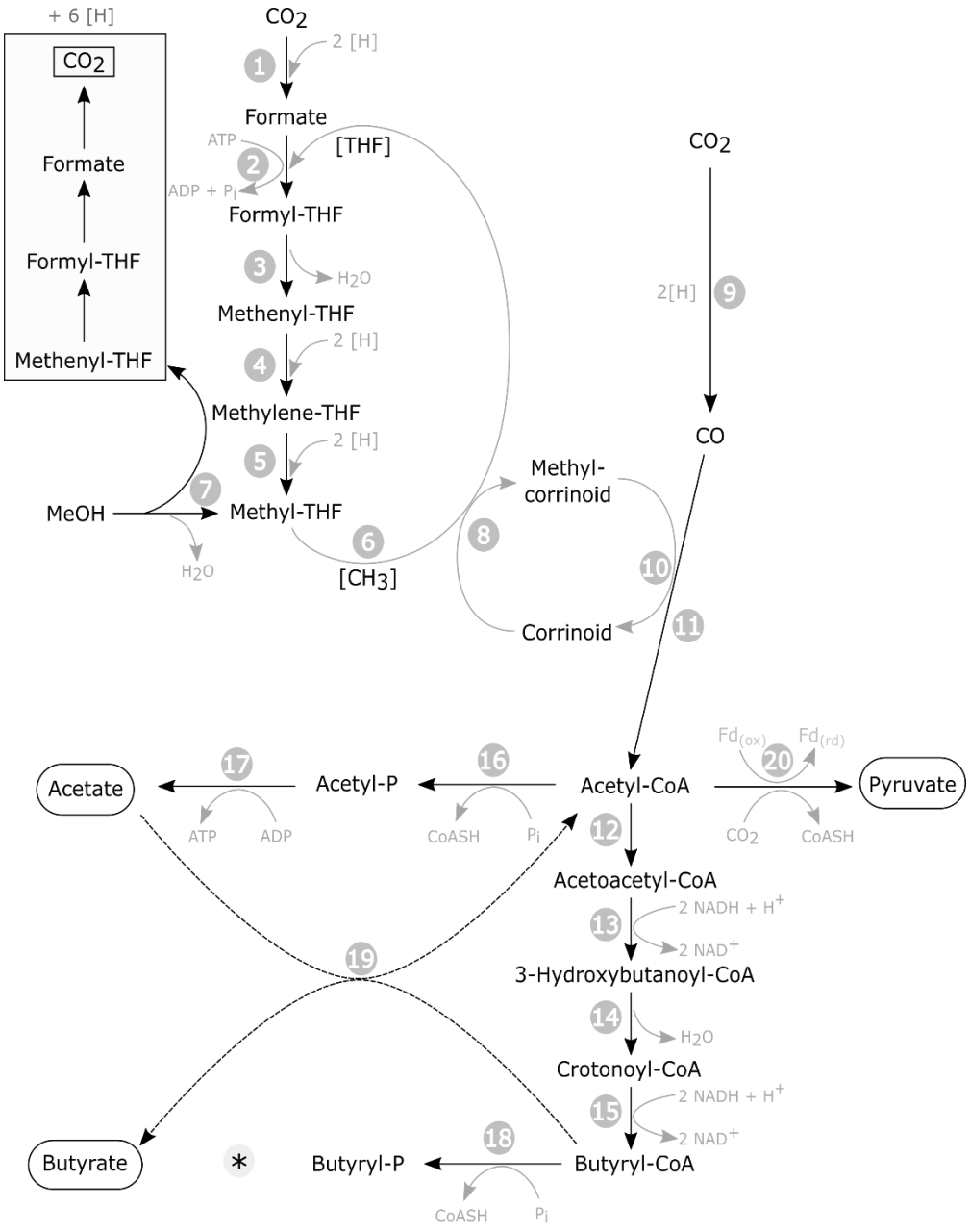


Figure 3.2.3 – Schematic of the Wood-Ljungdahl pathway and associated downstream pathways

Pathways required for core C1 metabolism, acidogenesis, and biomass synthesis. Table 3.2.6 (overleaf) contains details on enzymatic involvement, E.C. reaction identifiers, and relevant locus tags for homologues identified in the *E. limosum* ATCC 8486 genome.

Table 3.2.6 – Information table for Figure 3.2.3

Figure 3.2.3

#	Enzyme	Reaction E.C.	Locus ID (Ga0111251_)
1	Formate dehydrogenase	1.2.1.2	111919
2	Formate-THF synthetase	6.3.4.3	11342
3	Formyl-THF cyclohydrolase	3.5.4.9	11344
4	Methylene-THF dehydrogenase	1.5.1.5	11344
5	Methylene-THF reductase	1.5.1.20	11345
6	Methyltransferase	2.1.1.258	11341-11351, 11950
7	methanol: coenzyme M methyltransferase	2.1.1.90	112774
8	Corrinoid/iron-sulfur protein	2.3.1.169/1.2.7.4	111052-111055
9	CO dehydrogenase/acetyl-CoA synthase	1.2.7.4/2.3.1.169	111052-111055
10	Corrinoid/iron-sulfur protein	2.3.1.169/1.2.7.4	111052-111055
11	CO dehydrogenase/acetyl-CoA synthase	1.2.7.4/2.3.1.169	111052-111055
12	Thiolase	2.3.1.9	113833
13	3-hydroxybutyryl-CoA dehydrogenase	1.1.1.157	113831
14	Crotonase	4.2.1.17	113832
15	Butyryl-CoA dehydrogenase	1.3.8.1	113830
16	Phosphate acetyltransferase	2.3.1.8	11604
17	Acetate kinase	2.7.2.1	112277
18	Phosphate butyryltransferase	2.3.1.19	11734, 111765
19	Butyryl-CoA: acetate CoA – transferase	2.8.3.1/2.8.3.8	11892
20	Pyruvate oxidoreductase	1.2.7.1	1174

Genes and associated reactions for core C1 metabolism identified in the *E. limosum* ATCC 8486 genome. Assignments to locus tags determined during automatic annotation pipeline and confirmed by BLAST alignment. The genes of both core branches of the WL pathway are divided across three distinct loci in the *E. limosum* ATCC 8486 genome (Figure 3.2.3). Formate dehydrogenase (Figure 3.2.3: gene 1) is the only gene of the methyl-branch of the pathway that is located independently from an operon containing five remaining genes (Figure 3.2.3: genes 2-6). The genes responsible for the carboxyl branch of the pathway are similarly contained in a single operon (Figure 3.2.3: genes 8-11), encoding two bifunctional protein complexes ultimately responsible for coordinating the transfer of acetyl moieties to coenzyme-A. Genes encoding the subunits of the coenzyme-M methyltransferase responsible for the initial steps in methanol metabolism, are located separately from either of the three loci containing the core WL genes. Analysis of the type strain revealed genes encoding a Na+-dependent Rnf complex (Ga0111251_111050) and the flavin-based bifurcating enzyme e*tfAB* (Ga0111251_113828, Ga0111251_113829). Consistent with findings in *E. limosum* KIST612 the Rnf complex is encoded within the same operon as the carbonyl branch of the WL pathway, and *etfAB* in the gene cluster responsible for butyrate production.

The predicted pathway between acetyl-CoA and butyryl-CoA in *E. limosum* ATCC 8486 consists of four genes (Figure 3.2.3: genes 12-15). In the production of butyrate, it is anticipated that the CoA moiety of butyryl-CoA is transferred to acetate *via*butyryl-CoA:

acetate CoA-transferase (Figure 3.2.3: gene 19). The usage of this route to butyrate is particularly notable given that the genome encodes for two distinct homologues of the enzyme phosphate butyryltransferase (Figure 3.2.3: gene 18), which catalyses the conversion of butyryl-CoA to butyryl-phosphate. This finding is somewhat unusual because BLAST alignment revealed no strong candidates for a butyrate kinase which would be required to further convert butyryl-phosphate to butyrate.

To elucidate a possible mechanism(s) for ethanol production based on conventional pathways in clostridiales, a series of BLAST alignments were performed against the three genes, alcohol dehydrogenase (*adh*), acetaldehyde dehydrogenase (aor), and aldehyde dehydrogenase (*ald*). Aor homologues (Ga0111251_112608, Ga0111251_111919) were identified following pBLAST alignment against of *C. autoethanogenum* Aor subunits (CLAU_0099, CLAU_0081). Multiple Adh homologues were identified Ga0111251_111869, Ga0111251_112429, and Ga0111251_113461 following alignment to ALU35889 of *C. autoethanogenum*. However, no obvious candidates were found for an Ald following alignments to the bifunctional AdhE homologue in *C. autoethanogenum* (CLAU_3655/CLAU_3656) and AdhE1 from *C. acetobutylicum* (AAD04638.1). The lack of *ald* suggests that ethanol production most likely proceeds from acetate *via*acetaldehyde, initially mediated by an Aor catalysed reaction step requiring the oxidation of reduced ferredoxin, which is followed by an Adh catalysed dehydrogenation to yield ethanol. Analysis of a fourth Adh homologue identified in the genome indicates a potential butanol capable alcohol dehydrogenase, Ga0111251_114071. pBLAST alignment against the manually curated Swiss-Prot protein database revealed highest alignment score (e = 2e-97) with a butanol dehydrogenase (P13604) from *Clostridium saccharobutylicum* DSM 13864 (APPEDNIX – Table 7.1.2). This gene assignment is consistent with reports of butanol dehydrogenase activity in the original genome publication for *E. limosum* ATCC 8486 in 2015, however butanol production was not observed throughout the course of this project nor reported elsewhere in the published literature for *E. limosum* strains.

3.2.5. *E. limosum* strain comparison

In order to assess the relatedness of the newly sequenced type strain to other *E. limosum* strains in the literature, an initial genome comparison was undertaken. Two other strains, KIST612 and SA11, were selected based on the availability of complete and closed genomes[31,34]. Additionally, both comparison strains were subjects of numerous existing publications which, if the strains were adequately related, were expected to prove relevant to future work with our strain.

Mauve genome comparison software was used to align the complete *E. limosum* ATCC 8486 genome against the genomes of the other two strains (Figure 3.2.4). Alignment revealed 34 locally collinear blocks conserved across the three taxa. These blocks represent unbroken regions of genomic sequence which appeared to be free from recombination-based genome rearrangements. The 34 blocks indicate a reasonably significant evolutionary divergence between the strains by way of large genomic rearrangements. As expected, within these broader rearrangements exists even more significant sequence variation and dissimilarities at single base resolution.

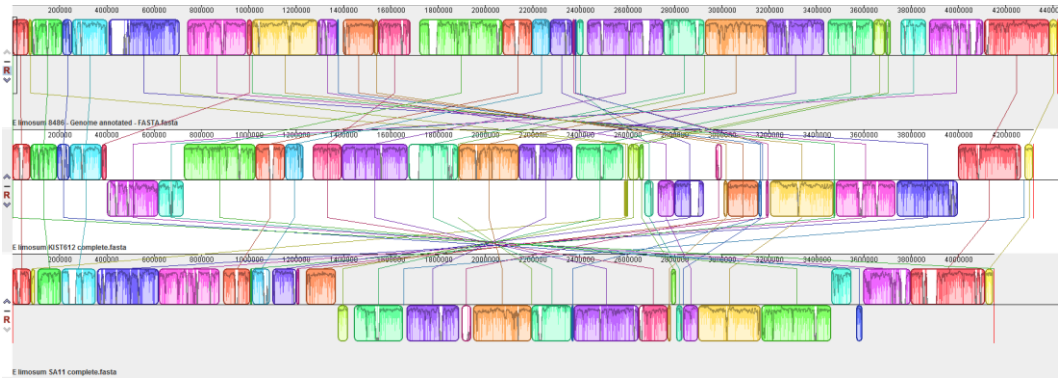


Figure 3.2.4 – Graphical alignment of three complete *E. limosum* genomes

Alignment was performed using Mauve genome comparison software. Chromosomes are represented linearly and displayed horizontally. Each coloured region represents a homologous region (block) in each genome. Blocks displayed below the centre axis are regions which occur in the inverse orientation relative to *E. limosum* ATCC 8486. Top = *E. limosum* 8486 (SBRC); Middle = *E. limosum* KIST612 (GCF_000152245.2); Bottom = *E. limosum* SA11 (GCF_001481725.1).

A pan-genome was generated from the three genomes using OrthoMCL on the Kbase analysis platform, revealing a core-genome consisting of 2951 conserved gene orthologues present across all three genomes. The core genome weighs in at approximately 75% of the total gene count for each strain, and indicated a total of 1268 singleton or non-core genes in the type strain (Appendix - Figure 7.1.3). Hypothetical proteins accounted for 677 of the genes which were not attributed to the core-genome. The remaining 591 singleton/non-core genes included a substantial number of phage related genes, in addition to a broad selection of genes encoding proteins of a wide variety of functionality, including transcriptional regulators, transporter proteins, and ATPases.

3.2.6. Type strain genome comparison

Within the last 6 months of the project an updated closed genome was published for the type strain (NCBI Reference Sequence: NZ_CP019962.1). It was deemed prudent to compare the sequence generated during this project (SBRC) to the newly published variant NZ_CP019962.1 (NCBI). Once again, Mauve was used to align the two genomes and to identify the nature of any differences (Figure 3.2.5). Mauve highlighted the location of each sequence variation between the two genomes and labelled them as 'gaps'. The nature of variations at each gap was analysed and collated into three categories based on gap sizes. The analysis reported 30 single nucleotide gaps, 13 gaps of 2-10 nucleotides, and 6 significant gaps of >75 base pairs (Table 3.2.7). The detailed gap summary is presented in appendix Table 7.1.3.

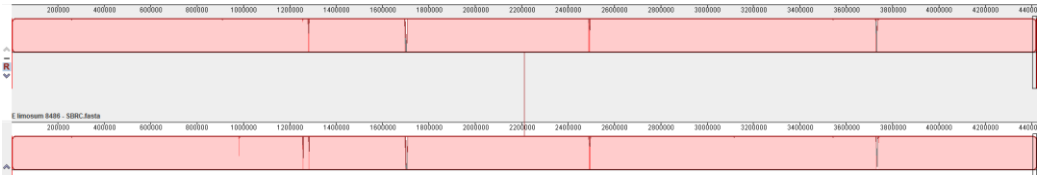


Figure 3.2.5 – Alignment schematic for NCBI and SBRC genomes for *E. limosum* ATCC 8486

Schematic generated by progressiveMauve for alignment of *E. limosum* ATCC 8486 (SBRC) with the corresponding recently closed genome from NCBI (accession number: NZ_CP019962.1), designated *E. limosum* ATCC 8486 (NCBI)[134].

Table 3.2.7 – Overview of variation between NCBI and SBRC genomes for *E. limosum* ATCC 8486

Variation type	Variation size (bp)	Variation frequency
Gap	1	30*
Gap	2-10	13
Gap	>75	6
	*Base substitution:	24
	*Base insertion/deletion:	6

Small Gap total	43
Large Gap total (>75)	6

Tabular overview of base variations identified by progressiveMauve between the published *E. limosum* ATCC 8486 (NCBI) genome and the *E. limosum* ATCC 8486 (SBRC).

Table 3.2.8 – Breakdown of major regions of genomic variation between NCBI and SBRC genomes

Region	A	B	C	D	Other	
Genome coordinates (SBRC) +/- 10 kb	1704409	3731999	1255346	2492519	257541 906907 980563	SNP Insertion (1 bp) Deletion (3 bp)
Gap length	8 kb	3.5 kb	1.7 kb	0.7 kb	113089 3540249	Insertion (1 bp) SNP
Loci Description	Phage Ple3	Phage Pfr2	Transposable element	Type I restriction system	Discrepancies (bp) at distinct genomic loci	

Table of inter-genome variation collated into four primary regions of difference (A-D) and a fifth group (other) containing the remaining loci of difference.

The majority of the variations identified between the two genomes can be broadly attributed to four loci, designated A, B, C, and D. Outside of A-D there are just five minor differences (SNPs and Indels) between the two genomes. An overview of the variations is presented in Table 3.2.8, with extended information available regarding the 'other' variations in Table 3.2.9.

Table 3.2.9 – Details of 'other' variations from Table 3.2.8

Coordinate	Locus Details	Result & Effect
257541	Intergenic non-coding between hypothetical proteins	Base substitution
906907	Polar amino acid transport system ATP-binding protein (Ga0111251_11840)	Single base deletion in SBRC – Premature stop codon
980563	Alanyl-tRNA synthetase (Ga0111251_11918)	Three base deletion in SBRC – frame maintained
3113089	Trk system potassium uptake protein TrkA	Single base deletion in NCBI – Premature stop codon
3540249	Membrane carboxypeptidase (penicillin-binding protein) (Ga0111251_113393)	Base substitution

Table contains further details on the discrepancies between SBRC and NCBI genome sequences which had been categorised as 'other' variations in
Table 3.2.8.

The three largest identified regions of variation A and B were associated with two prophages. Although the gaps called for regions A and B indicate several thousand bases of discrepancy between the two genomes, the actual disparity is the result of a relatively minor genomic repositioning of a selection of phage related genes. This included for example the inversion of a tape measure domain-containing protein (B2M23_11380, Ga0111251_111662) and the repositioning of a major capsid protein (B2M23_00015, Ga0111251_111654). Additionally, a predicted adenine-specific site-specific DNA-methyltransferase (Ga0111251_111648, B2M23_00050) was subject to repositioning between the two genomes. This methyltransferase shares 100% sequence identity to Ga0111251_113594 which exists within region B.

The next notable variation, designated region C, was the absence of a 1662 bp region preceding a sodium/phosphate symporter gene (B2M23_09030, Ga0111251_111181) at position 1255342 in the SBRC genome sequence. The missing fragment is annotated as a DDE transposase (B2M23_09025) in the NCBI sequence. The final large variation is a rearrangement of the specificity subunit of a type I restriction modification system by way of inversion of an internal portion of the open reading frame at the gene locus Ga0111251_112370. This inversion leads to premature termination of the coding sequence and truncation of the gene. The collective sum of the remaining

gaps/insertions/deletions identified in the alignment accounted exactly for the disparity in genome size reported, 4422837bp (NCBI) to 4421303 (SBRC). The full table of discrepancies is available at Appendix – Table 7.1.3.

3.2.7. Gene transfer

Establishing gene transfer in a laboratory setting requires success in three core concepts. First, getting transforming DNA into the organism. Second, ensuring intracellular retention of transformed DNA until a desired function, and thereby transformation, has occurred. Third, selection of transformed cells. Before an effective transformation methodology is known it is prudent to tackle each of these challenges in reverse. First establishing an appropriate selection method for identifying transformants of a target organism. Typically, this is achieved by harnessing the positive selection pressures of a suitable antibiotic coupled with a transformed resistance marker.

3.2.7.1. Antibiotic resistance

In order to establish a selection of suitable antibiotics for selection of successful gene transfer three antibiotics, each with an available resistance gene in the pMTL80000 vector series, were tested for efficacy against *E. limosum*. Antibiotics, thiamphenicol, clarithromycin, and spectinomycin, were tested over a concentration range broadly determined by a limited selection relevant of literature[135]. Each antibiotics efficacy for preventing measurable growth of *E. limosum* in broth cultures, over time courses of 48 and 144 hours, was assessed (Figure 3.2.6).

a)

b)

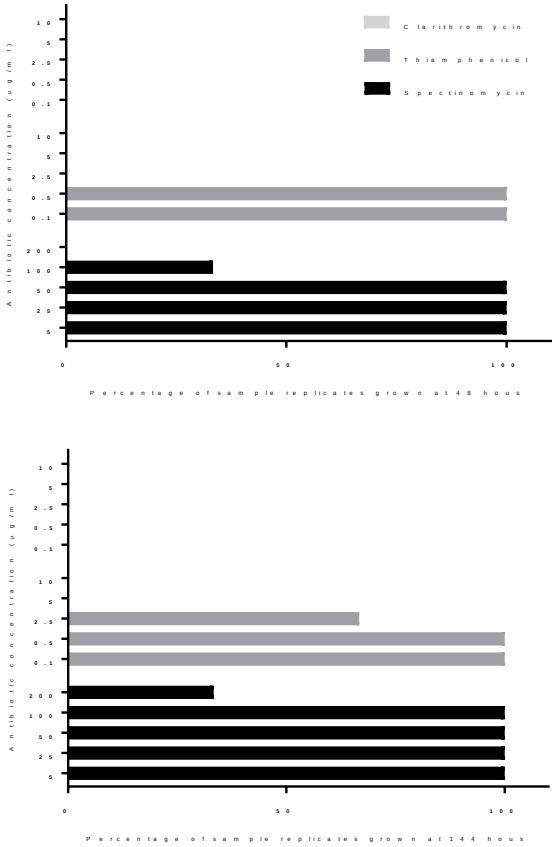


Figure 3.2.6 - Qualitative analysis of growth of *E. limosum* in PYG broth culture with three different antibiotics

Bars represent the percentage of samples which demonstrated growth at a given time point from three replicates of each condition. No growth was observed under any of the tested antibiotic concentrations of clarithromycin. Thiamphenicol prevented growth entirely at concentrations of 5 μg/ml and higher. The lower concentration of 2.5 μg/ml was initially sufficient prohibit growth, however the attenuating effects of the antibiotic appeared to have been overcome in two thirds of the replicates by 144 hours. Spectinomycin was notably less potent than the aforementioned antibiotics in terms of the concentrations required to inhibit growth. As for thiamphenicol, the potency of spectinomycin appeared to decrease over the 144 hour observational period.

3.2.7.2. Development and optimisation of electroporation methodology

With two suitable antibiotics identified in thiamphenicol and clarithromycin the next challenge was to establish an effective transformation method. Plasmid-borne gene transfer underpins the majority of tools used for the directed manipulation of microorganisms and is the obvious choice in early development of gene transfer methodologies. In order to fully exploit plasmid-based tools, a selection of viable

plasmid constructs were required on which to transport genetic payloads into *E. limosum*. Replicons perform a central role in defining the longevity of plasmids within cells following successful transformation, so it was important to identify a selection of replicons which were functionally active in *E. limosum*. To this end the initial transformation attempts were performed using a plasmid cocktail, enabling parallel testing multiple Gram-positive replicons in *E. limosum*.

Transformation was attempted with three *E. limosum* strains, DSM 2593, DSM 2594, and ATCC 8486 (DSM 29543) (Table 3.2.10), all ordered from the German collection of microorganisms and cell cultures (DSMZ). The preparation of electrocompetent cells was complicated by the extracellular polysaccharide (EP) produced (see Chapter 3.3.3), by *E. limosum* strains. The low density of the EP relative to the growth medium made the pelleting steps of the competent cell preparation difficult and in some cases completely ineffective. In the case of *E. limosum* DSM 2593, which produced perceptibly more EP than the other two strains during the preparation procedure, pelleting was not possible under the conditions tested (7000 x g, 4°C, 10-20 minutes). The remaining two strains were prepared and transformed by electroporation (Method 2.6.1.2) with a cocktail of eight different pMTL80000 derived plasmid constructs, each construct harbouring a different Gram-positive replicon. The cocktail transformation was performed using a range of different electroporation parameters to maximise the likelihood of success in the absence a previously established method. The electroporation procedure was successful in generating transformants in both DSM 2594 and ATCC 8486 *E. limosum* strains (Table 3.2.10).

Table 3.2.10 – Initial transformation attempts of *E. limosum* strains *via*electroporation

E. limosum strain DSM (ATCC)	Gene-transfer achieved	Notes
2593	No	Preparation not possible with standard method
2594	Yes	
29543 (8486)	Yes	

3.2.7.3. Replicon compatibility evaluation

Compatible replicons were identified through multiple colony PCRs of the transformants which resulted from the cocktail transformations into the type strain. Of the eight replicons tested just three demonstrated a capacity to permit plasmid persistence, pCB102, pIP404, and p19 (Table 3.2.11). Persistence of the plasmid was inferred from sustained growth on selective chloramphenicol plates, enabled by the *catP* resistance gene on the plasmid backbone. Using the proven electroporation parameters from the original cocktail transformation, each replicon was tested independently to confirm the results of the initial screen. Transformants were only observed where the three initially identified replicons were used in this second round of transformations, corroborating the initial results.

Table 3.2.11 – Replicon testing in *E. limosum* ATCC 8486

Modular plasmid (pMTL designation)	Gram-positive replicon	Organism source of replicon	Functionality in *E. limosum*
81151	pAMβ1	*Enterococcus faecalis*	-
82151	pBP1	*Clostridium botulinum*	-
83151	pCB102	*Clostridium butyricum*	Demonstrated
84151	pCD6	*Clostridium difficile*	-
85151	pIM13	*Bacillus subtilis*	-
86151	pIP404	*Clostridium perfringens*	Demonstrated
87151	p19	*Clostridium carboxidivorans*	Demonstrated
88151	pCAc	*Clostridium aceticum*	-

Plasmids 81151 to 86151 were sourced from the commercially available pMTL80000 vector series[119]. Plasmids 87151 and 88151 harbour Gram-positive replicon modules which were developed independently within the SBRC Nottingham[136].

3.2.7.4. Electroporation voltage optimisation

High transformation efficiencies were considered advantageous or essential in employing tools for advanced genetic manipulation. In an effort to improve efficiencies of gene transfer, optimisation of the electroporation voltage parameters were performed in order to maximise the number of transformants per sample. Colony counts were performed on selection plates following a series of transformations at voltages ranging between 0.5 and 2.5 kV in order to determine the optimal electroporation voltage (Figure 3.2.7). Voltage optimisation yielded approximately a 5-fold improvement in effectiveness at 2.1 kV over the initial testing voltage of 1.5 kV, increasing transformation rates from $5.2x10^3$ to $2.6x10^4$ CFU/µg plasmid.

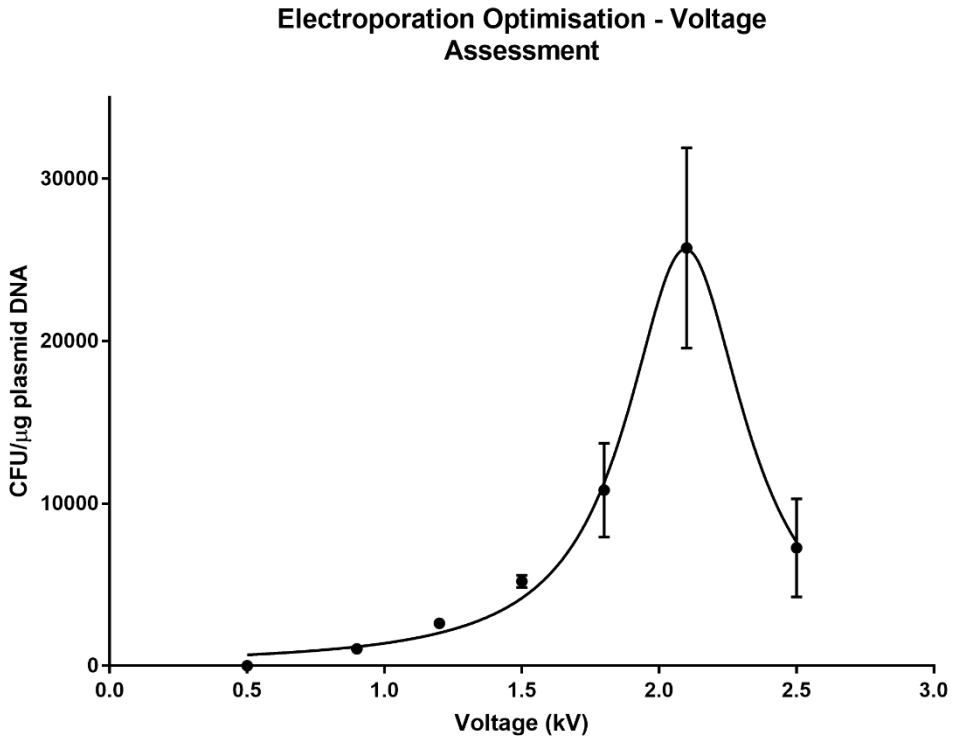


Figure 3.2.7 – Voltage optimization of electroporation of *E. limosum* ATCC 8486.

Colony forming unit (CFU) counts per µg of pMTL86151 plasmid DNA on selection plates following transformations at seven voltages between 0.5 and 2.5 kV. Each transformation was carried out in technical triplicate from a single batch of electrocompetent cells. Points represent the average CFU/ µg values of this triplicate and error bars reflect the standard deviation.

3.2.7.5. Replicon Characterisation

The stability of two replicons was tested in order to identify suitable candidates for different genetic modification applications. Analysis of plasmid retention over time was measured in serially passaged broth culture in the absence of sustained selective pressure (Figure 3.2.8). *E. limosum* harbouring plasmids containing a chloramphenicol resistance gene and one of either pCB102 or p19 Gram-positive replicons was serially passaged into fresh non-selective media once every 24 hours. Samples taken at each 24-hour time point were concurrently plated onto selective and non-selective media to assess for retention of the plasmid-based chloramphenicol resistance marker reflected by colony forming unit (CFU) counts across the two conditions. Each 24-hour period was calculated to represent an average of 15.7 (±1.05) generations of growth in non-selective media, giving an approximate total of 110 generations over the 7-day period.

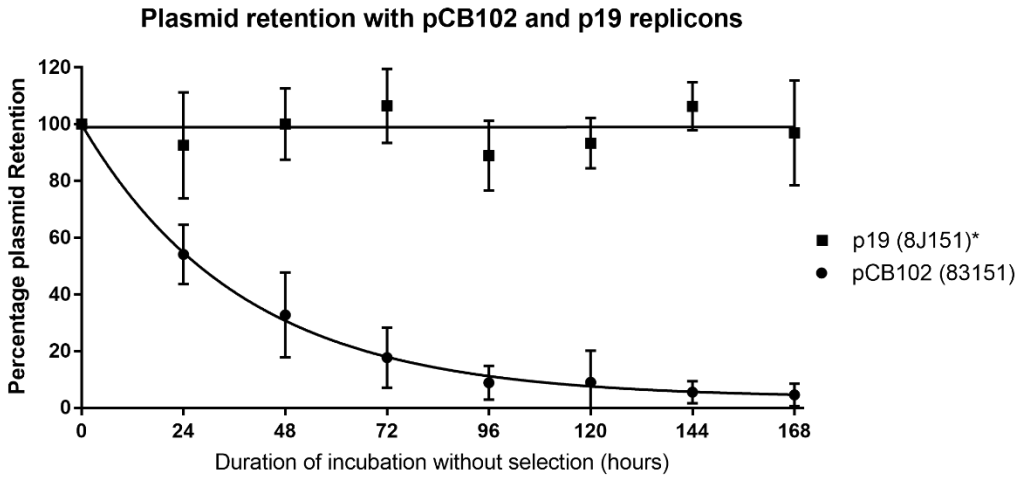


Figure 3.2.8 – Replicon stability assay data for replicons pCB102 and p19* in *E. limosum* ATCC 8486.

Plasmid containing cells were inoculated into non-selective media and incubated at 37°C cells and were serially passaged once every 24 hours for 7 days. The plasmids tested were pMTL83151 and pMTL8J151. Plasmid retention was assessed by colony forming unit count variations between selective and non-selective agar plates produced for each sample every 24 hours. Data points are representative of average values from four biological replicate samples for each replicon. Error bars represent the standard deviation within the quadruplicate.

3.2.8. Allele-coupled exchange (Lac-TcdR)

Following the establishment of transformation and identification of a suitable pseudo-suicide replicon, pCB102, it was possible to begin implementing advanced genome engineering tools based on manipulation of genomic targets identified within the genome. The roadmap suggests the implementation of an orthogonal promoter system as a platform for future overexpression studies. This requires the insertion of a gene cassette encoding a lactose inducible expression system controlling the expression of a TcdR transcription factor from *C. difficile*.

A Δ*pyrE* strain of *E. limosum* ATCC 8486 kindly provided by James Millard (University of Nottingham 2016) was used as a target strain for ACE with a Lac-TcdR gene cassette at the *pyrE* locus. The created Lac-TcdR+ strain was developed for use as an overexpression chassis later in the project (Chapter 4). Insertion of the Lac-TcdR cassette was coupled to the repair of the previously knocked-out (truncated) *pyrE* gene. The Δ*pyrE* phenotype of the target strain allowed for selection of successful ACE events by way of a restored uracil autotrophy. This was achieved by selection of transformants on uracil deficient media. Restoration of *pyrE* and the coupled insertion of the Lac-TcdR encoding region was confirmed by screening colony PCR using primers flanking the affected region of transformants following two re-streaks across uracil deficient agar plates in the absence of antibiotic selection (Figure 3.2.9).

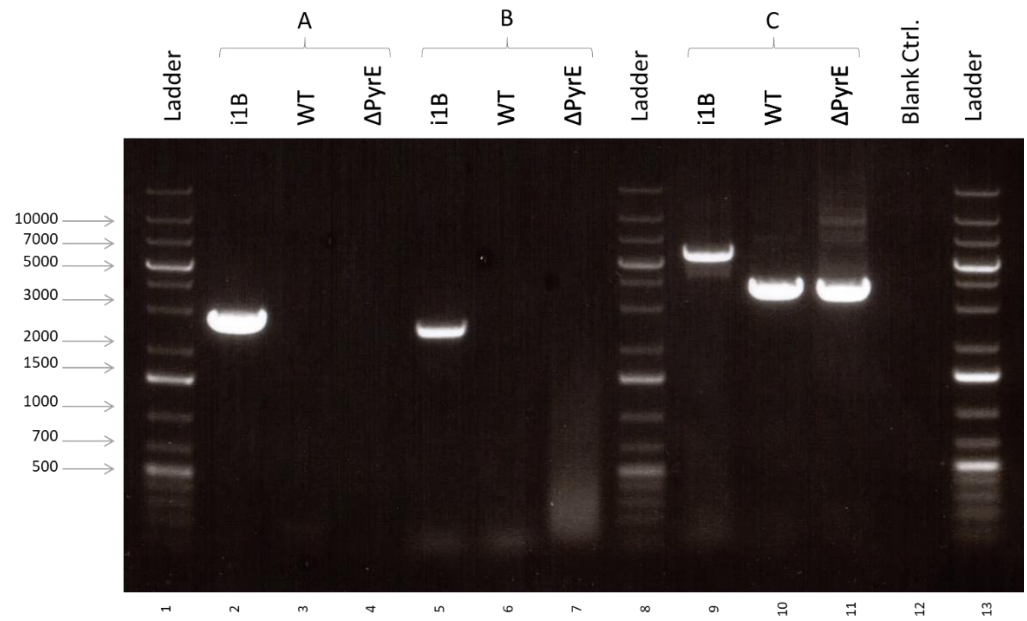


Figure 3.2.9 – Screening PCR confirmation of *pyrE* repair and Lac-TcdR knock-in by Allele-Coupled Exchange in *E. limosum* ATCC 8486

Screening of the Lac-TcdR strain of *E. limosum* 8486, designated mutant i1B. A) Screening from upstream of *pyrE* to an internal primer for the Lac-TcdR insert using primers: [*PyrE*_scr_F] and [*bgaR*_int_R2]. Lane 2-4 expected 2745 bp product if *LacTcdR* cassette is present at *pyrE* locus. B) Screening of downstream of the *pyrE* target region to an internal primer for Lac-TcdR insert using primers: [tcdR_int_F] and [*PyrE*_scr_Rv]. Lane 5-7 expected 2543 bp product if *LacTcdR* knock in is present at *pyrE* locus. C) complete screening of the targeted *pyrE* region using two flanking primers: [*PyrE*_scr_F] [*PyrE*_scr_R]. Lane 9-12 expected 5630 bp product if *LacTcdR* is present at *pyrE*, 3455 bp in WT, 3238 bp in Δ*pyrE* strain. O'generuler 1kb plus ladder (NEB).

3.2.9. CRISPR/Cas9 proof of concept

The CRISPR/Cas9 mediated knock out vector pSBRCas9_El_*Pbt*1 (Figure 3.2.10) was assembled as described in Method 2.3.2, for targeted KO of the *pbt1* gene (Ga0111251_111765) of *E. limosum* ATCC 8486. Following initial transformation of the vector into *E. limosum*, a selection of thiamphenicol resistant colonies were picked and re-streaked onto fresh antibiotic containing plates and subsequent growth was screened by way of colony PCR using primers flanking the homology arms at the *pbt*1 genomic locus to confirm knockout success. Following one re-streak ~50% of the colonies tested presented a KO genotype while the remaining colonies returned either no band or a WT sized band (Figure 3.2.11). The CRISPR plasmid was lost following 2 further re-streaks of the PCR confirmed KO mutants, evidenced by an inability to grow on thiamphenicol containing media.

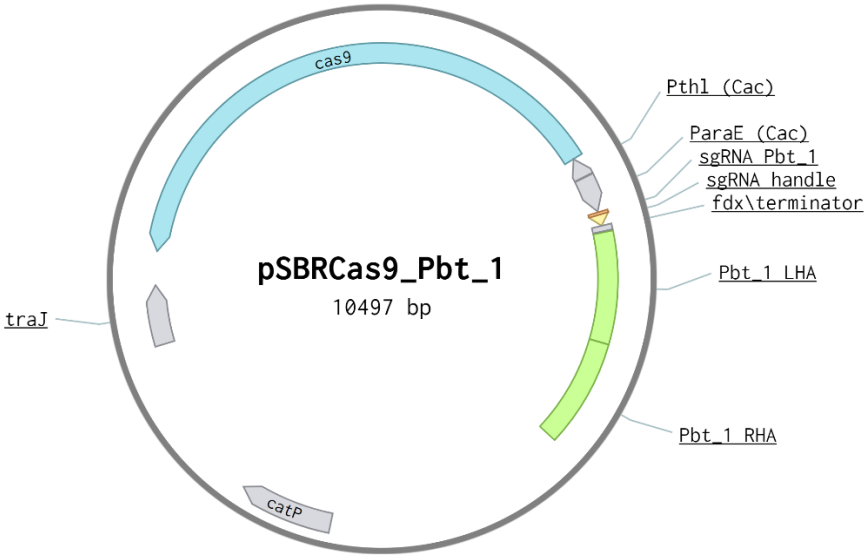


Figure 3.2.10 – Vector map for pSBRCas9_*Pbt*_1

A vector for targeted knockout of phosphate butyryltransferase 1 (Ga0111251_111764) by CRISPR-Cas9 mediated homologous recombination. Homology arms LHA and RHA (green) were derived from 750 bp up and downstream of the *pbt*1 gene in *E. limosum* ATCC 8486. The sgRNA handle (orange) is a 20 bp sequence within the *pbt*1 gene for targeting Cas9 cleavage (blue) activity. A detailed record of this vector is available at Figure 2.14.3.

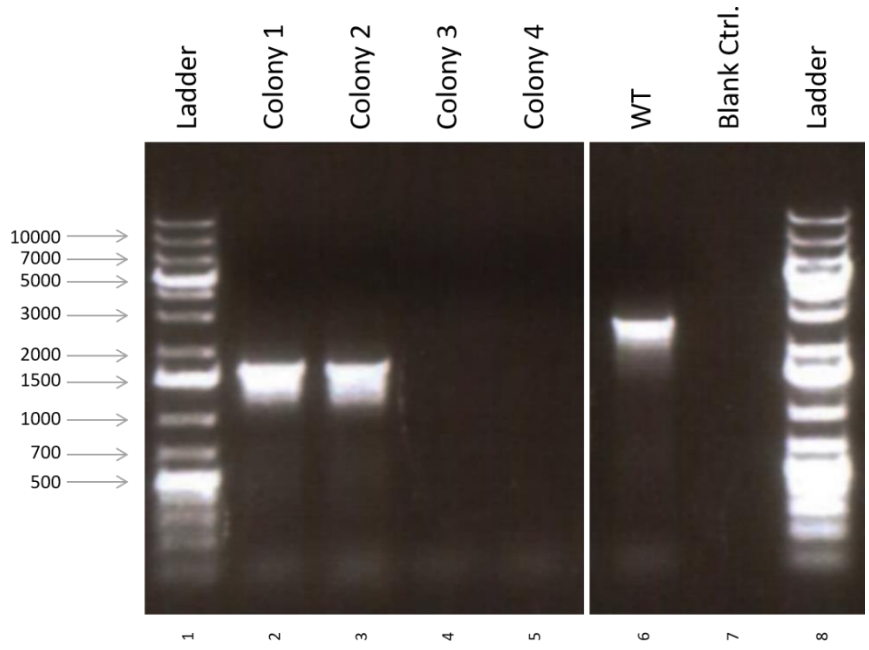


Figure 3.2.11 – Screening PCR for confirmation of *pbt*1 knock-out by CRISPR in *E. limosum* ATCC 8486

Screening the *pbt1* gene locus of *E. limosum* ATCC 8486 colonies to confirm gene deletion by CRISPR-Cas9 associated homologous recombination. Lanes 2-5: is a representative screen of four thiamphenicol resistant colonies following transformation of the WT strain with pSBRCas9_El_*Pbt*1. The desired CRISPR knockout genotype is observed in colonies 1 and 2 (lanes 2 and 3) indicated by the 0.9 Kb band, smaller than the 2.6 Kb band expected for the WT (lane 6). This reflects the 1.7 Kb deletion of the *pbt1* gene in the mutants. No band was generated from the colonies screened in colonies 3 and 4 (lanes 4 and 5), this may represent of failure of the PCRs or single-crossover plasmid integration at the target locus without CRISPR-Cas9 cleavage of the genome. Ladders: O'generuler 1kb plus (NEB). Primer pair: [pbt1_scr_F][Pbt1_scr_R].

3.3. Discussion

3.3.1. Target chassis

The initial focus of this project was the genetic development of an acetogenic bacterial species, selected based on attractive naturally-occurring phenotypes which were indicative of possible research and commercial value. A key phenotypic attribute which highlighted certain acetogens as possible subjects for this research was the ability to produce complex products, beyond acetate and ethanol, from C1 feedstocks. Two acetogenic species which fitted this criterion were *C. carboxidivorans* and *E. limosum* (Chapter 1).

C. carboxidivorans P7 was identified as an attractive target strain based on its ability to produce butyrate, butanol, hexanoate, and hexanol from syngas[39]. Encompassing one of the most extensive product ranges of all known acetogens. Previously, advanced genetic investigations of *C. carboxidivorans* had been precluded by a lack of effective gene transfer methodology in the organism. Preliminary results made available through an international collaboration with The Shanghai Institute of Biological Sciences, which indicated the successful development of a novel gene transfer methodology. On this basis *C. carboxidivorans* was selected as the initial target for strain development. Regrettably, following significant time and resource investment in attempting to recreate the preliminary findings, gene transfer in this organism was never ultimately established efficaciously. It is noteworthy that in spite of continued research and commercial interest globally, it appears that *C. carboxidivorans* remains untransformable, based on the publicly available literature at the time of writing.

In the search for an alternate interesting acetogenic chassis as the basis for study *E. limosum* was identified as an attractive target species for two foremost reasons. First, based on the acetogen's ability to naturally produce the C4 compound butyrate from C1 feedstocks. Second, the acetogen's ability to utilise a diverse range of C1 substrates, including CO, $CO_2 + H_2$, formate, and methanol. As for *C. carboxidivorans*, *E. limosum* had attracted notable research interest in its capacity as a biocatalyst prior to this study

and successful genetic transformation not been demonstrated. The principal focus of previous research had been fermentation characterisations and optimisations, generally using type strain, *E. limosum* ATCC 8486[137,138], or a Korean isolate, *E. limosum* KIST612[32,139]. These strains, along with a number or other reported *E. limosum* strains, harbour relatively similar growth profiles and characteristics[31]. As such, the prior research across each individual *E. limosum* strains was expected to be well applicable to the others, and therefore not an overwhelming factor in the decision for which strain to use.

The type strain, *E. limosum* ATCC 8486, was selected as the primary target for further study for a number of reasons. As well as it's amenability to transformation, the type strain was the easiest strain to handle in a practical setting. Specifically, the type strain cultures typically formed cell pellets more readily during centrifugation than *E. limosum* 2593 *and E. limosum* 2594, possibly linked to lower EP production. Prior research in the type strain was second only to the KIST612 strain, as measured by publication number on NCBI. This provided a solid point of reference for the project going forward. Legal restrictions surrounding research publication and development of the KIST612 strain represented a significant drawback which ultimately dissuaded further research on the strain in this project. This was particularly relevant in the context of future plans for commercial strain development.

3.3.2. Implementing the roadmap

The strain development roadmap was outlined in Chapter 1[55], and was loosely applied as a framework for the establishment of advanced genetic tools in *E. limosum* in this project. This began with the acquisition of a complete and annotated genome sequence. Following this, the development of an effective gene transfer processes was vital in enabling downstream targeted genetic manipulation techniques, ACE and CRISPR/Cas9.

3.3.3. Closing the Genome

Closed genomes offer a number of advantages over their draft counterparts, not least their significantly reduced error rate[140]. The complete closed genome sequence was thought to provide a solid basis for the project, allowing all future hypotheses testing with confidence in the genetic integrity of strain in which they would be evaluated. Crucially, such a genome provided the most robust foundation for numerous functional genomic applications, from basic primer design to the identification of metabolic pathways. Additionally, access to the 'whole picture' was valuable in performing inter-genome comparisons with closely related species and other *E. limosum* strains. Such comparisons, combined with evaluation of basic genomic organisation, can provide retrospective insights into the evolution of a given species or strain.

The dual sequencing approach which was implemented for the closure of the *E. limosum* ATCC 8486 genome involved an initial round of PacBio sequencing to generate a preliminary genome template, followed by error correction *via*mapping of Illumina sequencing read data. This approach harnessed PacBio sequencing data for assembly of a closed genome, where the short-read datasets of traditional NGS methods were sufficient only in the production of draft level (multi-contig) genome assembly[141].

The primary shortfall of Illumina datasets for genome closure are well documented and revolve primarily around the repeat sequences within bacterial genomes[61]. These regions cannot be resolved algorithmically due to the relatively short length of each Illumina read. The *de-novo* assembly of these Illumina reads results in the generation of numerous unresolvable contigs which must be mapped to an already complete reference genome for resolution of their order and thereby genome closure.

PacBio overcomes the issue of repeated sequences within genomes through significantly increased individual read lengths, whereby a single read can cover the entirety of a given repeat region. This allows the generation of fully resolvable overlapping reads, containing distinct sequences at either side of occurrences of

extended repeat motifs. An additionally benefit of PacBio over traditional NGS technologies for the project was the supplementary attainment of epigenetic information for the sequenced genome, in the form of base modifications[66]. This information was valuable in planning initial gene transfer approaches and provides a basis for further optimisation of established methods on the basis of RM system evasion. The higher coverage and comparatively lower error rate of traditional Illumina sequencing technology remained relevant for essential error correction of the PacBio-closed genome, prior to ultimate sequence finalisation[142].

Extraction of *E. limosum* genomic DNA of sufficiently high quality for the PacBio sequencing process was unexpectedly problematic at the project outset. This was due to the a somewhat unique phenotype exhibited by all *E. limosum* strains handled during the course of this project. The production of significant amounts of extracellular polysaccharide, glycogen, proved to be inhibitory to the application of a number of essential research procedures, namely DNA, RNA, and protein extractions. As a result, otherwise routine protocols were more complicated than was desirable. This glycogen producing phenotype had also been reported in a other *E. limosum* strains in the literature[31]. To address the inhibitory effects of the extracellular polysaccharide modified protocols for genomic DNA and protein extraction were developed, employing additional pre-treatment steps to negate the undesirable effects of the polysaccharides on the underlying protocols. More details for these process modifications can be found at Appendix 7.1.2.

The complete closed genome generated in this project provided a foundation for the experimental design and subsequent testing of both genome engineering techniques, CRISPR/Cas9 and ACE, in *E. limosum*. The ability to generate accurate homology arms at target loci is central to the application of both techniques. Although in many cases a draft genome is sufficient for achieving this, a complete genome provides a better platform for troubleshooting failures or identifying potential issues without needing to set foot in a lab. This includes full identification of duplicated regions and repeat

sequences which can disrupt cloning procedures through undesired off target effects. Such regions are indeed also the primary cause of draft genome assemblies, where overlapping fragments cannot be resolved, leaving genomes as multiple unassembled contigs[66].

An exciting future application of this complete genome is the development of a genome-scale model for *E. limosum* ATCC 8486. Production of such a model is one of the primary areas of focus for the continuation of this project. Well-designed models can provide unparalleled prediction tools for estimating the effects of genetic alterations, and can be used to guide the development of strains towards desired goals[90]. The informed decision making which modelling can enable, acts to streamline this process of strain modifications toward a desired outcome which can afford significant monetary and time savings in the lab[143].

Computer modelling applications benefit from the complete picture which a complete genome can provide, specifically the maximal enzymes identification, from which complete reaction pathways can be inferred. With specific regard to the implementation of *E. limosum* as a biological catalyst, carbon flux models generated to predict the conversion of C1 substrates into fermentation products, are arguably the most valuable tools for applying rational design to future strain development. Ultimately, computer modelling of microbial systems provides the most comprehensive bridge available to tie laboratory observations and genomic level bioinformatics.

3.3.4. Genome structure and features

The genes of the core WL pathway are split across three distinct loci in the genome of *E. limosum* ATCC 8486. This three part distribution of the methyl and carbonyl branches of the WL pathway aligns with previous observations made in the genomic structure of *E. limosum* KIST612 strain[35].

The predicted pathway for butyrate production in *E. limosum* ATCC 8486 proceeds *via*butyryl-CoA: acetate CoA-transferase. This is as opposed to the well documented butyrate kinase dependent mechanism for butyrate production which is reported in many butyrate producing microorganisms such as *C. acetobutylicum* ATCC 824[144]. This butyrate kinase-independent mechanism has previously been reported in a handful of other anaerobic microbes which typically constitute part of mammalian gut microflora[145], and is consistent with the reported genotype of *E. limosum* SA11[31]. The presence of two predicted butyryltransferases makes the lack of a butyrate kinase particularly unusual, as this suggests butyryl-phosphate generation without a means of ultimate conversion to butyrate. This may be indicative of an alternate role for butyryl-phosphate within *E. limosum* metabolism, though what exactly that might be is unclear. In the absence of obvious metabolic function, the phosphate butyryl-transferase genes were recognised as attractive targets for knockout, in the proof of principle application of the genome editing tool, CRISPR/Cas9.

The pathway for ethanol production in *E. limosum* has yet to be characterised fully in the literature. Based on established ethanol production pathways in other acetogenic organisms[146], the available gene selection in *E. limosum* is indicative of a single route for synthesis, proceeding *via*Aor. The alternate route, *via*aldehyde dehydrogenase, is unlikely primarily due to the absence of a good candidate homologues for this gene.

The capacity for *E. limosum* ATCC 8486 to produce butanol from syngas was noted in the original draft genome announcement of the strain in 2015, however neither evidence nor detailed analysis were provided[141]. Throughout this study butanol synthesis was never observed on any of the tested substrates, glucose, fructose, and methanol+CO_2. If the *E. limosum* Aor (Ga0111251_112608, Ga0111251_111919) possesses sufficient laxity in substrate specificity then the utilisation of butyrate, in addition to acetate, may be possible. Generation of butyraldehyde from butyrate *via*Aor in this manner, followed by an alcohol dehydrogenase conversion, is one possible pathway for butanol production in *E. limosum*. Such Aor flexibility reported in

C. autoethanogenum, where externally supplemented butyrate can be converted to butanol with over >80% (mol/mol) efficiency in CO grown batch cultures[146]. A possible explanation for the lack of observable butanol produced during fermentations over the course of this project is that *E. limosum* may have the capacity to assimilate the alcohol and utilise it as a substrate for growth. However, suggestion that butanol did not support growth of *E. limosum* in batch culture when used as the sole growth substrate in the initial characterisation of the strain report[137], provides some evidence in contradiction of this explanation.

A likely explanation to the limited alcohol titres observed during this project lie in the conditional nature of alcohol production in response to media related variables such as pH, salt concentrations, substrate concentrations, and organic nitrogen concentration[147]. Such variables impact product titres by affecting gene expression, through transcriptional modulation in response to real-time conditions, as well as by directly impacting underlying reaction kinetics[148]. The media formulations selected for the majority of work performed during this project, AWM1, prohibited detectable ethanol production, while ethanol was observed at low levels during some fermentations on AWM2 media (Chapter 4).

3.3.5. Genome Comparison

3.3.5.1. *E. limosum* Pan-genome

In order to assess the relatedness of the three *E. limosum* strains, ATCC 8486, KIST612, and SA11, basic genomic alignment and pan-genome analysis were performed. Initial alignments indicated reasonably significant divergence of the three strains from one another, resulting in a minimum of 34 locally collinear blocks. This would suggest at over 30 significant genomic rearrangement, horizontal transfer insertion, or large deletions, had occurred since the strains diverged from a common ancestor[149].

Pan-genome analysis revealed approximately 75% functional overlap between the three available complete *E. limosum* genome sequences by way of ortholog alignment. In reality the overlap might be even higher, with further ortholog assignments possible from the predicted non-core gene selection, particularly those currently assigned as proteins of unknown function or hypothetical proteins. A known weakness of OrthoMCL algorithm is the effective matching of hypothetical proteins, proteins of unknown function, and paralogues[150]. Significant variation between the three strains is not unusual however. Tettelin *et al*. reported an expected average of nearly 100 strain-specific genes per genome during intraspecies pan-genome analysis when comparing three given *Streptococcus agalactiae* stains[150]. Furthermore, the group reported that an intraspecies pan-genome accounted for ~80% of genes for each strain. A figure which aligns well with equivalent values based on the pan-genome generated for *E. limosum*. Gene orthologs for all genes identified as being central to core C1 metabolic pathways were found in the core genome (Figure 3.2.3), in each of the three genomes.

3.3.5.2. NCBI vs. SBRC genome sequences

A significant number of discrepancies were identified between the genome sequence generated in this study and a recently published closed genome for *E. limosum* ATCC 8486. However, the majority of these points of variation were isolated to four genomic loci. Two of these loci were related to prophages, the third related to a transposable element, and the fourth to a rearranging type I RM system.

The largest two 'gaps' called between the two strains, Regions A and B, both occurred within predicted prophages identified by PHAST the phage finding tool. In actuality both of these large gap calls were relatively minor apparent genomic repositioning of a small selection of genes at the ends of each prophage. Due to the natural flanking of prophages by insertion repeats[151], it seems likely that during the original read mapping during genome assembly and the subsequent alignment of genomes, would be problematic at these loci. If indeed these discrepancies are legitimate differences

between both strains, it remains difficult to predict the effects of these discrepancies due to limited knowledge of the prophages in question. How the activity of either/both discrepancies might interact with wider cellular metabolism, and ultimately how these effects might present themselves phenotypically, cannot be evaluated fully without practical data from a side by side comparison of the two strains.

The transposable element which gave rise to variation region C is reported as a DDE-motif containing transposase, a putative retroviral integrases[152]. Integrases and alike would ideally be avoided in strains destined for genetic development, with view to their negative impact on stability of downstream genetic manipulations. The absence of this feature from the SBRC strain is therefore seemingly positive, however further investigation revealed five sequence-identical loci elsewhere in the genome, one less than the NCBI sequence. This rearrangement is likely the result of activity of a shared dormant phage leading to excision and reinsertion of the relocated fragments. Transposon numbers tend to exist at copy number equilibriums controlled by regulatory mechanisms related to their transposition rates, affected by copying/cutting and (re)insertion frequencies, in a bid to minimise their deleterious effects on host cells[153]. It seems likely that the 5-6 copy number represents an equilibrium for this transposon within *E. limosum* and should be relatively stable on this basis. Whilst the wider implications of these rearrangements are currently limited to speculation, one could expect that polar transcriptional effects at the excision and reinsertion sites may have some bearing on neighbouring gene function.

Rearranging Type I systems have been described in a number of prokaryotic species[154]. The genomic shuffling is implicated in the modification of specificity subunits, which can lead to inactivation of RM systems or an adjustment in their motif specificity[87]. Rearrangements rates vary from system to system and can be regulated by a number of internal factors, some occurring so regularly that single colony populations will contain mixed configurations[155]. It is hypothesized that the inversion of the specificity of the S subunit (Ga0111251_112372) of the type I RM system (OrfD) identified in the

type strain is catalysed by a genomically encoded site-specific recombinase genes. The XerD type recombinase Ga0111251_112372 seems like the most likely candidate for this based on its immediate proximity to the invertible region.

Just five alignment discrepancies existed outside of regions A-D. Of these none were obviously implicated in affecting genes responsible for core metabolism, so it would be expected that their phenotypic impacts would be minimal. As noted for above however, without data from side-by-side comparison of the two strains it is impossible to say conclusively that these variations are unimportant.

3.3.6. Establishing gene transfer in *E. limosum*

Establishing gene transfer in prokaryotes according to the roadmap includes the identification and characterisation of libraries of suitable plasmids and, where necessary, optimisation of the process to improve transfer efficiencies.

3.3.6.1. Antibiotic testing

Antibiotic testing was required to establish a collection of effective selective markers for downstream genetic manipulations in *E. limosum*. Three antibiotics were selected based on the availability of resistance genes in the pMTL80000 vector series, which contains a number of interchangeable resistance markers, each well established in their effectiveness for cloning and in applied research with organisms of the clostridiales order. These were, clarithromycin, thiamphenicol and spectinomycin. Clarithromycin and thiamphenicol had expected efficacy in *E. limosum* based on literature of antibiotics of the same class, erythromycin and chloramphenicol respectively, however spectinomycin was previously unproven[135,156]. An effective working concentration for our applications, required that antibiotics maintained efficacy for at least a four-day time period in accordance with procedural requirements and growth rates.

Qualitatively assessment of antibiotic efficacy against *E. limosum* over 144 hours indicated that thiamphenicol and clarithromycin would be suitable as selective markers. Clarithromycin was selected as an alternative to erythromycin due to the pH sensitivity of the latter. Although the optimal media pH for *E. limosum* strains has been reported to occur in the relatively neutral range of 7.0-7.2[127], a target which is generally reflected in culture medium, acetate and butyrate production during fermentation can lead to significant pH reductions where appropriate buffering is not employed or available[157]. Clarithromycin was effective even at the lowest tested concentration of 0.1 µg/ml and showed no loss in efficacy over the observation period. Thiamphenicol was selected because unlike chloramphenicol it does not possess a reducible nitro group which leaves the antibiotic molecule open to inactivation by nitroaryl reductases which are present in a number of clostridiales[158,159 Pg.348].

Experimental observation from later work has suggested that the observed inhibitory concentrations in broth did not correlate exactly with growth on solid media conditions for thiamphenicol. Following successful transformation procedures, background colonies begun to appear on thiamphenicol agar at concentrations of 7.5 µg/ml after 144 hours. For this reason, working concentrations of 15 and 10 µg/ml thiamphenicol was selected for agar and broth cultures respectively. No background growth was observed using clarithromycin at 6 µg/ml on either broth or agar.

3.3.6.2. Development & Characterisation of Gene Transfer

The two most well-established methods of gene transfer in Gram-positive bacteria are electroporation and conjugation[82,160]. It was decided that, if achievable, electroporation was the preferable, primarily because electroporation is not limited by the basic requirements of mobilisable plasmids used for conjugation. This includes the necessity for an origin of transfer (OriT), most commonly RK2, on the transforming plasmid[40]. Conjugation also requires a suitable conjugal donor which is able to both establish effective conjugation and subsequently enable selective separation from the newly

transformed strains. However this is generally quite easy to achieve using *E. coli* conjugal donors and Gram-negative specific antibiotics such as trimethoprim or D-cycloserine[161]. Additionally, recombineering type transformation approaches such as *recT* mediated homologous recombination are not compatible with conjugative gene transfer, as these methods require direct transfer of linear oligos or dsDNA transformation[162]. Finally, electroporation would appear more amenable to scale-up and robotic automation than conjugal gene transfer, based on more simplistic selection processes.

Using a cocktail style approach to electroporating multiple plasmids simultaneously, viability of eight replicons were tested simultaneously. Three functional replicons were identified, pCB102 from *C. butyricum*, pIP404 from *C. perfringens*, and p19 from *C. carboxidivorans*. Interestingly these results sit in almost complete contrast with what has been observed in the close relative of *E. limosum*, *A. woodii*. Replicons pCD6 pBP1 and pAMβ1 have all been demonstrated as functional in *A. woodii* but showed no discernible efficacy in *E. limosum*[163,164].

With functional replicons determined, experimentation to optimise the transformation protocol was performed by way of electroporation voltage adjustments. Optimisations successfully resulted in a ~5-fold increase in efficiency up to $2.6x10^4$ CFU/μg at a shock voltage of 2.1 kV. At the time of writing this was almost three times that of the highest rate in reported literature for an acetogenic species, 4.5×10^3 CFU/μg of plasmid DNA reported for *A. woodii* by the Dürre lab[164]. Though this rate was clearly sufficient for standard genome engineering with ACE and CRISPR/Cas systems (demonstrated above), higher rates still would be preferable for techniques such as transposon mutagenesis where over one millions unique transformants might be required[72]. Future work might look to improve these efficiencies further through further method optimisation or manipulation of restriction modification systems (Chapter5).

3.3.6.3. Replicon Stability Assessment

Different plasmid replicons can have a range of impacts on transformed cells, whilst performing their role in enabling sustained plasmid replication. Certain properties, such as stability, copy number, and cellular fitness impact, can be exploited for different applications. Two replicons were identified for further investigation based on the contrasting phenotypes observed in *E. limosum* harbouring pCB102 and p19 containing plasmids. During initial transformation testing, colonies harbouring plasmid with the first replicon, pCB102, had a reduced growth rate on plates following transformation (visible at >48 hours) compared to those harbouring pIP404 and p19 vectors (visible 24-48 hours).

The colony growth rate observations encouraged the formal testing of pCB102 and p19 stabilities in *E. limosum*. A stark contrast was observed in the plasmid stability afforded by the two replicons tested (Figure 3.2.8). Results indicated losses of >40% per 24 hours (~15.7 generations) with the pCB102 replicon, and the pCB102 containing plasmid (pMTL83151) had been almost lost from the population entirely after 7 days. In contrast, no statistically significant loss was detected with the p19 replicon over the same time course. Identifying the stability disparity was essential in guiding our selection of either replicon for downstream applications. Instability of the pCB102 replicon was expected and indeed shown to be useful in genome engineering applications such as allele-coupled exchange (ACE), and CRISPR-Cas9 mediated editing. In both applications plasmid loss is essential to permit iterative transformations and plasmid cure cycles, required to perform multiple genomic mutations. Conversely, it was expected that replicon with such stability as observed with p19 would hold value in plasmid-based gene expression applications where loss was not desirable.

Pseudo-suicide replicons are those which have significant enough instability to reduce organism growth rates under selection and are easily lost from populations in the absence of such selective pressure. Such characteristics are well documented for pCB102 in other clostridiales and are central to allelic exchange based genome

engineering[98]. The slowed growth rate as a result of cellular burdening by the pseudo-suicide plasmid is required to select for single cross-over events (i.e., genomic integration) during homologous recombination. It is subsequently valuable in ensuring plasmid loss following successful genome modification regardless of the underlying engineering method. Replicon pCB102 has demonstrated a broad host range across clostridial species, however, its mechanism of replication remains unclear[119].

Plasmid-based expression offers the benefit of increased gene copy number where high gene expression is desirable. Multi-copy existence was confirmed for p19 in *E. limosum* by Millard, J. (SBRC Nottingham, Thesis 2017), where >20 copies and <10 copies of pCB102 and p19 containing plasmids respectively were reported in *E. limosum* at mid-stationary growth phase. Combining this data with the observed stability presented above, it would appear that in the case of these two replicons, greater plasmid retention does not correlate with higher intracellular copy number of the plasmid. This is contrary to what could be expected in terms of the statistical likelihood of plasmid loss occurring during cellular partitioning, at least when using random-distribution or non-regulated models for vertical plasmid transmission during cellular partitioning[78]. This is suggestive that other factors may be involved in the preferential retention of p19 containing plasmids.

Although there are no obvious canonical active partitioning systems present within either replicon, it is possible that the observed retention of p19 is a result of complementary interaction between currently unknown replicon features and chromosomally-encoded proteins. Such interactions can impose subcellular localisation of plasmids during cellular division to ensure plasmid retention in both daughter cells[78]. It is worth noting that the data from James Millard indicates that the copy numbers of the two plasmids may be closely aligned during the exponential growth phase. This means that the possible benefits of increased copy number on plasmid retention during cell division are not relevant to pCB102 during the time where cells are most actively dividing. Further investigation is required to elucidate the mechanisms of both

replicons if we are to fully understand the mechanisms for their stability. However, what is clear based on the growth analysis data, is that harbouring pCB102 based plasmids has the more negative impact on cellular fitness. For this reason, observed loss of plasmid from pCB102 harbouring populations would be expected at least in part due to the greater fitness of the wild type cells, which enables the faster-dividing plasmid-less cells to dominate over numerous generations.

Plasmid based expression systems are useful platforms for high- or over-expression systems as a result of increased gene copy-number and resultant increased protein expression. In comparison, equivalent chromosomally integrated genes which would occur at single-copy per cell. Leveraging p19 in construction of plasmid-based expression systems could alleviate the necessity for time consuming genomic integration steps during the engineering process. However, it is important to note that the reported stabilities in this study only pertain to the p19 replicon on an otherwise empty pMTL8J151 vector. If the impressive stability of p19-plasmids translates to larger or potentially deleterious payloads (chapter 4: Acetone & Isopropanol production cassettes), then it could alleviate the requirement for sustained selective pressure to ensure plasmid maintenance. This would provide an overexpression system which could be applied at scale without the need for costly antibiotic selection to ensure plasmid maintenance, or indeed the any required engineering of a more cost-effective retention means such as engineered auxotrophies.

3.3.7. Genome Engineering Proof of Application

Two genome engineering applications described above validate the success of the initial work in the production of a foundation for advanced genetic research in *E. limosum*. First ACE, a versatile method for knock-in and knock-out based on the creation and subsequent leveraging of auxotrophy phenotypes for selection of double-crossover homologous recombination events. Second CRISPR-Cas9 mediated genome editing, a three-part system capable of both knock-out and knock-in, using guiding RNA

molecules (gRNA) to target nuclease activity (Cas9) to target genomic sequences, thereby enforcing counter-selection at single-base resolution for homology directed repair (HDR) events.

Allele-coupled exchange in *E. limosum* has been exemplified in this study through its application to generate a lactose inducible TcdR expression strain, designated i1B. The product of the TcdR gene is a highly specialised RNA polymerase sigma-factor which belongs to a subgroup of such proteins, uniquely expressed by a selection of toxinogenic clostridia[165]. TcdR was utilised in combination with one of the sigma factor's target promoters, in this case P_{TcdB} to generate an orthogonal promoter system by Zhang *et al* (2015)[166]. Control of the TcdR expression, and thereby TcdR- P_{TcdB} interaction, *via*a lactose inducible module derived from *C. perfringens* allows for controlled high-level expression of heterologous genes and operons driven by the P_{TcdB} promoter. Importantly this expression is detached from the standard cellular transcriptional regulation processes due to strict specificity of the heterologous sigma factor. The system has significant benefits to cloning processes prior to plasmid transfer into *E. limosum,* because the expression of potentially toxic genes is mitigated by poor recognition of the P_{TcdB} promoter by the native *E. coli* RNA polymerase. This significantly reduces the fitness impacts otherwise associated with harbouring and expressing toxic cargo in cloning strains. Conventionally, the lowered fitness of cells expressing toxic cargo, drives selection for the inactivation of toxic pathways through SNP accumulation or more significant mutation. One would anticipate that this effect is only exacerbated where high copy number plasmids are employed.

The second genome editing technology demonstrated in this project based on CRISPR/Cas9 system adapted from native adaptive immune system of *S. pyogenes*[167]. This system has a number of significant advantages over previous technologies such as the targeted intron insertion methods ClosTron, and aforementioned ACE based editing. ClosTron and equivalent methods are useful for relatively quick generation of mutants, but with the drawback of potentially unwanted polar effects, lingering

antibiotic resistance marker insertion and is limited to knock-out applications. Conversely, ACE has the versatility to perform both knock-in and knock-out, and can achieve such mutants in a precise manner to maintain genomic fidelity outside of the targeted regions. However, the application of ACE is typically a lengthy procedure due to the necessity for the initial development of a compatible target strain by leveraging a counter-selectable auxotrophy. Furthermore, the subsequent screening for mutants is a relatively laborious procedure due to lack of a strong counterselection against wild type cells. Genome engineering with CRISPR/Cas9, however, is capable of both knock-in or knock-out, in a similar manner to ACE. However, engineering with CRISPR/Cas9 couples the finesse of homologous recombination events to powerful positive selective pressure for mutant genotypes. This selection is driven by persistent targeted genomic cleavage of un-modified target stains which containing active complexes, lethal if unrepaired. This enables relatively fast and precise generation of mutants. The most notable drawbacks in the application of CRISPR/Cas9 systems is the relatively large plasmid sizes, which are due primarily to the incorporation of Cas9 (>4000 bp) and the two homology arms (each 700-1000 bp). These requirements are on top of existing core plasmid features, for example Gram-positive replicon, Gram-negative replicon, and a selection marker[119]. This can result in vector sizes well in excess of 10 kb, where the entire CRISPR/Cas9 system is introduced in a single transformation. Reduced transformation efficiencies associated with increased plasmid size[168], are compounded by the inherent lethality of active CRISPR/Cas9 where native repair mechanisms fail to engage. Collectively this means that high transformation frequencies are often essential to reliably apply this method of genome editing.

CRISPR/Cas9 based genome editing technology represents a logical evolution of the strain development roadmap published by Minton *et al*. (2016)[55]. It improves on previous technologies in both its simplicity to implement as well as the efficiency with which it can generate genomic modifications in a scarless manner. In this study CRISPR/Cas9 based genome editing was used to knockout one of two phosphate

butyryltransferase (*pbt*) genes, designated *pbt1* and *pbt2* in *E. limosum*. The decision to target one of the *pbt* genes (*pbt1*) for the proof of principle application of CRISPR/Cas9 was made primarily for two reasons. First and foremost, it was expected that neither *ptb* individually was likely to be essential due to the presence of the other, a prerequisite for any successful gene knockout. Post knockout, it was anticipated that the remaining *ptb* would deliver sufficient enzymatic activity within the cell that, all be it at an overall reduced rate than where both homologues were present. Although it was acknowledged that these genes could be independently regulated and linked to quite distinct cellular processes. Secondly, the *ptb* genes were identified as genes of interest to advance understanding of the metabolic processes linked to product synthesis by *E. limosum*. The existence of *pbt* genes stood out as unusual in particular due to the non-existence of a butyrate kinase gene (*buk*). A *buk* would be required to convert the forward product of *pbt*, butyryl phosphate, into butyrate. The presence of a butyrate kinase in the *E. limosum* strain KIST612, has been assumed in a metabolic pathway proposal by Jeong *et al*. (2015), however no candidate homologues were reported[35]. Furthermore, no explanation for the mechanics of butyrate in the absence of this gene were proposed. Based on investigative comparison of the *E. limosum* 8486 genome with butyrate producing clostridia, we proposed a *pbt* independent mechanism for butyrate production.

3.4. Conclusion

The ability to successfully implement genome engineering tools is critical for future research efforts and rational strain development processes in *E. limosum*. The successful proof-of-concept application of ACE and CRISPR/Cas9 is evidence of the firm foundation generated in this chapter for *E. limosum* research, specifically genome acquisition, transformation establishment, and vector characterisations. Now that these genome-editing tools have been demonstrated, further advanced genetic research and strain development can be performed in *E. limosum*. Collectively these successes reflect the implementation and logical advancement of the strain development roadmap in *E. limosum*[55]

4. Extending the product spectrum of *E. limosum*

4.1. Introduction

Synthetic biology is a rapidly advancing field of research, which encompasses the design and construction of artificial biological parts, devices, and systems[169]. Arguably the primary source of obstacles to successful implementation of synthetic biology stems from a lack of complete understanding of biology at the most basic level in native contexts. This knowledge gap is due in part to the enormous scale of the problem of understanding how every component in a multi-million-part system operates, and how it interacts with every other component. This problem is compounded in many cases by a lack of appropriate investigative tools to fill the knowledge gaps fast enough to meet the ever-increasing demand to develop and implement ideas. In spite of this, progress in the field continues to demonstrate the promise of applied synthetic biology, with applications in the fields of biofuels, biomaterials, therapeutics, and biosensing[170]. Such applications are assisted by an ever-improving knowledge baseline and new technologies which enable the improved characterisation and design of biological systems, in particular in silico modelling. Microbes are a primary target for applied synthetic biology, particularly with a view towards rational strain engineering.

The broad diversity which exists in microbial life provides an extensive range of naturally evolved phenotypes which can be harnessed for industrial processes. A key area of industry in this regard is the production of renewable fuels and chemicals, and the recycling of waste materials. Such processes can engage the natural capacity of organisms to metabolise target feedstocks and convert them into useful products. Synthetic biology offers a potential avenue towards extending the range of said products beyond those which are naturally produced by the microorganisms, to enable the production of more interesting and commercially valuable products.

4.1.1. Isopropanol and Acetone

Isopropanol is a secondary alcohol which is naturally produced by select microorganisms[171]. Isopropanol is a versatile solvent with a wide range of uses, including cleaning applications, as an additive or replacement for existing petrochemical fuels, and as a precursor to polypropylene[172]. As a fuel bio-isopropanol represents an attractive alternative to bio-ethanol, due to the numerous compatibility issues associated with use of ethanol with existing petrochemical focused infrastructure. Specifically, ethanol use is hampered by the solvent's natural corrosive nature towards ferrous metals, incompatibility with existing fuel pipelines, and a relatively low energy content[173]. The primary sources of isopropanol are currently through the chemically derivation of acetone or propylene, which in turn are primarily sourced from fossil fuels.

Acetone is a water miscible organic compound which is also a product of metabolism of a number of microorganisms[174]. In addition to natural capacity as a solvent, the primary industrial use of acetone is in the production of precursors to polycarbonate plastics[175]. Currently acetone is produced primarily *via*a fossil fuel driven industrial process known as the Cumene process, coupled to a concomitant production of phenol[176]. With production volumes of 7 million tonnes per year and an annual market growth at >3%[177], acetone represents an attractive target chemical for production from more sustainable feedstocks.

Engineered acetone and isopropanol production has been demonstrated in a number of different microorganisms, including *A. woodii*[163], *Clostridium acetobutylicum*[144], *Escherichia coli*[115], and the cyanobacteria *Synechococcus elongatus*[178]. In general, the artificially generated routes to isopropanol in these studies have remained consistent, employing four enzyme groups to catalyse four distinct reactions to convert acetyl-CoA into isopropanol. First, an acetyl-CoA transferase; second, an acetoacetyl-CoA transferase; third, an acetoacetate decarboxylase; fourth, a secondary alcohol dehydrogenase (Figure 4.2 – page 167). The artificial introduction of each of these

genes was not essential where native homologues were already present, as is the case for in *C. acetobutylicum* for the first three reactions. This study describes the successful introduction and initial optimisation of isopropanol production in *E. limosum*, using readily transferrable plasmid-based systems for gene expression. Efficacy of the system and ease of transferability is also briefly demonstrated in establishing isopropanol in the acetogen *C. autoethanogenum*.

4.1.2. Pathway scaffolding

Enzyme scaffolding is a naturally evolved mechanism used by microbes to enhance the efficiency and productivity of reaction pathways, most notably in the context of cellulosomes. Cellulosomes are multienzyme complexes, which are employed by numerous microbial species to break down the complex compositions of cellulose, hemicellulose, and lignin in plant biomass[179]. Cellulosomes are broadly composed of two parts. First, a single high molecular weight scaffold protein (a scaffoldin), which is comprised of a number of distinct dockerin domains, linked together by short linker regions. Second, a collection of cohesin-domain-containing enzymes, each of which binds specifically to a corresponding dockerin. The result is a modularly assembled complex made from multiple unique interactions between enzymes and a scaffold, where spatial coordination of the enzymes and their activities aids in the efficient hydrolysis of complex target substrates (Figure 4.1).

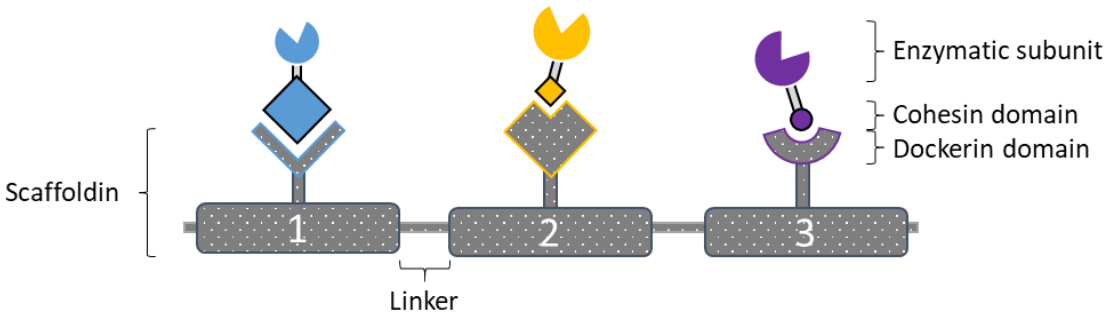


Figure 4.1 – Example dockerin scaffoldin schematic

Three enzymes (Ɑ) interacting *via* dockerin domains with an artificial three-part scaffoldin protein (grey spots) harbouring a single cohesin domain for each dockerin.

Adaption of naturally occurring cellulosomes has been used in the context of developing 'designer' cellulosomes, as an elegant solution to traditional enzyme cocktails often used in traditional lignocellulosic degradation processes[180]. Additionally, the synthetic design and production of mini-cellulosomes in *C. acetobutylicum*, where the self-assembly of these artificial constructs was effectively demonstrated[100]. The principle of enzyme scaffolding may be expandable beyond simply cellulosomes, and might represent a valuable tool in the arsenal of synthetic biology research, as a means of enforcing the co-localisation of proteins. In principle, co-localisation of otherwise stochastically distributed proteins can have a number of benefits, including improvements to the efficiencies of reaction and diffusion processes through enhanced substrate channelling in reaction cascades[181]. Enhanced cascade catalysis with synthetic scaffolds has been demonstrated in a number of microbial systems, including for improved glucaric acid and mevalonate titres in *E. coli*[182,183], and production of ethylene and 2,3-butandiol production in cyanobacteria and yeast respectively[184,185]. This study offers insight as to the impacts of scaffolding three enzymes (CtfAB, Adc, Sadh) involved in the artificially instigated production of isopropanol from *E. limosum* and *C. autoethanogenum*.

4.1.3. Aims of this Study

The aim of this study was to engineer *E. limosum* ATCC 8486 to produce an extended range of metabolic products. Acetone and isopropanol were selected as industrially relevant targets for this proof of concept study. Extension of the fermentation product profile was demonstrated through the addition of plasmid borne gene cassettes, operons of heterologous genes encoding metabolic pathways towards the desired products. Growth profile analysis was performed primarily in closed batch fermentations with fructose as the primary growth substrate.

4.2. Results

4.2.1. Development of acetone and isopropanol production cassettes

A number of different gene cassettes were created throughout the course of this project, through the iterative development of constructs pertaining to acetone and isopropanol production in *E. limosum*. Three distinct generations of production cassettes were constructed based on the underlying synthetic pathway for isopropanol production in *E. coli* published by Hanai *et al*. (2007)[115]. Each cassette consisted of a single promoter driving an operon containing a selection of four heterologous genes, encoding enzymes related to the conversion of acetyl-CoA and metabolic intermediates along a pathway to acetone and isopropanol (Figure 4.2). The initial three steps of the pathway from acetyl-CoA to acetone catalysed by an acetyl-CoA transferase, an acetoacetyl-CoA transferase, and acetoacetate decarboxylase. In the context of the developed gene cassettes, homologues for each of these genes was derived from the solventogenic clostridium *C. acetobutylicum*, *thl*, *ctfAB*, and *adc*. For the final pathway step, between acetone and isopropanol, a secondary alcohol dehydrogenase from the *C. beijerinckii* NRRL B-593 strain was employed, *sadh*. Generational variation between the developed gene cassettes occurred with the promoter employed to drive expression of the operon, as well as with the absence or inclusion of the acetyl-CoA acetyltransferase (Figure 4.3). A side development of the generation-I constructs, generation-IS, employed a novel concept of artificially co-locating the pathway's enzymes by binding them to a purpose-built scaffold protein.

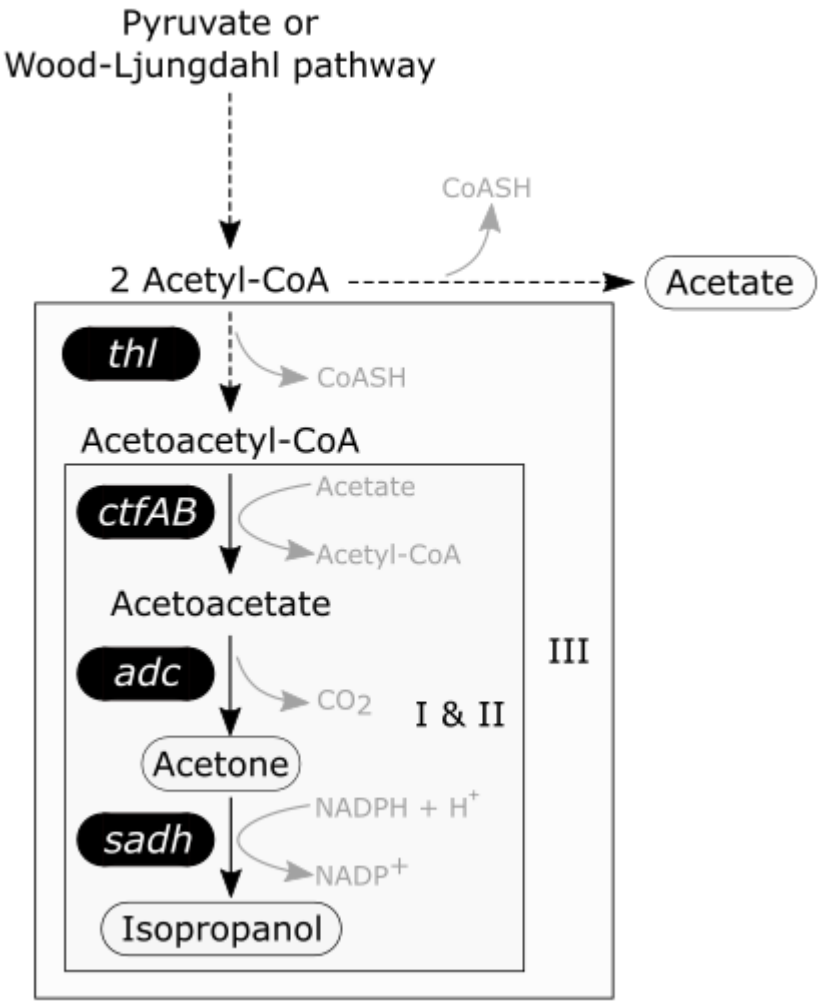


Figure 4.2 – Synthetic metabolic pathway for isopropanol production in *E. limosum*

Schematic based on publication by Hanai *et al*. (2007) demonstrating proof of concept isopropanol production in *E. coli*[115]. Dashed line arrows represent reaction pathways for which genomic analysis predicts that genes already exist in *E. limosum* ATCC 8486. Solid line arrows represent the new pathways to be added as a result of hetrologous gene expression. Acetyl-CoA can be generated through the oxidative decarboxylation of pyruvate or from Wood-Ljungdahl metabolism of C1 substrates. The enzymes responsible for catalysing four distinct enzymatically catalysed reactions for conversion of acetyl-coA to isopropanol are marked (black bubbles). Acetyl-CoA acetyltransferase, E.C. 2.3.1.9, catalysed by the natively occurring in *E. limosum* thiolase *(thl)* gene (Ga0111251_113833), or by a heterologous *thl* derived from *C. acetobutylicum* ATCC 824; CoA transferase, E.C. 2.8.3.9, activity provided by the *C. acetobutylicum* ATCC 824 *ctfAB* enzyme complex; Acetoacetate decarboxylase, E.C. 4.1.1.4, activity provided by the *adc* enzyme also of *C. acetobutylicum* ATCC 824; Secondary alcohol dehydrogenase, E.C. 1.1.1.80, activity provided by the *sadh* gene of *C. beijerinckii* NRRL B-593. Boxes marked 'I & II' and 'III' refer to the selection of genes contained by three iterations (Generation I-III) of isopropanol operon development. Generation I and II contained just the latter three genes, relying on the native *E. limosum thl* to catalyse the first reaction. Generation III was an extension of the Generation II, with the addition of a heterologous *thl* gene. Acetone producing operons created for each generation were identical to the isopropanol version, other than the absence of the *sadh* gene.

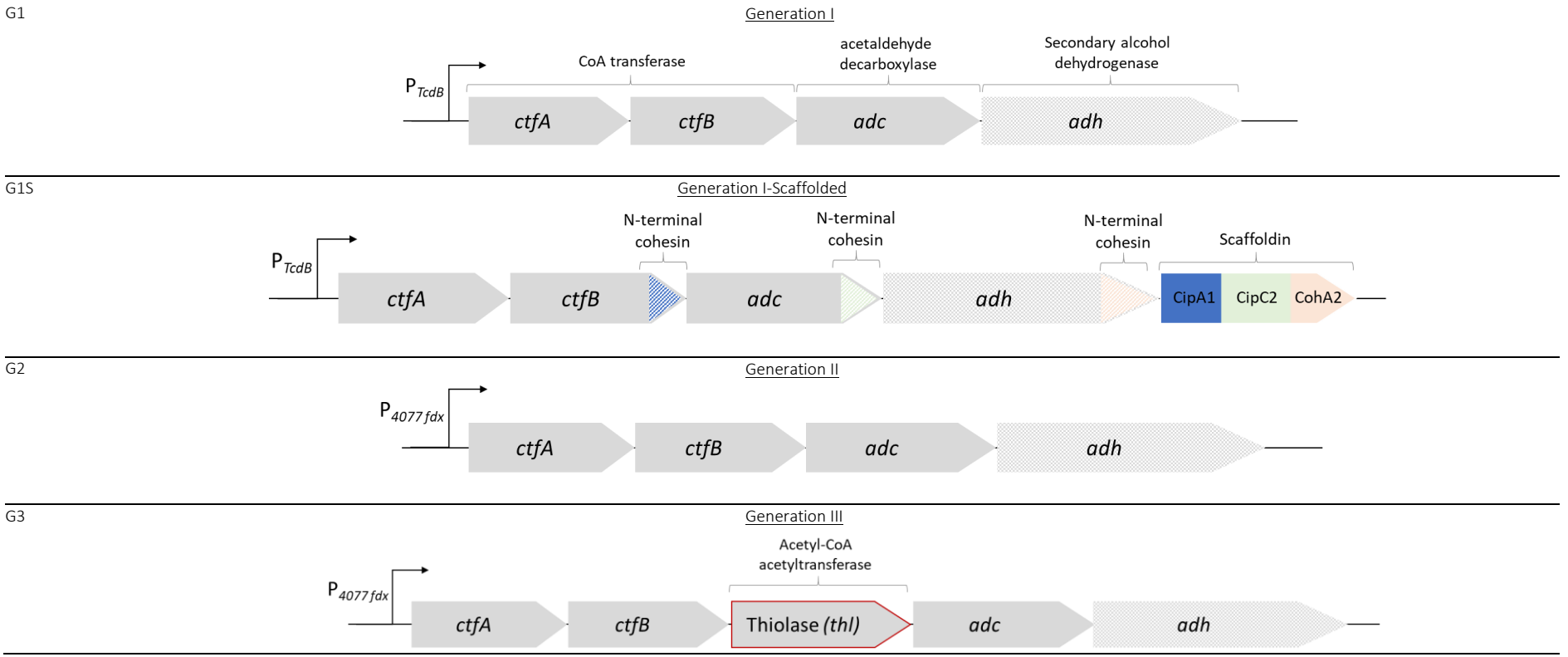


Figure 4.3 – Schematics of Acetone and Isopropanol production operon cassettes

Three generations of gene operons created during the development of plasmid driven acetone and isopropanol production in *E. limosum* ATCC 8486. Two operons were produced within each generation, one targeted to acetone production, the other to isopropanol. This figure depicts the isopropanol producing operons only, however these differ from the acetone operons only by the inclusion of secondary alcohol dehydrogenase (*adh* – transparent grey). Expression of Generation I (G1) and Generation I-Scaffolded (G1S) constructs is controlled by the *C. difficile* derived *tcdB* promoter (P_{TcdB}). G1s constructs are modified versions of their respective G1 counterparts, where a 5' sequence addition to the ORF of each gene (except *ctfA*) encodes for a unique N-terminal cohesin domain, which is complementary to a co-expressed artificial scaffoldin protein. G1s scaffold: CipA1-CipC2-CohA2. Generation II (G2) operons differ from G1 only in the promoter used, an exchange of the P_{TcdB} to the *C. perfringens* ferredoxin derived promoter, P_{4077}. Generation III (G3) constructs are extensions of the respective G2 operons, where a thiolase gene *(thl)* has been added to the pathway downstream of *ctfB*. Genes *ctfA* and *ctfB* (*ctfAB*), and *adc* from *C. acetobutylicum* ATCC 824 provide CoA transferase activity (E.C. 2.8.3.9) and acetacetate decarboxylase activity (E.C. 4.1.1.4) respectively. Secondary alcohol dehydrogenase activity is provided by the *C. beijerinckii* NRRL B593 derived adh. Acetyl-CoA acetyltransferase activity is conferred by the thiolase gene, thl, also derived from *C. acetobutylicum* ATCC 824.

4.2.2. Construct features

The development of the constructs centred around ease of cloning and downstream versatility. Constructs were broadly grounded in the standardised pMTL80000, four-quadrant, format. Figure 4.4 offers a labelled plasmid map for the G1 isopropanol construct pMTL8J15-iso-G1. This plasmid is characteristic in general structure of all of the finalised expression vectors used for acetone and isopropanol production. Firstly, the Gram-positive replicon p19 was selected based on the strength of its retention in *E. limosum* as described in Chapter 3 (Figure 3.2.8), as well as for its versatility in other organisms[136]. Second, the chloramphenicol selective marker was selected based on its efficacy as a selective marker in *E. limosum*, and again, its versatility across numerous industrially relevant organisms for which the plasmid could be transferred. Third, the high copy Gram-negative replicon ColE1 and the conjugative element, *traJ*. Fourth, the gene cassette, containing a promoter driving an operon of genes for acetone and isopropanol production, insulated from transcriptional readthrough in the rest of the plasmid by two flanking transcriptional terminators.

4.2.3. Generation I & IS

The first generation of expression cassettes (Figure 4.3: G1A, G1B) were built around expression of the genes *ctfA*, *ctfB*, *adc*, and adh, driven by the P_{TcdB} promoter from *C. difficile*. Alongside the basic G1 constructs, a second set of vectors was developed, the generation-I scaffolded (G1S) constructs (Figure 4.3: G1SA, G1SB). The scaffolded constructs contained the same promoter and genes as G1, but each gene was modified at the 3' end to add a dockerin domain, derived from naturally occurring cellulosomes, to the encoded protein. This scaffoldin domain enabled the modified proteins to bind to a co-expressed scaffoldin protein, also derived from cellulosomes, in order to hold enzymes in close proximity to one another within the cell. Activation of the P_{TcdB} promoter is mediated by a specialised class of RNA polymerase sigma-factor, TcdR. The *tcdR* gene is conditionally expressed in response to lactose in the *E. limosum* ATCC 8486 mutant, i1B, development of which is discussed in Chapter 3. G1 and G1S constructs were transformed into i1B alongside an empty pMTL8J151 vector as a negative control for acetone and isopropanol production.

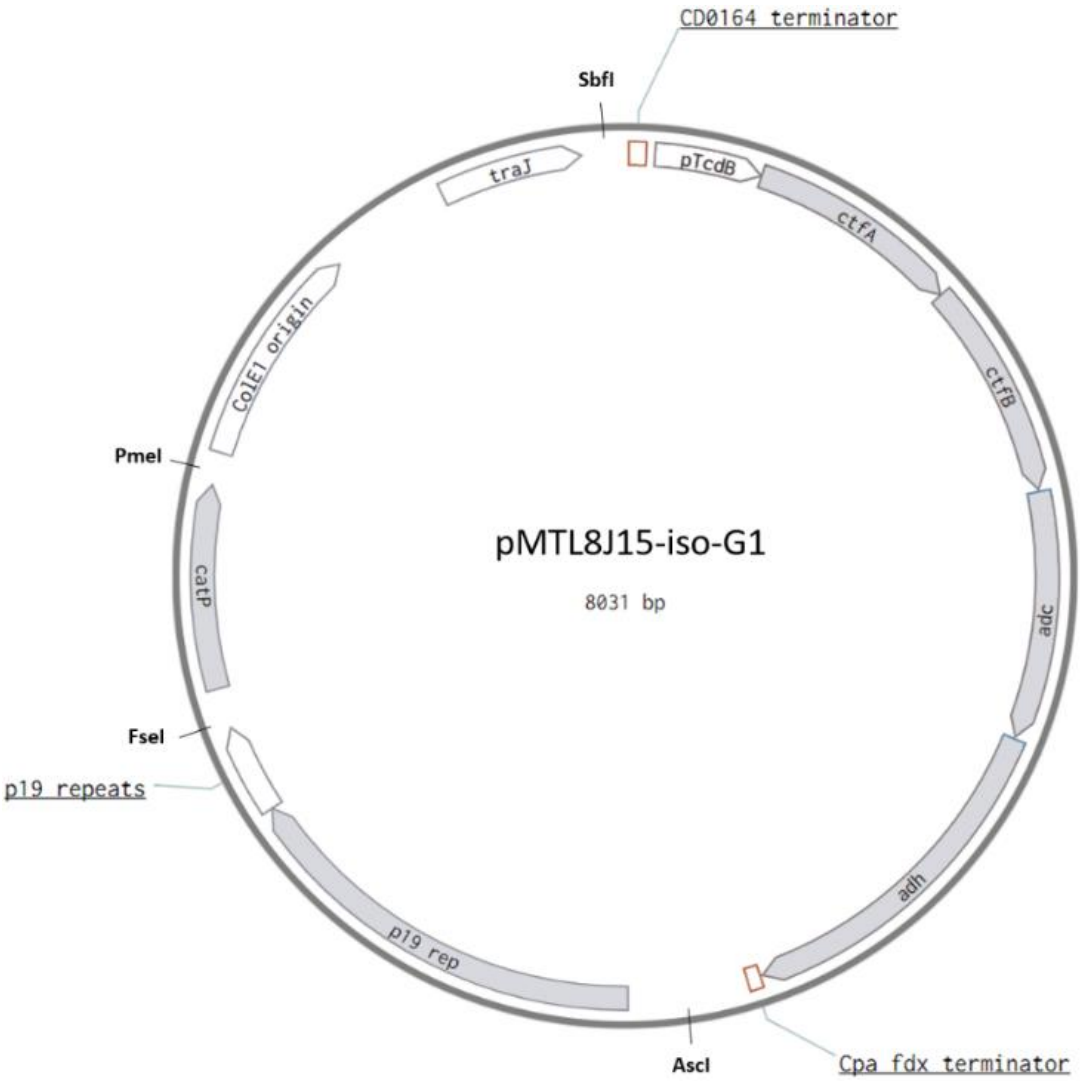


Figure 4.4 – Vector map of plasmid pMTL8J15-iso-G1

The basic format of this vector is representative of all of the tested constructs for acetone and isopropanol production in this study. Conforming to the underlying pMTL80000 modular vector series format[119], vectors are split into four discrete regions separated by unique restriction sites. Clockwise from the SbfI restriction site: variable region (to AscI); Gram-positive replicon (to FseI); selective marker (to PmeI); Gram-negative replicon (to SbfI). For all the production cassettes developed in this study, the variable region was defined by the chosen promoter (P_{TcdB} or P_{4077}) and gene pathways as described in Figure 4.3, while the remaining vector remains consistent between all tested vectors. For pMTL8J15-iso-G1 this is a P_{TcdB} promoter driving the expression of *ctfA, ctfB, adc*, and *adh*. Variable regions are insulated from the remaining plasmid by two flanking terminators, CD0164 and Cpa_fdx. Clockwise from the variable region: p19 rep, a Gram-positive replicon derived from a native plasmid in *C. carboxidivorans* P7; p19 repeats, a non-coding region of sequence repeats required for replicon functionality; *catP*, Chloramphenicol acetyltransferase gene conferring resistance to chloramphenicol; ColE1 origin, a Gram-negative replicon; traJ, a regulator or conjugation transfer.

Fermentation analysis was performed on biological triplicate batch cultures for each i1B transformant strain. Fermentations were carried out in AWM1 media with fructose as the primary substrate, under thiamphenicol selective pressure, and lactose was added to each sample following the first timepoint measuring OD_{660} >0.2. Sample supernatants for product profiling, and OD_{660} readings as a measure of biomass, were taken once every 24 hours for 5 days, followed by a final time point at 220 hours. Figure 4.5 presents the biomass and GC measured profile changes over the 220 hour time course of this growth analysis for each sample. Endpoint specific analysis for this dataset is available in Figure 4.6, and a value summary including estimated mol/mol yields can be found in Table 4.1.

GC analysis of the samples indicated no significant increases (>0.1mM) in the acetone and isopropanol titres of the G1 and G1S samples against the control at any timepoint across the 220 hours. The reported maximum concentrations for all of the tested samples (including the negative control) ranged from 0.02-0.05mM and 0.39-0.48mM for acetone and isopropanol respectively. Independent but comparable analysis of untransformed wild-type *E. limosum* produced comparable GC profiles. However, no acetone or isopropanol was detected by HPLC-UV/IR in wild type *E. limosum*, nor has it been reported in the published literature.

End point analysis (Figure 4.6/ Table 4.1) was selected for statistical analysis due to large intrasample variation in the early time points for a number of the constructs. Although the transformed G1 and G1S plasmids did not result in significant levels of acetone or isopropanol being produced, the profile of the other metabolites was in some cases slightly impacted where production cassettes were present. For the most part however, these impacts were not statistically significant to a confidence interval (CI) of >95% ($P < 0.05$), when evaluated by students t-test. Variation against the control sample with a CI >95% were observed most notably in butyrate titres. G1-ace, G1S-ace, and G1S-iso all demonstrated titres which were statistically different to the control ($P < 0.05$). G1-ace demonstrated a >1mM increase in butyrate, while both G1S constructs caused a >1mM decrease, when compared to the empty vector control. Acetate and ethanol readings were statistically consistent with the control ($P > 0.05$) in all samples except for G1S-iso, where an increase of >3mM acetate ($p < 0.02$) and a

drop of >1mM ethanol ($p < 0.015$) was observed. Ultimately, the G1 and G1S constructs were not effective in establishing detectable acetone and isopropanol in *E. limosum* under the conditions tested, and the profile changes which were observed in some of the other analytes require further experimentation to improve their statistical power and confidence in their validity.

The initial strategy for testing and development of isopropanol production constructs was to perform testing primarily on MeOH based media. However, this position was revaluated following preliminary data indicating a total absence of isopropanol with the G1 constructs on both MeOH and fructose. Figure 4.5 and Table 4.1 includes comprehensive analysis of the first-generation acetone and isopropanol constructs on fructose-based media.

Insignificant acetone and isopropanol titres in the G1 and G1S constructs indicated that multiple rounds of construct optimisation, analysis, and testing would likely be required, in order to first establish and then begin to optimise isopropanol production. The timeframe for testing any given construct in *E. limosum* is approximately two to three times longer for MeOH based growth when compared to fructose based, from the time of initial transformation. This is primarily a result of the reduced growth rates and significant lag phases which are observed where the C1 substrate is employed. It was therefore decided that the initial development and testing should centre around fructose-based media in an effort to speed up the process. Comparable systems have been successfully implemented to introduce acetone production to *A. woodii* strains in continuous fermentations of syngas[163]. In light of these successes with C1 in similar organisms, higher-throughput fructose-based fermentations were deemed an acceptable testing ground for gene operons in *E. limosum*. This was primarily in the interest of speeding up initial testing and development processes, proceeding with reasonable confidence that the constructs would also work with C1 substrates when appropriate.

Figure 4.5 legend (Page. 174): Samples were grown in batch culture on AWM1 media supplimented with 20mM fructose as the substrate and a 10 µg/ml working concentration of thiamphenicol. OD_{660} readings and samples were taken once daily for five days, followed by a final time point at 220 hours. A) i1B harbouring an empty pMTL8J151 plasmid as a baseline control for comparison; B) A strain expressing the acetone expression cassette from plasmid pMTL8J15_ace (G1); C) A strain expressing the isopropanol expression cassette from plasmid pMTL8J15_iso (G1); D) A strain expressing pMTL8J15_ace_DFS (G1S), a modified version of the acetone expression cassette, where the functional proteins were modified to contain a fused dockerin domain in order to enable the non-covalent attachment of each enzyme a coexpressed scaffoldin protein. E) A strain expressing pMTL8J15_iso_DFS (G1S), a modified version of the isopropanol expression cassette, where the functional proteins were modified to contain a fused dockerin domain in order to enable the non-covalent attachment of each enzyme a coexpressed scaffoldin protein. No statistically relevant levels of isopropanol or acetone were observed in any of the tested samples by comparison to the pMTL8J151 baseline. Data shown is the average values from biological triplicate samples, where error bars represent the standard error of the mean (SEM). Lag phase variability within triplicate repeats was accounted for where necessary by normalisation of the 24-hour timepoint for each sample, adjusted to the first reading of effective growth (OD_{660} 0.2 and above). Lactose was added to each sample to a concentration of 20mM at this normalised 24-hour timepoint.

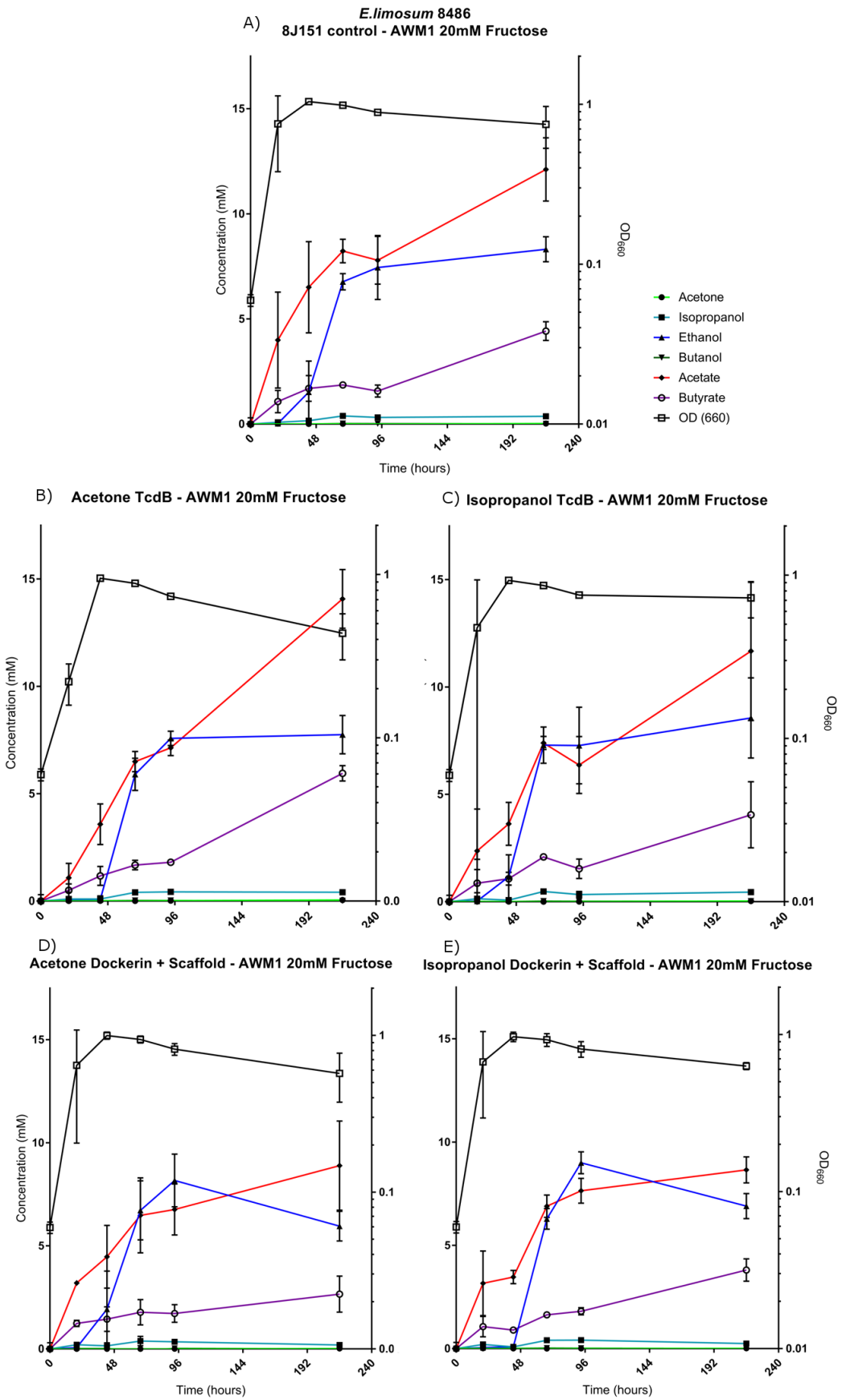


Figure 4.5 – Comprehensive growth curve and product profile data for five transformed variants of the *E. limosum* ATCC 8486 mutant i1B

Table 4.1 – Endpoint dataset for the testing of the first-generation Acetone and Isopropanol cassettes in five transformed variants of the *E. limosum* ATCC 8486 mutant i1B

Sample	8J151	8J151-Acetone G1	8J151-Isopropanol G1	8J151-Acetone G1S	8J151-Isopropanol G1S
Growth					
Max OD_{660}	1.030 (±0.17)	0.949 (±0.18)	0.929 (±0.08)	0.997 (±0.22)	0.967 (±0.25)
220hr OD_{660}	0.750 (±0.22)	0.437 (±0.14)	0.726 (±0.18)	0.572 (±0.20)	0.629 (±0.02)
Product conc. [mM]					
Acetone	0.03 (±0.02)	0.05 (±0.01)	0.03 (±0.003)	0.02 (±0.01)	0.01 (±0.01)
Isopropanol	0.37 (±0.02)	0.41 (±0.01)	0.44 (±0.11)	0.19 (±0.06)	0.24 (±0.05)
Ethanol	8.31 (±0.59)	7.74 (±0.89)	8.56 (±1.86)	5.95 (±0.75)	6.90 (±0.06)
Acetate	12.11 (±1.50)	14.07 (±1.37)	11.67 (±3.24)	8.89 (±2.17)	8.65 (±0.63)
Butyrate	4.43 (±0.44)	5.95 (±0.35)	4.05 (±1.54)	2.66 (±0.87)	3.08 (±0.54)
Calculated Yield [M/M]					
Acetone	0.002	0.003	0.002	0.001	0.001
Isopropanol	0.019	0.021	0.022	0.010	0.012
Ethanol	0.416	0.387	0.428	0.298	0.345
Acetate	0.606	0.704	0.584	0.445	0.433
Butyrate	0.222	0.298	0.203	0.133	0.154

Product yields were calculated by GC for acetone, isopropanol, ethanol, acetate, and butyrate. OD_{660} values were collected manually with each sample for biomass quantification. Data shown is based on the average of 220-hour biological triplicate samples (± standard error of mean), grown on AWM1 with 20mM fructose and 10 µg/ml thiamphenicol in closed batch fermentations. Calculated yields are based on conversion of the initial 40 mM fructose substrate.

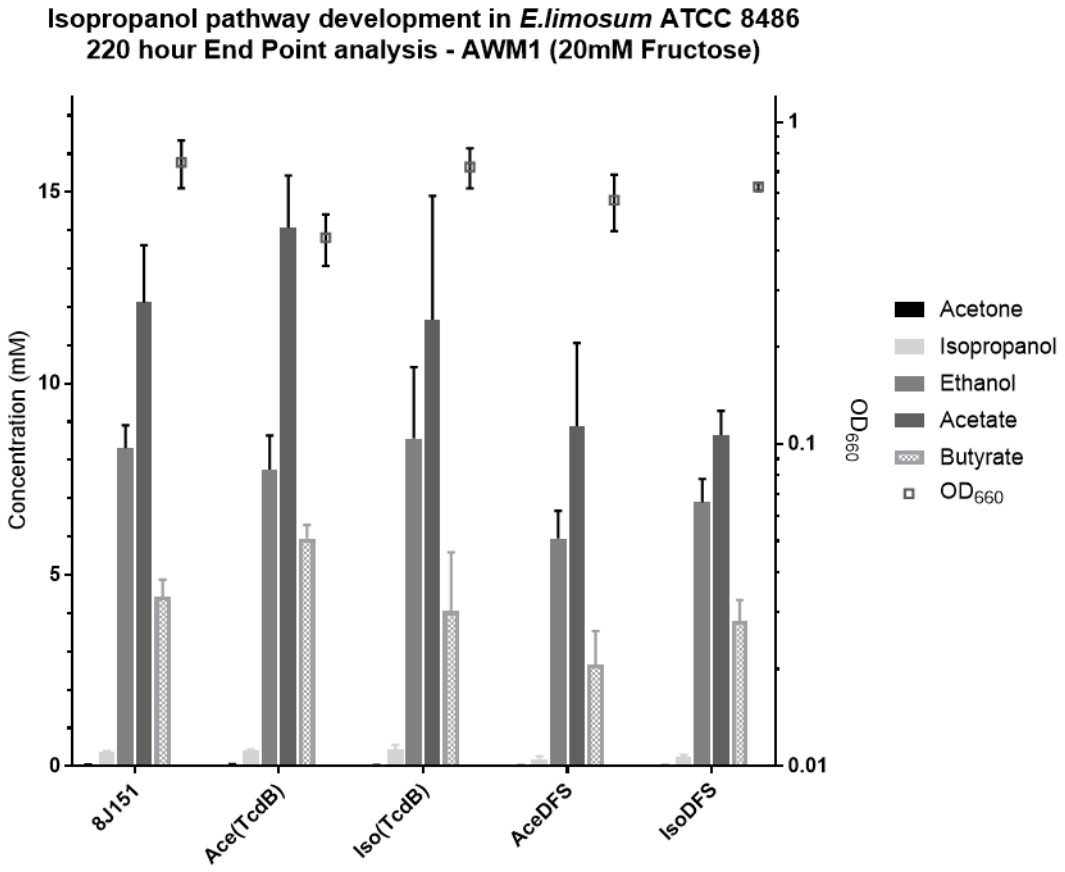


Figure 4.6 – Endpoint data for Generation I and Generation I-Scaffolded, Acetone and Isopropanol cassettes in five transformed variants of the *E. limosum* ATCC 8486 mutant i1B

Samples grown for 220 hours on AWM1 with 20mM fructose and 10 µg/ml thiamphenicol in closed batch fermentations. Bars represent the product yields for acetone, isopropanol, ethanol, acetate, butyrate in mM. Square points represent the biomass of samples as measured by OD_{660}. Data presented is based on biological triplicate samples, given as mean (± standard error of mean). Control: *E. limosum* ATCC 8486-i1B harbouring an empty pMTL8J151 plasmid as a baseline control for comparison; Ace-G1: A strain expressing the acetone expression cassette from plasmid pMTL8J15_ace Generation I (G1); Iso-G1: A strain expressing the isopropanol expression cassette from plasmid pMTL8J15_iso G1; Ace-G1S: A strain expressing pMTL8J15_ace_DFS Generation I-Scaffolded (G1S), a modified version of the acetone expression cassette, where the functional proteins were modified to contain a fused dockerin domain in order to enable the non-covalent attachment of each enzyme a coexpressed scaffoldin protein. Iso-G1S A strain expressing pMTL8J15_iso_DFS G1S, a modified version of the isopropanol expression cassette, where the functional proteins were modified to contain a fused dockerin domain in order to enable the non-covalent attachment of each enzyme a coexpressed scaffoldin protein.

4.2.4. GC method re-evaluation

The reported low-level isopropanol production (~0.4Mm) which was observed even in the control 8J151 control samples, conflicting with literature and other analytical methods, encouraged a brief reassessment of the GC methodology in order to gauge appropriate thresholds for limits of detection (LoD) of the analytes.

Analysis of the GC outputs for media, samples, spiked samples, and standards, indicated that isopropanol detection at sub 0.5mM concentrations was unreliable using the employed GC method. The primary issue, a slight overlap in the retention times of isopropanol (2.252 min) and ethanol (2.252 min), is highlighted in Figure 4.7. LoDs were defined in line with ICH recommendations, based on a combined formula for 'Standard Deviation of the Response and the Slope', and 'Signal-to-Noise'[186,187].

Combined formula:

$$LoD = \frac{(3 \cdot S_a)}{b} + 3 \cdot S_b$$

Equation 4 – Combined formula for limit of detection calculations of gas chromatography measured analytes

S_a is the standard deviation of the regression, calculated using triplicate samples from a standard calibration curve, over the low concentration range of 0–1.0 mM. b is the triplicate average gradient for the full calibration curve. S_b is the average standard deviation for triplicate blank samples.

In order to account for peak overlap issues, the more stringent threshold of $10 \cdot S_a$, normally reserved for Limit of Quantification (LoQ) calculations, was applied to the isopropanol and ethanol LoD calculations in place of $3 \cdot S_a$. The final LoD values calculated for GC analysis in this study are presented in Table 4.2.

Table 4.2 – Calculated LoD values for GC analysis

Analyte	Acetone	Isopropanol	Ethanol	Butanol	Acetate	Butyrate
LoD (mM)	0.06	0.56	0.41	0.36	0.27	0.10

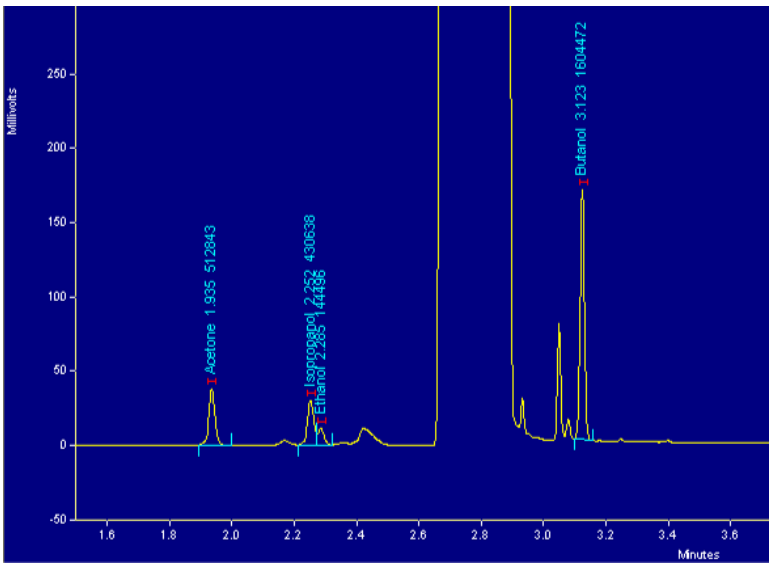


Figure 4.7 – Example of raw GC chromatagraph output showing ethanol-isopropanol peak overlap

Graph shows a representative peak profile from a standard sample made up to 10mM of each Acetone (1.935 min), Isopropanol (2.252 min), Ethanol (2.285 min), Butanol (3.123 min), Acetate (not shown), Butyrate (not shown) made up in AWM2 media and prepared using the typical GC sample preparation method used in this study (Method 2.5.7). Slight overlap of isopropanol and ethanol peaks is apparent between 2.2-2.4 minutes. Quantifications of analyte concentrations are made based on integrated peak areas as defined by automated peak calling by the GC software, ChromQuest, represented graphically by cyan tick marked line boarders.

4.3. Generation 2 & 3

The root cause of low acetone and isopropanol titres in the G1 constructs could not be conclusively deduced from the available data, however a number of possibilities causes were considered. The three primary areas of focus for the root cause of pathway ineffectiveness were as follows: first, the possibility that the intended pathways may have been non-functional because one or more of the desired reaction steps was thermodynamically unfavourable under the conditions of *E. limosum* growth; second, the possibility that there may have been insufficient enzyme expression of the heterologous genes in the cell; Third,

the possibility that upstream bottlenecks in the pathway precluded metabolic flux towards acetone and isopropanol.

The first focus area was addressed through brief evaluation of literature available on the selected enzymes, as well as assessment of the most closely comparable research demonstrating success, namely acetone production from *A. woodii*[163]. The findings of these searches reinstalled confidence that the pathway was at least in theory viable in *E. limosum*, and that it would be possible to overcome the failures associated with G1 constructs through appropriate genetic modifications.

The second possibility, lack of enzyme expression, was assessed by investigating the strength of the P_{TcdB} promoter utilised in the G1 constructs. This was achieved through the production of a promoter library for *E. limosum*, based on the plasmid borne expression of a reporter protein, *catP*, under the control of various promoters. Figure 4.8 presents the results of this promoter library.

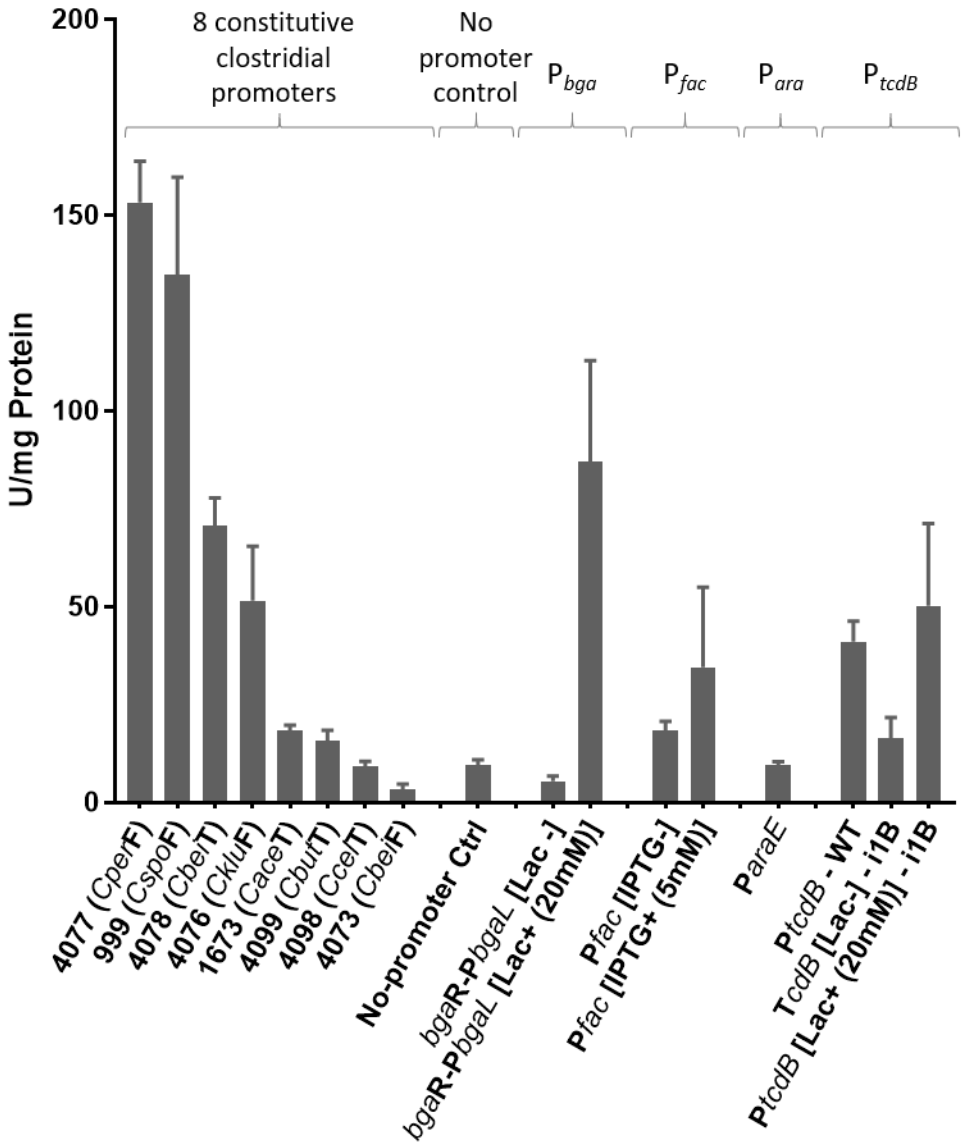


Figure 4.8 – Promoter strength library for mid-exponential growth phase *E. limosum* ATCC 8486

Enzymatic Assay of Chloramphenicol Acetyltransferase (CAT) as a measure of c*atP* protein levels within samples. A total of seventeen distinct conditions were tested, from a total of thirteen promoter systems. All systems were tested as plasmid borne c*atP* expression cassettes using a pMTL8J15x backbone, where 'x' constitutes the variable region consisting of a defined promoter driving expression of the c*atP* gene. From left to right: eight constitutive promoters derived from the ferredoxin or thiolase genes of different clostridial species; a promoterless c*atP* construct functioning as a negative control; a lactose inducible promoter system (*bgaR*–P_{bgaL}) derived from *C. perfringens*, with and without 20mM lactose induction; a synthetic IPTG inducible promoter system (P_{fac}) based on the *on the promoter of the C. pasteurianum* ferredoxin gene modified to include a synthetic lac operator region, with and without 5mM IPTG induction; the constitutive P_{araE} promoter from *C. acetobutylicum*; the TcdB promoter from *C. difficile*, tested in both WT *E. limosum* and in a *E. limosum* mutant (i1B) which conditionally expresses the heterologous TcdR transcription factor, under the control of the *bgaR*–P_{bgaL} system, with and without 20mM lactose induction. Data shown is the mean value of nine replicates (three biological samples assayed in triplicate), where error bars represent the standard error of the mean (SEM).

The expression levels of *cat*P from thirteen promoter systems and seventeen distinct conditions were tested. Reported expression levels ranged from >150 U/mg to <5.5 U/mg of protein, where the no promoter control reported an average 9.6 U/mg and a standard deviation of 4.1 U/mg. Eight constitutive promoters, derived from promoter regions from ferredoxin and thiolase genes of various clostridial species, were tested. A promoter derived from the ferredoxin gene of *C. perfringens* (P_{cperF}) was the strongest of all promoters tested, with a reported expression level of 153.4 U/mg. The P_{TcdB} promoter was found to deliver only one third the expression level, <51 U/mg, of the P_{cperF}. Additionally, indirect induction of P_{TcdB} *via*20mM lactose addition in transformant *E. limosum* ATCC 8486-i1B, for TcdR expression, did not result in significantly greater reporter feedback when compared to the relevant controls in a WT background or uninduced i1B background. However, lactose induction at 20mM was demonstrated to be effective in increasing the expression levels of the *bgaR*-P_{bgaL} promoter cassette by more than an order of magnitude, from 5.5 to 87.1 U/mg. Induction of the inducible P_{fac} promoter with 5mM IPTG, resulted in a near doubling the reported promoter activity, from 18.6 to 34.7 U/mg. The constitutive P_{araE} promoter derived from *C. acetobutylicum* ATCC 824, and used to drive *cas9* expression in the modular CRISPR vectors utilised in this study, reported the relatively low expression levels of 9.7 U/mg, comparable to that of the no promoter control.

In addition to confirming the efficacy of the P_{TcdB} promoter it was important to confirm the integrity of the gene expression cassettes, in order to determine whether zemblanitous mutations were responsible for the lack of acetone or isopropanol in G1 constructs. Figure 4.9 presents an example of initial size-based screening of cassettes by colony PCR, for G1 and G1S constructs with appropriate controls. Size estimates for all constructs were as expected, suggesting that no major mutations had occurred in these constructs. Subsequently, a representative band for each construct was extracted and the entire operon analysed with Sanger sequencing. No SNPs were observed that would account for the lack of efficacy with regard to acetone and isopropanol in the samples under scrutiny.

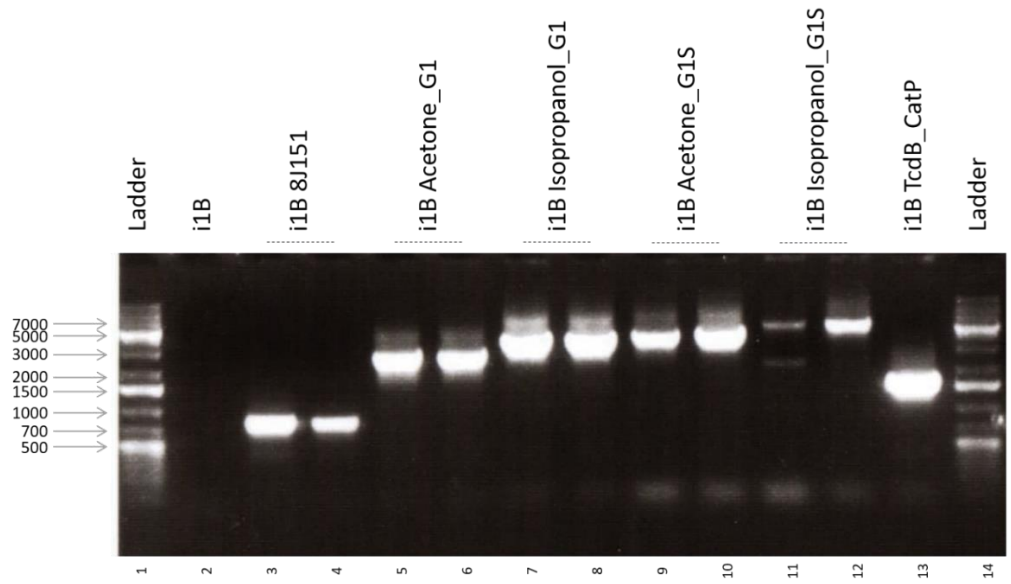


Figure 4.9 – Colony PCR to confirm the integrity of plasmids used for testing of the Generation I and Generation I-Scaffolded constructs for acetone and isopropanol production in *E. limosum* ATCC 8486-i1B (i1B)

The target samples for analysis were the same transformed *E. limosum* strains which were used in the experiments for growth profiling. Primers targeted external to the cassette region of the plasmid [ColE1+Tra F2] and [P19 R1] were used to provide full coverage of the cassette region of the tested plasmids. Lane order from left to right: 1. O'generuler 1kb plus ladder; 2. Untransformed i1B as a negative control; 3,4. i1B containing pMTL8J151; 5,6 i1B containing pMTL8J15_ace_G1; 7,8 i1B containing pMTL8J15_iso_G1; 9,10 i1B containing pMTL8J15_ace_G1S; 11,12 i1B containing pMTL8J15_iso_G1S; 13. i1B containing pMTL8J15_ c*atP*; 14. O'generuler 1kb plus ladder.

Based on the results of the promoter assay and the confirmation of genetic integrity of the cassettes in *E. limosum*, two subsequent generations of acetone and isopropanol cassette were produced. Second generation (G2) constructs were slight adaptions of the G1 constructs, where the P_{TcdB} promoter was replaced with the strongest tested promoter, P_{cperF} (Figure 4.3: G2A, G2B). Third generation (G3) constructs further developed the G2 construct, to also include an acetyl-CoA transferase gene, the *thl* gene of *C. acetobutylicum* ATCC 824 (Figure 4.3: G3A, G3B).

A direct end point comparisons of *E. limosum* ATCC 8486 transformed with plasmids for G1-G3 isopropanol cassettes and grown for 120-hour are presented in Figure 4.10 (product titres) and Figure 4.11 (carbon distribution). Both second and third generation cassettes demonstrated increased isopropanol over G1 and control samples. The average values for acetone, isopropanol, acetate and butyrate concentrations is available in Table 4.3.

Table 4.3 – End point analysis (120 hours) of all tested isopropanol cassettes and the pMTL8J151 plasmid in *E. limosum* ATCC 8486

Sample	Acetone	Isopropanol	Acetate	Butyrate
8J151 (Control)	0.04 (±0.02)	0.35 (±0.34)	29.73 (±1.92)	8.3 (±1.16)
Iso-TcdB (G1)	0.21 (±0.07)	0.58 (±0.44)	14.03 (±3.16)	14.84 (±5.19)
Iso-77 (G2)	0.27 (±0.07)	1.18 (±0.35)	26.98 (±7.85)	10.04 (±4.04)
Iso-77-Thl (G3)	0. 64 (±0.07)	2.93 (±0.86)	24.69 (±3.95)	10.19 (±3.15)
Iso-DFS (G1S)	0.2 (±0.05)	0.18 (±0.05)	18.73 (±0.03)	17.21 (±1.99)

Values represent the average of triplicate samples (± standard deviation). Iso-TcdB and Iso-DFS samples were in transformants of the i1B strain of *E. limosum* ATCC 8486, the remaining three were transformants in a wildtype background.

The statistical significance of acetone and isopropanol titres in each of the samples was measured by students t-test against the control sample (8J151). A statistically relevant increase in acetone was observed in all samples ($P < 0.01$) within a 99% CI. Variation in isopropanol titre between the control sample and both G1 (Iso-TcdB) and G1S (Iso-DFS) samples were $P > 0.1$, and therefore not statistically significant. Although it should be noted that, the mean concentration value of 0.58 mM from the Iso-TcdB samples exceeded the LoD calculated for this study (0.52, Table 4.2), which was not the case for Iso-DFS or the control. Isopropanol levels in G2 (Iso-77) and G3 (Iso-77-Thl) samples were significant to >95% ($P < 0.042$) and 99% ($P < 0.008$) confidence respectively. The reported range for acetone and isopropanol concentrations across triplicate G3 samples was 0.56-0.68mM and 2.23-3.86mM respectively.

Reductions in acetate concentration were statistically significant to a 99% CI in G1 ($P = 0.001$) and G1S ($P < 0.001$) samples when compared to the control. Similarly, both samples demonstrated statistically relevant differences in butyrate concentration, each higher than the control sample. G1S was statistically different with a CI of 99% ($P = 0.003$), however G1 was only relevant to a 90% CI due to large variation within sample replicates ($P < 0.1$). The changes in acetate and butyrate reflect a shift in the acidogenic profile of both G1 strains which appears to be independent of the transformed plasmids as it is not observed with G2 or G3. None of the GC analysis for this sample batch reported an ethanol titre over the established LoD (0.41 mM). As such, no ethanol values have been included in the presented data.

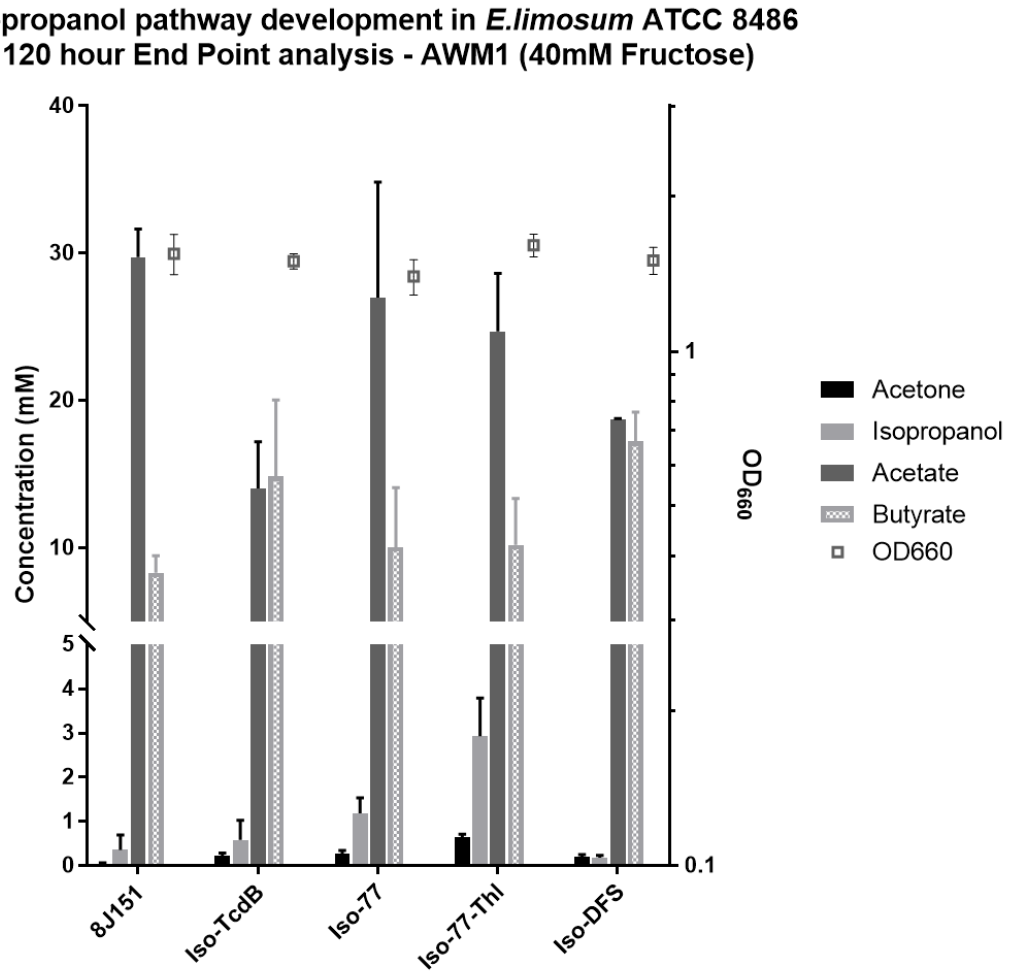


Figure 4.10 – Endpoint data for five transformed *E. limosum* ATCC 8486 variants, including the four isopropanol cassettes developed in this study

Samples were grown for 120 hours on AWM1 with 40mM fructose and 10 µg/ml thiamphenicol in closed batch fermentations. Bars represent the product yields for acetone, isopropanol, ethanol, acetate, butyrate in mM. Square points represent the biomass of samples as measured by OD_{660}. Data presented is the average value from triplicate samples, where error bars represent the standard error of the mean (SEM). Control: WT *E. limosum* ATCC 8486 harbouring an empty pMTL8J151 plasmid as a baseline control for comparison; Iso-G1: *E. limosum* ATCC 8486-i1B harbouring the pMTL8J15_iso_G1; Iso-G2: WT *E. limosum* ATCC 8486 harbouring the pMTL8J15_iso_G2 plasmid; Iso-G3: WT *E. limosum* ATCC 8486 harbouring the pMTL8J15_iso_G3 plasmid; Iso-G1S: *E. limosum* ATCC 8486-i1B harbouring the pMTL8J15_iso_G1S.

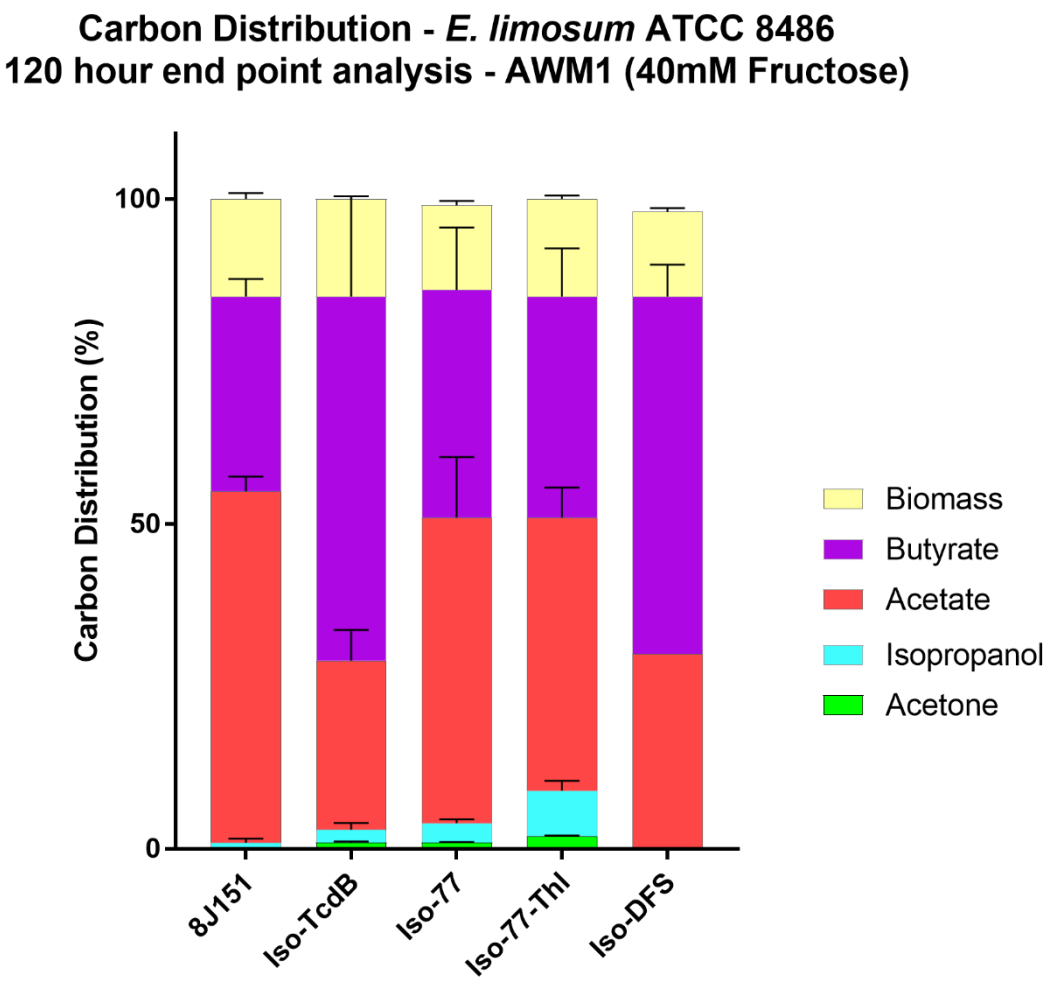


Figure 4.11 – Endpoint carbon distribution for five transformed *E. limosum* ATCC 8486 variants

Transformants include four isopropanol cassettes developed in this study and a plasmid control sample. Samples were grown for 120 hours on AWM1 with 40mM fructose and 10 µg/ml thiamphenicol in closed batch fermentations. Bars represent the carbon distributions for acetone, isopropanol, acetate, butyrate and biomass. Biomass calculations were based on the calculations: $1OD_{660}$ = 0.27 g/L[139,188], biomass mW = 100.4 ($C_4H_8O_2N$)[189]. Data presented is the average value from triplicate samples, where error bars represent the standard error of the mean (SEM). Control: WT *E. limosum* ATCC 8486 harbouring an empty pMTL8J151 plasmid as a baseline control for comparison; Iso-G1: *E. limosum* ATCC 8486-i1B harbouring the pMTL8J15_iso_G1; Iso-G2: WT *E. limosum* ATCC 8486 harbouring the pMTL8J15_iso_G2 plasmid; Iso-G3: WT *E. limosum* ATCC 8486 harbouring the pMTL8J15_iso_G3 plasmid; Iso-G1S: *E. limosum* ATCC 8486-i1B harbouring the pMTL8J15_iso_G1S.

4.3.1. Testing the effects of pathway scaffolding

Testing the effects of enzyme scaffolding in *E. limosum* had initially proved problematic due to low titres of target products, acetone and isopropanol. In light of the positive improvements seen with the G2 and G3 constructs, it was clear that both a stronger promoter and the inclusion of a heterologous acetyl coA-transferase *(thl)* were necessary to achieve the acetone and isopropanol titres required for testing scaffolded constructs in *E. limosum*. In order to avoid significant cloning steps involved in reengineering the G1S constructs, it was decided that testing in a more appropriate organism might be prudent. To this end, *C. autoethanogenum* was identified as an appropriate platform organism from which to test, due to its strong knowledgebase, availability of an appropriate TcdR strain, and ability to grow on (i.e., test on) C1 gas. Additionally, preliminary evidence from a modified G1 construct, assigned G1.2, showed promising preliminary evidence of efficacy with regard to basic acetone isopropanol production from *C. autoethanogenum*. G1.2 differed from G1 only in the addition of the *C. acetobutylicum* derived *thl* gene inserted after *ctfB* in the original cassette.

The *C. acetobutylicum* ATCC 824 thiolase gene was subsequently added in the same manner to the G1S constructs in order to create G1S.2 (Figure 4.15), respectively. The G1S.2 cassettes were conjugated into a *C. autoethanogenum* Lac-TcdR mutant, CH20. CH20 is the *C. autoethanogenum* equivalent of *E. limosum*-i1B, in the addition of the lactose inducible TcdR module at the *pyrE* locus. G1S.2 harbouring strains were added to a collection of appropriate control transformants, to make a 5-strain collection for testing (Table 4.4). Conjugation of the aforementioned vectors into *C. autoethanogenum*, and subsequent sampling of the transformants, was performed by Dr. Chris Humphreys and Simona Pentassuglia (SBRC, Nottingham. 2017).

Table 4.4 – *C. autoethanogenum* strain list for scaffold testing

Strain	Plasmid
C. autoethanogenum Lac-TcdR CH20	No plasmid
C. autoethanogenum Lac-TcdR CH20	pMTL-ace-G1.2
C. autoethanogenum Lac-TcdR CH20	pMTL-iso-G1.2
C. autoethanogenum Lac-TcdR CH20	pMTL-ace-G1S.2
C. autoethanogenum Lac-TcdR CH20	pMTL-iso-G1S.2

Four *C. autoethanogenum* Lac-TcdR CH20 transformants harbouring scaffolded and unscaffolded cassettes for acetone and isopropanol and an untransformed control.

The 120-144 hour timepoint dataset for the five *C. autoethanogenum* samples grown on PETC media with 40mM fructose are presented in Figure 4.13. Product calculations were performed for acetate, ethanol, acetone, isopropanol and lactate using GC/UV-vis, while biomass generation was monitored spectrophotometrically *via* OD_{600}. Full growth curves can be found at Figure 4.13. All plasmid harbouring constructs presented notably different product profiles than the wildtype strain. In all cases carbon flux was directed away from acetate production and towards the other measured products. G1.2_ace and G1.2_iso resulted in increased acetone and isopropanol production as expected. The G1S.2 constructs, however, did not generate either product at levels close to the non-scaffolded variants. The most significant change in profile in the samples containing scaffolded constructs was a substantial increase in lactate levels, coupled with a distinct reduction in culture OD_{600}, and thereby biomass of the samples.

Table 4.5 – Product yields (mM) for PETC Fructose *C. autoethanogenum* transformants

Sample	Acetate	Ethanol	Acetone	Isopropanol	Lactate	OD_{600}
WT	52.8 (-)	5.5 (-)	0.1 (-)	0.0 (-)	0.4 (-)	1.02 (-)
G1.2_Ace	44.0 (5.2)	10.1 (1.1)	1.8 (0.1)	5.6 (0.6)	1.8 (0.3)	1.12 (0.2)
G1.2_Iso	35.2 (5.5)	2.5 (0.6)	1.8 (0.3)	10.7 (1.9)	0.3 (0.03)	0.96 (0.04)
G1S2_Ace	25.5 (1.4)	13.9 (1.9)	0.3 (0.04)	1.2 (0.06)	7.8 (0.2)	0.66 (0.03)
G1S2_Iso	20.6 (0.6)	21.0 (0.5)	0.1 (0.01)	0.1 (0.01)	14.1 (0.2)	0.59 (0.03)

Values represent the average of 120- and 144-hour time point concentrations for each sample (± standard deviation), measured by GC/UV-vis. Samples were grown in closed batch cultures on PETC media with 40mM fructose.

In order to quantify biomass more accurately as a product of metabolism, for evaluation of product ratios, the following conversion formula was used:

$$1\ OD_{600} = 0.317\ g/L\ biomass$$

Equation 5 – OD_{600} to biomass conversion formula for *C. autoethanogenum*

Formula based on reported cell density to dry weight interconversion published by Bruno-Barcena *et al*. (2013).

Table 4.6 – Product molar ratio for PETC Fructose grown *C. autoethanogenum* transformants

Sample	Acetate	Ethanol	Acetone	Isopropanol	Lactate	Biomass
WT	85	9	0	0	1	5
G1.2_Ace	66	15	3	8	3	5
G1.2_Iso	66	5	3	20	1	6
G1S2_Ace	50	27	1	2	15	4
G1S2_Iso	36	36	0	0	24	3

Samples were taken at mid stationary phase (120-144 hours growth) grown in closed batch cultures on PETC media with 40mM fructose. Biomass calculations were based on the calculations: $1OD_{600}$ = 0.317 g/L[190], biomass mW = 100.4 ($C_4H_8O_2N$)[189]. Values were rounded to their nearest integer.

In order to take the basic molar ratio of the measured products a step futher, the ratio of carbon flux towards each products was evaluated. The molecular formula for each products

is presented in Figure 4.12, where biomass composition was approximated using the generic formula presented by Fast *et al*. (2012)[189]:

Biomass: $C_4H_8O_2N$ (mW = 100.4)

Equation 6 – The general molecular formula for acetogen biomass as described by Fast *et al*. (2012)[189].

Biomass values were used as approximations of carbon flux towards solid products within the fermentations, based on the OD measuremnts of cultures taken at each sample time point. This included cell matter and extracellular-polysaccharide which were isolated from supernatant GC samples during centrifugation stages of analysis preparation.

Table 4.7 – Product carbon ratio for *C. autoethanogenum* transformants

Sample	Acetate	Ethanol	Acetone	Isopropanol	Lactate	Biomass
WT	81	8	0	0	1	10
G1.2_Ace	59	13	4	11	4	9
G1.2_Iso	56	4	4	26	1	10
G1S2_Ace	44	24	1	3	20	7
G1S2_Iso	31	31	0	0	32	6

Samples were taken at mid stationary phase (120-144 hours growth) grown in closed batch cultures on PETC media with 40mM fructose. Biomass calculations were based on the calculations: $1OD_{600}$ = 0.317 g/L[191], biomass mW = 100.4 ($C_4H_8O_2N$)[189]. Values were rounded to their nearest integer.

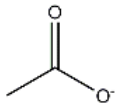
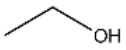
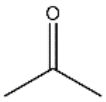
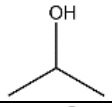
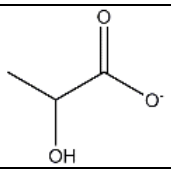
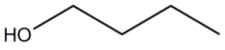
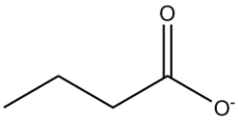
Chemical	Formula	Structure
Acetate	CH_3COO^-	
Ethanol	C_2H_4OH	
Acetone	CH_3CO CH_3	
Isopropanol	C_3H_7OH	
Lactate	$C_3H_5O_3^-$	
Butanol	$C_4H_{10}O$	
Butyrate	$C_4H_7O_2^-$	
Biomass	$C_4H_8O_2N$	-

Figure 4.12 – The molecular structure and formula for the analysed products.

Biomass is represented as an approximate molecular formula based on work described by Fast *et al*. (2012).

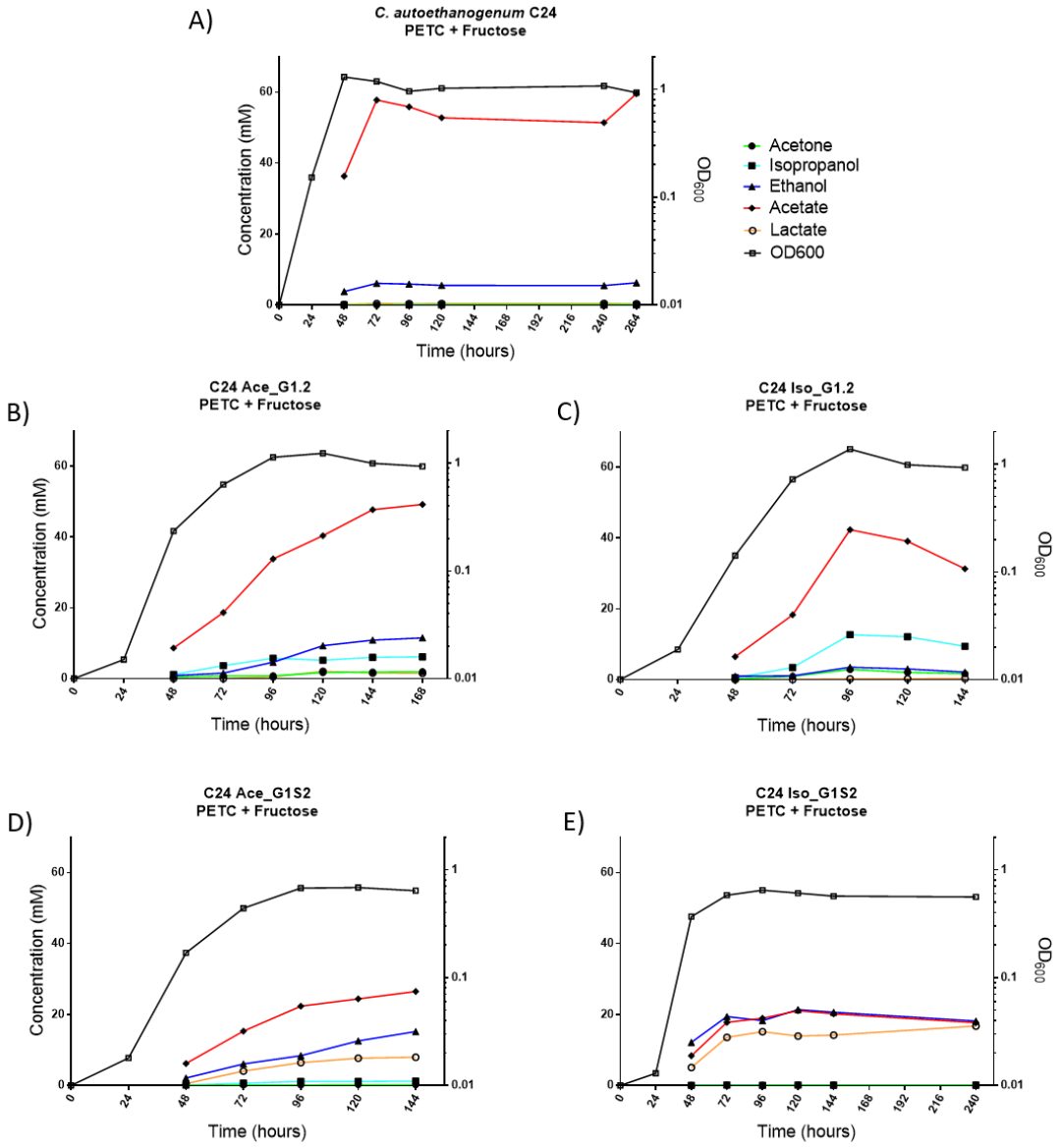


Figure 4.13 – Comparative analysis of scaffolded and unscaffolded acetone/isopropanol production cassettes in *C. autoethanogenum* CH20 grown on PETC media with 40mM fructose

All samples were tracked over a period of at least one week (144 hours) taking samples every 24 hours from the 48-hour time point. A) untransformed *C. autoethanogenum* C24; B) *C. autoethanogenum* C24 transformed with pMTL8515_Ace_G1.2; C) *C. autoethanogenum* C24 transformed with pMTL8515_Iso_G1.2; D) *C. autoethanogenum* C24 transformed with pMTL8515_Ace_G1S2; E) *C. autoethanogenum* C24 transformed with pMTL8515_Iso_G1S2.

In parallel with the testing of the constructs on fructose substrate, C1 growth analysis was performed with growth on PETC + CO (1 bar overpressure)(Figure 4.15). The transformed samples appeared to reach a stabilised product profile roughly 24-96 hours after reaching their respective maximal OD_{600}, and these stabilised values were used as the primary values for comparison (Table 4.8). As was expected from the lower energy carbon source, less overall product turnover, and thereby carbon turnover, was observed on CO grown samples, compared to fructose counterparts. Isopropanol titres exceeding the LoD were only observed in the G1.2_iso harbouring strain (1.8mM). Scaffold containing samples behaved in an unusual manner in that both appear to have switched between two distinct phases of acetate and ethanol production during the first 144 hours of their growth. G1S.2_ace produced ethanol as high as 53.6mM, surpassing the maximal 43.6mM acetate produced by the WT strain within the first 192 hours. After this time the product concentrations and ratios were broadly reflective of the final time point readings, up to 264 hours after inoculation.

Table 4.8 – Product yields (mM) for *C. autoethanogenum* transformants grown on PETC media with CO

Sample	Acetate	Ethanol	Acetone	Isopropanol	Lactate	Biomass
WT	43.6 (0.4)	0.9 (<0.1)	0.0 (<0.1)	0.0(<0.1)	0.0 (<0.1)	1.8 (0.5)
G1.2_Iso	28.9 (0.2)	3.9 (1.8)	0.2(<0.1)	1.8 (0.1)	0.0 (<0.1)	1.1 (0.2)
G1S2_Ace	3.8 (0.3)	53.6 (2.2)	0.0(<0.1)	0.2 (<0.1)	0.8 (<0.1)	1.2 (0.1)
G1S2_Iso	38.3 (1.9)	4.6 (0.2)	0.0(<0.1)	0.1 (<0.1)	1.0 (<0.1)	1.6 (0.1)

Values represent the average concentrations for the two lattermost timepoints taken of each sample within the first 192 hours (± standard deviation), measured by GC/UV-vis. Biomass calculations were based on the calculations: $1OD_{600}$ = 0.317 g/L[191], biomass mW = 100.4 ($C_4H_8O_2N$)[189].

Scaffolded pathways did not positively impact the acetone and isopropanol titres on fructose or CO substrates in *C. autoethanogenum*. The preliminary data indicates that the scaffolds instead reduced the titres of both solvents. The most notable effects of scaffolding appear to be increased lactate in both scaffolded samples on fructose, and notably high ethanol titres produced by G1S.2 harbouring cells grown on CO.

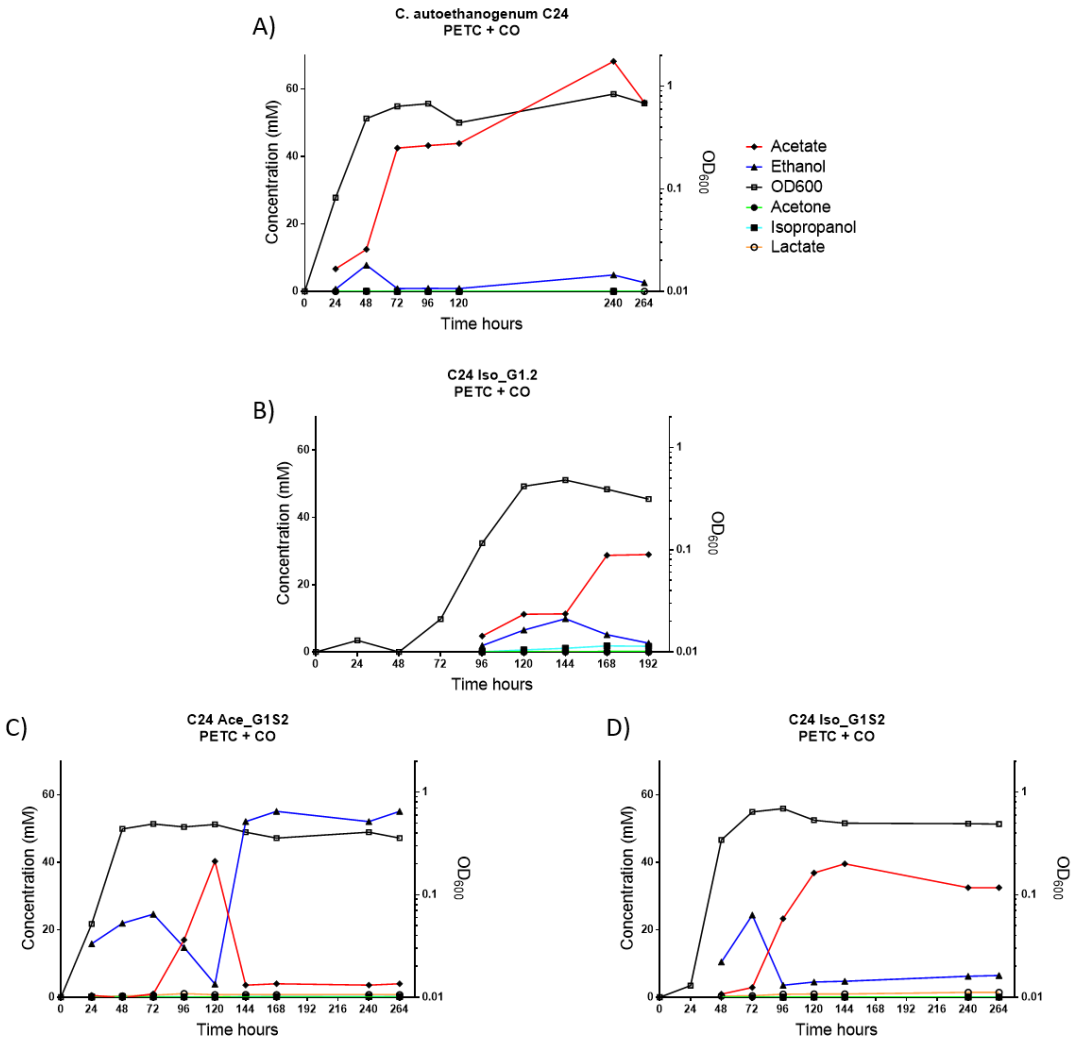


Figure 4.14 – Comparative analysis of scaffolded and unscaffolded acetone/isopropanol production cassettes in *C. autoethanogenum* CH20

Samples grown on PETC media with CO (1 bar overpressure) and tracked over a period of at least one week (144 hours) taking samples every 24 hours from the 48-hour time point. A) untransformed *C. autoethanogenum* C24; B) *C. autoethanogenum* C24 transformed with pMTL8515_Iso_G1.2; C) *C. autoethanogenum* C24 transformed with pMTL8515_Ace_G1S2; D) *C. autoethanogenum* C24 transformed with pMTL8515_Iso_G1S2.

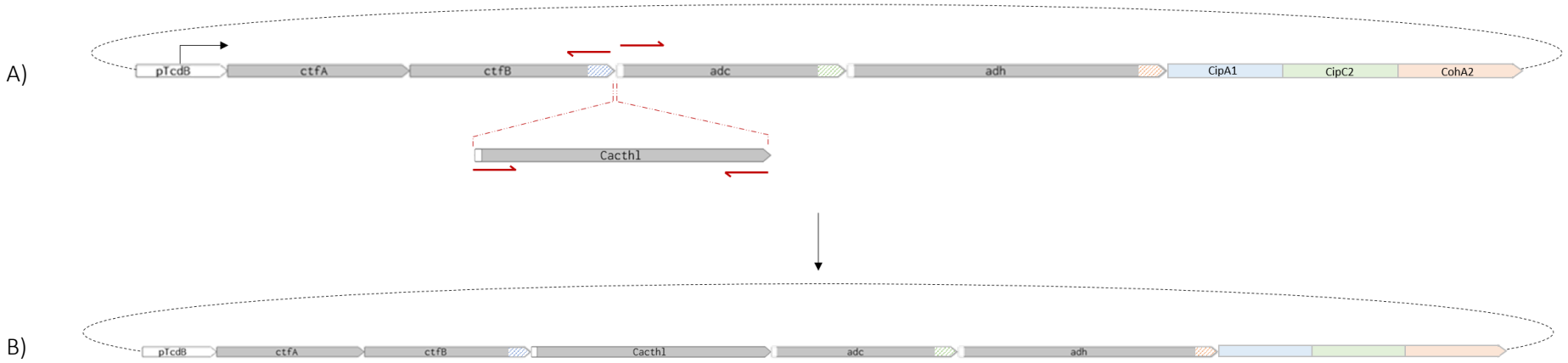


Figure 4.15 - Schematics for the assembly of G1S2, generation-I scaffolded acetone and isopropanol cassettes

G1S.2 was generated with the insertion of acetyl coA-transferase derived from *C. acetobutylicum* ATCC 824 (Cac*thl*) into a G1S cassette. Insertion of Cac*thl* was achieved using the HiFi cloning methodology with overlapping PCR fragments. Primers are indicated by red directional half arrows. Dashed black line represents the plasmid backbone. A) Two PCR amplifications: first, the G1S plasmid backbone with divergent primers at the *ctfB-adc* junction; second, the amplification of Cac*thl* from *C. acetobutylicum* gDNA. B) The resulting plasmid, complete P_{TcdB} driven cassette including Cac*thl*.

4.4. Discussion

In this study, isopropanol production has been established in *E. limosum* through the development and application of plasmid borne synthetic metabolic pathways. Isopropanol titres from *E. limosum* strains, transformed with the final iterative development of the production cassette (G3), were as high as 3.86mM in closed batch fermentation with fructose as the primary carbon source. Additionally, comparable synthetic metabolic pathway cassettes were tested in *C. autoethanogenum* and shown to be effective in generating isopropanol titres in excess of 10mM on fructose and 1.8mM on CO, in closed batch cultures. Testing in *C. autoethanogenum* presented a platform to analyse the effects of introducing enzyme scaffolds to the synthetic pathways in order to coordinate enzyme proximities with a view to possibly improving the rate of reaction along the pathway. Scaffolding showed no improvement in isopropanol production, and in fact led to a marked reduction in yields in all cases where scaffolds were employed.

4.4.1. General Discussion

The development and implementation of synthetic metabolic pathways is an exciting area of research which has opened the door to novel modes of strain development, particularly with regard to enhancing the capacity of organisms to produce more valuable products. This concept, coupled to organisms which have evolved with a natural propensity for metabolising compounds that are readily available in human waste streams or from renewable sources, presents an exciting means for development of novel, green, and commercially viable bioprocesses. The effective transformation methodology outlined in chapter 3 has enabled the application of a number of molecular tools in *E. limosum* 8486 with the goal of strain development through genetic engineering. The introduction of isopropanol biosynthesis in *E. limosum*, extending the product spectrum beyond the naturally occurring acetate, butyrate, and ethanol, is an

industrially interesting proof of the robustness and efficacy of the tools now available for engineering *E. limosum*.

4.4.2. Methodology limitations

Limited success was achieved with the first-generation cassettes in *E. limosum*, as measured by resultant acetone and isopropanol yields. However, the results generated from these constructs were valuable in laying the foundations for the remainder of the project. In particular, early results directed the development of cassettes, encouraged fine tuning the analysis methods, and forced re-evaluation of fermentation medium conditions.

4.4.2.1. Sample Analysis

The GC analysis for G1 constructs reported low-level isopropanol (0.19-0.37 mM) in all of the tested samples, including the 8J151 control, but no acetone of note. This result was treated with a degree of suspicion due to the absence of isopropanol reports in the literature or indeed from inhouse qualitative HPLC-UV and HPLC-IR analysis (Appendix - Figure 7.2.1). This suspicion resulted in a brief revaluation of the GC/UV-vis analysis method.

The observed levels are expected to be a consequence of a slight overlap of the elution times of ethanol and isopropanol with the GC parameters used (Figure 4.7). No record of isopropanol production from *E. limosum* has been reported in the literature or was observed in previous samples measured by HPLC-UV. This results in slight peak overlap which cannot be reliably resolved by either manual or automated peak area calling. The effects of these peak overlaps were observed in the GC standards and spiked samples where both isopropanol and ethanol were present. Pending a more appropriate method of analysis, quantification of isopropanol yields was treated with caution going forward.

Due to a reasonably significant variation within triplicate samples it was difficult to draw robust conclusions as to whether or not the transformed plasmids had a tangible effect on cell growth and product profile. This is particularly problematic during the early stages of growth where the low concentrations combine poorly with the notable variability within sample triplicates. However, later time points, where concentrations are generally higher and the wayward impact of intra-sample variation is therefore proportionally reduced, can provide some insight into the more general effects of these constructs on the cells.

Peak overlap in the manner evident in Figure 4.7 would likely lead to incorrect concentration estimates, especially in instances where the concentrations of each molecule are not balanced, and therefore not directly comparable to the standards used. Additionally, where low concentrations of either ethanol or isopropanol are present in samples, its limit of detection (LoD) is directly affected by the concentration of the other. Where concentrations of just one is high, the overlapping tail of the its peak exceed the maximal peak height of the other. This effect is more pronounced for low ethanol concentrations as the relative overall peak areas are smaller and therefore more prone to be overlapped.

This lack of statistical power was compounded by inconsistencies in growth profiles demonstrated by *E. limosum* ATCC 8486 throughout the project. This variation occurred between samples grown under seemingly identical conditions within single experiments, and was particularly evident between biological replicates. The exact causes of the variability remain unclear; however, it seems likely that *E. limosum* is either highly sensitive to the handling or conditions it was subjected to, or was affected by high population heterogeneity. Additionally, the extracellular glycopolymer production made hampered the consistency with which sample inoculations could be made.

As an initial proof of concept study, to demonstrate the viability of isopropanol production by *E. limosum*, minor quantification inaccuracies were considered to be

tolerable. Particularly since the inter-sample comparison was the most relevant output, and comparable samples were affected in a reasonably consistent manner. Of the available analytical methods for this project, GC represented the most convenient and reliable way of processing large sample numbers, and for the reasons discussed above was assessed to be sufficient to confirm the success or failure of meeting the project aims. Beyond this proof of concept study however, the accuracy of these analysis methods will be desirable and likely necessary, particularly in the latter stages of strain/process optimisation.

4.4.2.2. Re-evaluation of Fermentation Media

One particular discrepancy which was noted during the testing of the G1 constructs was inter-fermentation variability. In particular, a loss or significant reduction in ethanol titres was observed in select fermentations without obvious cause. The inconsistency of ethanol production was believed to be related to variability introduced during the preparation of the media. Specifically, adjustments in the buffering capacity of the media broth as a result of high temperature exposure during autoclave sterilisation. Varying levels of precipitation were observed following autoclaving of AWM2 media. This precipitation was coincided with a drop in the pH of the media, indicating a reduced buffering capacity of the media. Variation in solvent production in response pH buffering and variability is reflective of a number of solventogenic species, most well studied in *C. acetobutylicum*[192]. In these species natural acetate and butyrate production is balanced late in the growth cycle with acid re-uptake processes central to production of ethanol and to a lesser extent butanol. As a means of resolution to this issue, all future fermentations were performed on AWM1 media sterilised by filter-based sterilisation method.

4.4.3. Isopropanol Operon Developments

The lack of success with the G1 IPOs in establishing detectable levels of isopropanol and acetone was initially a puzzling setback to the project. Similar attempts to establish acetone and/or isopropanol had been successfully applied in a diverse range of species with varying degrees of success[144,163,115,178]. A small scale metabolic model, developed in collaboration with this study, demonstrated the efficacy of the pathway in silico[193]. As a result, the identified point of failure was expected to lie with the underlying components of the IPO cassettes. Specifically, the P_{TcdB} promoter system and the genetic integrity of the cassettes were queried.

4.4.3.1. Promoter Identification

The conditional overexpression system, a genomic lac-inducible TcdR sigma-70 factor working in combination with the P_{TcdB} promoter driving the IPOs, was used for two reasons. First, it was expected to enable relatively stable cloning of potentially toxic genes due to poor recognition of the promoter by the native *E. coli* transcriptional machinery[194]. Second, it had been shown to generate extremely high levels of expression of downstream proteins in a number of clostridial species, including *C. acetobutylicum*, *C. autoethanogenum*, *C. ljungdahlii* and *C. pasteurianum* (SBRC Nottingham, unpublished). Overexpression is particularly useful in early stage development and other scenarios where ensuring that expression levels are not a significant limitation on the to achieving a desired engineering result[195], in this case isopropanol production. Importantly, no instances of the TcdR- P_{TcdB} interaction failing to generate high expression had previously been observed within our research group. Interestingly, however, the early reports which indicated a total lack of detectable P_{TcdB} expression in the absence of TcdR are at odds with the data published in a more recently reported study[196]. Grosse-Honebrink (2017) indicated that expression from P_{TcdB} in the absence of TcdR in *E. coli,* had a relative strength of approximately one third that of a known strong promoter derived from the ferredoxin gene of *C. sporogenes*.

This necessitated a review of the promoter system, and the identification of an appropriate alternative system, achieved through the generation of a promoter library.

A CAT assay was performed to generate a library based on eight clostridial ferredoxin or thiolase derived promoters, as well as a selection of other useful promoter systems. Promoter modules were taken from constructs used for equivalent promoter studies carried out previously within the SBRC. The difference between the weakest and strongest promoters was roughly 30-fold with a reasonable distribution of strengths in between. Access to such a library is valuable for fine tuning gene expression during downstream strain engineering and optimisation processes[197]. A library such as this one has advantages over many of the libraries generated through rational design or random mutation of a single promoter[197,198], in that the promoters do not share any particular homology to one another so can be used in parallel without fear of undesired homologous recombination events occurring.

4.4.3.1.1. TcdR Failure

As a sigma factor, it is possible that TcdR may directly affect gene regulation in *E. limosum*. This would of course require both the inducibility of the system to be leaky enough to express *tcdR* in the absence of lactose, and for there to be sufficient interaction between the TcdR protein and the *E. limosum* genome. Putative TcdR consensus sequences were generated manually based on published -35 and -10 sequences from native *tcdA* and *tcdB* promoter regions[199](Binding site identification was based on the -35 and -10 boxes identified in Martin-Verstraete (2016)[199]. Calculations for the expected number of occurrences is based on a double stranded 4421303bp genome where there is a 1 in 4 chances of each base, A, C, T, G, at each position.Table 7.2.1). The number of possible binding sites varied between 0 and 83 depending on the stringency applied to the more variable -10 sequences. Looking at the issue from the reverse direction, pBLAST analysis of TcdR against *E. limosum* genomes indicated low but appreciable levels of homology to the C-terminal portion

(inclusive of the specificity defining sub-region 4.2) of two putative sigma-70 factors native to *E. limosum*, Ga0111251_111823 and Ga0111251_11633 (Table 7.2.2). Though homology between sigma factors is not in itself unusual, alignment scores were higher than those reported for an equivalent alignment of TcdR against *C. acetobutylicum* and *C. autoethanogenum* genomes. Cross reactivity between the sigma factors of different clostridiales has previously been demonstrated by Dupuy *et al*.[165], so it is possible that TcdR activity is not limited to the expression of just the P_{TcdB} promoter in *E. limosum*. Equally, TcdR may be inadvertently subject to unknown post-translational regulatory mechanisms when expressed in *E. limosum*, precluding its activity on the *tcdB* promoter. In both events, the expected effect of TcdR on the P_{TcdB} promoters in G1 and G1S constructs, may be significantly altered.

4.4.3.2. Expression levels and Bottleknecks

The idea to introduce acetone and isopropanol production *via*plasmid borne gene cassettes was decided for a number of reasons. One, the speed at which strain development and testing can be carried out is significantly faster on plasmids than where genome integration, and the associated additional selection and screening steps stages, are required. Additionally, the discovery of the stability of p19 replicon revealed a possible stable means of introducing genes into *E. limosum* with multi-copy number, and thereby increased gene expression over traditional integrative methods. Heightened expression levels are valuable at the stage of testing new pathways where limited enzyme production might otherwise cause bottlenecks in an implemented metabolic pathway.

The successful incremental improvements in isopropanol yields from the G1 to the G2 and G3 IPOs highlighted two fundamental limiting factors of the G1 constructs. First, the insufficient expression levels which were overcome to some extent by the replacement of the promoter in G2. Second, the lack of sufficient activity of the native acetyl-CoA transferase (thiolase) and therefore a shortage of acetoacetyl-CoA, which

was countered with the addition of the heterologous thiolase from *C. acetobutylicum* in G3.

In-spite of the multi-copy plasmid-based approach to expression, pathway flux limitations due to inadequate enzyme availability, appears to have been one of the fundamental issues with the G1 constructs. Simply by increasing the level of enzyme availability, through a stronger promoter alone, the G2-iso constructs resulted in a ~2-fold increase in isopropanol production over the respective G1 counterpart (Figure 4.10).

Interestingly no obvious change in butyrate titres was observable with the G3 constructs. This was somewhat unexpected because in addition to the effects on isopropanol pathways, the addition of the heterogeneous thiolase gene might result in increased flux toward butyrate, as a result of increased intracellular acetoacetyl-CoA. Instead, the only statistically significant differences between G2 and G3 constructs were the increased acetone and isopropanol yields in the latter. It is interesting to note, however, that there appeared to be decreased acetate levels, and slightly increased butyrate levels in both of the first-generation constructs, G1 and G1S, in comparison to the other three samples. This variation might be indicative that the introduction of the lac inducible TcdR expression system at the repaired *pyrE* locus i1B mutants, in which G1 and G1S constructs are hosted, has some tangible impact on growth profile. Without further research, we can only speculate as to the possible mechanisms through which the Lac-TcdR system might impact the wider genome.

4.4.4. Scaffolded Pathways

The initial hypothesis surrounding scaffolding of the solvent pathways to acetone and isopropanol was that by forcing co-localisation of the appropriate enzymes, rate limitations imposed by intracellular diffusion of substrates could be reduced (Figure 4.16). By binding the enzymes in close proximity, as in the scaffolded model, the

distance of this movement could be reduced, thereby potentially reducing rate limitations which might have occurred as a consequence. This could result in improves rates of product formation by enhanced substrate channelling along reaction cascades[181]. Scaffolded pathways may also generate micro-localised concentration gradients to encourage less favourable thermodynamic reaction steps in the broader cellular environment[200]. An additional expectation for the effect of enzyme scaffolding was that the if a rapid turnover of intermediates could be generated, this might limit the intercellular concentration of intermediates which might otherwise be toxic to the cell. Although in the case of the synthetic isopropanol pathways used in this study, the isopropanol end product has a higher toxicity than its precursor acetone[201].

The ineffectiveness of the unscaffolded G1 constructs in establishing a baseline for acetone and isopropanol production in *E. limosum* made initial evaluations of the scaffolded variants problematic. No significant change in acetone or isopropanol yield was observed using the G1S constructs over the G1 constructs (Figure 4.5.b.c.d.e). Furthermore, there was no statistical significance between the active constructs and the negative control sample (Figure 4.5.a). It is reasonable that scaffolding alone would be insufficient to resolve the undetectably low functionality in the unscaffolded system, where the primary limiting factors are not resolvable by simple co-localisation of the enzymes. The effect of co-localisation is expected to primarily increase in rate along the pathway, so a discernible baseline rate in the unscaffolded constructs is a prerequisite for comparative purposes. Alternate analytical methods with greater sensitivity would be necessary to draw further conclusions from these samples with a reasonable degree of confidence, due to the observed acetone and isopropanol yields being so close to or below the lower limits of detection (LoD). Ultimately, the generation of a robust baseline from the unscaffolded constructs was required before further time investment into scaffolded constructs was deemed worthwhile.

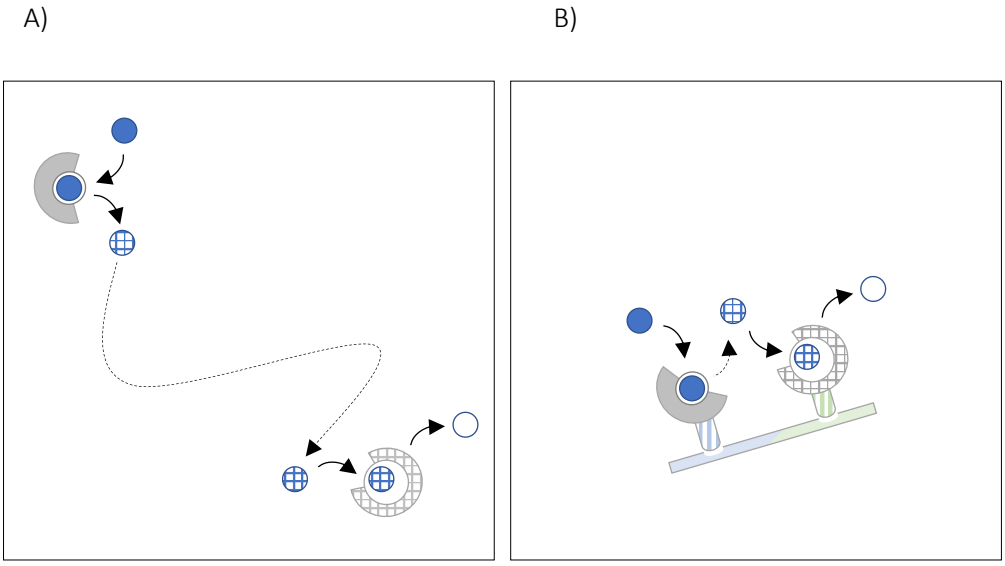


Figure 4.16 – A basic illustration of the enzyme scaffolding principle for a hypothetical metabolic pathway containing two enzymatically catalysed reactions.

The filled grey shape represents the first enzyme (enzyme 1) in this hypothetical pathway, which catalyses the conversion of metabolite 1 (filled blue circle) to metabolite 2 (hashed blue circle). The hashed grey shape represents the second enzyme (enzyme 2) in the hypothetical pathway, which catalyses conversion of metabolite 2 to metabolite 3 (empty blue circle). A) depicts the metabolic pathway based on a model of free floating enzymes. B) depicts a model of the pathway where enzymes are held in close proximity by interactions between a scaffold protein (solid blue/green bar) and the unique dockerin domains (blue and green dashed shapes) fused to each enzyme. Arrows represent the physical intracellular movement of metabolites and substrates with regard to the two enzymes. The dashed arrows depict the molecular movement of metabolite 2, from the product of enzyme 1 catalysis, to the substrate for enzyme 2.

In light of the *E. limosum* promoter assay findings, and the improved isopropanol yield resulting from the promoter modification in the iso-G2 construct, the low expression levels of the P_{TcdB} promoter were identified as one of limiting factors to isopropanol. To avoid the cloning and screening required to replace the promoters of G1S constructs, with no guarantee of desirable outcomes afterwards, it was decided that migration to a better suited testing host was desirable. *C. autoethanogenum* was identified as a candidate for further testing of the scaffolded constructs on the basis of its strong research foundation within our group and the wider literature as a model acetogen[202,203].

The efficacy of the P_{TcdB} promoter system had previously been demonstrated in *C. autoethanogenum*, as well as compatibility of the p19 replicon (SBRC, Nottingham. unpublished). As a result, the previously constructed G1 and G1S constructs required just a single cloning step to make them appropriate for transfer. The addition of the *C. acetobutylicum* derived thiolase gene to account for the lack of a native acetyl-CoA transferase (CacThl). The addition of CacThl to G1 and G1S was able to be performed in parallel with equivalent addition to the non-scaffolded G2 constructs as the iterative improvement to G3. The resulting constructs (G1.2 and G1S.2) were comparable to those used in engineered acetone production reported in *C. ljungdhalii*[204], a strain closely related to *C. autoethanogenum*[142]. This knowledge, combined with a patent published in 2012 pertaining to the production of isopropanol directly from *C. autoethanogenum* with *C. acetobutylicum* derived *thlA-ctfAB-adc* genes[205], provided sufficient confidence to proceed in this host. The two key drawbacks for testing in *C. autoethanogenum* were the lack of a native acetyl-CoA acetyltransferase, and the presence of a native secondary alcohol dehydrogenase. The issue of acetyl-CoA acetyltransferase activity was resolved by the introduction of a *C. acetobutylicum* derived thiolase gene (Cac*thl*) into the constructs (Figure 4.15). The presence of a native secondary alcohol dehydrogenase meant that the conversion from acetone to isopropanol could occur independent of the heterologous *Sadh* in the isopropanol production cassettes. Although this was not ideal for testing the scaffolding effect of the full isopropanol pathway, this enabled narrow focus on the effect of scaffolding for a single reaction pair, *ctfAB* to *adc*.

Assuming a stochastic distribution model for the localisation of substrate within the cell at any one time it was reasonable to predict that the rate of enzyme substrate interaction could be improved by co-locating the production site of a given substrate with the appropriate enzyme for its subsequent utilisation. If this effect is indeed present, it does not appear to outweigh unforeseen negative effects of the scaffold on cellular metabolism. The data from *C. autoethanogenum* indicates that the scaffolding

approach implemented in this project was not effective in increasing isopropanol production titres in batch culture. It is, however, noteworthy that the acidogenic and solventogenic titres were significantly affected in the presence of scaffolded enzymes in contrast to the non-scaffolded counterpart, specifically with regard to ethanol and lactate titres. Both ethanol and lactate were significantly increased in the scaffold containing mutants. These increases appear to occur at the expense of acetate and isopropanol production by comparison to the unscaffolded isopropanol mutant. The explanation for these observations is possibly that the scaffolded constructs result in significant degree of stress in the host cells. The addition of the C-terminal dockerin domains may have been detrimental to the normal activity of one or all of the enzymes involved in the pathway. Additionally, it may have resulted in secondary interactions resulting in unintended function resulting in cell stress.

The lack of isopropanol and acetone observed in the *C. autoethanogenum* scaffold containing samples, combined with increased lactate and unusual ethanol profiles, suggest that the scaffoldin expression, or the cohesin modified enzymes are interfering with other metabolic pathways in some manner. While small terminal additions to translated enzymes sequences do not typically have a significant impact on activity, in some cases even small tags can be detrimental[206]. These effects are generally attributed to tags impeding the correct protein folding or where active sites are sterically hindered. It is not unlikely that the sizable ~75 AA cohesin additions could preclude the functionality of one or more of the tagged enzymes. The efficacy of the underlying system used in the cassettes is demonstrated by the successful production of acetone and isopropanol in the unscaffolded constructs. Following the addition of the dockerin domains, however, no acetone or isopropanol was observed. The sequence integrity of scaffolded constructs was confirmed by PCR and Sanger sequencing so it seems likely that the constructs were expressed in the cell.

The desired scaffolding effects may have failed at a number of fundamental levels. Results of Monte Carlo simulations of enzyme cascade reactions on a 2D surface

scaffold, published by Chado *et al.* (2016), indicated that cascade throughput was at its highest when enzymes were most closely positioned (1-10 nm)[207]. Additionally, their simulations confirmed the previously reported importance of specifically enzyme-pair positioning and orientation for enhancing cascade throughput[208,209]. Appropriate pair positioning is reflected in the scaffold designs used in this study, however no specific control of the orientation of the active sites was attempted. Even operating under the assumption that dockerin-scaffoldin interactions were able to occur within the cell it is quite possible that improper orientation may have precluded the desired substrate channelling effects. Protein crystallography of the scaffolded complexes will be required in order to assess the positioning of the active sites as they exist in the current constructs before such fine tuning is possible. The optimisation gains achieved through orientation, however, would appear to be inferior to the improvements achievable by minimising the distance between the scaffold bound enzymes[207].

4.4.5. Future work and improvements

Finding alternate homologues for each enzyme in the IPO pathway may be a route to further increasing yields. Hanai *et al*. (2007) reported an increased isopropanol concentration using a *E. coli* derived coA-transferase, *ato*AD, in place of *ctfAB* from *C. acetobutylicum*[115]. This is likely reflective of the significantly higher K_m value for acetate of *ctfAB* over *ato*AD, 1200mM and 53.1mM respectively[210]. The lower affinity to acetate of c*tfAB* was presumed to be the primary reason to the lower overall productivity of the *ctfAB* containing pathways. Alternately, the expression levels of these enzymes could have had a significant effect on their perceived activities. It is well established that both codon recognition and GC content play significant roles in protein expression[211]. In the cases of these both *E. coli* (50.8%) and *E. limosum* (47.2%) the GC content is significantly higher than that of *C. acetobutylicum* (30.9%) and *C. beijerinckii* (30%), and codon usage varies significantly between the former and latter pair[212]. Interestingly, genomic analysis of *E. limosum* in this study (Chapter 3) indicated a codon

usage closely comparable to that of *E. coli*. This will likely be useful for future research in terms of the direct application of *E. coli* genes in *E. limosum*, and in making accessible the wealth of previously optimised heterologous genes which have been already been generated for expression in *E. coli*.

An alternate route to improved acetone and isopropanol yields might be achievable through adjustment of the underlying synthetic pathway, using instead a non-acetate coupled acetone production which has been reported in *E. coli*[213]. This was achieved by replacing the thiolase catalysed reaction step in the *C. acetobutylicum* derived pathway, with the thioesterases TEIIsrf (*B. subtilis*) or YbgC (*H. influenzae*). However, this pathway was found to be only applicable in a narrow set of *E. coli* hosts, and is likely unsuitable for C1 growth due to the concomitance of acetate production and ATP generation for Wood-Ljungdahl growth.

Clostridial thiolases such as the heterologous *C. acetobutylicum* employed in the Iso-77-Thl constructs have greater feedback inhibition from reduced-CoA levels, compared to *E. coli* variants[210,214]. Employment of the *E. coli* variant may enhance isopropanol production. Also, the *ctfAB* used in this work has a relatively high Km so it may be prudent to employ the EtoDC variant from *E. coli*. Furthermore, the unknown effects of pH on CtfAB dimerization, specifically the effect on the required cysteine bridge formation, could be mitigated by the use of *E. coli* derived equivalents. The EtoDC variant is more likely to have evolved to operate closer to the neutral pH values used for *E. limosum* culturing, as opposed to the low pH environments enjoyed by *C. acetobutylicum*.

4.4.5.1. Optimism for Methanol

Preliminary evidence from a final experiment carried at the conclusion of the project indicated that the improvements to the isopropanol production cassettes translated to samples grown with methanol as a growth substrate. Although project time constraints

precluded full experimental analysis on methanol, a pilot test demonstrated isopropanol titres of over 2mM isopropanol at a final 220-hour timepoint for AWM1-methanol (62mM) grown *E. limosum* harbouring the improved cassettes. No isopropanol was observed using the G1 cassette or cassette-less control samples. With one of the ultimate goal of this research being the develop *E. limosum* as a potential industrial chassis, an obvious next step will be to establish and optimise the isopropanol producing strains in the setting of continuous methanol fermentation.

4.5. Conclusion

The highest level of isopropanol reported in this study equated to slightly below 0.25 g/L (3.86 mM) during batch fermentation of fructose. Such titres would be insufficient for industrial fermentation processes, but the developments to date represent a solid foundation for future process optimisation towards a commercially viable isopropanol producing strain.

5. Chapter 5 – Building methylation donor strains to overcome transformation barriers

5.1. Abstract

This chapter discusses the development of an approach to enable transformation or enhance their efficiencies in recalcitrant microorganisms which contain active restriction modification (RM) systems. The method develops upon existing concepts of mimicking DNA methylation patterns, by harnessing lambda-red recombineering to support the creation of fully tailored methylation donor strains of *E. coli* for preparation and protection of plasmids, compatible with both electroporation and conjugation procedures.

5.2. Project background

The initial focus of this study was the research and development of acetogenic species which had natural capacity for C4 production from C1 feedstocks. The primary candidate for this work was *C. carboxidivorans*. This butanol producing strain was of selected for its extensive product profile, and had already attracted commercial interest from Coskata, as a chassis for industrial C1 fermentation. Although no methodology had been published for gene-transfer in the strain, a collaborating partner, SIBS, claimed to have successfully achieved electroporation. The project initially centred around optimisation of the collaborators protocol, however following many months of trial and error the approach transpired to be wholly ineffective. The scale of RM system protection in *C. carboxidivorans* was identified as the primary barrier, however no method existed to overcome such a defensive arsenal. *E. limosum* was selected as a fall back, also a C4 producer (butyrate) and with significantly fewer RM systems to contend with.

The developed methodology described herein was developed alongside the *E. limosum* development project, with an end goal of creating a transformation method which would be effective in *C. carboxidivorans*. The method was not ultimately applied fully in *C. carboxidivorans* during this project as a result of time restrictions. The demonstrated proofs applied in other organisms, however, are extremely positive indications that the method will provide the much-needed solution to *C. carboxidivorans* in the near future.

5.3. Introduction

Gene transfer is cornerstone of modern microbial research. The targeted genetic manipulations which are enabled by gene transfer are central to top-down research efforts in understanding microbial genomes, and open up new avenues for strain developments in strains destined for industrial application. Two significant barriers to horizontal gene transfer, CRISPR/Cas systems and RM systems, stem from evolved mechanisms which protect microorganisms from the potentially catastrophic effects of foreign DNA. CRISPR/Cas systems operate with high specificity to target endonucleolytic activity against long target sequences, typically in the region of 20 bases[107]. Libraries of acquired target sequences are created at CRISPR loci, derived from previous exposures to foreign DNA, in a constantly evolving adaptive immune system[215]. In contrast, RM systems operate at a more basic level, targeting relatively short recognition sequences for cleavage by restriction endonucleases, determined by the specific epigenetic modification status of target DNA bases[216].

5.3.1. RM systems

Restriction modification (RM) systems are observed with near ubiquity across prokaryotic species, with over 95% of species harbouring at least one[133]. The systems have evolved to protect host cells from foreign DNA elements, such as that from

bacteriophage, which could prove harmful if transcribed. RM machinery distinguishes foreign from host DNA based on the identification of chemically modified nucleotides within recognised sequences (motifs). Specifically, modification of cytosine and/or adenine DNA bases by chemical methylation are recognised by enzymes in RM systems.

Typically, active RM systems perform three functions. First, the methylation of host DNA following cellular replication to mark it as self. Second, the identification of foreign DNA. Third, the destruction of DNA for which interrogation of methylation pattern does not match host modification status at the target sequence. Where these patterns differ from the 'expected' host pattern, RM systems engage endonuclease activity to cleave, and thereby inactivate, DNA molecules.

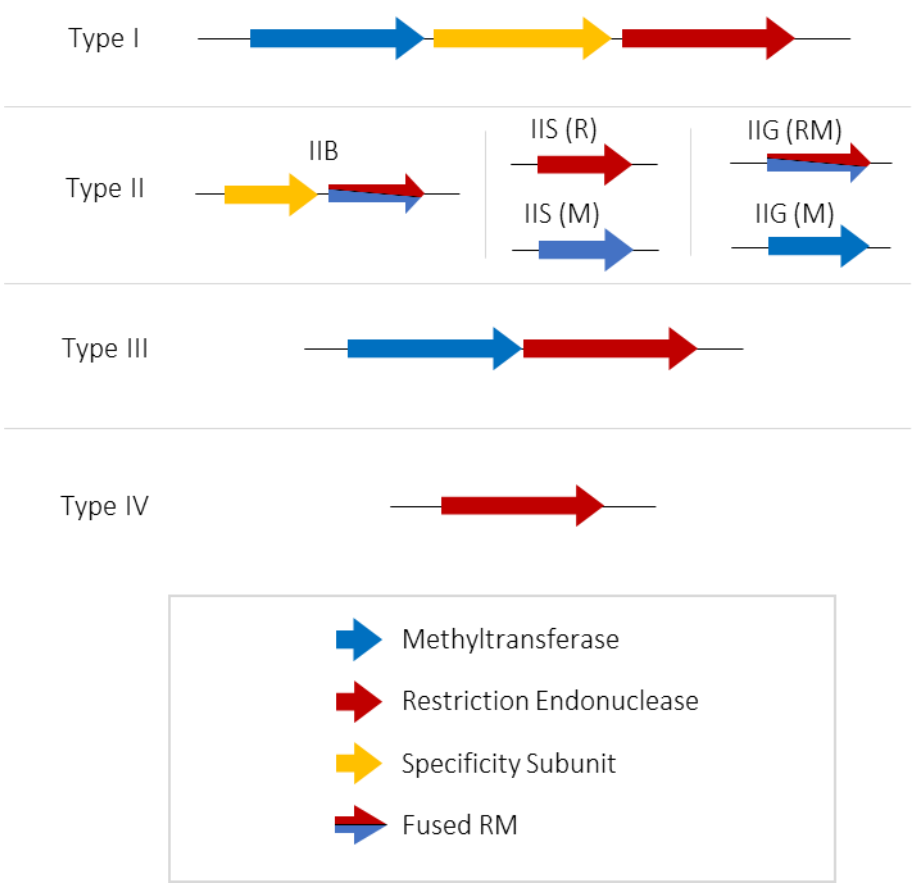


Figure 5.2.1 – Schematic of representative genomic loci for each RM system Type I-IV.

Methyltransferase encoding ORFs represented by blue arrow. Restriction endonuclease encoding ORFs represented by red arrow. Specificity determining subunits genes represented in yellow. Fused restriction endonuclease-methyltransferase genes are signified by the half red/blue arrow.

5.3.1.1. Types of RM system

RM systems are categorised into four distinct classes, Types I-IV. Type classification of a given system is based on subunit composition, target recognition motif, proximity of cleavage relative to target sequence, and the modification status of target substrates. Systems from all types recognise short specific DNA sequences, and perform endonucleolytic cleavage where the base modification status at target sequences fails to match system expectation. Figure 5.2.1 presents a schematic overview of a typical gene operon encoding for each of the four system types.

Type I systems are the most complex of the RM system types. These systems exist as multi-subunit complexes comprised of combinations of restriction, methylation, and specificity subunits. Together these subunits form functional enzymes which methylate at specific target sequences, but cleave DNA at random, far from the recognised sequences. Cleavage reactions require ATP and S-adenosyl-L-methionine (SAM). The three core activities of RM systems are most clearly exemplified by hetero-trimeric type I RM systems which are typically made up of three distinct subunits: a DNA methyltransferase (M subunit) for methylation; a specificity subunit (S subunit) for DNA sequence recognition and binding; and a restriction endonuclease (R subunit) for DNA cleavage.

Type II systems are a collection of largely unrelated enzyme systems, which differ from one-another to such an extent, that their rapid evolution exemplifies involvement in host-parasite interaction[217]. The majority of type II systems recognise palindromic sequences as a function of homodimeric binding, however others recognise continuous or discontinuous (gapped) sequences. Cleavage by Type II systems occurs at DNA within or nearby recognition sequences.

Type III systems exist as large multi-subunit RM enzyme complexes. Cleavage by these systems requires two unmodified recognition sequences in opposite orientation within a single DNA molecule, and cleave at defined points outside of their recognised target sequences[218]. This stringent requirement for two inversely oriented recognition

sequences ensures that type III systems do not cleave newly replicated, hemi-methylated, DNA.

The final class of RM system, Type IV, cannot be considered true RM systems because they do not possess methyltransferase components. These unique systems are comprised of one or two genes which recognise target sequences and, unlike for the previous classes, targets cleavage where methylation is present at these sites[86]. Type IV systems protect against methylated foreign DNA as well as maintaining the integrity of host methylation, self-targeting host DNA where expression of foreign MTases is detected[219].

5.3.2. Overcoming gene transfer barriers

The highly sequence specific nature of CRISPR/Cas systems means that they are rarely an issue in the initial establishment of gene transfer using standard plasmid-based transformations. In contrast the barriers posed by RM systems can be significant, and in some cases can be insurmountable without an engineered workaround. The shorter recognition sequences, generally conferring 4-8 bp in specificity[216], naturally occur through random chance with relatively high frequencies within typical DNA sequences. As a result, there is a reasonable likelihood that transforming DNA could contain one or more recognition sequences for host RM systems which would result in its destruction upon entry to the cell.

Most strains possess a unique methylation fingerprint as a result of high specificity and targeting diversity, which exist between different RM systems. The complexity of methylation patterns is compounded by the possibility that a given strain can harbour multiple RM systems simultaneously. This uniqueness is central to the functionality of the systems, ensuring that the offered protections are not bypassed by DNA originating from organisms presenting the corresponding methylation patterns. However, strain individuality has historically made evading said systems for the purpose of

transformation problematic, and means that there is no one size fits all approach to overcoming these barriers.

Broadly speaking, there are two choices available for overcoming RM barriers. First 'evasion', where DNA is camouflaged prior to transfer into the organism so as to be indistinguishable from native DNA, and thereby avoid engaging the RM systems[220]. Second 'inactivation', where the RM systems are actively disabled or removed from target organisms prior to gene transfer[221].

5.3.2.1. RM system evasion

Evasion allows research to be performed in an unmodified (WT) host strain, which is often preferable. However, it requires a thorough understanding of each RM system in the target strain and it is often time consuming to create the necessary tools/strains to generate effective DNA camouflage. Commercially available in-*vitro* methylation kits can similarly offer a basic level of methylation protection. This approach is appropriate in instances where targeting motifs of RM systems are known and can be blocked by blanket methylation of short and regularly occurring sequences[222]. However, in some cases excessive methylation can result in undesirable engagement of type IV RM systems if present in target organisms. Ultimately, any methylation method which relies on the introduction of imperfect and excessive methylation patterns, has the potential to inadvertently clash with other RM systems outside of those which it attempts to protect against.

5.3.2.2. RM system inactivation

Inactivation can be achieved simply by knocking out the restriction component of a relevant RM systems. This can be achieved relatively quickly in most organisms, but requires some level of DNA transfer prior to enable KO. Alternately inactivation can be achieved in a transient manner through methods such as brief heat-shock prior to

electroporation based transformations[223]. However, such transient inactivations are not applicable to all RM systems, nor are they inherently transferrable to new organisms.

5.3.3. Methods of Gene Transfer

5.3.3.1. Electroporation

Gene-transfer in modern Gram-positive research fields is predominantly achieved through either electroporation or conjugation. The efficacy of both methods is highly dependent on the target organisms, with neither method being universally effective. For example, for the 100% 16S homology bearing *C. autoethanogenum* and *C. ljungdahlii*, electroporation of the former was almost wholly ineffective while the latter is transformed in this manner routinely[224]. For these reasons, the method design was centred around building an approach which could accommodate both methods.

Electroporation based transformation approaches are typically the more popular method of plasmid transfer in Gram-positive bacteria. This is due primarily to higher efficiencies of plasmid transfer and independence from donor species compatibility issues[82]. However, electroporation is inhibited in a number of organisms, particularly in Gram-positive species as a result of thick peptidoglycan walls. In some cases this recalcitrance can be overcome through cell wall weakening[83]. Optimisation and finetuning of electroporation procedures is highly organism specific, and can represent a significant research burden where initial transformation efficiencies are low.

5.3.3.2. Conjugation

Conjugation is the horizontal transfer of DNA elements between cells *via*direct cell-cell contact mediated by pili[225]. Successful conjugation has a number of prerequisites with regard to the target organism, donor organism, and the plasmid. The donor cells must contain a fertility factor (F^+), containing all of the functional elements needed for

conjugal gene transfer. F^+ gene operons are typically located on large cryptic plasmids, such as the *P. mirabillis* derived pR702 used in this study[226]. In nature conjugation is used to transfer F^+ factor elements between cells, however in a research setting this machinery can be hijacked to transfer other plasmids present in the donor cells. Conjugal transfer of DNA occurs through a single stranded intermediate, generated by formation of a nick at defined origins of be transfer (*oriT*)[227]. This mechanism means that conjugation is attempted with any plasmid containing *oriT* in the F^+ donor strain. One major benefit to conjugation, in some cases, is evasion of RM systems as a result of the single stranded DNA intermediate used in transfer[228]. In such instances conjugation alone provides resistance to certain RM systems which would otherwise cleave transforming DNA. This protection is a result of the single stranded DNA intermediate which is transferred into target cells during transformation. Many, although not all, RM systems do not recognise either single stranded intermediates or subsequent hemi-methylated DNA as targets clevage. One of the key draw backs to conjugation is donor-target compatibility. The full extent of factors which effect this compatibility are not well studied, but include factors such as necessity for common permissive growth conditions required for efficacy.

5.3.4. Aims of this study

The aim of this study was to develop an improved method for establishing and/or improving gene transfer in recalcitrant microorganisms through the development of bespoke plasmid donor strains. The proposed method harnesses lambda-red mediated recombineering for the rapid engineering of *E. coli* donor strains, for the inducible expression of DNA methylation systems derived from the ultimate target strain. Methylation systems are employed in order to mimic the methylation patterns of a target strain in *E. coli* methyl donor strains to protect transforming DNA against host restriction systems.

Herein is described an improved workflow for establishing transformation in RM protected strains. The method centres around the development of bespoke methylation donor (MD) strains for camouflaging transforming DNA (Figure 5.2.2).

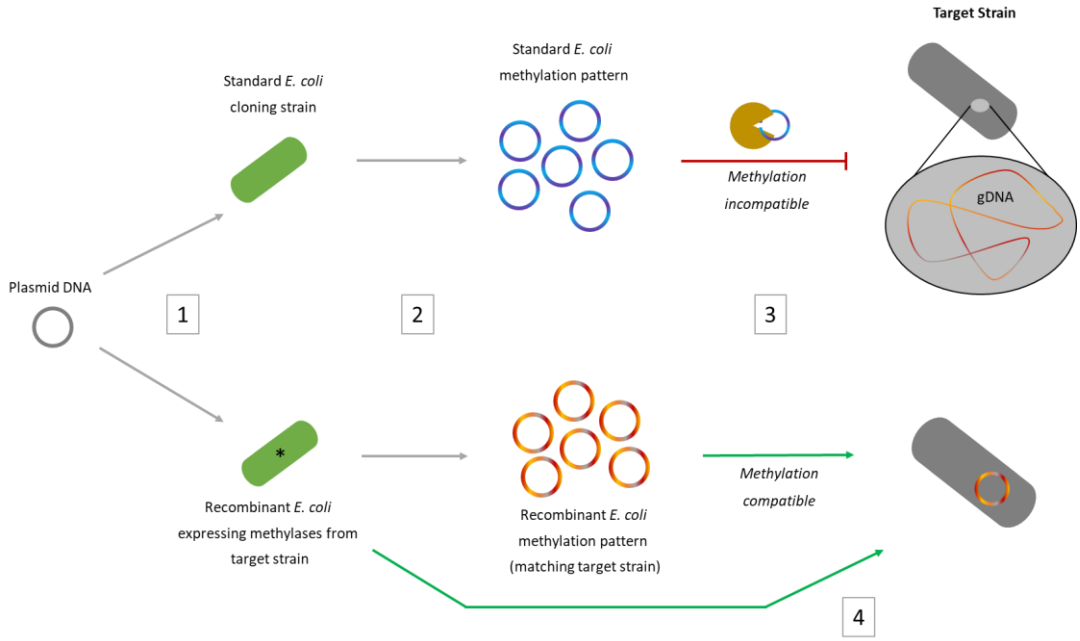


Figure 5.2.2 –The effects of methylation compatibility on transformation efficacy

A schematic overview of the stages of electro-transformation of target strains using plasmids prepped from standard cloning strains (top) and engineered methylation donor strains (lower). 1) Plasmid DNA is transformed into *E. coli* for amplification. 2) *E. coli* methylated DNA is extracted by miniprep. 3) miniprepped plasmids are transformed into a target strain. Where methylation patterns are incompatible with target RM systems (i.e., standard *E. coli* methylation patterns) transformation is blocked or efficiencies reduced as a result of endonuclease cleavage by RM systems. 4) Direct transfer from *E. coli* to target strain by conjugation. RM systems are not engaged with camouflaged plasmids extracted from engineered strains, so efficiencies are not impacted.

5.4. Results

5.4.1. Transformation Workflow

A six-part workflow for establishing or improving gene transfer into target microorganisms containing one or more active RM systems was developed (Figure 5.2.3). The workflow relies first on a complete annotated genome sequence (Figure 5.2.3.1), used in the identification of RM systems with the online tools and databases available at ReBase[133] (Figure 5.2.3.2). The methyltransferase enzymes, and where necessary specificity subunits, of relevant RM systems are then assembled into

'methylation cassettes' (Figure 5.2.3.3). This is achieved by PCR amplification of the appropriate genes from host genome and cloning them into holding vectors such as pRBV1 (Figure 5.2.6). Modified pRBV-x vectors are subsequently used as a template for a second PCR, generating a linear DNA fragment for lambda-red mediated recombineering into a purpose-engineered *E. coli* strain (Figure 5.2.3.4). The resultant recombinant *E. coli* can then be transformed with plasmids which are modified in *vivo* by the gene products of the recombineered methylation cassette (Figure 5.2.3.5). Following this transformation, transformation attempts into the target strain can be made by either direct conjugation, or by electroporation following appropriate plasmid extraction (Figure 5.2.3.6).

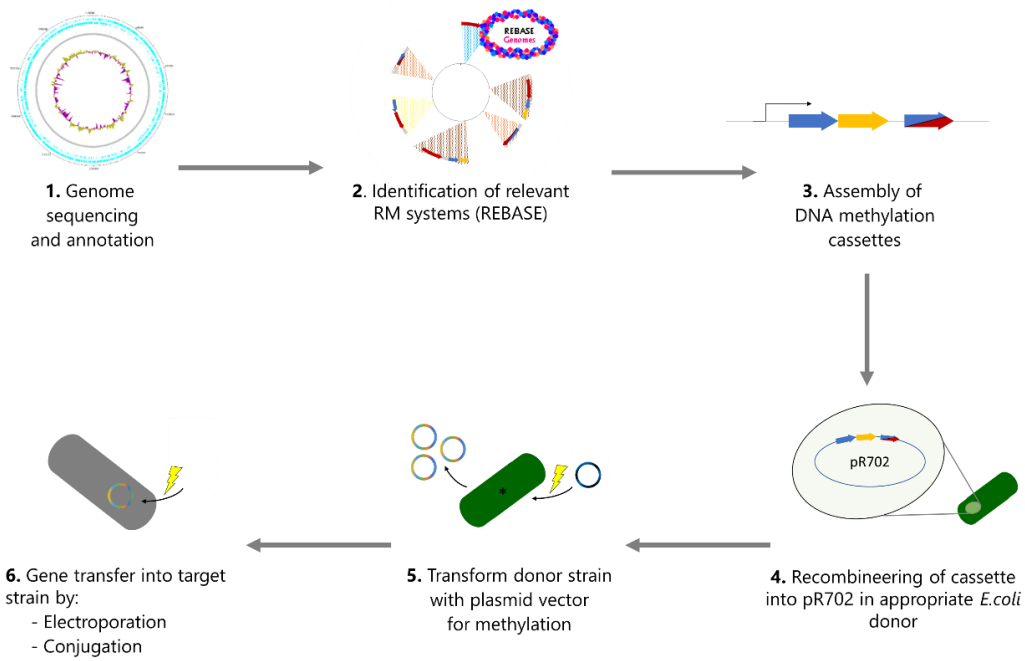


Figure 5.2.3 – Six-part workflow for the generation of methylation donor strains of *E. coli*

5.4.2. Genome RM System Identification

Prokaryotic genome sequences uploaded to NCBI are automatically analysed by ReBase and RM systems identified. RM data for four target organisms was located within the ReBase genome database (Table 5.2.1). Details were taken directly from the information pages of each organism, including where desired genome coordinates of

the appropriate genes against the genome sequence files available from NCBI. Relevant RM systems were identified as those containing both a restriction endonuclease and a methyltransferase. Orphan methytransferases, those not coupled to a REase, were deemed to have no relevance to transformation efforts and therefore disregarded during the development of methylation donor strains.

Four target organisms were initially selected as desirable targets for gene transfer establishment or efficiency improvements. *C. botulinum* 19397 and *C. carboxidivorans* P7 had never been transformed before within our group nor the published literature. *C. difficile* DH1916 had been transformed previously by conjugation, but at undesirably low efficiency. *E. limosum* ATCC 8486 had not been transformed prior to this study, but following successful development of an electroporation-based method, it was a strong target for improving transformation efficiencies.

The target strain of *C. difficile* in this study, *C. difficile* DH1916, did not have a published genome sequence in the NCBI database so could not be searched directly *via*ReBase. A partial genome sequence data, provided by Patrick Ingle (Ingle, P. SBRC Nottingham, Personal communication 2017), was used to align RM systems of DH1916 against that of R20291. BLAST alignment revealed 100% homology (Query cover: 100%, Identity: 100%, E-value: 0) in the DH1916 genome to the single Type I system highlighted by ReBase in R20291. Further BLAST alignment against the ReBase gold-standard enzyme database revealed no additional, REase containing, Type I-III systems in the DH1916 partial genome.

Two active RM systems, designated Orf717 and Orf 2811, were identified in ReBase for the *C. botulinum* strain 19397. Three additional Type II RM systems were highlighted in the ReBase genome analysis, however, none contained candidate REase components. As such, these systems were not deemed relevant to this project.

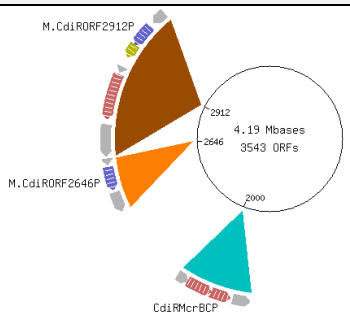
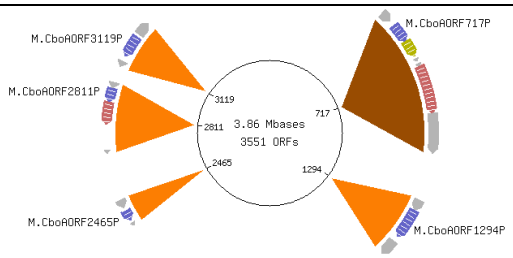
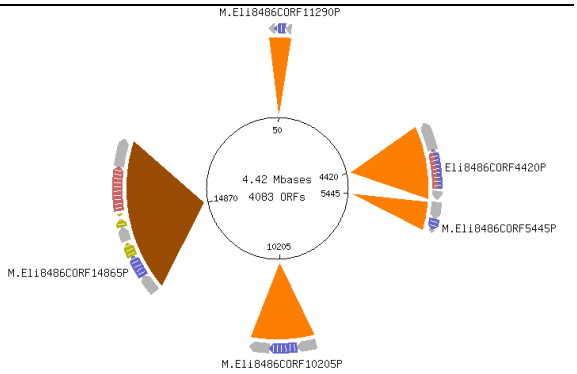
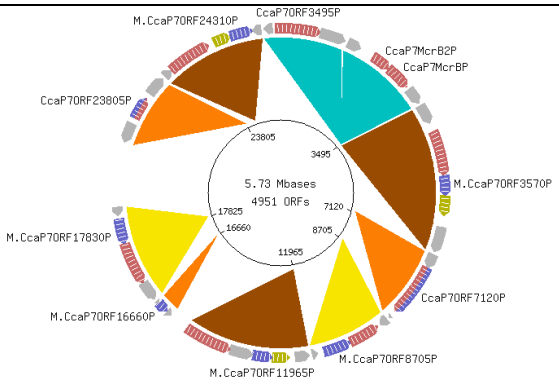
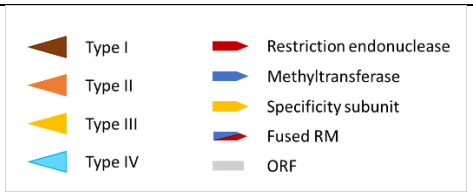
ReBase identified five RM systems in *E. limosum* ATCC 8486 including two REase containing systems. The first non-orphan system, designated Elim8486ORFA (OrfA),

was a fused Type II RM system. The second, designated Elim8486ORFD (OrfD), was a characteristic three-part Type I RM system, HsdRSM (Restriction endonuclease, Specificity subunit, and Methyltransferase).

A total of ten RM systems were identified in the genome of *C. carboxidivorans* P7. This included three Type I, three Type II, two Type II, and two Type IV RM systems. Only one of these systems was devoid of a putative REase containing subunit, Orf16660P.

Additional data was made available for analysis of the *E. limosum* and *C. carboxidivorans* genomes as a result of access to PacBio sequencing data. *E. limosum* ATCC 8486 PacBio data was generated as discussed in Chapter 3. The equivalent data for *C. carboxidivorans* P7 was kindly provided by Professor Yang Gu (SIBS, CAS, China).

Table 5.2.1 – RM system details from ReBase for four target organisms

Organism	RM genome atlas
C. difficile R20291 4 RM systems – active (orphan MTase) T I 1 T II 0 (1) T IV 2	
C. botulinum 19397 5 RM systems – active (orphan MTase) TI 1 TII 1 (3)	
E. limosum ATCC 8486 5 RM systems – active (orphan MTase) T I 1 T II 1 (3)	
C. carboxidivorans P7 10 RM systems – active (orphan MTase) T I 3 T II 3 (1) T III 2 T IV 2	
	

RM system summary for *C. difficile* R20291, *C. botulinum* 19397, *E. limosum* ATCC 8486, and *C. carboxidivorans* P7. RM system identification and atlas map generation was performed by ReBase[133].

5.4.3. Application of PacBio Data

Motif Summaries had been generated as standard output of the PacBio sequencing method, and these files were analysed for *E. limosum* ATCC 8486 and *C. carboxidivorans* P7. These files contained Motif-Strings, which represent repeatedly occurring patterns of base modification observed during the SMRT sequencing. Each identified motif was reflective of methylation activity by at least one RM system within the genome. Motif Summaries were analysed both manually and *via*the online ReBase pipeline, in order to assign each Motif-String specific RM systems.

Both manual and ReBase assignments relied on the same underlying principles. First, BLASTp alignment of RM systems against libraries of RM systems of known motif. Second, assignment of motifs to RM system classes, performed automatically by ReBase. Finally, process of elimination matching of RM systems to motifs where sufficient information was available. Tables for the analysis of Motif Summary files and putative RM system assignment are available in Table 5.2.2 (*E. limosum*) and Table 5.2.5 (*C. carboxidivorans*).

5.4.3.1. *E. limosum*

All base modifications recognized by the PacBio sequencing of the *E. limosum* ATCC 8486 genome were methylation of adenosine bases, resulting in N6-methyldeoxyadenosine (6mA). No 5-methylcytosine (5mC) or other expected modifications were observed, and no methyl-directed Type IV RM systems were identified. A comprehensive summary for the five identified RM systems is presented in Table 5.2.3.

Eli8486ORFAP (OrfA) was expected to be responsible for the recognition and methylation of the asymmetric GCGRAG sequence based on its similarity to the GCCRAG recognising Awo1030IV from *A. woodii*. Alignment of OrfA to this Type II-G methyltransferase revealed 71% identity from 99% coverage for the total length of both proteins. The system's REase Ga0111251_11257, was expected to operate in

conjunction with the larger methyltransferase-domain containing protein Ga0111251_11256, to complete a functional RM system. However, inverse orientation of the two genes ([–] Ga0111251_11256, [+] Ga0111251_11257) indicates that each may be subject to independent transcriptional regulation.

Table 5.2.2 - Methylated motifs detected in *E. limosum* ATCC 8486 with best candidate RM systems identified for each

Motif String (+ strand): Recognition Sequence	Type of RM system	% of occurrences methylated	Gene OID (IMG/M)	Locus Tag (Ga0111251_)	Assigned gene Nomenclature
CAAAAAR	II	98.3	**2653374577**	111421	Eli8486OrfB
CACNNNNRTANNG	I	99.7/99.7	**2653375521**	112370	Eli8486OrfD
GCGRAG	II	99.4	**2653373418**	11256	Eli8486OrfA

Assignment of Eli8486ORFBP (OrfB) as a type II category RM system was based initially on its sequence similarity to the *C. autoethanogenum* methyltransferase, M.Cau10061I (CLAU_1258). The BLASTp alignment revealed 50% identity over 98% coverage (E= $6e^{-178}$). ReBase reports M.Cau10061I Type II N6-adenine DNA methyltransferase, with a targeting specificity for CAAAAAR. This made OrfB the most likely candidate for the CAAAAAR methylation observed in the PacBio base modification dataset for the *E. limosum* ATCC 8486 genome. A N-6 methyltransferase domain was highlighted following alignment of the amino acid sequence to multiple databases, including UniProt, Expasy (Prosite) and EMBL, however no domains for endonuclease activity were reported for this protein.

From the three remaining methyltransferases which were identified, M.Eli8486ORFDP (OrfD) was anticipated to be the best match for CACNNNNRTANNG methylation. The remaining two genes, Eli8486ORFCP (OrfC) and Eli8486ORFGP (OrfG), both occur at prophage associated genomic loci. Due to this prophage association, as well as a lack of remaining unassigned methylation patterns, these two systems were expected to be inactive during standard cell cycles.

In the *E. limosum* ATCC 8486 genome sequence generated during this study (Chapter 3), OrfD was annotated as five adjacent ORFs. This five-part format was the result a somewhat unusual three-part specificity (S) subunit arrangement. The three identified S subunits (S1, S2, S3) appeared to be a result of an interruption causing rearrangement involving the S2 subunit. The rearrangement followed an internal inversion of the S2 region coupled to interruption by phage integrase gene Ga0111251_112373. This rearrangement is not present in the previously published partial genome sequence of *E. limosum* ATCC 8486.

Table 5.2.3 – RM systems in the *E. limosum* ATCC 8486 genome

Type	Subunit names	Gene	Methylation type	Recognition sequence*	Locus Tag (Ga0111251_)	Size (AA)
II-G gamma	M.Eli8486ORFAP	M	6mA	GCGRAG	11256	929
	R.Eli8486ORFAP	R			11257	213
II gamma	M.Eli8486ORFBP	M	6mA	CAAAAAR	111421	672
II alpha	M.Eli8486ORFCP	M		-	11459	251
I gamma	M.Eli8486ORFDP	M			112370	522
	S1.Eli8486ORFDP	S1			112371	153
	S2.Eli8486ORFDP	S2	6mA	CACNNNNNRTANNG	112372	237
	S3.Eli8486ORFDP	S3			112374	411
	R.Eli8486ORFDP	R			112375	1045
II beta	M.Eli8486ORFGP	M		-	113594	217

Details of five RM systems identified within the *E. limosum* ATCC 8486 genome. *Best candidate for the three observed modified motifs in *E. limosum*. Where a defined sequence is not reported (-), data was insufficient to link activity of the RM system to a modification.

In the *E. limosum* ATCC 8486 genome sequence generated during this study (Chapter 3), OrfD was annotated as five adjacent ORFs. This five-part format was the result a somewhat unusual three-part specificity (S) subunit arrangement. The three identified

S subunits (S1, S2, S3) appeared to be a result of an interruption causing rearrangement involving the S2 subunit. The rearrangement followed an internal inversion of the S2 region coupled to interruption by phage integrase gene Ga0111251_112373. This rearrangement is not present in the previously published partial genome sequence of *E. limosum* ATCC 8486.

A PCR screen was performed across the region in question to confirm the nature of this rearrangement (Figure 5.2.4). Screening was performed on samples of transformed *E. limosum* ATCC 8486, with the expectation that inactivation of the system, as a result of inversion, might be favourable for transformation. Screening indicated that even in transformed cells, both orientations of this internal S2 fragment were generally present within a population at any one time. This was suggestive of genetic heterogeneity at the OrfD locus within working populations of *E. limosum*. The three-part arrangement made initial BLASTp alignment of OrfD to the RM databases for the assignment of the CACNNNNRTANNG motif problematic. Following PCR confirmation of the presence of both OrfD configurations, the uninterrupted single ORF specificity subunit configuration (S4) was used for future alignments.

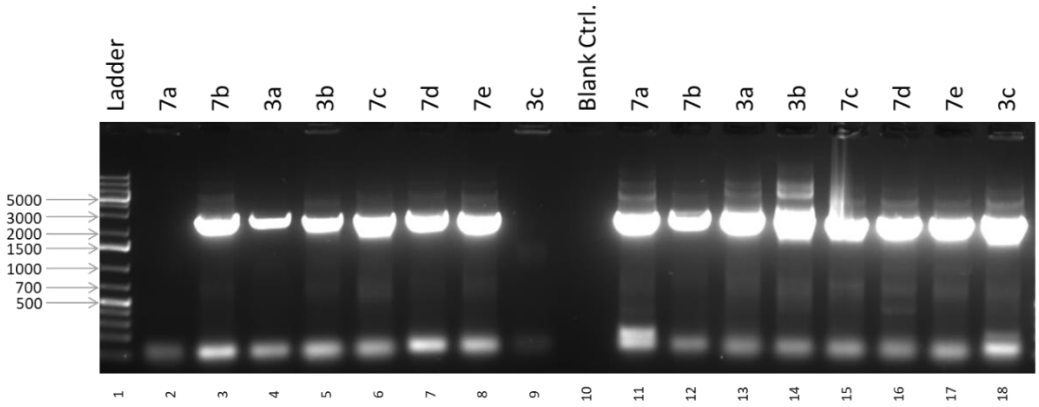


Figure 5.2.4 –PCR screening of OrfD specificity subunit rearrangements

Colony PCR was performed on eight transformed *E. limosum* ATCC 8486 samples. Primers [332][333] were used to screen for a single continuous OrfD gene (lanes 2-9). Primers [332][334] were used to screen for the interrupted variant of the RM system containing an internal inversion (lanes 11-18). Ladder: O'Generuler 1Kb plus. Both orientations were confirmed to be present in six of eight samples.

OrfD S4 was aligned using BLASTp against subunits with previously assigned/characterised recognition sequences in the ReBase Type I S subunit database (Table 5.2.4). None of the top hits from this alignment recognised motif sequences closely comparable to the unassigned

CACNNNNRTANNG motif of *E. limosum*. Nonetheless, OrfD was cautiously assigned to this motif as the most likely candidate. This was principally due to OrfD being the only Type I RM system and the motif being typical for bipartite target sequences of Type I systems[229].

Table 5.2.4 – BLASTp alignment of Elim8486ORFD S subunit against ReBase Type I S subunit database

Assignment (ReBase)	Organism	Recognition Motif	Identity	Coverage	E value
S.SspRAC02I	*Sinorhizobium* species RAC02	TCANNNNNGTGG	37%	75%	$E=2e^{-68}$
S.LhoZ1II	*Laribacter hongkongensis*	CCAYNNNNNTGA	37%	75%	$E=7e^{-65}$
S.Eco51542I	*Escherichia coli* CFSAN051542	TAACNNNNNTGA	51%	76%	$E=7e^{-64}$

5.4.3.2. C. carboxidivorans P7

A total of 10 distinct RM systems were identified within the *C. carboxidivorans* P7 genome, including just one putative orphan-methylase system (

Figure 5.2.5). All base modifications recognized by the PacBio sequencing of the *C. carboxidivorans* P7 genome were methylation of adenosine bases, resulting in N6-methyldeoxyadenosine (6mA) (Table 5.2.5). No 5-methylcytosine (5mC) modifications were observed. Two distinct methyl-directed Type IV RM systems, consisting of three separate open reading frames were identified. A comprehensive summary for the five identified RM systems is presented in Table 5.2.5 and Table 5.2.6. Assignment of RM systems was performed in the same manner as for *E. limosum*. A summary of the significant alignments used in these assignments is available in Table 5.2.7.

Table 5.2.5 – Methylated motifs detected in *C. carboxidivorans* P7 with best candidate RM systems identified for each

Recognition sequence - Motifs (+ strand)	Type of RM system	% of occurrences methylated	Genome coordinates - Methyltransferase	Locus Tag (Ccar_)	Nomenclature
CRAAAAR	II	97.0	5299217-5300845	23805	CcaP7ORF23805P
CAYNNNNNCTGC	I	99.7/99.5	2728619-2730115	11965	CcaP7ORF11965P
CCANNNNNNNNTCG	I	99.2/98.8	830425-831846	03570	CcaP7ORF3570P
AGAAGC	II	99.7	-	-	-
GATAAT	II	91.9	-	-	-
GCANNNNNNNTNNCG	I	99.6/100.0	5416106-5417695	24310	CcaP7ORF24310P

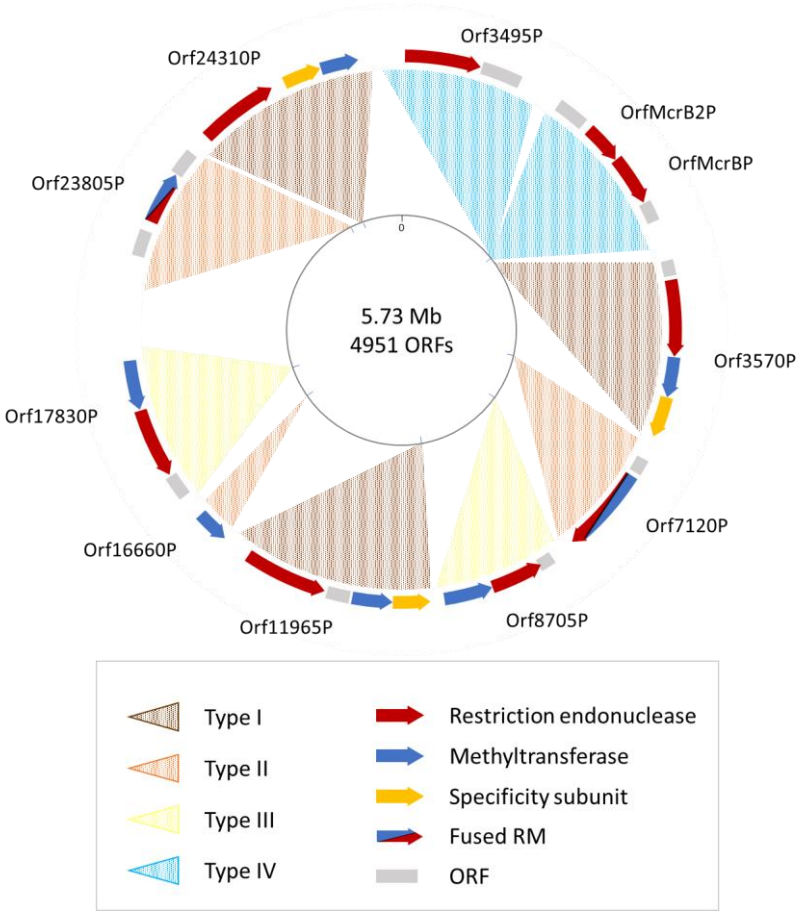


Figure 5.2.5 – *C. carboxidivorans* RM system genome schematic

A total of ten distinct RM systems were identified in the ReBase analysis. This includes three Type I systems, three Type II systems, two Type III systems, and Two Type IV systems. Schematic based on the automatically generated ReBase RM system RM system genome atlas.

Table 5.2.6 – RM systems in the *C. carboxidivorans* P7 genome

Type	Name	Gene	Methylation type	Recognition Sequence *	Locus Tag (Ccar_)	Size (AA)
II-G alpha	CcaP7ORF23805P	RM	6mA	CRAAAAR	23805	542
II-G alpha	CcaP7ORF7120P	RM	6mA	-	7120	1241
I gamma	R.CcaP7ORF11965P	R	6mA	CAYNNNNNCTGC	11975	1102
	M.CcaP7ORF11965P	M			11965	498
	S.CcaP7ORF11965P	S			11960	407
II	M.CcaP7ORF16660P	M	6mA	-	16660	254
III beta	M.CcaP7ORF17830P	M	6mA	-	17830	650
	R.CcaP7ORF17830P	R			17825	980
I gamma	M.CcaP7ORF24310P	M	6mA	GCANNNNNNNTNNCG	24310	529
	S.CcaP7ORF24310P	S			24305	402
	R.CcaP7ORF24310P	R			24300	1069
I gamma	R.CcaP7ORF3570P	R	6mA	CCANNNNNNNNTCG	03565	1090
	M.CcaP7ORF3570P	M			03570	473
	S.CcaP7ORF3570P	S			03575	447
III beta	M.CcaP7ORF8705P	M	6mA	-	08705	625
	R.CcaP7ORF8705P	R			08700	752
IV Methyl-directed	CcaP7McrB2P	R*	N/A	-	03520	490
IV Methyl-directed	CcaP7McrBP	R*	N/A	-	03525	568
IV Methyl-directed	CcaP7ORF3495P	R*	N/A	-	03495	1059

*Best candidate for four of the six observed modified motifs in *C. carboxidivorans* P7. Where a defined sequence is not reported (-), data was insufficient to link activity of the RM system to a specific modification.

RM system CcaP7ORF24310P is predicted to be responsible for the observed base modifications at GCANNNNNNNTNNCG sequences. This hypothesis is based on a strong homology between the system's specificity-delivering targeted recognition domain

(TRD), with the S.Spy8198I and S.Spy322I, responsible for modification of GCANNNNNNTTAA and GCANNNNNNRTTG sequences respectively. Similarity of the TRD of CcaP7ORF3570P to S.BceSVI (CCANNNNNNNCTTA) and S.BmeD34I (CCANNNNNNNCTC) indicates that this system is the best candidate for modifications at CCANNNNNNNNTCG sequences. CcaP7ORF11965P has been assigned to CAYNNNNNCTGC modifications primarily through process of elimination, based on being the last unassigned Type I system and characteristic Type I motif modification. However, no closely related S subunits could be used to confirm this assignment at the time of writing.

CcaP7ORF23805P assignment to CRAAAAR modifications was similarly based on alignment against the ReBase 'Gold Standard' library. Alignment yielded closest matches with three enzymes associated with CAAAAA motif modification, M.CpaYL7II; M.Cdi630V; M.Pdi9689I. Each of the closely aligned proteins contain a structural domain which is characteristic of the TaqI-C superfamily. It is this domain that is expected to confer specificity to the analogous recognition sequences for these enzymes.

Assignment of the RM systems responsible for the remaining GATAAT and AGAAGC modifications proved problematic primarily because of an inability to generate further strong alignments against pre-characterised RM systems. This issue was exacerbated by the highly variable and non-intuitive relationships between protein sequences and targeting domains of RM systems, which precluded further assignments being made with a high degree of confidence. The best deduction which could be made for base modifications at GATAAT sequences was that one or both of the type III RM systems is responsible. Weak alignment was observed between the M.CcaP7ORF8705P, and the methyltransferases M.HindVI (CGAAT) and M.Mha183IV (GTTAAT). It should be noted that the confidence scores for these systems were significantly poorer than was desirable for confident assignment RM systems. Alignment of the other Type III system, M.CcaP7ORF17830P, yielded reasonably strong homology to M.Btr190II (RGTAAT). As

such, putative assignment of the GATAAT motif was given to the M.CcaP7ORF17830P system.

Table 5.2.7 – Relevant alignment scores for the assignment of RM systems to modified DNA motifs observed in the *C. carboxidivorans* P7 genome.

	Score	% coverage	E	% identity
C car_23805 - II G alpha (CRAAAAR)				
M.CpaYL7II CAAAAA 775 aa	179	0.85	1E-49	0.31
M.Cdi630V CAAAAA 577 aa	143	0.96	1E-37	0.28
M.Pdi9689I CAAAAA 577 aa	142	0.96	1E-37	0.28
C car_11965 - I gamma (Probably: CAYNNNNNCTGC)				
S.SmaF1I CAYNNNNNNTCA 463 aa	167	0.99	5E-37	0.3
C car_24310 - I gamma (GCANNNNNNNTNNCG)				
S.Spy8198I GCANNNNNNTTAA 399 aa	303	0.99	3E-101	0.42
S.Spy360I GCANNNNNNTTAA 399 aa	298	0.99	2E-99	0.41
C car_3570 - I gamma (CCANNNNNNNNTCG)				
S.BceSVI CCANNNNNNNCTTA 476 aa	320	0.95	2E-106	0.4
C car_08705 - III beta (Possibly (weak): GATAAT)				
M.HindVI CGAAT 748 aa	253	0.69	9E-76	0.37
M.Mha183IV GTTAAT 711 aa	143	0.67	5E-37	0.29
C car_17830 - III beta (Possibly (strong): GATAAT)				
M.Btr190II RGTAAT 600 aa	442	0.97	1E-149	0.41

Alignments were based specificity subunits where available, and methyltransferase containing ORFs where these were not. Each of the displayed alignments arose within the top six best aligned as ranked by E value, against the ReBase Golden Standard RM library.

5.4.4. Cassette and Plasmid Design

5.4.4.1. R702 recombineering target

The *aadA1* gene conferring spectinomycin and streptomycin resistances of pR702 (Figure 7.3.1), was selected as the recombineering target locus for cassettes integration. A genome sequence for the pR702 containing *E. coli* strain CA434, was not readily available in published literature, so relevant pR702 sequences had to be inferred

from the reported origin species for the plasmid, *P. mirabilis*[226]. The sequence of the target locus (i.e., *aadA1* and surrounding region) was derived from the sequences of an *aadA1* containing plasmid pIV10[230], and the published genome sequence for *P. mirabillis* PmCHE[231]. The sequence of the locus was verified with colony PCR of CA434 using primer pair [338][339], designed as flanking primers to the prospective *aadA1* gene. The resultant PCR products were sequenced (Method 2.1.12) and a consensus of the target locus established.

5.4.5. Plasmid engineering

A recombineering base vector (RBV), pRBV1, was designed for the initial assembly and storage of recombineering cassettes (Figure 5.2.6). The vector was assembled through a series HiFi based assembly steps as described in Method 2.3.3, and modified with appropriate gene expression cassettes as described in Method 2.3.4. RBVs contained two homology arms (HAs) targeted to the 500 bp up and downstream of the aada1 gene locus in pR702. Between the HAs pRBV vectors contained an arabinose inducible promoter system (araC-P_{BAD}), a multiple cloning site (MCS) or inserted gene(s) for expression, and an antibiotic resistance marker. Plasmids were designed in a modular fashion to maximise ease of cloning and manipulation for flexibility in downstream applications.

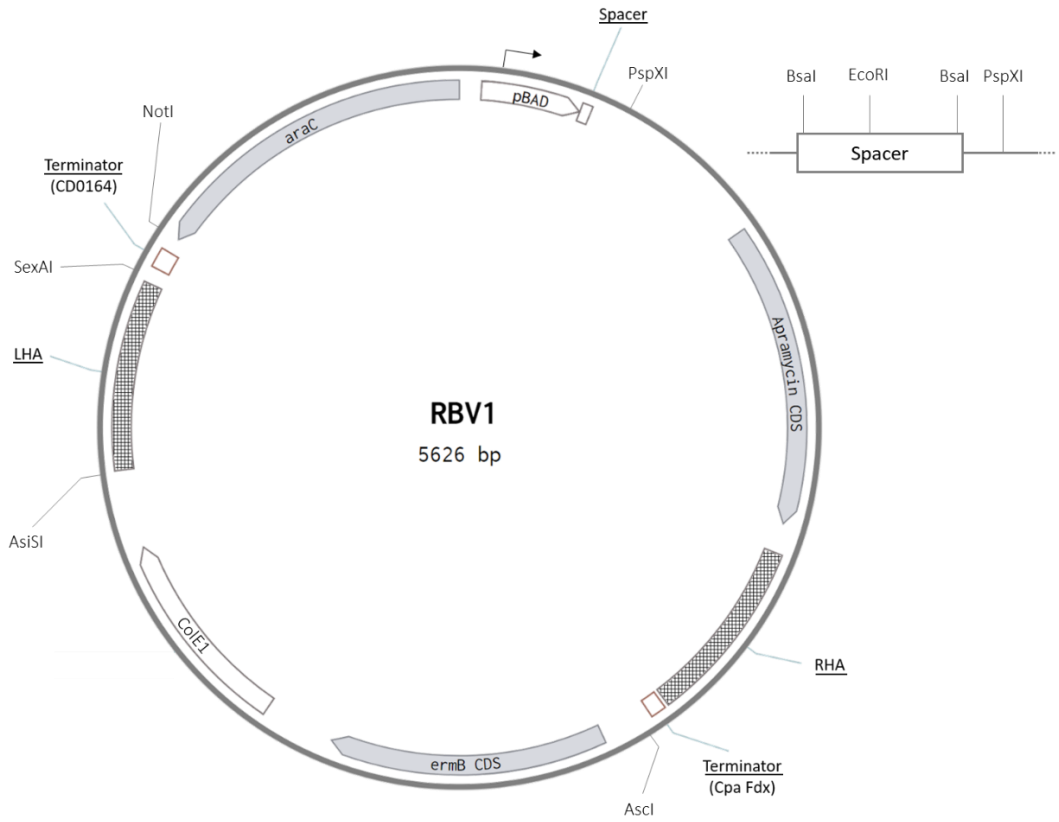


Figure 5.2.6 – pRBV1 holding vector map.

The RBV1 spacer region is the target insertion site for clockwise insertion of methylation cassettes, compatible with traditional restriction-ligation cloning, or GoldenGate and HiFi type assembly methods. Genes inserted at the spacer are controlled by an arabinose inducible promoter system (araC-P_{BAD}) and is insulated from transcriptional read through from the plasmid backbone by terminator CD0164. Two homology arms (LHA & RHA) are derived from flanking regions of AadA1 of the pR702 conjugation helper plasmid.

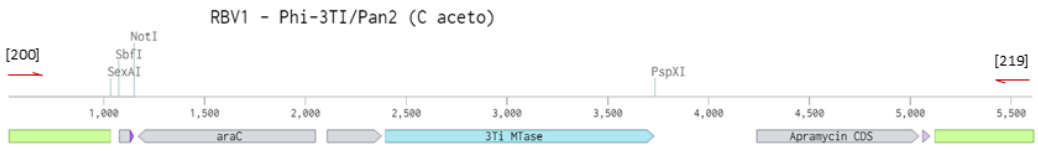


Figure 5.2.7 – Cassette example schematic (RBV-pAN2)

Example cassette amplified with primers [200][219] containing a methyltransferase gene (blue) derived from the *B. subtilis* 3TI phage. Cassette genes are flanked by left and right homology arms (green). Cassette expression is driven by araC-P_{BAD} (grey) and carries an apramycin resistance gene (grey) for selection.

5.4.6. Cassette design

Recombineering cassettes were PCR amplified from target RBV vectors using primer pair [200][219] (Figure 5.2.7). PCR amplified linear cassettes were transformed by electroporation into recombineering primed *E. coli* strains containing pR702, and pKD46cm. *E. coli* strains were primed for recombineering with L-arabinose induction of pKD46cm borne Lambda-red genes during competent cell preparation procedure.

Sixteen pRBV1 vectors were created in the course of this study. A summary of these vectors is presented in Table 5.2.8. This included two vectors for testing the induction of the *araC*-P_{BAD} expression system, using c*atP* and red-fluorescent-protein (RFP) as reporter genes.

Table 5.2.8 – Summary of recombineering vectors

Recombineering vectors	Description	Reference
pRBV1	Basic empty recombineering vector, no payload added at the RBV spacer (RBV1 control)	This study
pRBV-catP	*catP* gene inserted for promoter expression assessment at the RBV spacer locus	This study
pRBV-RFP	RFP gene inserted for promoter expression assessment at the RBV spacer locus	This study
pRBV-pAN2	RBV spacer carrying pAN2 derived methyltransferase gene, originally from Φ3TI of *Bacillus subtilis* for improved transformation of *C. acetobutylicum* ATCC 824	This study
pRBV-Cdif-HsdSM1	RBV spacer carrying methyltransferase and specificity subunits HsdSM1 derived from *C. difficile* DH1916	This study
pRBV-Orf717-2811	RBV spacer carrying methyltransferases Orf717-Orf2811 derived from *C. botulinum* A ATCC 19397	This study (Cloning and testing by Groothuis, D and Rodrigues, R. SBRC Nottingham)
pRBV-Eli-OrfA	RBV spacer carrying methyltransferase OrfA from *E. limosum* ATCC 8486	This study
pRBV-Eli-OrfD	RBV spacer carrying methyltransferase OrfD from *E. limosum* ATCC 8486	This study
pRBV-Ccar-30	RBV spacer carrying methyltransferase 30 from *C. carboxidivorans* P7	This study

pRBV-Ccar-32	RBV spacer carrying methyltransferase 32 from *C. carboxidivorans* P7	This study
pRBV-Ccar-34	RBV spacer carrying methyltransferase 34 from *C. carboxidivorans* P7	This study
pRBV-Ccar-36	RBV spacer carrying methyltransferase 36 from *C. carboxidivorans* P7	This study
pRBV-Ccar-38	RBV spacer carrying methyltransferase 38 from *C. carboxidivorans* P7	This study
pRBV-Ccar-40	RBV spacer carrying methyltransferase 40 from *C. carboxidivorans* P7	This study
pRBV-Ccar-42	RBV spacer carrying methyltransferase 42 from *C. carboxidivorans* P7	This study
pRBV-Ccar-44	RBV spacer carrying methyltransferase 44 from *C. carboxidivorans* P7	This study

5.4.7. Selection and Screening for Recombinant *E. coli*

Recombination events at the *aadA1* locus of pR702 caused a concurrent deletion of *aadA1* and insertion of recombinant gene expression cassettes. Thereby, putative recombinants could be positively selected for using apramycin selection, as well as negatively screened for loss of *aadA1* conferred resistances (spectinomycin). Where both screens met expectation, insertions at the desired locus could be inferred. Colony PCR-based screening of apramycin resistant colonies was performed as the primary confirmatory screen, using external flanking primer pair [338][339]. An example screen of recombineered *E. coli* rIS with seven different cassettes is shown in Figure 5.2.8.

Recombineering was attempted in parallel into three different *E. coli* hosts, each containing pR702 and pKD46cm. The three strains were, *E. coli* re-InterStellar, *E. coli* re-SExpress, and *E. coli* re-CA434. Recombination of cassettes was only successful in *E. coli* re-InterStellar, as measured by the production of colonies on apramycin containing selection plates, or lack thereof.

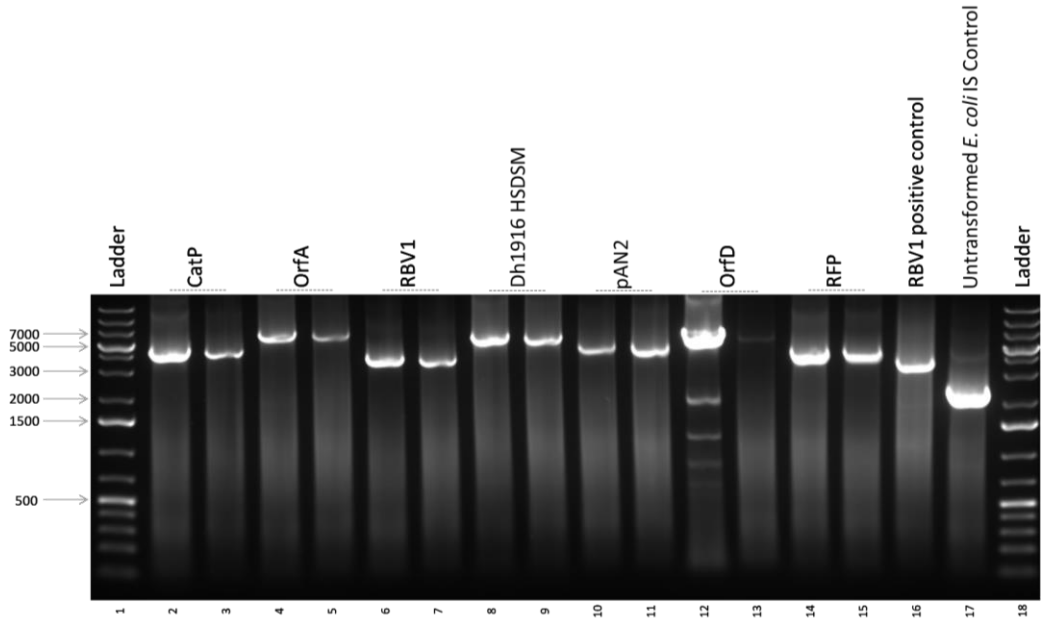


Figure 5.2.8 – PCR screening of recombineered *E. coli* re-InterStellar

Colony PCR screen using primer pair [338][339] of 16 colonies. Duplicate colonies were screened for 7 distinct recombineering transformations using cassettes: *catP*, OrfA, RBV1, DH1916 HsdSM, pAN2, OrfD, RFP. A positive control size marker was created by PCR using primers [200][219] with template plasmid pRBV1 (lane 16). An untransformed control was created with the [338][339] primer pair with an untransformed *E. coli* re-InterStellar colony as template. All band sizes matched the expected amplification lengths.

5.4.8. Proof of concept with *C. acetobutylicum* ATCC 824

The recombinant strain re-InterStellar-pAN2 (rIS-pAN2) was selected for the primary proof of principle application of the developed method (Figure 5.2.9). Transfer of plasmid pMTL83151 extracted from rIS-pAN2 methylation donor strain was >30 fold more efficient than comparable non-methylating strains. Furthermore, the rIS-pAN2 strain enabled equivalent transformation efficiencies to those achieved using the traditional dual plasmid approach, with *E. coli* Top10 harbouring the pAN2 plasmid[94]. Transformation improvements using rIS-pAN2 and Top10-pAN2 over rIS-RBV1 were statistically significant to a 99.99% CI ($P < 0.0001$). No statistically significant difference was found between Top10-pAN2 and rIS-RBV1 transformation efficiencies.

The efficiency of transformation at different concentrations was tested using the rIS-pAN2 plasmid (Figure 7.3.2). Even in the absence of induction (0 mM arabinose) transformation rates were significantly higher than the rIS-RBV1 control ($P = 0.00015$).

The difference between uninduced and 0.1 mM was statistically relevant ($P < 0.001$). However, the difference between the efficiencies observed at concentrations 0.1, 1, and 10 mM was not statistically significant when measured by One-way ANOVA ($P = 0.2218$).

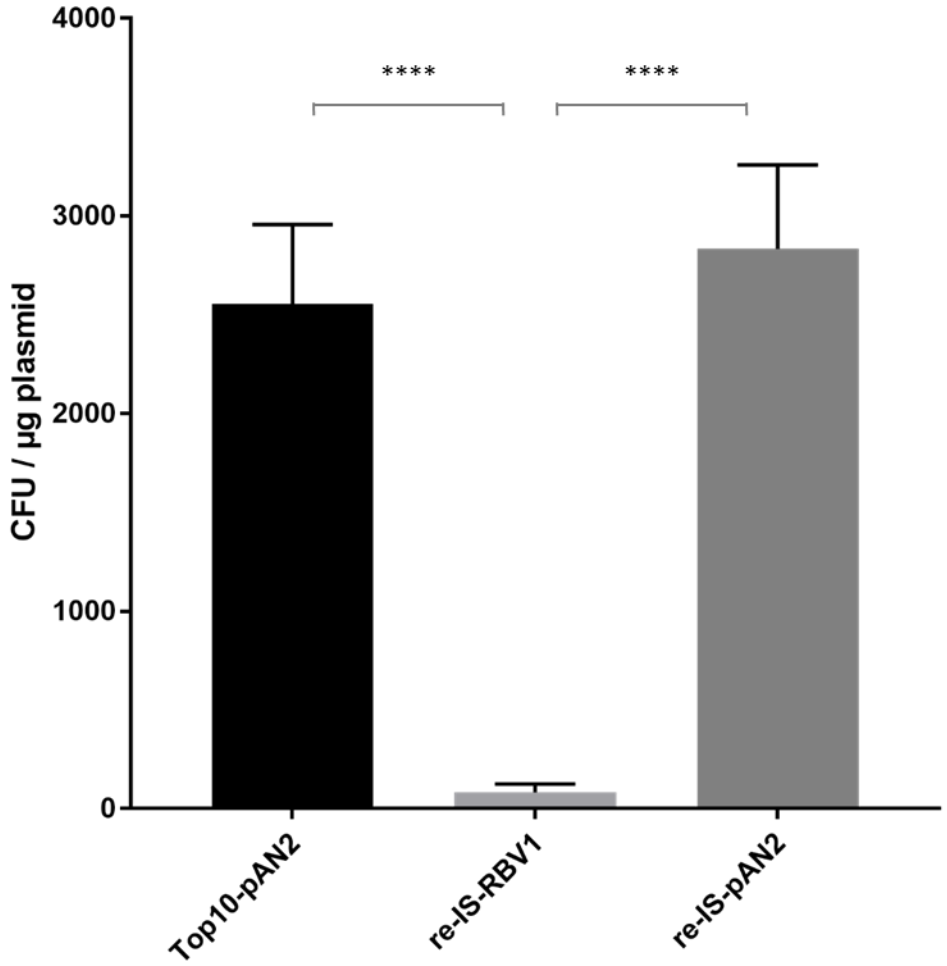


Figure 5.2.9 – Engineered methylation donor to *C. acetobutylicum* ATCC 824 electroporation efficiency comparison

Transformation efficiency of pMTL83151 prepared and extracted from three different *E. coli* strains and transformed into *C. acetobutylicum* by electroporation. *E. coli* strains were grown in parallel in LB media under maintained chloramphenicol selective pressure. Both recombinant-InterStellar (rIS) strains (rIS-RBV1 and rIS-pAN2) were cultured in the presence of 1mM L-arabinose for activation of the heterologous methylation cassette driven by the araC-PBAD promoter system. Error bars represent the standard error of the mean (SEM) for each sample. Data for Top10-Pan2 prepared plasmids is from triplicate transformation of *C. acetobutylicum*, while data for each rIS strains represents the respective average values from four distinct triplicate transformations of *C. acetobutylicum*. **** represents statistical significance of $P < 0.0001$ between indicated samples.

5.4.9. RFP promoter assay

An RFP based promoter assay was performed to confirm the efficacy of the araC-P_{BAD} promoter system used in the recombineered cassettes. Testing was performed using the recombineered rIS-RFP strain with differing levels of L-arabinose induction. In order to align testing with the conditions most relevant to the project, the strength of fluorescence was measured at 15.5 hours. This time interval reflected the RFP concentration following an overnight growth cycle, as would be the case for methylase expressing cassettes prior to plasmid extraction or conjugation initiation procedures.

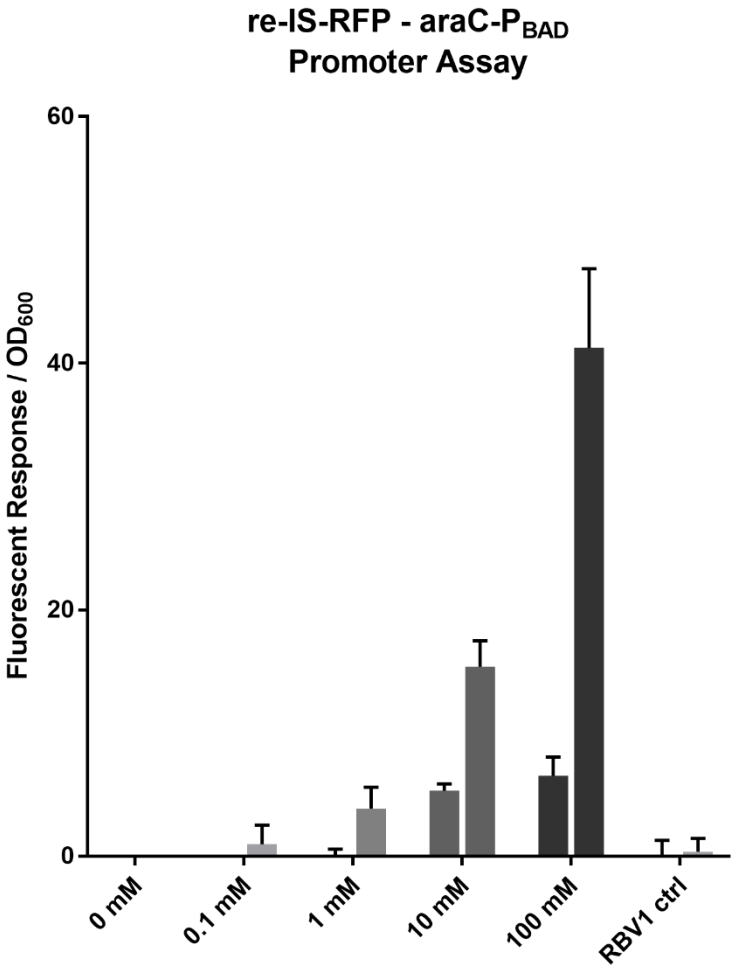


Figure 5.2.10 – RFP promoter assay for recombinant araC-P_{BAD} cassette

Re-IS-RFP samples cultured in 96 well plate format with constant shaking at 30 °C with varying levels of L-arabinose induction. Bar pairs represent the average fluorescent response adjusted for OD_{600} for triplicate samples at 15.5 hour (left) and 28 hour timepoints (right) for each sample. The RBV control values are representative of six replicates. Error bars represent the standard deviation of fluorescent response from triplicate samples of induced rIS-RFP and the six replicates of the RBV1 control sample.

For all samples containing less than 10 mM L-arabinose, fluorescent response was negligible and not statistically significant in relation to control samples at the 15.5-hour timepoint. Both 10 and 100 mM induced samples were higher than control and relevant to a 99.9% CI ($P < 0.001$). The difference between 10 and 100 mM was not statistically relevant at this timepoint. By 28 hours, 1 mM induction also showed a statistically relevant response ($P < 0.01$) over the RBV control sample. The disparity between 10 mM and 100 mM samples had notably increased by the later timepoint and became statistically relevant to a 99.5% CI ($P < 0.005$).

5.4.10. Transforming *C. difficile* DH1916

As a first complete test of the complete pipeline *C. difficile* DH1916 was selected as the target organism. *C. difficile* DH1916 contained two relevant RM systems, a Type I system designated HsdRSM1, and a Type IV system, *McrBC*1. Due to the targeting nature of Type IV systems in actively cleaving incorrectly methylated DNA, this system could not be actively protected against with additional methylation. However, the natural background of *E. coli* methylation donors was critical. Although the precise target motif of Mcr*BC*1 was not known, the system was known to be cytosine specific in nature, by designation as a Modified-cytosine-restriction (Mcr) system. As such, strains with DCM^+ genotypes, such as *E. coli* CA434, were not to be deemed appropriate backgrounds for this strain. An active Type I system was identified in the strain, designated HsdRSM1. The Type I system HsdRSM1 was putatively assigned to the targeting motif CANNNNNNNTAAAG. This assignment was based on ReBase analysis records available for an equivalent system with 100% homology identified in the PacBio sequenced *C. difficile* G46[232]. The methyltransferase and specificity subunits of the system, HsdS and HsdM were HiFi assembled into pRBV-HsdSM1 (Table 5.2.8).

The methylation cassette was amplified from pRBV-HsdSM1 with primers [200][219], and recombineered into *E.coli* rIS-HsdSM1. The recombinant pR702-HsdSM1 was transferred from *E. coli* rIS-HsdSM1 into *E. coli* Express, to create rSExpress-HsdSM1.

Transfer was performed by conjugation (Method 2.13.3) followed by apramycin selection on minimal media plates lacking proline to select against InterStellar. An equivalent *E. coli* SExpress-RBV1 strain was produced in the same manner as a recombineered pR702 control. To assess the efficacy of the HsdSM1 cassette in SExpress-HsdSM1, three *E. coli* SExpress strains were transformed with pMTL83151 for subsequent conjugations into *C. difficile* DH1916. *E. coli* SExpress and *E. coli* SExpress-RBV1 were negative controls for conjugation, for a standard pR702 and an RBV1 cassette (empty) recombineered pR702, respectively.

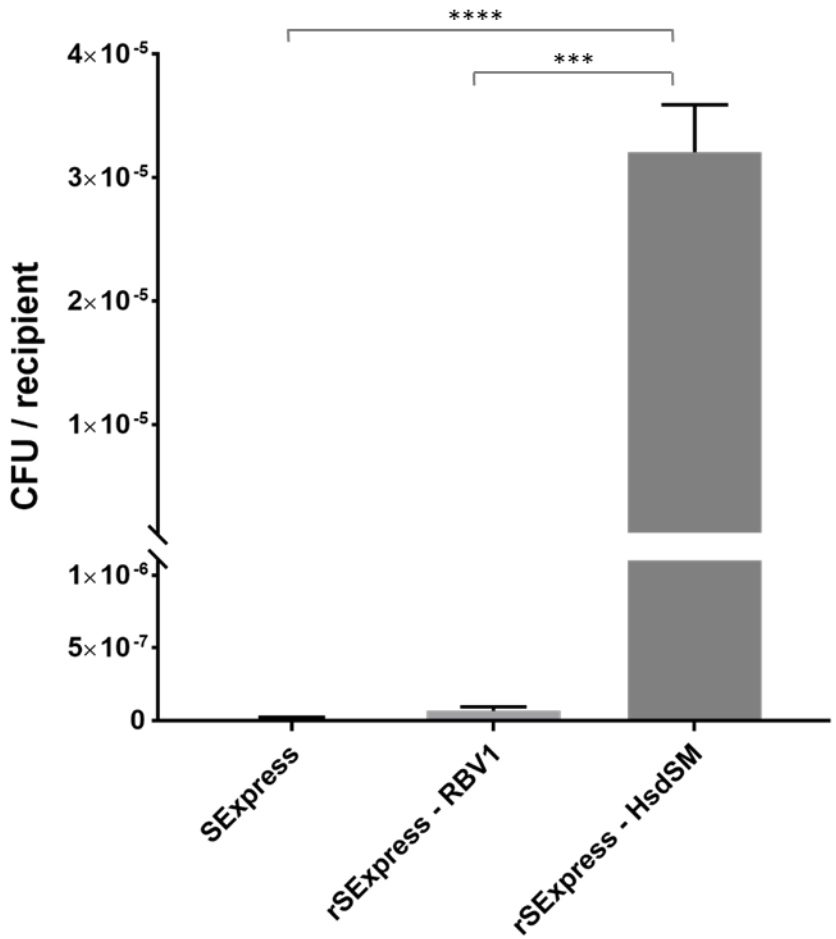


Figure 5.2.11 – Engineered methylation donor to *C. difficile* DH1916 trans-conjugation efficiency comparison

Methylation donor strain rSE-HsdSM produced significantly higher conjugation efficiencies into *C. difficile* DH1916 over SExpress or a recombinant control strain strain rSE-RBV1. Columns represent the average number of trans-conjugants per recipient cell from triplicate biological samples. Error bars represent the standard deviation of the triplicate values. *** and **** represents statistical significance of $P < 0.001$ and $P < 0.0001$ respectively between indicated samples respectively.

Triplicate conjugations into *C. difficile* DH1916 for each pMTL83151 harbouring *E. coli* strain were performed. *E. coli* replicates were biologically independent from the point of triplicate overnight starter cultures, which were mated with *C. difficle* recipient from a single starter culture.

Conjugation efficiencies were measured as a function of colony forming units (CFU) on thiamphenicol and D-cycloserine selection plates, against neat *C. difficile* counts with only D-cycloserine selection. Efficiencies from *E. coli* rSExpress-hsdSM1 were >450 times greater than those seen with either control strain, and significant to >99.98% CI ($P < 0.0002$). In practical terms, this improvement was an increase in CFU from <75 to >38000 CFU per conjugation. The difference between control strains suggested a 5.7-fold increase in the efficiency observed using the recombineered pR702 in rSExpress-RBV1 over standard SExpress ($P = 0.03$). However, low actual CFU data from selection plates of both controls should be acknowledged when interpreting the relevance of the inter-control variability, 0-10 colonies per conjugation.

5.4.11. Evidence for methyltransferase activity

In order to demonstrate that the observed efficiency improvements correlated with adenine methylation by HsdSM1 cassettes a South-Western blot analysis was performed. Two methyltransferase expressing cassettes were tested, RBV-pAN2 and RBV-HsdSM1, alongside two control samples. A positive control for adenine methylation was pMTL83151 plasmid extracted from *E. coli* CA434 (Dam^+/Dcm^+). A negative control was pMTL83151 extracted from the methylation deficient *E. coli* InterStellar.

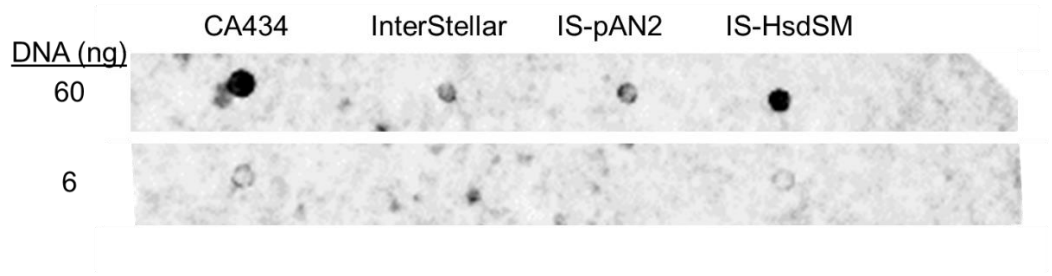


Figure 5.2.12 - Dot blot for methyltransferase activity in recombinant *E. coli* InterStellar-HsdSM

DNA was extracted and probed for adenine methylation with antibodies raised against methylated adenine bases (6mA). 6mA methylated DNA (lane 1) was extracted from the DAM+/DCM+ genotyped *E. coli* CA434. Unmethylated plasmid DNA (lane 2) was extracted from the DAM–/DCM– *E. coli* InterStellar. Plasmids extracted from recombinant *E. coli* IS-pAN2 (lane 3). Plasmids extracted from recombinant *E. coli* IS-HsdSM (lane 4).

Plasmid DNA extracted from *E. coli* InterStellar harbouring pRBV-HsdSM1 generated more than a 2-fold greater fluorescent response over plasmids extracted from non-6mA methylating negative control strains, InterStellar and IS-pAN2 (Figure 5.2.12 & Table 5.2.9). Non-specific binding to unmethylated DNA and non-6mA methylated DNA was observed in all 60 µg/ml samples to a baseline of 0.0233 (± 0.0014).

Table 5.2.9 – Fluorescence intensity values for South-Western Dot Blot of extracted plasmids

Strain	Fluorescent Response Concentration (µg/ml)	
	60	6
CA434	0.0851	0.0225
InterStellar	0.0223	-0.001
IS-pAN2	0.0244	-0.003
IS-HsdSM	0.0509	0.003

DNA extracted from four different *E. coli* background strains was probed with mouse-antibodies raised against 6mA nucleotides. A secondary fluorescent goat anti-mouse antibody was used to quantify the binding of the primary antibody, detected using the LiCor Odyssey FC imaging system with 120 second exposure at a wavelength of 800 nm. Fluorescent Values were automatically generated at manually defined dot locations using LiCor Image Studio software.

5.4.12. Transforming *C. botulinum* 19397

Following the significant improvement to the conjugation efficiencies of *C. difficile* with the engineered *E. coli* rIS-HsdSM1 methylation donor strain, the same approach was attempted in the previously untransformed strain *C. botulinum* 19397. Two relevant RM systems were identified in the target strain. A Type II system designated Orf2811, and a Type I system designated Orf717. The methylation and specificity subunits of Orf717, and the fused Type II system Orf2811 were assembled into the pRBV1 vector by three-part HiFi assembly. The methylation cassette (Figure 5.2.13) was subsequently amplified with primer pair [200][219] and transformed into recombineering primed re-InterStellar. Recombinant *E. coli* rIS-Orf717/2811 was then transformed with pMTL83151 and tested for conjugation into *C. botulinum*.

All cloning and conjugation steps for *C. botulinum* 19397 were carried out by Raquel Rodrigues and Daphne Groothuis (SBRC Nottingham, 2017).

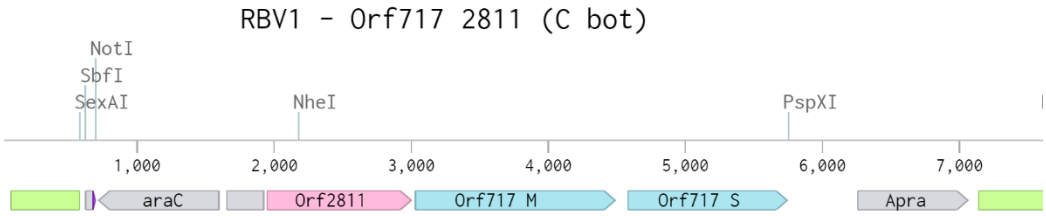


Figure 5.2.13 – Methylation cassette for *C. botulinum* ATCC 19397

Standard RBV cassette containing methylation components from two RM systems. A fused Type II system, Orf2811 (pink). Methyltransferase and specificity subunits of Type I system, Orf717 (blue). Cassette genes are flanked by left and right homology arms (green). Cassette expression is driven by araC-P_{BAD} (grey) and carries an apramycin resistance gene (grey) for selection.

Transformation was made possible for the first time in *C. botulinum* 19397 using the engineered rIS-Orf717/2811 methylation donor strain (Figure 5.2.14). Conjugation efficiency was statistically relevant to a >99% CI (P = 0.006). Neither conjugation using the *E. coli* strains, CA434 and rIS-RBV1 resulted in the formation of colonies on thiamphenicol + D-cycloserine selection plates. The methyl donor strain resulted in a triplicate average of 188 (± 50) CFU per conjugation.

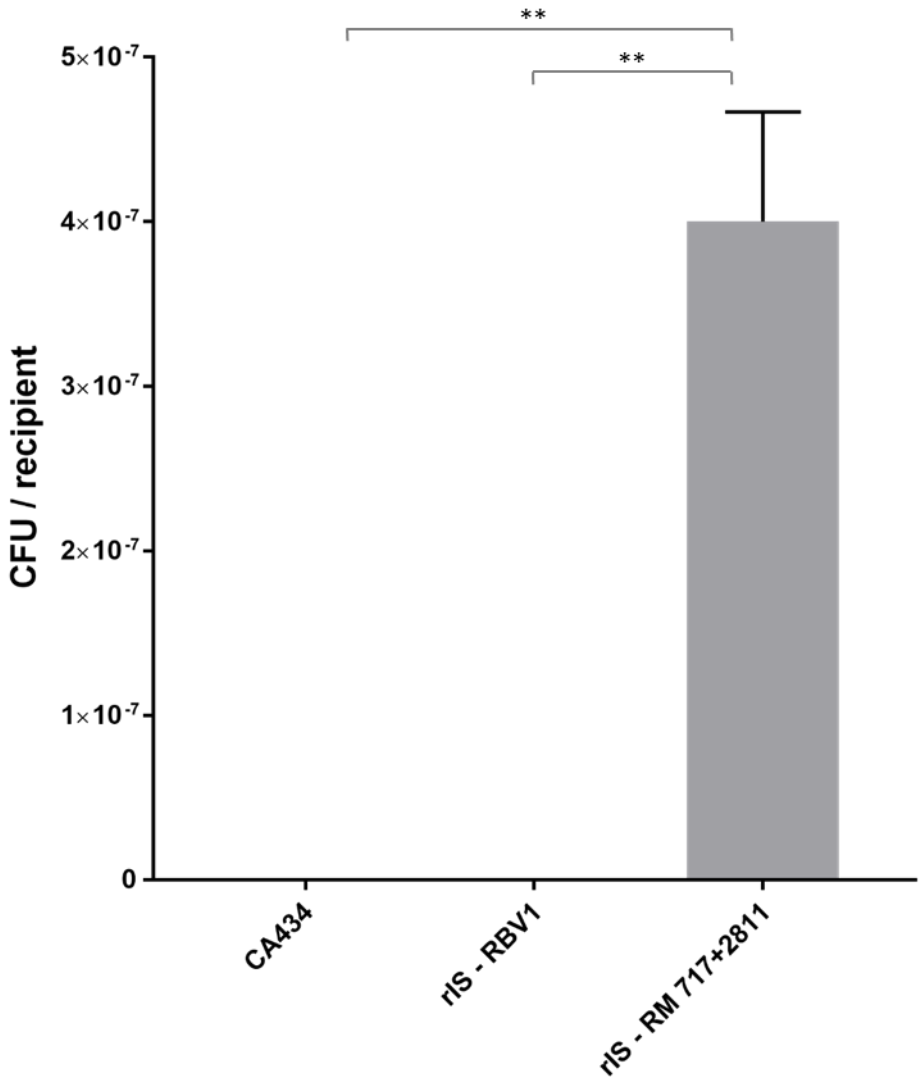


Figure 5.2.14 – Engineered methylation donor to *C. botulinum* ATCC 19397 trans-conjugation efficiency comparison

Methylation donor strain harbouring cassette expressing MTases from two distinct RM systems enabled DNA transfer into *C. botulinum* 19397. Trans-conjugation efficiency using three different conjugal strains: Left – *E. coli* CA434; Middle – recombinant InterStellar with empty cassette; Right – recombinant Interstellar expressing two MTase systems, Orf717 and Orf2811. Columns represent the average number of trans-conjugants per recipient cell from triplicate samples. Error bars represent the standard deviation of the triplicate values. ** represents statistical significance of P < 0.01 between indicated samples.

5.4.13. Improving *E. limosum* ATCC 8486 transformation efficiency

Two methylation donor strains were created for *E. limosum* ATCC 8486. First, for the Type I RM system, OrfD. Second, for the Type II RM system, OrfA. Following recombineering, both pR702 systems were transferred into *E. coli* Express for plasmid preparation stages. The process for testing the efficacy of these donor strains was the same as that which was used for testing rIS-pAN2 prepared pMTL83151 for electroporation of *C. acetobutylicum*.

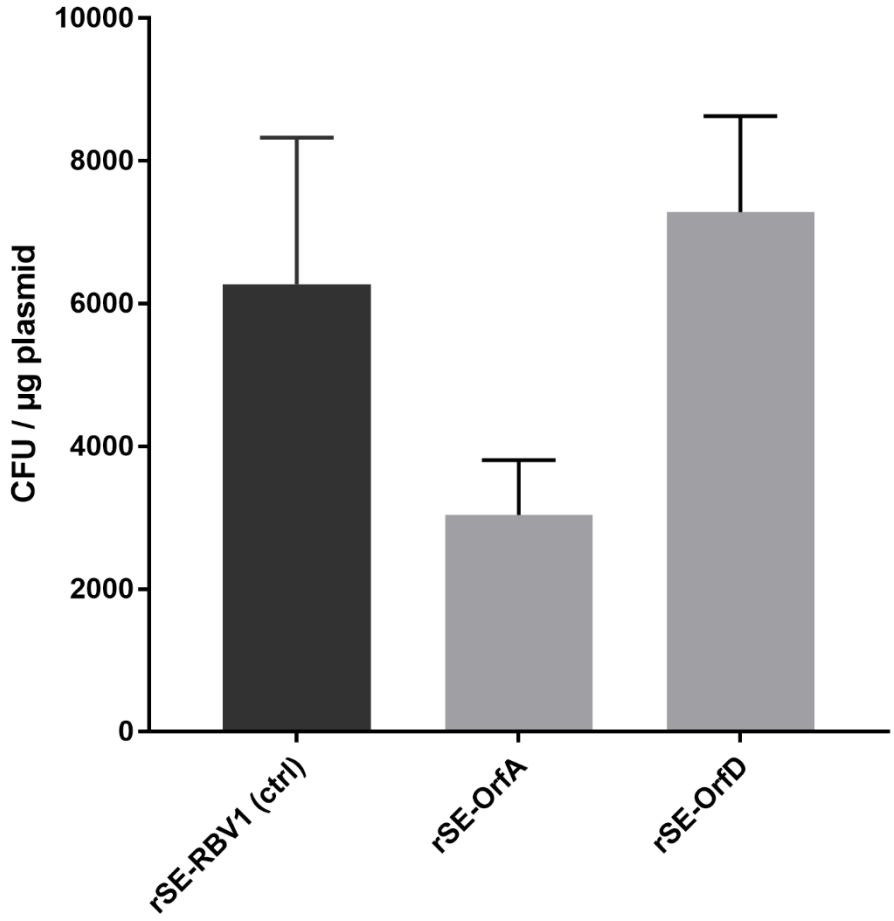


Figure 5.2.15 – Engineered methylation donor to *E. limosum* ATCC 8486 electroporation efficiency comparison

Transformation efficiency of pMTL83151 prepared and extracted from three different *E. coli* strains and transformed into *E. limosum* by electroporation. Methyltransferase containing Express strains harbouring recombinant pR702 (rSE-OrfA and rSE-OrfD) were cultured in the presence of 10mM L-arabinose for induction of the methylation cassette. Each column represents the average values from two distinct transformations of *E. limosum*, each performed in triplicate. Error bars represent the standard error of the mean (SEM) for each sample.

Transformation efficiencies into *E. limosum* ATCC 8486 were not improved as a result of expression of either OrfA or OrfD methyltransferase cassettes in plasmid preparation strains. The presence of OrfA appeared to have a negative effect on the efficiency of transformation with pMTL83151, however the difference against the rSE-RBV1 control was not statistically significant to a CI >95% (P > 0.095).

Alignment of the PacBio derived RM target motifs against transformed plasmids was performed. These alignments indicated that for pMTL83151 it should be expected that 2 sites would be protected with rSE-OrfD and 5 with rSE-OrfA.

Table 5.2.10 – *E. limosum* RM targeting motif alignment to pMTL83151 and pMTL86151

Sites in pMTL83151	Count	System
CAAAAAR	5	OrfB
CACNNNNRTANNG	2	OrfD
GCGRAG	5	OrfA
Total:	12	

Sites in pMTL86151	Count	System
CAAAAAR	8	OrfB
CACNNNNRTANNG	2	OrfD
GCGRAG	5	OrfA
Total:	15	

PCR amplification of OrfA and OrfD from the methylation donor cells harbouring pMTL83151 did not indicate any loss in sequence integrity of the MTases which might explain their lack of efficacy. However, gel separation of the miniprepped samples indicated variability in the content of the miniprepped samples which was not detectable by nanodrop during normalisation to 40 ng/µl performed prior to electroporation. The gel revealed a different banding pattern in OrfA, with increased intensity at higher weight bands. This may be indicative of reduced pMTL83151 extraction and increased extraction of the pR702 vector.

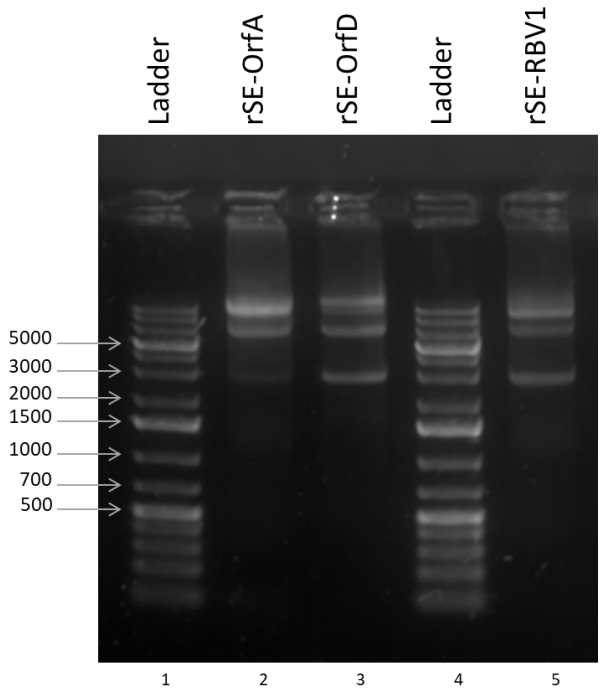


Figure 5.2.16 – Gel analysis of miniprep variability from methylation donors expressing OrfA and OrfD

Undigested miniprep product from three methylation donor strains run on a 1% agarose gel. 5 μl of miniprep normalised to 40 ng/μl was run for 40 minutes at 90 volts. Ladder: O'Generuler 1 kb plus (lanes 1 and 4) 2 μl and 5 μl. Plasmid pMTL83151 miniprepped from rSE-OrfA (lane 2), rSE-OrfD (lane 3), rSE-RBV1 control (lane 5).

5.4.14. Groundwork for *C. carboxidivorans* transformation

The scale of RM system protection in *C. carboxidivorans* P7 was unparalleled in this study, and notable even in the wider context of all available prokaryotic genomes. In total nine RM systems required avoidance, including two Mcr Type IV systems. The seven remaining non-orphan Types I-III systems accounted for 6mA methylation at six recognition motifs. There was not enough information available to confidently assign six systems to six motifs, so expression of MTases from all systems was deemed the most appropriate course of action.

To narrow down the functional systems, eight methyltransferases designated 30, 32, 34, 36, 38, 40, 42, and 44 were cloned into the pRBV1 spacer region. Plasmids were transferred into the methylation deficient *E. coli* InterStellar strain, and grown overnight with 10 mM L-arabinose induction. Plasmids were subsequently extracted and probed with SouthWestern blot as performed previously.

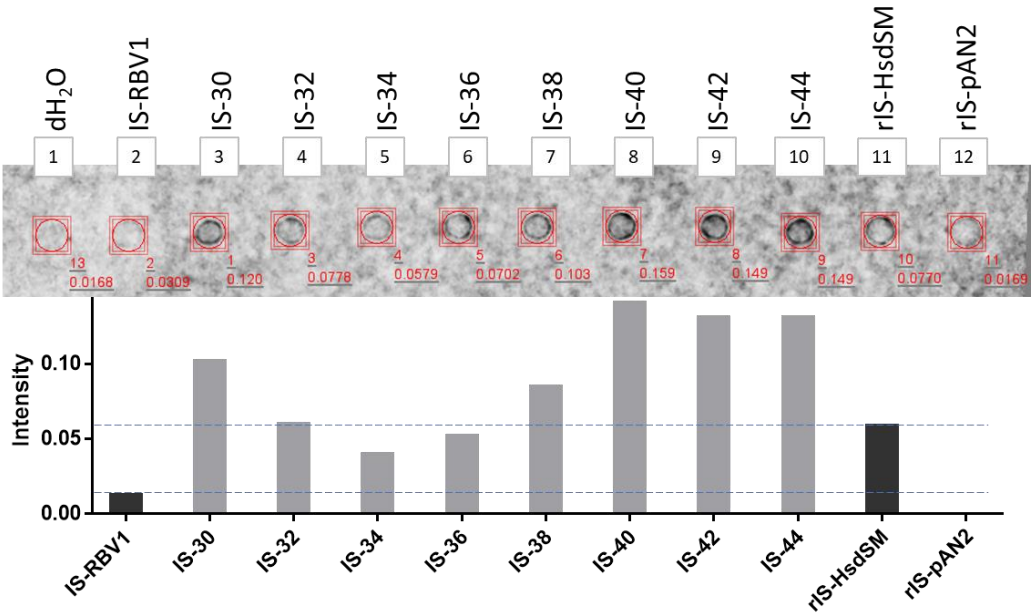


Figure 5.2.17 – Dot blot for methyltransferase by recombinant *E. coli* InterStellar-HsdSM

Plasmids were extracted from eleven modified *E. coli* InterStellar strains and probed for adenine methylation. Primary mouse antibodies raised against methylated adenine bases (6mA) and secondary goat-anti-mouse antibodies harbouring infrared fluorescent tag. Negative control, pRBV1 extracted from InterStellar (lane 2). Samples of pRBV-x plasmids where x is methylases 30, 32, 34, 36, 38, 40, 42, 44 from *C. carboxidivorans* (lanes 3-10)(Method 2.15.3). Positive control, pMTL83151 extracted from rIS-HsdSM (lane 11). Negative control #2, pMTL83151 extracted from rIS-pAN2. Dashed lines represent the control values, positive (upper) and negative (lower).

All eight methyltransferases reported a higher fluorescent response than the two negative control samples which did not contain 6mA methylation. Methyltransferases 34 and 36 both reported lower secondary anti-body fluorescence than the positive control rIS-HsdSM.

5.4.15. RM system evasion

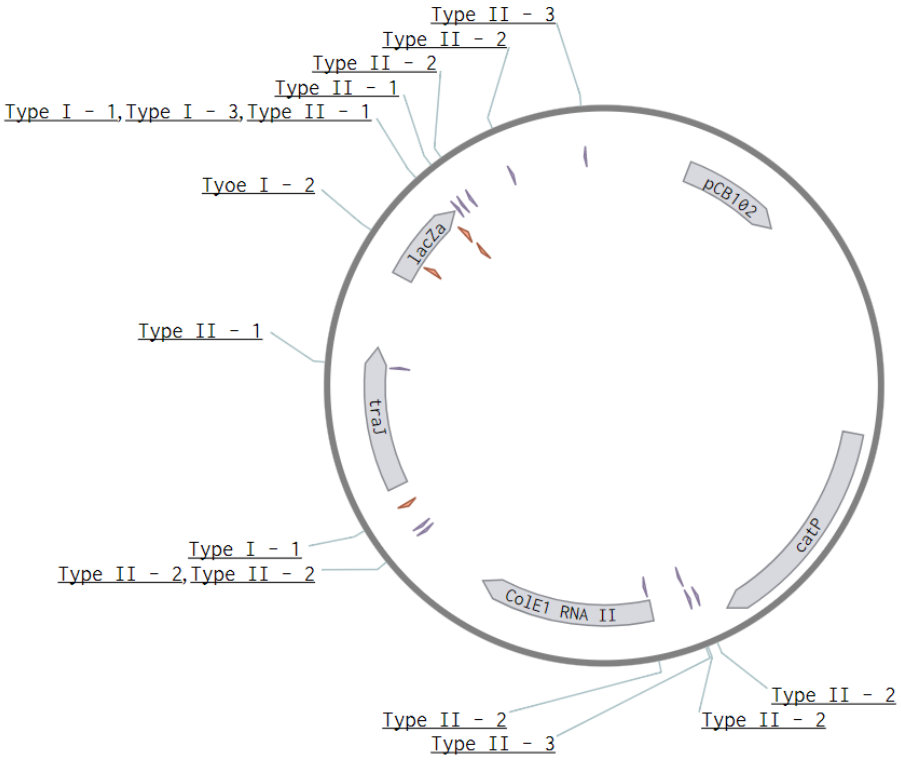


Figure 5.2.18 – Analysis of pMTL83151 with *C. carboxidivorans* RM motif sequences

Sixteen RM motifs were found in the pMTL83151 plasmid. This included four putatively assigned to Type I systems and twelve assigned to the remaining Type II and III systems.

Based on the PacBio base modification data, a reverse engineered plasmid, designed for the removal of all but four *C. carboxidivorans* P7 targeting motifs was produced (Figure 5.2.18). The pCB102 replicon (rep: pMTL83xxx) was selected on the basis of its broad host range in different clostridiales. Additionally, this replicon contained the fewest sites of any of the tested replicons Table 5.2.11. This plasmid, pMTL8315y, was created in an attempt to quickly evade RM system evasion at the DNA sequence level. The remaining sites were left in the Gram-negative module, ColE1-traJ. Although removal of the remaining four sites would have been preferable, it would have incurred a significant additional cloning burden, in addition to possibility that the plasmids replicative and transfer functions might be damaged.

Transformation of *C. carboxidivorans* was attempted with the engineered plasmid *via*electroporation. Competent cells were produced in the same manner as for *E. limosum*, save for a change in growth medium from AWM to YTF. No colonies were observed on the selection plates after 4 days following electroporation at a range of different electroporation voltages (0.9, 1.2, 1.5, and 1.8 kV) and a recovery time of 8 hours.

Table 5.2.11 – *C. carboxidivorans* RM targeting motif alignment to pMTL83151 and pMTL86151

Sites in pMTL83151	Count	System*
AGAAGC	3	CcaP7ORF23805P
CAYNNNNNCTGC	2	CcaP7ORF11965P
CCANNNNNNNNTCG	1	CcaP7ORF3570P
CRAAAAR	7	-
GATAAT	2	-
GCANNNNNNNNTNNCG	1	CcaP7ORF24310P
Total:	16	

Sites in pMTL86151	Count	System*
AGAAGC	3	CcaP7ORF23805P
CAYNNNNNCTGC	2	CcaP7ORF11965P
CCANNNNNNNNTCG	1	CcaP7ORF3570P
CRAAAAR	10	-
GATAAT	8	-
GCANNNNNNNNTNNCG	1	CcaP7ORF24310P
Total:	25	

* Motif-system assignment for the four systems which could be attributed with confidence.

5.5. Discussion

5.5.1. Overview

Genetic research into microorganisms can be hindered immensely by an absence of effective gene-transfer methodologies. This chapter reports the development of a six-step workflow for the establishing and/or improving gene-transfer in organisms which are protected by one or more RM systems Figure 5.2.3. The workflow is applicable to any organism with an available genome sequence which can be examined for RM system content using ReBase analysis and database tools. The relevant RM systems which are identified are then assembled into an inducible expression cassette and integrated into the large conjugal assistance plasmid pR702. In a penultimate pathway step, *E. coli* strains carrying the modified pR702 are transformed to camouflage plasmid DNA appropriately for the evasion of target organism RM systems. Gene-transfer can then be performed *via*direct conjugation from the methylation-donor strain, or by electroporation of extracted camouflaged plasmid.

The efficacy of the engineered system was demonstrated first in the transformation of *C. acetobutylicum* ATCC 824. This proof of concept was based on plasmid protection by *B. subtillis* phage derived ϕ3TI MTase, published in 1993 by Mermelstein and colleagues[220]. Plasmids extracted from engineered methylation donor strain, expressing the ϕ3TI MTase from a recombinant pR702 expression cassette, resulted in significant gene-transfer efficiency improvements over unmethylated controls (Figure 5.2.9). Efficiency improvements were equivalent to that seen using existing plasmid protection methods, specifically those employing the pAN2 plasmid[94].

The workflow was subsequently successfully applied to two organisms of significant research interest in the field of human health. Application of the workflow with the hyper-virulent *C. difficile* isolate, *C. difficile* DH1916[233], resulted in a >450 fold improvement in conjugation efficiency (Figure 5.2.11). Application with the previously untransformable type-A *C. botulinum* strain, *C. botulinum* A ATCC 19397[234], enabled gene-transfer for the first time in this organism. Both of these applications of the

method demonstrated for the first time the coupling of conjugation-based gene transfer to in-*vivo* plasmid methylation, from an engineered methylation-donor.

Attempts to apply the methodology against two RM systems, OrfA and OrfD, in *E. limosum* ATCC 8486 was unsuccessful. These results highlighted possible areas for improvement in the method, as well as further areas of investigation into incompatibilities of heterologous RM systems with selected methylation-donor strains.

Preliminary work was carried out in evaluating the current RM based barriers to gene transfer in *C. carboxidivorans* P7. Four out of six modified sequence motifs, determined from PacBio sequencing data, were assigned with confidence to RM systems. The remaining two motifs were also putatively assigned, however in these cases further evidence is still required for confirmation in the absence of sufficient bioinformatic data at this time.

5.5.2. Concept

The concept of protecting transforming DNA has been successfully used in the transformation of a number of microorganisms. Possibly the most notable example of this in the field of clostridiales is the plasmid borne expression of the phage derived ϕ3TI methyltransferase employed to enhance the transformation of *C. acetobutylicum* ATCC 824[220]. The use of RM systems derived directly from transformation target strains has since provided a more widely adaptable route to protecting transforming DNA, through the creation of bespoke methyl-donor strains. Recently a mimicking-of-DNA-methylation-patterns (MoDMP) pipeline was published by Zhang and colleagues[235]. Using this pipeline, the group demonstrated transformation of four difficult-to-transform strains through expression of heterologous MTases from the target strains in *E. coli*. The method presented in this work attempted to improve upon and fine-tune the previous approaches. The primary areas of development centred around employing

more extensive preliminary bioinformatic analysis of target strains, as well as the design of a quick, stable, and conjugation compatible method.

5.5.3. Bioinformatics

5.5.3.1. Basic Bioinformatics

The first level of genome analysis was characterisation of all RM systems identified in a target genome through ReBase. Type classifications and characterisation of the RM systems was based on a number of aspects. Namely, subunit composition, predicted recognition sequences, cofactor requirement, and cleavage position. In this study, this was accomplished through combining the DNA base modification outputs derived from PacBio datasets, the annotated genome files, and amino-acid sequence alignment against previously characterised and published RM subunits in the restriction enzyme database (ReBase)[133]. There were two key take home messages from this analysis. First, the identification of active Type I-III systems, ignoring any orphan-methylase containing systems. Second, identification of relevant Type IV systems, crucial for selecting an appropriate *E. coli* background strain. Mcr-type Type IV systems are active against 5mC methylated DNA, and were identified in both the *C. difficile* and *C. carboxidivorans* genomes. Even in the absence of precise targeting motifs it was clear that avoiding the use of Dcm^+ *E. coli* strains (i.e., CA434) as methylation donors was prudent. Improper choice of host strains with both Dam^+ and Dcm^+ genotypes has been shown to prevent gene transfer in a number of species[236].

5.5.3.2. Advanced

A second level of bioinformatics was performed for each RM system for which recognition motif sequences had been assigned in ReBase. This required either prior characterisation of a system, or access to PacBio data to make the assignment. For enzymes which had not undergone such investigative studies, prediction of the

recognition motifs was still possible in some cases. This was achieved through BLASTp alignment of the amino-acid sequences of relevant specificity determining subunits, against libraries of enzyme with known motif. In this study alignment to the ReBase Gold Standard database revealed the targeting motif of the *C. difficile* DH1916 Type I RM system HsdRSM1 to be CANNNNNNNTAAAG. PacBio™ whole genome sequencing (WGS) provided information on the modification status of sequenced DNA of *C. carboxidivorans* and *E. limosum*. Knowledge of the modified motifs assisted in the assignment of sequence specificity to four of the eight methyltransferase genes in *C. carboxidivorans* and all three *E. limosum* RM systems.

Motif targeting information was further used to supplement the basic bioinformatic analysis, and identify which systems were active against transforming plasmid sequences. In this way the 'relevant' RM systems could be further refined. Time spent correctly identifying active RM systems could afford significant time savings during the workflow. Specifically, reducing amount of cloning required during *E. coli* engineering stages.

HsdRSM1 recognition sequence was determined to account for just one cleavage site in the pMTL83151 plasmid sequence, within the *catP* open reading frame. Ironically, the only antibiotic selection modules from the pMTL80000 vector series[119] which did not contain a recognition sequence match was *ermB* and *tetA*. Neither erythromycin nor tetracycline are typically used for genetic work with *C. difficile* due to natural resistance to both antibiotics observed in heavily researched and original *C. difficile* genome sequenced, *C. difficile* 630[237].

Equivalent motif analysis of *E. limosum* and *C. carboxidivorans* PacBio data was performed in relation to a selection of pMTL8x151 vectors. *C. carboxidivorans* contained in excess of 15 sites with typical pMTL8x151 plasmid, while *E. limosum* had in excess of 11. In principle the removal of the motif-sequences from the plasmid represents a means of achieving plasmid transfer into hard to transform strains. Indeed, work in cyanobacterium has demonstrated the removal of targeted sites improving

conjugation efficiencies by over an order of magnitude per single site in some cases[228]. However, in practical terms the removal of all sites represents a significant workload. As such it is not a realistic solution for every plasmid that one might want to transformed into the strain in the future.

5.5.4. System design

Recombinant methyl donor strains were built through targeted integration of methylation cassettes at a defined locus on the conjugative plasmid pR702. pR702 is a large IncPβ group plasmid derived from *Proteus mirabilis*, which confers resistance to streptomycin, tetracycline, kanamycin, sulphonamides and mercury salts[238,239]. Additionally, R702 possessing the capacity to mobilise compatible plasmids (and itself) *via*conjugative transfer[238]. The targeting of this R-factor for cassette integration was decided for a number of reasons. First and foremost, building into the conjugative plasmid enables relatively quick and simple transfer of all finalised constructs into different *E. coli* backgrounds. This is primarily relevant where Dam^+ or Dcm^+ phenotypes could be beneficial in evading RM systems in a target host. In such instances, using Dam^+/Dcm^+ backgrounds would limit the number of heterologous methylases required, thereby potentially reducing the total number of necessary cloning stages.

The production of a tailored methylation donor strain required the expression of heterologous MTases in donor strains, thereby exposing them to heterologous methylation patterns. If the produced methylation is detrimental to core cellular processes or is otherwise incompatible with the donor, then the cells will not survive or methylating genes will be inactivated through mutation. DNA methylation plays important roles in prokaryotic gene regulation[229,240], so it is to be expected that in some cases heterologous RM system expression will effect donor strains beyond simply methylating their DNA.

5.5.5. Background genotypes for donors

Three donor strains were selected for the project, each with different background methylation. *E. coli* CA434 (*E. coli* HB101 carrying pR702) as the Dam^+/Dcm^+ background, already had a long history in establishing conjugation into clostridiales[241]. *E.coli* SExpress (*E. coli* NEB Express carrying pR702) as the Dam^+/Dcm^- background, had demonstrated conjugation efficiency improvements over *E.coli* CA434 into the Type IV RM system-harbouring *C. autoethanogenum*[242]. *E. coli* InterStellar (*E. coli* Stellar carrying pR702) was selected as the Dam^-/Dcm^- strain, a clean unmethylated platform from which to build.

Use of the InterStellar for all applications was tempting, however, as a result of the Dam-/Dcm- genotype the strain suffered from a significantly elevated background mutation rate. This stems from inactivation of the mismatch repair mechanisms of the cell[243]. The usually desirable RecA inactivation which is typical of laboratory cloning strains was not possible in the Dam^- strain, as it would result in further increase mutability to unsustainably high levels[244]. Collectively this meant that InterStellar was inappropriate for storage of expression cassettes with confidence in their genetic fidelity. For this reason, recombinant R702 vectors were generally transferred into Express or HB101. Express in particular was useful in this project as no Mrr type Type-IV, which could target 6mA of Dam methylated DNA, were detected in any of the strains.

The effects of improper donor strain selection were observed following the cloning of methylases from *C. carboxidivorans* P7 in the *E. coli* strain DH5α. Heterologous MTases expressed in DH5α were found in all cases to contain SNPs (data not shown). These mutations typically occurred in critical locations within gene operons which were expected to cause gene inactivation. Loss of sequence fidelity included base changes within RBS regions, as well as frame shift inducing indel mutations at varying locations within genes. The incompatibility of DH5α with the expressed MTases was most likely due to the presence of a type IV RM system, Mrr. Where MTases created 'foreign' methylation patterns in the donor strains Mrr systems would have engaged resulting in

self-genome-cleavage and cell death. This represents strong pressure to avoid expression of foreign RM systems, a driving force for mutation.

5.5.6. Recombineering backgrounds

Integration of methylation cassettes into pR702 was mediated by Lambda-red recombineering[245]. Recombineering was initially attempted in all three pR702 containing *E. coli* strains. Even using the smallest cassette fragments (RBV1), recombineering was only successful into the InterStellar strain. These findings are believed to stem from the $RecA^+$ and Dam^- genotypes of InterStellar, which is not shared by the either CA434 or SExpress. $RecA^+$ enhancement of lambda red recombineering is well documented[246], and Dam knockout, *Mrr*-deficient strains, report >100 fold increases in recombineering efficiency[247]. Recombineering of methyltransferase cassettes was successfully performed with fragments in excess of 8 Kb in length. The sizes align close to the upper limits of reported chromosomal integration by Lambda-red recombineering in published literature[248,249].

The production of a tailored methylation donor strain requires the expression of heterologous MTases and thereby heterologous methylation patterns in an appropriate donor. If the produced methylation is detrimental to central cellular processes as a result of improper regulation, or is otherwise incompatible with the host, then either the donor strain will die or the methylation genes will be inactivated through mutation. In either case the methylation donor strain will not be able to perform its function in aiding gene transfer. This effect was readily observed when expressing methylases from *C. carboxidivorans* P7 in the *E. coli* strain DH5α. Heterologous MTases expressed in DH5α were regularly observed to contain SNPs (data not shown). These mutations typically occurred in critical locations within operon which were expected to lead to gene inactivation. Loss of sequence fidelity included base changes within RBS regions, as well as frame shift inducing indel mutations at varying locations within genes. The incompatibility in this instance is most likely due to the presence of type-IV RM systems

in DH5α, which would result in self genome cleavage where foreign methylation patterns were detected.

5.5.7. Proof of concept

The proof of principle strain rIS-pAN2 was designed around the established methodologies for the transformation of the model clostridial species *C. acetobutylicum* ATCC 824, centred around a *B. subtillis* phage derived methyltransferase, ϕ3TI[220]. The ϕ3TI methyltransferase possesses two target recognition domains which results in the concurrent methylation of two distinct sequences, at the internal cytosine residue of GGCC and GCNGC[250]. Methylation of the latter protects DNA against cleavage by the Cac824I restriction endonuclease of *C. acetobutylicum* ATCC 824, the primary RM barrier to transformation of the strain which recognises and cleaves at GCNGC sequences in the absence of cytosine methylation[251,220]. It is fortuitous that in the case of *C. acetobutylicum* ATCC 824 the unnecessary ϕ3TI methylation activity at GGCC sequences does not have a deleterious effect on transformation efficiency[220], due to the lack of type-IV restriction systems in this strain.

Transformation efficiencies using plasmids extracted from rIS-pAN2 were significantly improved over unmethylated controls. This success demonstrated the efficacy of construct and method design as applied to electroporation. Induction levels over 0.1 mM did not correlate with any statistically relevant improvement in plasmid protection with rIS-pAN2 (Appendix - Figure 7.3.2). Improved transformation was also seen with uninduced rIS-pAN2, although not as significant as 0.1 mM and higher. This is most likely a reflection of laxity in expression control by the promoter system.

The fluorescence reporter assay data indicated weak expression levels of the araC-P_{BAD} promoter system relative to previous studies[120]. This was anticipated as a result of the single copy R-Factor by comparison to multicopy plasmids which are typically tested.

Expression was, however, significant after 15.5 hours at 10 and 100 mM L-arabinose induction. Interestingly although, reporter data suggested that uninduced expression was undetectable even after 48 hours, rIS-pAN2 still managed to enhance transformation without induction (data not shown). This indicated that the level of expression required to afford plasmid protection is markedly low, but that high intracellular concentrations of ϕ3TI MTase is not essential to confer protection. Whether this translates to the other MTases relevant to this study is unclear, although kinetic studies of *E. coli* MTases suggest that rates can vary significantly between systems[252].

Marginal improvements were expected over the dual cloning plasmid approach (Top10-pAN2) as a result of improved plasmid yields from miniprepped cells. Specifically, avoidance of incidental extraction of the methylating plasmid when using the >69 Kb R702, which should not be captured by standard miniprep protocols. However, evidence from gel electrophoresis of miniprepped pMTL83151 (Figure 5.2.16) suggests that the specificity of the kits did not entirely exclude pR702 in reality.

5.5.8. Applications

5.5.8.1. *C. difficile* DH1916

Transformation of *C. difficile* has significant relevance to clinical research, however, gene-transfer had proven problematic. Conjugation approaches are the preferred method of gene-transfer into *C. difficile*, where reports of electroporation based techniques have been unreproducible and questionable in validity[241,253]. Lack of electroporation may stem from the *C. difficile's* distinctive highly deacetylated peptidoglycan cell wall and S-layer composition[254], preventing appropriate surface contact of exogenously added plasmid DNA. Between one and nine RM systems were found in each of the 39 *C. difficile* genomes in the ReBase database[133], with an average of three per strain. Poor conjugation, lack of electroporation, and harbouring of RM systems, made *C. difficile* a textbook candidate for application of the pipeline developed

in this study. The conjugation from the engineered donor into *C. difficile* DH1916 strain produced transformation efficiencies >450 higher than observed in control samples. This improvement resulted from the protection of just a single recognition site on transformed plasmids.

The improvements observed in conjugation with the rSE-HsdSM1 methylation donor are comparable to that of a genomic KO of the HsdRSM1 REase in *C. difficile* DH1916 strain (Ingle, P. SBRC Nottingham, Thesis 2018). The KO strain allowed conjugation improvements of over two orders of magnitude. The importance of selecting the appropriate background *E. coli* strain in this study, NEB Express, should not be overlooked. Data from the SBRC Nottingham indicated conjugation rates improved roughly one order of magnitude using SExpress over CA434 (Ingle, P. SBRC Nottingham, Thesis 2018).

5.5.8.2. *C. botulinum* A ATCC 19397

C. botulinum A ATCC 19397 was another clinically relevant organism from the clostridiales order. Much like *C. difficile*, *C. botulinum* genetic research has primarily centred on conjugation-based transfer[255–257]. Previous to this study transformation had been achieved in *C. botulinum* A ATCC 19397, and so inactivation of the RM systems was not an option. Expression of both active RM systems Orf717 and Orf2811 from rIS-Orf717-2811 enabled transformation at an effectiveness of ~200 CFU per transformation. Although the effective transformation counts are relatively low, these numbers represent an unoptimized conjugation procedure for this particular strain from which to build.

5.5.8.3. *E. limosum* ATCC 8486

The results from the *E. limosum* transformations using plasmids prepared in two engineered donor strains rSE-OrfA and rSE-OrfD were unexpected. Neither methylation

donor strain resulted in improved efficiency of gene-transfer, moreover rSE-OrfA actually resulted in a relative decrease in efficiency. The reduction in efficiency could be attributed in part to unusual plasmid preparation outcomes (Figure 5.2.16). Seemingly increased non-specific extraction of cellular DNA, presumably pR702, was increased in the rSE-OrfA strain. Further analysis is required to confirm the exact nature of the differences observed in miniprep profile from the donor strains. Though the reason for the differences is not immediately clear, one possibility is the interference of OrfA methylation with regulatory processes in the donor cell[240]. In opposition to this hypothesis, the GCGRAG methylation motif effected by OrfA, does not appear to have any overlap with the methylation of *E. coli* Express, specifically the Dam system (GATC)[258].

Lack of transformation improvements with rSE-OrfD was particularly unexpected in light of results from a KO of the system which were performed in parallel within the SBRC (Millard, J. SBRC, Nottingham. 2017). The parallel study reported a 10-fold increase in transformation efficiency in the KO over WT strain. This points to ineffective expression or activity of the OrfD enzyme pair in the donor strain. No inability to express heterologous MTase genes was reported by Zhang and colleagues following the cloning 24 heterologous MTases in *E. coli*[235], so the issue may instead lie in expression levels or failed induction. Further testing is required to identify the precise nature of the issue for each construct.

5.6. A future for *C. carboxidivorans*

The scale of the RM system problem in *C. carboxidivorans* was unparalleled in this project. The high number of system which required manipulation was compounded by the practical difficulties amplifying and cloning genes of such low GC content (G: C 29.7%)[259]. As a result, assembly and testing of a multi-methyltransferase expressing *E. coli* methylation donor strain *C. carboxidivorans* was not possible within the time constraints of the project. However, a reverse engineered plasmid, designed for the

removal of all but four *C. carboxidivorans* P7 targeting motifs was produced in an attempt to demonstrate RM system evasion at the DNA sequence level.

The removal of all (sixteen) but four sites was achieved in a single assembly step. Although the remaining four sites would have ideally been removed also, this would have required a more involved cloning process, as well as downstream testing to confirm the effects of the base changes. The primary concern with these four loci was their existence within RNA based ColE1 replicon and origin of transfer OriT[241]. Mutations in these regions might have led to significant instability or complete disfunction of the replicon. Transformation of the four-site plasmid by electroporation under a number of different conditions was attempted. No transformants were observed under any of the electroporation conditions tested. At the close of the project an SBRC colleague (Redfern, Beth. SBRC Nottingham, 2018) successfully demonstrated conjugation-based transfer of pMTL8315y into *C. carboxidivorans* from *E. coli* SExpress.

Due to the significant number of MTase genes which may be required to protect DNA in *C. carboxidivorans*, multiple rounds of recombineering will be required to build the final expression cassette in pR702. This will be possible through antibiotic cycling between *apra* and *aada1* resistance markers, and appropriate changes to the vector's left homology arms. This allows sequential assembly of more extensive methylation operons beyond the size limitations of a single round of recombineering.

5.7. Conclusion

This chapter has demonstrated the development and application of an improved pipeline for overcoming RM systems in the transformation of recalcitrant microorganisms. The method utilises bespoke methylation donor strains for the in-*vivo* methylation and protection of transforming DNA. The approach improves on previously published methods through the application of lambda-red recombineering for rapid methylation donor strain construction. Additionally, the described method offers

compatibility with conjugation-based transformation protocols, and allows for stable assembly of extensive multigene methylation cassettes.

5.8. Chapter acknowledgements:

Testing of the developed pipeline was aided immensely through the assistance and collaboration of a number of friends and colleagues within the SBRC, Nottingham. Specific thanks and acknowledgements go to the following: Dr. Zak Bean, for performing transformations of *C. acetobutylicum* 824 in the testing of rIS-pAN2; Patrick Ingle, for performing *C. difficile* DH1916 conjugations in the testing of rSE-DH1916_HsdSM1; Raquel Rodrigues and Daphne Groothuis, for the cloning and testing of rIS-Orf717-2811 in the transformation of *C. botulinum* A ATCC 19397; Georgina Bates for supporting in the testing of rSE-OrfA and rSE-OrfD for *E. limosum* 8486 transformation.

6. Original contributions and recommendations

6.1. Project aims

The primary aim of this project was to establish and develop genetic manipulation tools in novel acetogenic microorganisms, in particular *E. limosum*. This required first the development of a gene-transfer method, and subsequently the establishment of genetic engineering tools. The next aim was to showcase the efficacy of the established methods in the genetic development of *E. limosum* ATCC 8486 to produce acetone and isopropanol. Finally, this study set about to create an improved pipeline to establishing gene transfer in recalcitrant organisms.

6.2. Original contributions

6.2.1. Genome assembly

A complete closed genome sequence was produced and annotated for *E. limosum* ATCC 8486, using PacBio and Illumina based sequencing technologies. This genome adds to a closed genome sequence which was published by Korea Advanced Institute of Science and Technology during the course of this study.

6.2.2. Gene transfer toolkit

Gene-transfer was demonstrated for the first time to *E. limosum* ATCC 8486 and *E. limosum* DSM 2594. Following optimisation of electroporation parameters, transformation efficiencies in excess of 2.5 x10^4 CFU/µg plasmid, the highest efficiency reported for any acetogenic species. We present a genetic toolkit for *E. limosum* which includes three Gram-positive replicons (pIP404 from *C. aceticum*, p19 from *C. carboxidivorans*, and pCB102 from *C. butyricum*), two functional antibiotic selection markers (*ermB* and *catP*), and a library of thirteen characterised promoter systems.

6.2.3. Strain engineering

The established toolkit was used to demonstrate two genome engineering tools, Allele-coupled exchange (ACE) and CRISPR/Cas9, in *E. limosum* ATCC 8486. Application of these tools, for both gene knock-in and knock-out, represents the first examples of targeted genome modification in *Eubacterium* genus. Subsequently, the toolkit was applied for to introduce a number of synthetic metabolic pathways to introduce acetone and isopropanol production to *E. limosum*.

6.2.4. Gene transfer pipeline

Inability to transformation the initial target organism for the project, *C. carboxidivorans* P7 prompted the development of an improved methodology for the transformation of recalcitrant microorganisms. Lambda-red recombineering was harnessed in the creation of methylation-donor strains of *E. coli*, expressing target-strain-derived methyltransferase genes, for the protection of transforming DNA. The method built on previously published techniques in using *P. mirabilis* derived pR702 plasmid as the target for integration into methylation gene cassettes. In addition to stability improvements over using typical multi-copy cloning vectors for methytransferase expression, the proposed method represents the first conjugation compatible methyl-donor approach of its kind.

The proposed method was successfully applied first in a proof-of-concept with *C. acetobutylicum* ATCC 824, and subsequently in two clinically relevant clostridial species. Trans-conjugation efficiencies were improved >450 fold into *C. difficile* DH1916, while in *C. botulinum* A ATCC 19397 the method enabled gene-transfer for the first time.

A review of available PacBio derived base modification data for *E. limosum* ATCC 8486 and *C. carboxidivorans* P7 was performed in conjunction with the method. This allowed assignment of four of six identified modification motifs to four of eight RM systems in *C. carboxidivorans*, and assignment of all three motifs in *E. limosum* to three RM systems.

6.3. Impacts

The findings of this study provide a firm platform for further research and genetic development of *E. limosum* as an industrially capable biocatalyst for C1 fermentation. Additionally, the development of a universally applicable pipeline for improved gene transfer in bacterial species where restriction modification systems are limiting to transfer, has potential for widespread application across microbiological research. Beyond the obvious benefit of establishing gene transfer in previously intractable strains, the method is of value in numerous situations where increased transformation frequencies are required.

6.4. Future Studies

The *E. limosum* ATCC 8486 genome sequence generated during this study is currently being applied to the development of a genome scale model. Genome modelling will provide a useful platform for predicting the appropriate strain modifications required to achieve maximal production titres of target chemicals. Combining this knowledge with the genetics toolkit and engineering techniques reported in this study, future research can implement rational strain engineering approaches, toward development of an industrially applicable chassis.

A genetic target of particular note for further study are the glycogen synthesis genes (glgA, glgB, glgC). These genes are expected to be responsible for the significant amount of extracellular polysaccharide (EP) which has been observed and reported in *E. limosum* species[31]. If knockout of these genes correlates with a loss of the aforementioned EP phenotype then this would likely make strain handling in the lab distinctly less laborious. More importantly, the reduced EP would likely have a significant impact on the carbon-flux efficiency of the strain towards certain target products. Such efficiency improvements will be critical in the development of the strain from a biocatalytic perspective.

Following the demonstration of engineered acetone and isopropanol production by *E. limosum* grown on fructose and glucose, future strain research should focus on pathway optimisations and further study on methanol and other C1 feedstocks. Additionally, future testing of the engineered strains produced in this study under continuous fermentation conditions is required to properly evaluate the efficacy of the synthetic pathways. As one of the few C4 producing acetogens, appropriate engineering towards increased butyrate or butanol production from C1 would seem a strong avenue for further study. Enhancing metabolic flux through thiolase and towards butyryl-CoA will be central to achieving both goals. The genome model will be a useful tool in determining the most appropriate way to achieve this, whether through gene overexpression, introduction of heterologous genes, or targeted knockout in competing pathways. With gene-transfer now established, high-throughput research techniques such as Transposon Directed Insertion-site Sequencing (TraDIS) can be applied in *E. limosum*[260]. TraDIS can provide comprehensive information regarding conditionally essential genes, which would be a valuable supplement the ongoing modelling efforts and future targeted engineering.

The other significant avenue for future development of this study exists in the application of the proposed gene-transfer methodology. The wide scope for application of the method means that it could be useful across the breadth of microbiology research. Where the ability to perform the genetic manipulations afforded by gene-transfer, is valuable from both basic research and strain development standpoints. With the preliminary work for method application already performed, establishing gene-transfer in *C. carboxidivorans* is the obvious next step. *So, back to square one…*

7. Appendices

7.1. Chapter 3 supplementary

7.1.1. gDNA extraction protocol development

E. limosum species generates a significant quantity of glycopolymer which interferes with many standard preparatory techniques which are required for genomic, transcriptomic, and proteomic analysis. Extraction of RNA and genomic DNA are most significantly impacted because, as for nucleic acids, the glycopolymers are highly polar and will remain in the same 'phase' as DNA during standard separation DNA from cell lysates. This made the acquisition of sufficient DNA, to meet the more significant thresholds required for PacBio sequencing. As such, it was necessary to modify and optimise a genomic DNA extraction to extract high quality DNA from *E. limosum*.

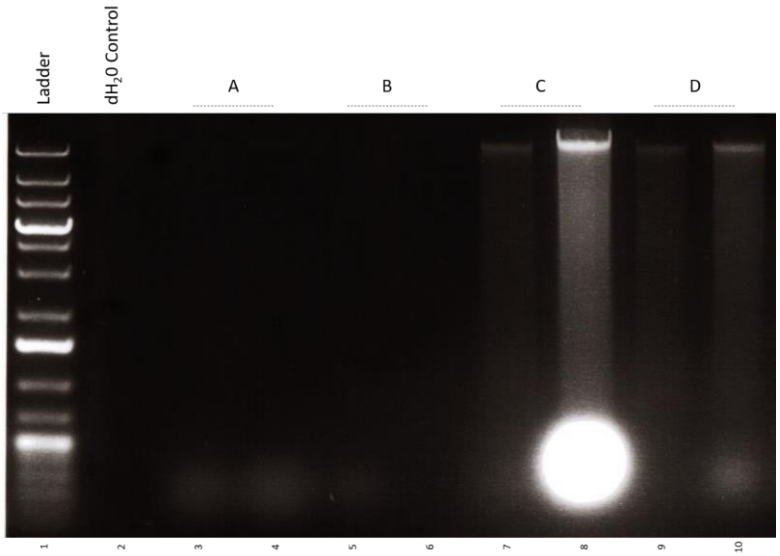


Figure 7.1.1 – Optimization of genomic DNA (gDNA) extraction in *E. limosum*

A modified phenol chloroform extraction methodology (Method 2.9.5) was developed to improve genomic extractions in *E. limosum* and other species that produce significant amounts of extra cellular glycopolysaccharide. Figure is representative of multiple distinct gDNA extractions with methods A, C, and D. Method B was only performed once for comparison. Each method was carried out with from both 2µl and 4 µl (left and right lane of each pair). Lane 1 - O'Generuler 1Kb-plus ladder; Lane 2 – dH2O control (blank); Lane 3,4 - Method A – Standard phenol chloroform extraction; Lane 5,6 – Method B - Std. plus di-ethyl ether treatment; Lane 7,8 – Method C – Std. plus bead lysis treatment; Method D – Std. plus bead lysis and di-ethyl ether treatment.

Extraction Method	Culture volume used (ml)	Nanodrop measured Conc. (ng/μl)	Qubit measured Conc. (ng/μl)
1	2	89.4	19.00
1	4	133.9	39.40
2	2	107.9	15.75
2	4	15.8	2.705
3	2	159.3	96.05
3	4	1374.9	522.50
4	2	88.9	98.25
4	4	290.8	295.00

Table 7.1.1 – DNA concentrations measurements from Nanodrop and Qubit for four methods of DNA extraction from *E. limosum* ATCC 8486.

The combination method, variant 4, utilising both physical cell lysis (variant 3) and cell wall treatment by diethyl ether (variant 2), was the most effective in producing high quality DNA with minimal RNA contamination. All samples prepared using method variants 1-3 produced notably higher concentrations when measured by the Nanodrop in comparison to the Qubit values. Nanodrop calculates total nucleic acid content whereas the Qubit quantifies specifically double stranded DNA. Using the discrepancy between Qubit and Nanodrop readings as a guide to RNA contamination, each method 1-3 resulted in the inadvertent extraction of significant quantities of RNA in spite of multiple RNase treatment steps in the standard method.

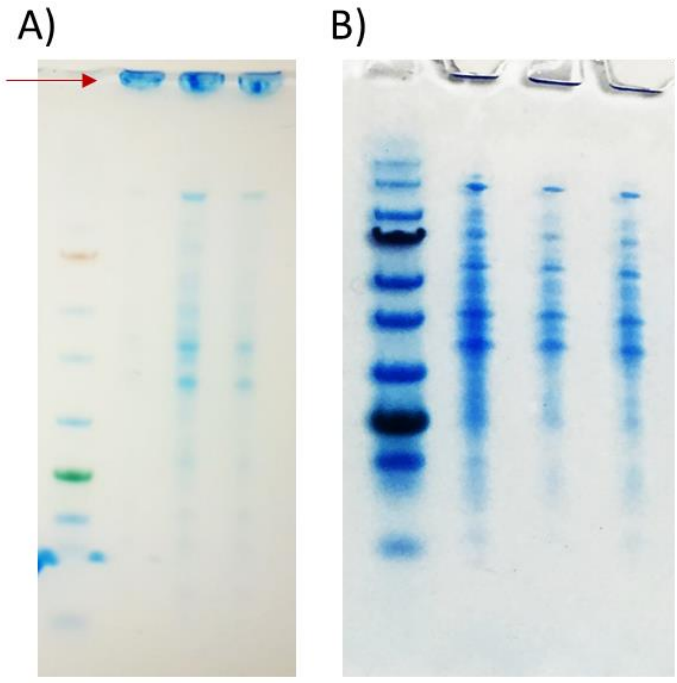


Figure 7.1.2 – Protein extraction enhancement with salt wash pre-treatment

A salt wash prior to extraction yielded improved protein extraction. A) Representative Coomassie stain of SDS-page of three *E. limosum* samples following protein extraction without pre-treatment. Arrow (red) indicates aggregated protein trapped in sample wells, unable to travel through gel. B) Representative SDS-page of three *E. limosum* samples following protein extraction with two 1.5 M NaCl wash steps.

7.1.2. Recommended protein and gDNA extraction method (*E. limosum*)

7.1.2.1. Pre-treatment

- Pellet cells (max speed, 4°C, 10 mins)
- Pour off supernatant (as much as possible without losing biomass)
- Re-suspend in 1.5 M NaCl (equal volume to initial sample)
- Pellet cells (max speed, 4°C, 10 mins)
- Remove supernatant
    - Pour off until final 2-3 ml
    - Pipette off remaining
- Repeat NaCl resuspension and pelleting until consistent pellet is observed (no further polysaccharide 'strings' should be observed when pipetting)

7.1.2.2. Protein extraction (Scale as appropriate)

- Salt wash
- BugBuster™ + Protease inhibitor resuspension
- Bioruptor™ Sonication (Method 2.1.1)

7.1.2.3. gDNA extraction (4 ml of overnight culture)

- Salt wash
- Phenol Chloroform extraction (Method 2.9.5)

Table 7.1.2 – pBLAST of Ga0111251_114071 alignment against Swiss-Prot database

Description	Max score	Total score	Query cover	E value	Ident	Accession
NADPH-dependent butanol dehydrogenase;	288	288	97%	2e-93	41%	P13604.1
Aldehyde-alcohol dehydrogenase 2;	245	245	98%	2e-72	36%	Q24803.1
Alcohol dehydrogenase 4;	232	232	92%	2e-71	38%	Q09669.1
Alcohol dehydrogenase 4;	230	230	98%	3e-71	35%	P10127.3
Alcohol dehydrogenase 4;	230	230	98%	5e-71	35%	A6ZTT5.2
Aldehyde-alcohol dehydrogenase;	226	226	98%	2e-65	36%	P0A9Q8.2

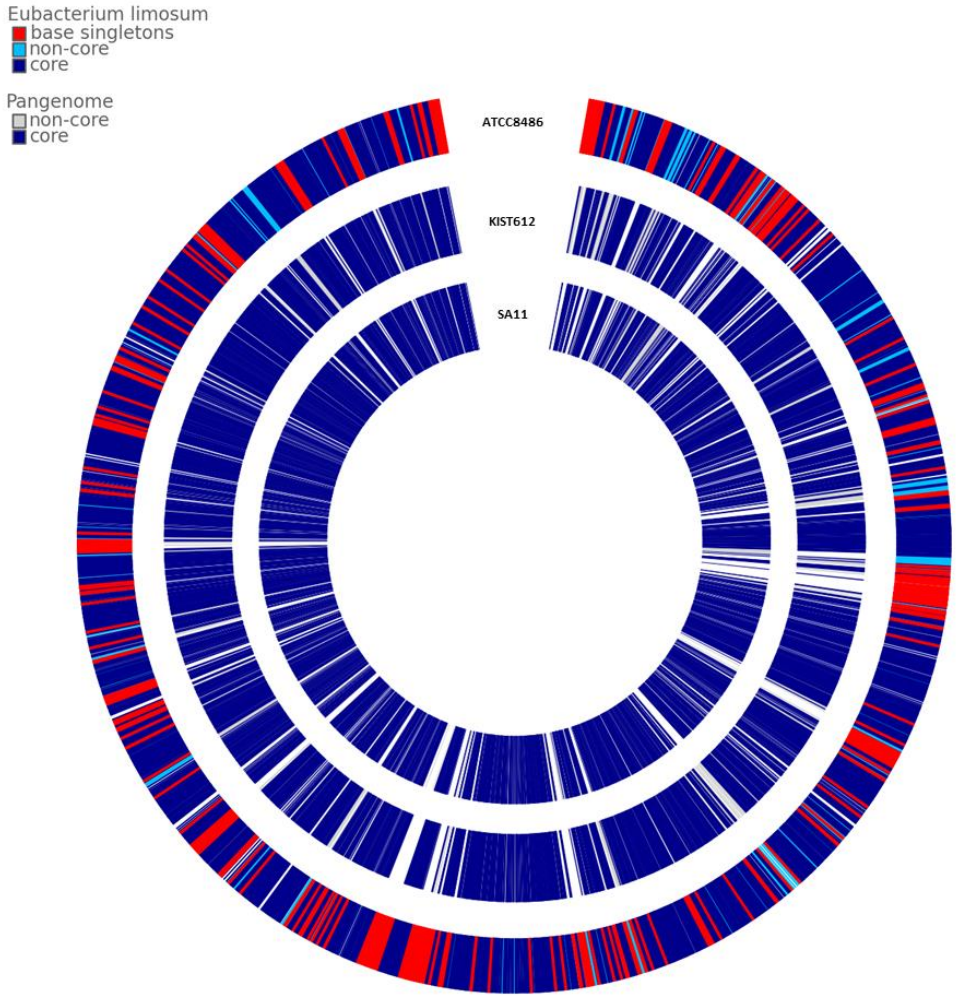
Table 7.1.3 - Full alignment gap summary for NCBI vs SBRC genome sequences

Genome	Contig	Position in Contig	GenomeWide Position	Length
SBRC sequence	CP019962.1	1596018	1596018	1
SBRC sequence	CP019962.1	1946117	1946117	4
SBRC sequence	CP019962.1	2387122	2387122	1
SBRC sequence	CP019962.1	2387183	2387183	1
SBRC sequence	CP019962.1	2387212	2387212	1
SBRC sequence	CP019962.1	2395174	2395174	7959
SBRC sequence	CP019962.1	3181291	3181291	6
SBRC sequence	CP019962.1	3182922	3182922	8
SBRC sequence	CP019962.1	3183051	3183051	1
SBRC sequence	CP019962.1	3183100	3183100	2
SBRC sequence	CP019962.1	3183127	3183127	10
SBRC sequence	CP019962.1	3183152	3183152	132
SBRC sequence	CP019962.1	3183517	3183517	1
SBRC sequence	CP019962.1	4417411	4417411	1
SBRC sequence	CP019962.1	4417483	4417483	1
SBRC sequence	CP019962.1	4417955	4417955	1
SBRC sequence	CP019962.1	4417999	4417999	1
SBRC sequence	CP019962.1	4418284	4418284	1
SBRC sequence	CP019962.1	4418301	4418301	1
SBRC sequence	CP019962.1	4418595	4418595	1
SBRC sequence	CP019962.1	4419018	4419018	1
SBRC sequence	CP019962.1	4419509	4419509	3407
CP019962.1	SBRC sequence	980563	980563	3
CP019962.1	SBRC sequence	1255342	1255342	1662
CP019962.1	SBRC sequence	1696450	1696450	7956
CP019962.1	SBRC sequence	2489866	2489866	5
CP019962.1	SBRC sequence	2489867	2489867	1
CP019962.1	SBRC sequence	2492511	2492511	86
CP019962.1	SBRC sequence	2492883	2492883	6
CP019962.1	SBRC sequence	2492884	2492884	2
CP019962.1	SBRC sequence	3113089	3113089	1
CP019962.1	SBRC sequence	3726803	3726803	1
CP019962.1	SBRC sequence	3727061	3727061	2
CP019962.1	SBRC sequence	3727751	3727751	1
CP019962.1	SBRC sequence	3727863	3727863	1
CP019962.1	SBRC sequence	3728206	3728206	1
CP019962.1	SBRC sequence	3728252	3728252	1
CP019962.1	SBRC sequence	3728294	3728294	1
CP019962.1	SBRC sequence	3728384	3728384	1
CP019962.1	SBRC sequence	3728419	3728419	4
CP019962.1	SBRC sequence	3728429	3728429	1
CP019962.1	SBRC sequence	3728494	3728494	2
CP019962.1	SBRC sequence	3728558	3728558	2
CP019962.1	SBRC sequence	3728623	3728623	1
CP019962.1	SBRC sequence	3728644	3728644	1

CP019962.1	SBRC sequence	3728666	3728666	1
CP019962.1	SBRC sequence	3728698	3728698	1
CP019962.1	SBRC sequence	3728730	3728730	1
CP019962.1	SBRC sequence	3728759	3728759	1

1704409	3731999	1255346	2492519
Phage Loci	Phage Loci	Transposable element	Type I restriction system
Phage 2 (ple3)	Phage 3 (pfr2)	DDE transposase	OrfD
8 Kb	3.5 Kb	1.7 Kb	0.7 Kb

Figure 7.1.3 – Pan genome for *E. limosum* strains

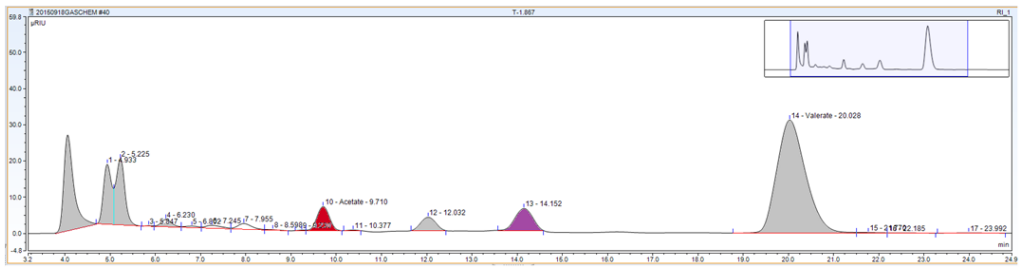


Genome track atlas, comparing *E. limosum* ATCC 8486 against a pan-genome created from three *E. limosum* strains, ATCC 8486, SA11, and KIST612.

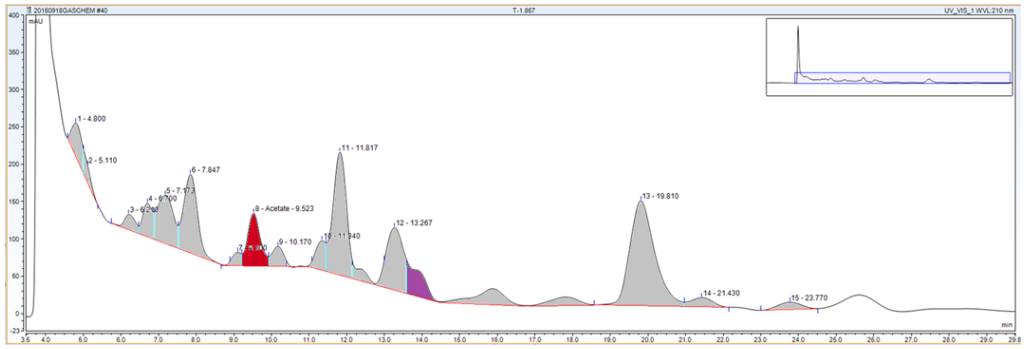
7.2. Chapter 4 supplementary

7.2.1. HPLC-UV/IR

A)



B)



C)

Analyte	UV-vis RT (min)	RI RT (min)
Glucose	5.69	5.87
Lactate (1)	7.63	7.85
Succinate	7.71	7.92
Lactate (2)	8.03	8.24
Formate	8.71	8.92
Acetate	9.5	9.71
Mes	-	11
Acetoin	11.29	11.5
Ethanol	-	13.01
Acetone	-	13.65
Butyrate	13.9	14.1
Prapan-2-ol	-	16.27
Valerate	19.91	20.12
Butanol	-	22.39

Figure 7.2.1 – HPLC-UV/IR analysis of *E. limosum* ATCC 8486 fermentation

Example HPLC analysis chromatogram for *E. limosum* ATCC 8486. Samples taken from OD_{600}: 1.867 culture. A) HPLC-IR chromatogram. B) HPLC-UV chromatogram. Peaks for acetate (red) and butyrate (purple) were not apparent in relevant media blank samples. Peaks marked in grey were also observed in the media blank. C) Table of analyte retention times based on control standards.

7.2.1. TcdR off site target analysis

Table 7.2.1 – Potential binding sites for TcdR sigma factor in *E. limosum* ATCC 8486.

Promoter Source	-35 -10 box sequences	Observed matches in genome	Matches expected due to chance
TcdA	TTTACAnnnnnnnnnnnnnnnnnCTCCTTA	0	0.1
TcdB	TTTACAnnnnnnnnnnnnnnnnnGTCTTTA	0	0.1
TcdR consensus (TcdA+B)	TTTACAnnnnnnnnnnnnnnnnnSTCYTTA	1	1.1
Loosened TcdR -10 box consensus (11bp fixed)	TTTACAnnnnnnnnnnnnnnnnnNTCNTTA	3	2.1
Loosened TcdR -10 box consensus (9bp fixed)	TTTACAnnnnnnnnnnnnnnnnnNNNNTTA	83	33.7

Binding site identification was based on the -35 and -10 boxes identified in Martin-Verstraete (2016)[199]. Calculations for the expected number of occurrences is based on a double stranded 4421303bp genome where there is a 1 in 4 chances of each base, A, C, T, G, at each position.

Table 7.2.2 – pBLAST alignment of TcdR amino acid sequence against genomes of *E. limosum*, *C. acetobutylicum*, and *C. autoethanogenum*.

Organism	Score	Cover	E value	Ident
E. limosum ATCC 8486	38.5	32%	$6e^{-4}$	43%
E. limosum ATCC 8486	37.7	26%	0.001	45%
C. acetobutylicum ATCC 824	28.1	53%	2.6	33%
C. autoethanogenum DSM 10061	27.7	31%	4.7	29%

Parameter: DELTA-BLAST; All non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF excluding environmental samples from WGS projects; BLASTP 2.7.0+.

7.3. Chapter 5 supplementary

7.3.1. Recombineering target locus

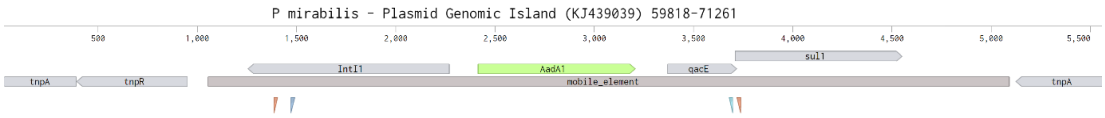


Figure 7.3.1 – *P. mirabilis* pKJ439039 (R702) genomic island

Target locus for recombineering integration events at the spectinomycin/streptomycin resistance gene *aada1*. Primers binding sites marked below annotations, from left to right: [339] forward strand; [200] forward strand; [219] reverse strand; [338] reverse strand.

7.3.2. Effect of L-arabinose concentration on gene-transfer

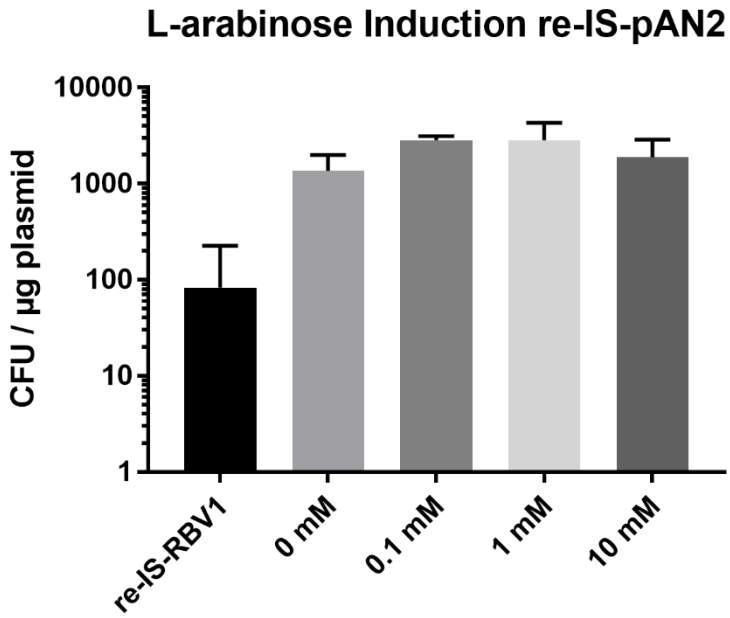


Figure 7.3.2 – Arabinose induction test

Transformation efficiencies were calculated at four different L-arabinose induction concentrations with the recombinant *E. coli* rIS-pAN2. Control sample rIS-RBV1 (left). All pAN2 containing constructs resulted in significantly higher transformation of plasmid pMTL83151 over the RBV1 control ($P \leq 0.00015$). Induction at 0.1 mM resulted in statistically significant improvement over uninduced (P = 00038). One-way ANOVA analysis revealed no significant difference between 0.1, 1, and 10 mM induction.

Electronic genome sequence data and plasmid maps generated through the course of this study are available on request from Nottingham SBRC (robert.mansfield2@nottingham.ac.uk, nigel.minton@nottingham.ac.uk).

8. Bibliography

1. Chemsystems, N. Biochemical Opportunities in the United Kingdom. *NNFCC Proj. Rep.* (2008).

2. BP. BP Statistical Review of World Energy 2017. *Br. Pet.* 1–52 (2017). doi:http://www.bp.com/content/dam/bp/en/corporate/pdf/energy-economics/statistical-review-2017/bp-statistical-review-of-world-energy-2017-full-report.pdf

3. EIA. Total Petroleum Consumption (Thousand Barrels Per Day). *International Energy Statistics* (2015). Available at: http://www.eia.gov/cfapps/ipdbproject/iedindex3.cfm?tid=5&pid=5&aid=2&cid=regions&syid=1980&eyid=2013&unit=TBPD.

4. U.S. EPA. *Climate change indicators in the United States 2016*. *Climate Change Indicators* (2016). doi:10.1016/j.ajodo.2007.08.016

5. UN. Paris Agreement (7.d). *Treaties UN* **21–7D,** 1–7 (2017).

6. Kittner, N., Lill, F. & Kammen, D. M. Energy storage deployment and innovation for the clean energy transition. *Nat. Energy* **2,** 17125 (2017).

7. Mohr, A. & Raman, S. Lessons from first generation biofuels and implications for the sustainability appraisal of second generation biofuels. *Energy Policy* **63,** 114–122 (2013).

8. Eusebtraut, A. *Sustainable Production of Second-Generation Biofuels*. **34,** (OECD Publishing, 2010).

9. Holmgren, J., Chung, A., Abdul, D., Hashim, R., Holmgren, J., Messina, J., Rieschel, G., Simpson, S., Ryoso, T. & Wyse, R. *Executive Summary: LanzaTech (2014)*. (2014).

10. Dahiya, S., Kumar, A. N., Shanthi Sravan, J., Chatterjee, S., Sarkar, O. & Mohan, S. V. Food waste biorefinery: Sustainable strategy for circular bioeconomy. *Bioresour. Technol.* **248,** 2–12 (2018).

11. Kircher, M. Sustainability of biofuels and renewable chemicals production from biomass. *Curr. Opin. Chem. Biol.* **29,** 26–31 (2015).

12. Mohammadi, M., Najafpour, G. D., Younesi, H., Lahijani, P., Uzir, M. H. & Mohamed, A. R. Bioconversion of synthesis gas to second generation biofuels: A review. *Renew. Sustain. Energy Rev.* **15,** 4255–4273 (2011).

13. Higman, C. & der Burgt, M. *Gasification*. (Gulf professional publishing, 2011).

14. Beychok, M. R. *Process and environmental technology for producing SNG and liquid fuels*. (U.S. Environmental Protection Agency, Office of Research and Development, National Environmental Research Center, 1975).

15. Min, F., Kopke, M. & Dennis, S. Gas Fermentation for Commercial Biofuels Production. *Liq. Gaseous Solid Biofuels - Convers. Tech.* (2013). doi:10.5772/52164

16. Bertau, M., Offermanns, H., Plass, L. & Schmidt, F. *Methanol : The Basic Chemical and Energy Feedstock of the Future*. doi:10.1007/978-3-642-39709-7

17. Xin, J., Cui, J., Niu, J., Hua, S., Xia, C., Li, S. & Zhu, L. Production of methanol from methane by methanotrophic bacteria. *Biocatal. Biotransformation* **22,** 225–229 (2004).

18. Anthony, C. How half a century of research was required to understand bacterial growth on C1 and C2 compounds; the story of the serine cycle and the ethylmalonyl-CoA pathway. *Sci. Prog.* **94,** 109–137 (2011).

19. Yasin, M., Park, S., Jeong, Y., Lee, E. Y., Lee, J. & Chang, I. S. Effect of internal pressure and gas/liquid interface area on the CO mass transfer coefficient using hollow fibre membranes as a high mass transfer gas diffusing system for microbial syngas fermentation. *Bioresour. Technol.* **169,** 637–643 (2014).

20. Munasinghe, P. C. & Khanal, S. K. Evaluation of hydrogen and carbon monoxide mass transfer and a correlation between the myoglobin-protein bioassay and gas chromatography method for carbon monoxide determination. *RSC Adv.* **4,** 37575–37581 (2014).

21. Küsel, K. & Drake, H. L. Acetogens. in *Encyclopedia of Geobiology* (eds. Reitner, J. & Thiel, V.) 1–5 (Springer Netherlands, 2011). doi:10.1007/978-1-4020-9212-1_2

22. Müller, V. Energy conservation in acetogenic bacteria. *Appl. Environ. Microbiol.* **69,** 6345–53 (2003).

23. Deamer, D. W. The first living systems: a bioenergetic perspective. *Microbiol. Mol. Biol. Rev.* **61,** 239–261 (1997).

24. Kaufmann, M. On the free energy that drove primordial anabolism. *Int. J. Mol. Sci.* **10,** 1853–1871 (2009).

25. Martin, W. F. Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. *FEBS Lett.* **586,** 485–493 (2012).

26. Wieringa, K. T. Over het verdwijnen van waterstof en koolzuur onder anaerobe voorwaarden. *Antonie Van Leeuwenhoek* **3,** 263–273 (1936).

27. Schuchmann, K. & Müller, V. Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nat. Rev. Microbiol.* **12,** 809–821 (2014).

28. Lanzatech. Facilities: China Waste Gas to Fuel Demonstration Plant. (2013). Available at: http://www.lanzatech.com/facilities/. (Accessed: 1st September 2017)

29. Cottingham, R. W. The DOE systems biology knowledgebase (KBase). in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics - BCB '15* 510–510 (ACM Press, 2015). doi:10.1145/2808719.2811433

30. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** (2010).

31. Kelly, W. J., Henderson, G., Pacheco, D. M., Li, D., Reilly, K., Naylor, G. E., Janssen,

P. H., Attwood, G. T., Altermann, E. & Leahy, S. C. The complete genome sequence of *Eubacterium limosum* SA11, a metabolically versatile rumen acetogen. *Stand. Genomic Sci.* **11,** 26 (2016).

32. Chang, I. S., Kim, B. H., Kim, D. H., Lovitt, R. W. & Sung, H. C. Formulation of defined media for carbon monoxide fermentation by *Eubacterium limosum* KIST612 and the growth characteristics of the bacterium. *J. Biosci. Bioeng.* **88,** 682–685 (1999).

33. Tanner, R. S., Stackebrandt, E., Fox, G. E. & Woesell, C. R. A Phylogenetic Analysis of *Acetobacterium woodii, Clostridium barkeri, Clostridium butyricum, Clostridium lituseburense, Eubacterium limosum*, and *Eubacterium tenue*. *Curr. Microbiol.* **5,** 35–38 (1981).

34. Roh, H., Ko, H.-J., Kim, D., Choi, D. G., Park, S., Kim, S., Chang, I. S. & Choi, I.-G. Complete genome sequence of a carbon monoxide-utilizing acetogen, *Eubacterium limosum* KIST612. *J. Bacteriol.* **193,** 307–308 (2011).

35. Jeong, J., Bertsch, J., Hess, V., Choi, S., Choi, I.-G., Chang, I. S. & Müller, V. Energy Conservation Model Based on Genomic and Experimental Analyses of a Carbon Monoxide-Utilizing, Butyrate-Forming Acetogen, *Eubacterium limosum* KIST612. *Appl. Environ. Microbiol.* **81,** 4782–90 (2015).

36. Liou, J. S.-C. Clostridium carboxidivorans sp. nov., a solvent-producing clostridium isolated from an agricultural settling lagoon, and reclassification of the acetogen *Clostridium scatologenes* strain SL1 as *Clostridium drakei* sp. nov. *Int. J. Syst. Evol. Microbiol.* **55,** 2085–2091 (2005).

37. Bruant, G., Lévesque, M. J., Peter, C., Guiot, S. R. & Masson, L. Genomic analysis of carbon monoxide utilization and butanol production by Clostridium carboxidivorans strain P7^{T}. *PLoS One* **5,** 1–12 (2010).

38. Liou, J. S. C., Balkwill, D. L., Drake, G. R. & Tanner, R. S. *Clostridium carboxidivorans* sp. nov., a solvent-producing clostridium isolated from an agricultural settling lagoon, and reclassification of the acetogen Clostridium scatologenes strain SL1 as Clostridium drakei sp. nov. *Int. J. Syst. Evol. Microbiol.* **55,** 2085–2091 (2005).

39. Phillips, J. R., Atiyeh, H. K., Tanner, R. S., Torres, J. R., Saxena, J., Wilkins, M. R. & Huhnke, R. L. Butanol and hexanol production in Clostridium carboxidivorans syngas fermentation: Medium development and culture techniques. *Bioresour. Technol.* **190,** 114–121 (2015).

40. Shin, J., Song, Y., Jeong, Y. & Cho, B. K. Analysis of the core genome and pan-genome of autotrophic acetogenic bacteria. *Front. Microbiol.* **7,** (2016).

41. Li, N., Yang, J., Chai, C., Yang, S., Jiang, W. & Gu, Y. Complete genome sequence of Clostridium carboxidivorans P7T, a syngas-fermenting bacterium capable of producing long-chain alcohols. *J. Biotechnol.* **211,** 44–45 (2015).

42. Drake, H., Küsel, K. & Matthies, C. Acetogenic Prokaryotes. in *The Prokaryotes SE - 61* 3–60 (2013). doi:10.1007/978-3-642-30141-4_61

43. Ljungdahl, L. & Wood, H. G. Incorporation of c14 from carbon dioxide into sugar phosphates, carboxylic acids, and amino acids by *Clostridium thermoaceticum*. *J. Bacteriol.* **89,** 1055–64 (1965).

44. Schidlowski, M. A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature* **333,** 313–318 (1988).

45. Fuchs, G. CO_2 fixation in acetogenic bacteria: Variations on a theme. *FEMS Microbiol. Lett.* **39,** 181–213 (1986).

46. Ragsdale, S. W. & Kumar, M. Nickel-Containing Carbon Monoxide Dehydrogenase/Acetyl-CoA Synthase. *Chem. Rev.* **96,** 2515–2540 (1996).

47. Ragsdale, S. W. Enzymology of the Wood-Ljungdahl pathway of acetogenesis. *Ann. N. Y. Acad. Sci.* **1125,** 129–136 (2008).

48. Pierce, E., Xie, G., Barabote, R. D., Saunders, E., Han, C. S., Detter, J. C., Richardson, P., Brettin, T. S., Das, A., Ljungdahl, L. G. & Ragsdale, S. W. The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). *Environ. Microbiol.* **10,** 2550–2573 (2008).

49. Orita, I., Sato, T., Yurimoto, H., Kato, N., Atomi, H., Imanaka, T. & Sakai, Y. The ribulose monophosphate pathway substitutes for the missing pentose phosphate pathway in the archaeon Thermococcus kodakaraensis. *J. Bacteriol.* **188,** 4698–4704 (2006).

50. Bogorad, I. W., Chen, C.-T., Theisen, M. K., Wu, T.-Y., Schlenz, A. R., Lam, A. T. & Liao, J. C. Building carbon–carbon bonds using a biocatalytic methanol condensation cycle. *Proc. Natl. Acad. Sci.* **111,** 15928–15933 (2014).

51. Sun, A. Y., Ljungdahl, L. & Wood, H. G. Total synthesis of acetate from CO2. II. Purification and properties of formyltetrahydrofolate synthetase from *Clostridium thermoaceticum*. *J. Bacteriol.* **98,** 842–844 (1969).

52. Liew, F.-M., Kopke, M. & Dennis, S. Gas Fermentation for Commercial Biofuels Production. in *Liquid, Gaseous and Solid Biofuels - Conversion Techniques* (InTech, 2013). doi:10.5772/52164

53. Köpke, M., Mihalcea, C., Liew, F., Tizard, J. H., Ali, M. S., Conolly, J. J., Al-Sinawi, B. & Simpson, S. D. 2,3-butanediol production by acetogenic bacteria, an alternative route to chemical synthesis, using industrial waste gas. *Appl. Environ. Microbiol.* **77,** 5467–5475 (2011).

54. Higashide, W., Li, Y., Yang, Y. & Liao, J. C. Metabolic engineering of *Clostridium cellulolyticum* for production of isobutanol from cellulose. *Appl. Environ. Microbiol.* **77,** 2727–2733 (2011).

55. Minton, N. P., Ehsaan, M., Humphreys, C. M., Little, G. T., Baker, J., Henstra, A. M., Liew, F., Kelly, M. L., Sheng, L., Schwarz, K. & Zhang, Y. A roadmap for gene system development in *Clostridium*. *Anaerobe* **41,** 104–112 (2016).

56. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & Merrick, J. M. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).

57. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).

58. Najafi, A., Bidkhori, G., Bozorgmehr, J. H., Koch, I. & Masoudi-Nejad, A. Genome scale modeling in systems biology: algorithms and resources. *Curr. Genomics* **15,**

130–59 (2014).

59. Van Dijk, E. L., Lè Ne Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** 418–426 (2014).

60. (LSAC), L.-S. G. S. and A. C. DNA Sequencing Costs: Data. (2017). Available at: https://www.genome.gov/27541954/dna-sequencing-costs-data/. (Accessed: 1st October 2017)

61. Liao, Y. C., Lin, S. H. & Lin, H. H. Completing bacterial genome assemblies: Strategy and performance comparisons. *Sci. Rep.* **5,** (2015).

62. Prakash, T. & Taylor, T. D. Functional assignment of metagenomic data: Challenges and applications. *Brief. Bioinform.* **13,** 711–727 (2012).

63. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** (2012).

64. Treangen, T. J., Abraham, A.-L., Touchon, M. & Rocha, E. P. C. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* **33,** 539–571 (2009).

65. Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H. & Phillippy, A. M. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14,** R101 (2013).

66. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* **13,** 278–289 (2015).

67. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).

68. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 75 (2008).

69. Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N. & Kyrpides, N. C. IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40,** (2012).

70. Richardson, E. J. & Watson, M. The automatic annotation of bacterial genomes. *Brief. Bioinform.* **14,** 1–12 (2013).

71. Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., Bower, L., Browne, P., Chan, W. M., Dimmer, E., Eberhardt, R., Fazzini, F., Fedotov, A., Foulger, R., Garavelli, J., *et al.* Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39,** 214–219 (2011).

72. Langridge, G. C., Phan, M. D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J. & Turner, A. K. Simultaneous assay of every *Salmonella Typhi* gene using one million

transposon mutants. *Genome Res.* **19,** 2308–2316 (2009).

73. Zhong, J.-J. *Biomanufacturing*. **87,** (Springer Berlin Heidelberg, 2004).

74. Leader, B., Baca, Q. J. & Golan, D. E. Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov* **7,** 21–39 (2008).

75. Aune, T. E. V. & Aachmann, F. L. Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed. *Appl. Microbiol. Biotechnol.* **85,** 1301–1313 (2010).

76. Yu, D., Ellis, H. M., Lee, E. C., Jenkins, N. A., Copeland, N. G. & Court, D. L. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 5978–5983 (2000).

77. Zhang, Y., Buchholz, F., Muyrers, J. P. & Stewart, A. F. A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat. Genet.* **20,** 123–128 (1998).

78. Million-Weaver, S. & Camps, M. Mechanisms of plasmid segregation: Have multicopy plasmids been overlooked? *Plasmid* **75,** 27–36 (2014).

79. Heap, J. T., Ehsaan, M., Cooksley, C. M., Ng, Y. K., Cartman, S. T., Winzer, K. & Minton, N. P. Integration of DNA into bacterial chromosomes from plasmids without a counter-selection marker. *Nucleic Acids Res.* **40,** (2012).

80. Judson, N. & Mekalanos, J. J. Transposon-based approaches to identify essential bacterial genes. *Trends Microbiol.* **8,** 521–526 (2000).

81. Reysset, G. Transformation and Electrotransformation in Clostridia. in *Genetics and Molecular Biology of Anaerobic Bacteria* (ed. Sebald, M.) 111–119 (Springer New York, 1993). doi:10.1007/978-1-4615-7087-5_7

82. Pyne, M. E., Moo-Young, M., Chung, D. A. & Chou, C. P. Development of an electrotransformation protocol for genetic manipulation of *Clostridium pasteurianum*. *Biotechnol. Biofuels* **6,** 50 (2013).

83. Trevors, J. T., Chassy, B. M., Dower, W. J. & Blaschek, H. P. 17 - Electrotransformation of Bacteria by Plasmid {DNA}. in *Guide to Electroporation and Electrofusion* (eds. Chang, D. C., Chassy, B. M., Saunders, J. A. & Sowers, A. E.) 265–290 (Academic Press, 1992). doi:https://doi.org/10.1016/B978-0-08-091727-6.50020-4

84. Dupuis, M. È., Villion, M., Magadán, A. H. & Moineau, S. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat. Commun.* **4,** 1–7 (2013).

85. Ran, F. A., Hsu, P. D. P., Wright, J., Agarwala, V., Scott, D. a & Zhang, F. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).

86. Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., Blumenthal, R. M., Degtyarev, S. K., Dryden, D. T. F., Dybvig, K., Firman, K., Gromova, E. S., Gumport, R. I., Halford, S. E., Hattman, S., Heitman, J., Hornby, D. P., Janulaitis, A., Jeltsch, A., *et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31,** 1805–1812 (2003).

87. Johnston, C. D., Skeete, C. A., Fomenkov, A., Roberts, R. J. & Rittling, S. R.

Restriction-modification mediated barriers to exogenous DNA uptake and incorporation employed by *Prevotella intermedia*. *PLoS One* **12,** 1–30 (2017).

88. Altenburg, E. The Limit of Radiation Frequency Effective in Producing Mutations. *Am. Nat.* **62,** 540–545 (1928).

89. AUERBACH, C. Chemical Mutagenesis. *Biol. Rev.* **24,** 355–391 (1949).

90. Esvelt, K. M. & Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.* **9,** 641–641 (2014).

91. Selle, K. & Barrangou, R. Harnessing CRISPR-Cas systems for bacterial genome editing. *Trends Microbiol.* **23,** 225–232 (2015).

92. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164,** 29–44 (2016).

93. Lambowitz, A. M. & Zimmerly, S. Group II introns: Mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* **3,** 1–19 (2011).

94. Heap, J. T., Pennington, O. J., Cartman, S. T., Carter, G. P. & Minton, N. P. The ClosTron: A universal gene knock-out system for the genus Clostridium. *J. Microbiol. Methods* **70,** 452–464 (2007).

95. Chen, Y., McClane, B. A., Fisher, D. J., Rood, J. I. & Gupta, P. Construction of an alpha toxin gene knockout mutant of *Clostridium perfringens* type A by use of a mobile group II intron. *Appl. Environ. Microbiol.* **71,** 7542–7 (2005).

96. Heap, J. T., Kuehne, S. A., Ehsaan, M., Cartman, S. T., Cooksley, C. M., Scott, J. C. & Minton, N. P. The ClosTron: Mutagenesis in Clostridium refined and streamlined. *J. Microbiol. Methods* **80,** 49–55 (2010).

97. Kuehne, S. A., Heap, J. T., Cooksley, C. M., Cartman, S. T. & Minton, N. P. ClosTron-mediated engineering of *Clostridium*. *Methods Mol. Biol.* **765,** 389–407 (2011).

98. Ng, Y. K., Ehsaan, M., Philip, S., Collery, M. M., Janoir, C., Collignon, A., Cartman, S. T. & Minton, N. P. Expanding the Repertoire of Gene Tools for Precise Manipulation of the *Clostridium difficile* Genome: Allelic Exchange Using pyrE Alleles. *PLoS One* **8,** (2013).

99. Schwarz, K., Zhang, Y., Kuit, W., Ehsaan, M., Kovács, K., Winzer, K. & Minton, N. P. *Chapter 13 - New Tools for the Genetic Modification of Industrial Clostridia A2 - Himmel, Michael E. BT - Direct Microbial Conversion of Biomass to Advanced Biofuels*. *Direct Microbial Conversion of Biomass to Advanced Biofuels* (Elsevier B.V., 2015). doi:http://dx.doi.org/10.1016/B978-0-444-59592-8.00013-0

100. Kovács, K., Willson, B. J., Schwarz, K., Heap, J. T., Jackson, A., Bolam, D. N., Winzer, K. & Minton, N. P. Secretion and assembly of functional mini-cellulosomes from synthetic chromosomal operons in *Clostridium acetobutylicum* ATCC 824. *Biotechnol. Biofuels* **6,** 117 (2013).

101. Chua, K. L. & Oliver, P. Intramolecular homologous recombination of linearized plasmids in *Escherichia coli* K12. *Mol. Gen. Genet. MGG* **232,** 199–205 (1992).

102. Jones, E. W. & Fink, G. R. Regulation of amino acid and nucleotide biosynthesis in yeast. (1982).

103. Dubeau, M. P., Ghinet, M. G., Jacques, P. É., Clermont, N., Beaulieu, C. & Brzezinski, R. Cytosine deaminase as a negative selection marker for gene disruption and replacement in the genus *Streptomyces* and other actinobacteria. *Appl. Environ. Microbiol.* **75,** 1211–1214 (2009).

104. Nariya, H., Miyata, S., Suzuki, M., Tamai, E. & Okabe, A. Development and application of a method for counterselectable in-frame deletion in *Clostridium perfringens*. *Appl. Environ. Microbiol.* **77,** 1375–1382 (2011).

105. Zhang, X. Z., Yan, X., Cui, Z. L., Hong, Q. & Li, S. P. *mazF*, a novel counter-selectable marker for unmarked chromosomal manipulation in Bacillus subtilis. *Nucleic Acids Res.* **34,** (2006).

106. DiCarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J. & Church, G. M. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Research* **41,** 4336–4343 (2013).

107. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* **109,** E2579–E2586 (2012).

108. Mojica, F., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. *Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements*. *Journal of molecular evolution* **60,** (2005).

109. Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J. & Charpentier, E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602 (2011).

110. Symington, L. S. End Resection at Double-Strand Breaks: Mechanism and Regulation. *Cold Spring Harb. Perspect. Biol.* **6,** a016436–a016436 (2014).

111. Thurtle-Schmidt, D. M. & Lo, T.-W. Molecular biology at the cutting edge: A review on CRISPR/CAS9 gene editing for undergraduates. *Biochem. Mol. Biol. Educ.* n/a-n/a doi:10.1002/bmb.21108

112. Pitcher, R. S., Wilson, T. E. & Doherty, A. J. New insights into NHEJ repair processes in prokaryotes. *Cell Cycle* **4,** 675–678 (2005).

113. Edwards, J. S. & Palsson, B. O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274,** 17410–17416 (1999).

114. Tomar, N. & De, R. K. Comparing methods for metabolic network analysis and an application to metabolic engineering. *Gene* **521,** 1–14 (2013).

115. Hanai, T., Atsumi, S. & Liao, J. C. Engineered synthetic pathway for isopropanol production in *Escherichia coli*. *Appl. Environ. Microbiol.* **73,** 7814–7818 (2007).

116. Green, M. R. (Michael R., Sambrook Joseph, (author.) & Laboratory, C. S. H. *Molecular cloning : a laboratory manual*. (Cold Spring Harbor, N.Y. Cold Spring Harbor Laboratory Press, 2012).

117. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).

118. Engler, C., Kandzia, R. & Marillonnet, S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS One* **3,** e3647 (2008).

119. Heap, J. T., Pennington, O. J., Cartman, S. T. & Minton, N. P. A modular system for Clostridium shuttle plasmids. *J. Microbiol. Methods* **78,** 79–85 (2009).

120. Bi, C., Su, P., Müller, J., Yeh, Y.-C., Chhabra, S. R., Beller, H. R., Singer, S. W. & Hillson, N. J. Development of a broad-host synthetic biology toolbox for *Ralstonia eutropha* and its application to engineering hydrocarbon biofuel production. *Microb. Cell Fact.* **12,** 107 (2013).

121. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 3746–3750 (1986).

122. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

123. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).

124. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).

125. Chiaromonte, F., Yap, V. B. & Miller, W. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* **126,** 115–126 (2002).

126. Sabathé, F. & Soucaille, P. Characterization of the cipA scaffolding protein and in vivo production of a minicellulosome in *Clostridium acetobutylicum*. *J. Bacteriol.* **185,** 1092–1096 (2003).

127. Genthner, B. R. S. & Bryant, M. P. Growth of *Eubacterium limosum* with carbon monoxide as the energy source. *Appl. Environ. Microbiol.* **43,** 70–74 (1982).

128. Genthner, B. R. S. & Bryant, M. P. Additional characteristics of one-carbon-compound utilization by *Eubacterium limosum* and *Acetobacterium woodii*. *Appl. Environ. Microbiol.* **53,** 471–476 (1987).

129. Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., Benders, G. A., Montague, M. G., Ma, L., Moodie, M. M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E. A., Young, L., Qi, Z.-Q., Segall-Shapiro, T. H., *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–56 (2010).

130. Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., *et al.* Design and synthesis of a minimal bacterial genome. *Science (80-.).* **351,** (2016).

131. Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R. & Church, G. M. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460,** 894–898 (2009).

132. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **36,** 52–57 (2008).

133. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE-a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res.* **43,** D298–D299 (2015).

134. Song,Y. and Cho, B. K. *Eubacterium limosum* strain ATCC 8486 chromosome, complete genome. *Unpublished* (2017). Available at: https://www.ncbi.nlm.nih.gov/nuccore/1174896550. (Accessed: 20th July 2017)

135. Fulghum, R. S., Beverly, B., Williams, P. P. & Baldwin, B. B. Antibiotic Susceptibility of Anaerobic Ruminal Bacterial. *Appl. Microbiol.* **16,** 301–307 (1968).

136. Millard, J. The Development of Forward and Reverse Genetic Tools in *Eubacterium limosum* - PhD thesis. (University of Nottingham, 2018).

137. Genthner, B. R., Davis, C. L. & Bryant, M. P. Features of rumen and sewage sludge strains of *Eubacterium limosum*, a methanol- and H_2-CO_2-utilizing species. *Appl. Environ. Microbiol.* **42,** 12–19 (1981).

138. Loubiere, P. & Lindley, N. D. The use of acetate as an additional co-substrate improves methylotrophi growth of the acetogenic anaerobe Eubacterium limosum when CO_2 fixation is rate-limiting. *J. Gen. Microbiol.* **137,** 2247–2251 (1991).

139. Chang, I. S., Kim, D., Kim, B. H. & Lovitt, R. W. Use of an industrial grade medium and medium enhancing effects on high cell density CO fermentation by *Eubacterium limosum* KIST612. *Biotechnol. Lett.* **29,** 1183–1187 (2007).

140. Fraser, C. M., Eisen, J. A., Nelson, K. E., Ian, T., Salzberg, S. L. & Paulsen, I. T. The Value of Complete Microbial Genome Sequencing (You Get What You Pay For). *J. Bacteriol.* **184,** 6403–6405 (2002).

141. Song, Y. & Cho, B.-K. Draft Genome Sequence of Chemolithoautotrophic Acetogenic Butanol-Producing *Eubacterium limosum* ATCC 8486. *Genome Announc.* **3,** e01564-14 (2015).

142. Brown, S. D., Nagaraju, S., Utturkar, S., De Tissera, S., Segovia, S., Mitchell, W., Land, M. L., Dassanayake, A. & Köpke, M. Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnol. Biofuels* **7,** 40 (2014).

143. Kim, T. Y., Sohn, S. B., Kim, Y. Bin, Kim, W. J. & Lee, S. Y. Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr. Opin. Biotechnol.* **23,** 617–623 (2012).

144. Lee, J., Jang, Y. S., Choi, S. J., Im, J. A., Song, H., Cho, J. H., Seung, D. Y., Terry Papoutsakis, E., Bennett, G. N. & Lee, S. Y. Metabolic engineering of *Clostridium acetobutylicum* ATCC 824 for isopropanol-butanol-ethanol fermentation. *Appl. Environ. Microbiol.* **78,** 1416–1423 (2012).

145. Duncan, S. H., Barcenilla, A., Stewart, C. S., Pryde, S. E. & Flint, H. J. Acetate utilization and butyryl coenzyme A (CoA): Acetate-CoA transferase in butyrate-producing bacteria from the human large intestine. *Appl. Environ. Microbiol.* **68,** 5186–5190 (2002).

146. Liew, F., Henstra, A. M., Köpke, M., Winzer, K., Simpson, S. D. & Minton, N. P. Metabolic engineering of *Clostridium autoethanogenum* for selective alcohol production. *Metab. Eng.* **40,** 104–114 (2017).

147. Adnan, N. A. A., Suhaimi, S. N., Abd-Aziz, S., Hassan, M. A. & Phang, L. Y. Optimization of bioethanol production from glycerol by *Escherichia coli* SS1. *Renew. Energy* **66,** 625–633 (2014).

148. Nikerel, E., Berkhout, J., Hu, F., Teusink, B., Reinders, M. J. T. & de Ridder, D. Understanding regulation of metabolism through feasibility analysis. *PLoS One* **7,** (2012).

149. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve. 1394–1403 (2004). doi:10.1101/gr.2289704

150. Angiuoli, S. V., Dunning Hotopp, J. C., Salzberg, S. L. & Tettelin, H. Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics* **12,** (2011).

151. Hertel, R., Rodríguez, D. P., Hollensteiner, J., Dietrich, S., Leimbach, A., Hoppert, M., Liesegang, H. & Volland, S. Genome-based identification of active prophage regions by next generation sequencing in *Bacillus licheniformis* DSM13. *PLoS One* **10,** 1–18 (2015).

152. Guérillot, R., Siguier, P., Gourbeyre, E., Chandler, M. & Glaser, P. The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. *Genome Biol. Evol.* **6,** 260–272 (2014).

153. Munoz-Lopez, M. & Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* **11,** 115–128 (2010).

154. Lees, J. & Gladstone, R. A. R-M systems go on the offensive. *Nat. Rev. Microbiol.* **13,** 131 (2015).

155. Croucher, N. J., Coupland, P. G., Stevenson, A. E., Callendrello, A., Bentley, S. D. & Hanage, W. P. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.* **5,** 1–12 (2014).

156. Sutter, V. L. & Finegold, S. M. Susceptibility of anaerobic bacteria to 23 antimicrobial agents. *Antimicrob. Agents Chemother.* **10,** 736–752 (1976).

157. Maddox, I. S., Steiner, E., Hirsch, S., Wessner, S., Gutierrez, N. a, Gapes, J. R. & Schuster, K. C. The cause of 'acid-crash' and 'acidogenic fermentations' during the batch acetone-butanol-ethanol (ABE-) fermentation process. *J. Mol. Microbiol. Biotechnol.* **2,** 95–100 (2000).

158. O'Brien, R. W. & Morris, J. G. The Ferredoxin-dependent reduction of chloramphenicol by *Clostridium acetobutylicum*. *J. Gen. Microbiol.* **67,** 265–271 (1971).

159. Goldman, E. & Green, L. H. *Practical Handbook of Microbiology: Second Edition*. *Journal of Chemical Information and Modeling* **53,** (2009).

160. Williams, D. R., Young, D. I. & Young, M. Conjugative plasmid transfer from *Escherichia coli* to *Clostridium acetobutylicum*. *Microbiology* **136,** 819–826

(1990).

161. Herbert, M., O'Keeffe, T. A., Purdy, D., Elmore, M. & Minton, N. P. Gene transfer into *Clostridium difficile* CD630 and characterisation of its methylase genes. *FEMS Microbiol. Lett.* **229,** 103–110 (2003).

162. Dong, H., Tao, W., Gong, F., Li, Y. & Zhang, Y. A functional *recT* gene for recombineering of Clostridium. *J. Biotechnol.* **173,** 65–67 (2014).

163. Hoffmeister, S., Gerdom, M., Bengelsdorf, F. R., Linder, S., Flüchter, S., Öztürk, H., Blümke, W., May, A., Fischer, R. J., Bahl, H. & Dürre, P. Acetone production with metabolically engineered strains of *Acetobacterium woodii*. *Metab. Eng.* **36,** 37–47 (2016).

164. Stratz, M., Sauer, U., Kuhn, A. & Durre, P. Plasmid Transfer Into the Homoacetogen *Acetobacterium woodii* By Electroporation and Conjugation. *Appl. Environ. Microbiol.* **60,** 1033–1037 (1994).

165. Dupuy, B. & Matamouros, S. Regulation of toxin and bacteriocin synthesis in *Clostridium* species by a new subgroup of RNA polymerase sigma-factors. *Res. Microbiol.* **157,** 201–205 (2006).

166. Zhang, Y., Grosse-Honebrink, A. & Minton, N. P. A Universal Mariner Transposon System for Forward Genetic Studies in the Genus *Clostridium*. *PLoS One* **10,** 1–21 (2015).

167. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).

168. Hanahan, D. Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166,** 557–580 (1983).

169. Roberts, M. A. J., Cranenburgh, R. M., Stevens, M. P. & Oyston, P. C. F. Synthetic biology: biology by design. *Microbiology* **159,** 1219–1220 (2013).

170. Khalil, A. S. & Collins, J. J. Synthetic Biology : Applications Come of Age. **11,** 367–379 (2010).

171. Roehr, M. *Biotechnology: Products of Primary Metabolism*. *VCH Publishers Inc.* **6,** (1996).

172. Kibby, C. L. & Hall, W. K. Studies of acid catalyzed reactions: XII. Alcohol decomposition over hydroxyapatite catalysts. *J. Catal.* **29,** 144–159 (1973).

173. Bruno, T. J., Wolk, A. & Naydich, A. Composition-Explicit Distillation Curves for Mixtures of Gasoline with Four-Carbon Alcohols (Butanols). *Energy & Fuels* **23,** 2295–2306 (2009).

174. Gerischer, U. & Dürre, P. Cloning, sequencing, and molecular analysis of the acetoacetate decarboxylase gene region from *Clostridium acetobutylicum*. *J. Bacteriol.* **172,** 6907–6918 (1990).

175. NIIR. *Modern Technology of Petroleum, Greases, Lubricants & Petro Chemicals (2nd Revised Edition)*. (NIIR, 2017).

176. Zakoshansky, V. M. The cumene process for phenol-acetone production. *Pet. Chem.* **47,** 273–284 (2007).

177. Rajeev M. Pandia. The phenol-acetone value chain: prospects and opportunities. (2013). Available at: https://pdfs.semanticscholar.org/.../3d2ff91e455aae3e417f5e1b7b572b9c2ca3.pdf%0A%0A.

178. Kusakabe, T., Tatsuke, T., Tsuruno, K. & Hirokawa, Y. Engineering a synthetic pathway in cyanobacteria for isopropanol production directly from carbon dioxide and light. *Metab. Eng.* **20,** 101–108 (2013).

179. Artzi, L., Bayer, E. A. & Moraïs, S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15,** 83–95 (2016).

180. Vazana, Y., Morais, S., Barak, Y., Lamed, R. & Bayer, E. A. Designer cellulosomes for enhanced hydrolysis of cellulosic substrates. *Methods Enzymol.* **510,** 429–452 (2012).

181. Wheeldon, I., Minteer, S. D., Banta, S., Barton, S. C., Atanassov, P. & Sigman, M. Substrate channelling as an approach to cascade reactions. *Nat. Chem.* **8,** 299–309 (2016).

182. Moon, T. S., Dueber, J. E., Shiue, E. & Prather, K. L. J. Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *Escherichia coli*. *Metab. Eng.* **12,** 298–305 (2010).

183. Dueber, J. E., Wu, G. C., Malmirchegini, G. R., Moon, T. S., Petzold, C. J., Ullal, A. V, Prather, K. L. J. & Keasling, J. D. Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* **27,** 753–759 (2009).

184. Jindou, S., Ito, Y., Mito, N., Uematsu, K., Hosoda, A. & Tamura, H. Engineered Platform for Bioethylene Production by a Cyanobacterium Expressing a Chimeric Complex of Plant Enzymes. *ACS Synth. Biol.* **3,** 487–496 (2014).

185. Kim, S. & Hahn, J.-S. Synthetic scaffold based on a cohesin-dockerin interaction for improved production of 2,3-butanediol in *Saccharomyces cerevisiae*. *J. Biotechnol.* **192 Pt A,** 192–196 (2014).

186. ICH GUIDELINE. *ICH Topic Q2 (R1) Validation of Analytical Procedures : Text and Methodology*. *International Conference on Harmonization* **1994,** (1996).

187. Shrivastava, A. & Gupta, V. Methods for the determination of limit of detection and limit of quantitation of the analytical methods. *Chronicles Young Sci.* **2,** 21 (2011).

188. Chang, I. S., Kim, D. H., Kim, B. H., Shin, P. K., Sung, H. C. & Lovitt, R. W. CO fermentation of *Eubacterium limosum* KIST612. *J. Microbiol. Biotechnol.* **8,** 134–140 (1998).

189. Fast, A. G. & Papoutsakis, E. T. Stoichiometric and energetic analyses of non-photosynthetic CO_2 fixation pathways to support the production of fuels and chemicals Stoichiometric Model of Carbon Fixation. **158,** 19716 (2012).

190. Bruno-Barcena, J. M., Chinn, M. S. & Grunden, A. M. Genome Sequence of the Autotrophic Acetogen *Clostridium autoethanogenum* JA1-1 Strain DSM 10061, a Producer of Ethanol from Carbon Monoxide. *Genome Announc.* **1,** e00628-13-e00628-13 (2013).

191. Bruno-Barcena, J. M., Chinn, M. S. & Grunden, A. M. Genome Sequence of the

Autotrophic Acetogen *Clostridium autoethanogenum* JA1-1 Strain DSM 10061, a Producer of Ethanol from Carbon Monoxide. *Genome Announc.* **1,** e00628-13-e00628-13 (2013).

192. Zhao, Y., Tomas, C. A., Rudolph, F. B., Papoutsakis, E. T. & Bennett, G. N. Intracellular butyryl phosphate and acetyl phosphate concentrations in *Clostridium acetobutylicum* and their implications for solvent formation. *Appl. Environ. Microbiol.* **71,** 530–537 (2005).

193. Emerson, M. Analysis of the metabolic pathways in the gas fermenting *Eubacterium limosum* - Masters thesis. (University of Nottingham, 2017).

194. Moncrief, J. S. & Barroso, L. a. Positive regulation of *Clostridium difficile* toxins. **65,** 1105–1108 (1997).

195. Terpe, K. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **72,** 211 (2006).

196. Grosse-Honebrink, A. Forward and Reverse Genetics in Industrially Important Clostridia. (2017).

197. Alper, H., Fischer, C., Nevoigt, E., Stephanopoulos, G., Demasi, J., Huh, K., Nakatani, Y. & Mu, K. For the article ‘“Tuning genetic control through promoter engineering,”’ by Hal Alper, Curt Fischer, Elke Nevoigt, and Gregory Stephanopoulos, which appeared in issue 36, September 6, 2005, of. *PNAS* **103,** 3006–3007 (2006).

198. Yang, G., Jia, D., Jin, L., Jiang, Y., Wang, Y., Jiang, W. & Gu, Y. Rapid Generation of Universal Synthetic Promoters for Controlled Gene Expression in Both Gas-Fermenting and Saccharolytic *Clostridium* Species. *ACS Synth. Biol.* **6,** 1672–1678 (2017).

199. Martin-Verstraete, I., Peltier, J. & Dupuy, B. The regulatory networks that control *Clostridium difficile* toxin synthesis. *Toxins (Basel).* **8,** 1–24 (2016).

200. Castellana, M., Wilson, M. Z., Xu, Y., Joshi, P., Cristea, I. M., Rabinowitz, J. D., Gitai, Z. & Wingreen, N. S. Enzyme clustering accelerates processing of intermediates through metabolic channeling. *Nat. Biotechnol.* **32,** 1011–1018 (2014).

201. Collas, F., Kuit, W., Clément, B., Marchal, R., López-Contreras, A. M. & Monot, F. Simultaneous production of isopropanol, butanol, ethanol and 2,3-butanediol by *Clostridium acetobutylicum* ATCC 824 engineered strains. *AMB Express* **2,** 45 (2012).

202. Humphreys, C. M., McLean, S., Schatschneider, S., Millat, T., Henstra, A. M., Annan, F. J., Breitkopf, R., Pander, B., Piatek, P., Rowe, P., Wichlacz, A. T., Woods, C., Norman, R., Blom, J., Goesman, A., Hodgman, C., Barrett, D., Thomas, N. R., Winzer, K., *et al.* Whole genome sequence and manual annotation of *Clostridium autoethanogenum*, an industrially relevant bacterium. *BMC Genomics* **16,** 1085 (2015).

203. Nagaraju, S., Davies, N. K., Walker, D. J. F., Köpke, M. & Simpson, S. D. Genome editing of *Clostridium autoethanogenum* using CRISPR/Cas9. *Biotechnol. Biofuels* **9,** 219 (2016).

204. Jones, S. W., Fast, A. G., Carlson, E. D., Wiedel, C. A., Au, J., Antoniewicz, M. R., Papoutsakis, E. T. & Tracy, B. P. CO_2 fixation by anaerobic non-photosynthetic mixotrophy for improved carbon conversion. *Nat. Commun.* **7,** 12800 (2016).

205. Lindsay, A. F., Stamper, B. L., Koepke, M., Simpson, S., Liew, F. & Chen, W. Fermentation process for producing isopropanol using a recombinant microorganism. **1,** (2012).

206. Wu, J. & Filutowicz, M. Hexahistidine (His6)-tag dependent protein dimerization: A cautionary tale. *Acta Biochimica Polonica* **46,** 591–599 (1999).

207. Chado, G. R., Stoykovich, M. P. & Kaar, J. L. Role of Dimension and Spatial Arrangement on the Activity of Biocatalytic Cascade Reactions on Scaffolds. *ACS Catal.* **6,** 5161–5169 (2016).

208. Fu, J., Liu, M., Liu, Y., Woodbury, N. W. & Yan, H. Interenzyme Substrate Diffusion for an Enzyme Cascade Organized on Spatially Addressable DNA Nanostructures. *J. Am. Chem. Soc.* **134,** 5516–5519 (2012).

209. Idan, O. & Hess, H. Origins of activity enhancement in enzyme cascades on scaffolds. *ACS Nano* **7,** 8658–8665 (2013).

210. Wiesenborn, D. P., Rudolph, F. B. & Papoutsakis, E. T. Thiolase from *Clostridium acetobutylicum* ATCC 824 and Its Role in the Synthesis of Acids and Solvents. *Appl Env. Microbiol* **54,** 2717–2722 (1988).

211. Elena, C., Ravasi, P., Castelli, M. E., Peir, S. & Menzella, H. G. Expression of codon optimized genes in microbial systems: Current industrial applications and perspectives. *Front. Microbiol.* **5,** 1–8 (2014).

212. Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33,** 1141–1153 (2005).

213. May, A., Fischer, R. J., Maria Thum, S., Schaffer, S., Verseck, S., Dürre, P. & Bahl, H. A modified pathway for the production of acetone in *Escherichia coli*. *Metab. Eng.* **15,** 218–225 (2013).

214. Ithayaraja, M., Janardan, N., Wierenga, R. K., Savithri, H. S. & Murthy, M. R. N. Crystal structure of a thiolase from *Escherichia coli* at 1.8 A resolution. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* **72,** 534–544 (2016).

215. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117,** 119–128 (2015).

216. Wilson, G. G. & Murray, N. E. Restriction and Modification. *Annu. Rev. Genet.* **25,** 585–627 (1991).

217. Kusano, K., Naito, T., Handa, N. & Kobayashi, I. Restriction-modification systems as genomic parasites in competition for specific sequences. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 11095–11099 (1995).

218. Meisel, A., Bickle, T. a, Kriiger, D. H. & Schroeder, C. Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature* **355,** 467–469 (1992).

219. Fukuda, E., Kaminska, K. H., Bujnicki, J. M. & Kobayashi, I. Cell death upon epigenetic genome methylation: a novel function of methyl-specific

deoxyribonucleases. *Genome Biol.* **9,** R163–R163 (2008).

220. Mermelstein, L. D. & Papoutsakis, E. T. In vivo methylation in *Escherichia coli* by the *Bacillus subtilis* phage phi 3T I methyltransferase to protect plasmids from restriction upon transformation of *Clostridium acetobutylicum* ATCC 824. *Appl. Environ. Microbiol.* **59,** 1077–1081 (1993).

221. Dong, H., Zhang, Y., Dai, Z. & Li, Y. Engineering *Clostridium* strain to accept unmethylated DNA. *PLoS One* **5,** 1–8 (2010).

222. Kwak, J., Jiang, H. & Kendrick, K. E. Transformation using in vivo and in vitro methylation in *Streptomyces griseus*. *FEMS Microbiol. Lett.* **209,** 243–248 (2002).

223. Edwards, R. A., Helm, R. A. & Maloy, S. R. Increasing DNA transfer efficiency by temporary inactivation of host restriction. *Biotechniques* **26,** 892–4, 896, 898 passim (1999).

224. Liew, F. Metabolic Engineering of *Clostridium autoethanogenum* - PhD thesis. (University of Nottingham, 2015).

225. Holmes, R. K. & Jobling, M. G. *Genetics*. *Medical Microbiology* (1996).

226. Hedges, R. W. R factors from *Proteus mirabilis* and *P. vulgaris*. *J. Gen. Microbiol.* **87,** 301–311 (1975).

227. Kenneth, J. R. & Ray, C. G. *Sherris Medical Microbiology*. *Sherris Medical Microbiology* (McGraw-Hill Medical, 2004). doi:10.1036/0838585299

228. Elhai, J., Vepritskiy, A., Muro-Pastor, A., Flores, E. & Wolk, C. Reduction of conjugal transfer efficiency by three restriction activities of *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* **179,** 1998–2005 (1997).

229. Blow, M. J., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., Froula, J., Kang, D. D., Malmstrom, R. R., Morgan, R. D., Posfai, J., Singh, K., Visel, A., Wetmore, K., Zhao, Z., Rubin, E. M., Korlach, J., Pennacchio, L. A. & Roberts, R. J. The Epigenomic Landscape of Prokaryotes. *PLoS Genet.* **12,** 1–28 (2016).

230. Márquez, A. J., Stougaard, J., Udvardi, M., Parniske, M., Spaink, H., Saalbach, G., Webb, J. & Chiurazzi, M. *Lotus japonicus handbook*. (Springer, 2005).

231. Siebor, E. & Neuwirth, C. *Proteus* genomic island 1 (PGI1), a new resistance genomic island from two Proteus mirabilis French clinical isolates. *J. Antimicrob. Chemother.* **69,** 3216–3220 (2014).

232. Gaulton, T., Misra, R., Rose, G., Baybayan, P., Hall, R., Freeman, J., Turton, J., Picton, S., Korlach, J., Gharbia, S. & Shah, H. Complete Genome Sequence of the Hypervirulent Bacterium *Clostridium difficile* Strain G46, Ribotype 027. *Genome Announcements* **3,** (2015).

233. Burns, D. A., Heeg, D., Cartman, S. T. & Minton, N. P. Reconsidering the sporulation characteristics of hypervirulent *Clostridium difficile* BI/NAP1/027. *PLoS One* **6,** e24894 (2011).

234. Smith, T. J., Hill, K. K., Foley, B. T., Detter, J. C., Munk, A. C., Bruce, D. C., Doggett, N. A., Smith, L. A., Marks, J. D., Xie, G. & Brettin, T. S. Analysis of the neurotoxin complex genes in *Clostridium botulinum* A1-A4 and B1 strains: BoNT/A3, /Ba4

and /B1 clusters are located within plasmids. *PLoS One* **2,** e1271 (2007).

235. Zhang, G., Wang, W., Deng, A., Sun, Z., Zhang, Y., Liang, Y., Che, Y. & Wen, T. A Mimicking-of-DNA-Methylation-Patterns Pipeline for Overcoming the Restriction Barrier of Bacteria. *PLoS Genet.* **8,** (2012).

236. Kolek, J., Sedlar, K., Provaznik, I. & Patakova, P. Dam and Dcm methylations prevent gene transfer into *Clostridium pasteurianum* NRRL B-598: development of methods for electrotransformation, conjugation, and sonoporation. *Biotechnol. Biofuels* **9,** 14 (2016).

237. Stabler, R. A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T. D., Sebaihia, M., Quail, M. A., Rose, G., Gerding, D. N., Gibert, M., Popoff, M. R., Parkhill, J., Dougan, G. & Wren, B. W. Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol.* **10,** R102 (2009).

238. Hedges, R. W. & Jacob, A. E. Transposition of Ampicillin Resistance from RP4 to Other Replicons. *Mol. Gen. Genet.* **132,** 31–40 (1974).

239. Macherey-Nagel. Plasmid DNA Purification User Manual. 1–27 (2010).

240. Casadesús, J. & Low, D. Epigenetic Gene Regulation in the Bacterial World. *Microbiol. Mol. Biol. Rev.* **70,** 830–856 (2006).

241. Purdy, D., O'Keeffe, T. A. T., Elmore, M., Herbert, M., McLeod, A., Bokori-Brown, M., Ostrowski, A. & Minton, N. P. Conjugative transfer of clostridial shuttle vectors from *Escherichia coli* to *Clostridium difficile* through circumvention of the restriction barrier. *Mol. Microbiol.* **46,** 439–452 (2002).

242. Woods, C. Forward Genetics in Clostridial Acetogens as a Route to Process Improvements - PhD thesis. (University of Nottingham, 2018).

243. Schlagman, S. L., Hattman, S. & Marinus, M. G. Direct role of the *Escherichia coli* Dam DNA methyltransferase in methylation-directed mismatch repair. *Journal of Bacteriology* **165,** 896–900 (1986).

244. Wang, T. C. & Smith, K. C. Inviability of dam recA and dam recB cells of *Escherichia coli* is correlated with their inability to repair DNA double-strand breaks produced by mismatch repair. *J. Bacteriol.* **165,** 1023–5 (1986).

245. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 6640–5 (2000).

246. Wang, J., Sarov, M., Rientjes, J., Fu, J., Hollak, H., Kranz, H., Xie, W., Stewart, A. F. & Zhang, Y. An improved recombineering approach by adding RecA to lambda Red recombination. *Mol. Biotechnol.* **32,** 43–53 (2006).

247. Costantino, N. & Court, D. L. Enhanced levels of Red-mediated recombinants in mismatch repair mutants. *Proc. Natl. Acad. Sci.* **100,** 15748–15753 (2003).

248. Ng, C. Y., Farasat, I., Maranas, C. D. & Salis, H. M. Rational design of a synthetic Entner-Doudoroff pathway for improved and controllable NADPH regeneration. *Metab. Eng.* **29,** 86–96 (2015).

249. Wang, Y. & Pfeifer, B. A. 6-Deoxyerythronolide B production through chromosomal localization of the deoxyerythronolide B synthase genes in

Escherichia coli. *Metab. Eng.* **10,** 33–38 (2008).

250. Tran-Betcke, A., Behrens, B., Noyer-Weidner, M. & Trautner, T. A. DNA methyltransferase genes of *Bacillus subtilis* phages: comparison of their nucleotide sequences. *Gene* **42,** 89–96 (1986).

251. Mermelstein, L. D., Welker, N. E., Bennett, G. N. & Papoutsakis, E. T. Expression of cloned homologous fermentative genes in *Clostridium acetobutylicum* ATCC 824. *Biotechnology. (N. Y).* **10,** 190–5 (1992).

252. Malygin, E. G., Zinoviev, V. V., Evdokimov, A. A., Lindstrom, W. M., Reich, N. O. & Hattman, S. DNA (cytosine-N4-)- and -(adenine-N6-)-methyltransferases have different kinetic mechanisms but the same reaction route: A comparison of M.BamHI and T4 Dam. *J. Biol. Chem.* **278,** 15713–15719 (2003).

253. Ackermann, G., Tang, Y. J., Henderson, J. P., Rodloff, A. C., Silva, J. J. & Cohen, S. H. Electroporation of DNA sequences from the pathogenicity locus (PaLoc) of toxigenic *Clostridium difficile* into a non-toxigenic strain. *Mol. Cell. Probes* **15,** 301–306 (2001).

254. Kirk, J. A., Banerji, O. & Fagan, R. P. Characteristics of the *Clostridium difficile* cell envelope and its importance in therapeutics. *Microbial Biotechnology* **10,** 76–90 (2017).

255. Marshall, K. M., Bradshaw, M. & Johnson, E. A. Conjugative Botulinum Neurotoxin-Encoding Plasmids in *Clostridium botulinum*. *PLoS One* **5,** 1–15 (2010).

256. Dineen, S. S., Bradshaw, M. & Johnson, E. A. Cloning, Nucleotide Sequence, and Expression of the Gene Encoding the Bacteriocin Boticin B from *Clostridium botulinum* Strain 213B. *Appl. Environ. Microbiol.* **66,** 5480–5483 (2000).

257. Bradshaw, M., Goodnough, M. C. & Johnson, E. A. Conjugative Transfer of the *Escherichia coli – Clostridium perfringens*. *Plasmid* **40,** 233–237 (1998).

258. Marinus, M. G. & Morris, N. R. Isolation of deoxyribonucleic acid methylase mutants of *Escherichia coli* K-12. *J. Bacteriol.* **114,** 1143–1150 (1973).

259. Paul, D., Austin, F. W., Arick, T., Bridges, S. M., Burgess, S. C., Dandass, Y. S. & Lawrence, M. L. Genome sequence of the solvent-producing bacterium *Clostridium carboxidivorans* strain P7T. *J. Bacteriol.* **192,** 5554–5555 (2010).

260. Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J. & Turner, A. K. Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants. *Genome Research* **19,** 2308–2316 (2009).