Development of molecular tools for optimisation of C1 gas fermentation in acetogens



Peter N Rowe

Synthetic Biology Research Centre
School of Life Sciences
University of Nottingham

Primary Supervisor: Prof. Nigel P Minton

Dissertation submitted for the degree of
Doctor of Philosophy

Nottingham, UK September 2017

Declaration

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

Peter N. Rowe

September 2017

Abstract

Access to renewable energy and reduction of carbon emissions represent two major issues facing humankind in the twenty first century and beyond. The underlying driving forces behind both are multi-faceted and often intrinsically connected, ranging from environmental concerns over climate change to improving economic security through self-sustaining energy production.

Possible solutions to reliance on non-renewable, carbon-emitting fossil fuels have been explored over recent decades, with significant interest placed on biofuels. Due to ease of integration into liquid-based petrochemical fuel infrastructure, these renewable alternatives have been a consistent topic of both industrial and academic interest. Despite offering renewable energy, conventional crop-based biofuel production has faced criticism due to consumption of land, water and other resources associated with agriculture.

Acetogens provide a solution to conventional biofuel production due to their utilisation of carbon monoxide and carbon dioxide gas as carbon and energy sources, rather than plant matter. This allows generation of a range of chemical products from a broad range of sources, including industrial waste gases and gasified solid waste. Acetogens offer the double benefit of both renewable energy production, and carbon emission sequestation.

This study outlines the development of genetic tools to provide a foundation for using synthetic biology approaches to improve performance of acetogens as industrial chassis. Specifically, development of tools and techniques for the acetogen *Clostridium autoethanogenum* are described, with further applications of such technology to other clostridia.

Acknowledgements

Many thanks firstly to Professor Nigel Minton for his marvellous support and advice over the course of this study. Thank you also to Associate Professor Klaus Winzer and Dr Anne Henstra in their capacity as secondary supervisors, and to the BBSRC for their funding, without any of which this research would not be been possible.

I would also like to express my graditute specifically to Dr Christopher Humphreys, who provided endless advice and support. In addition, a special thank you is due to fellow PhD candidates and Bench A inhibitants Craig Woods, Robert Mansfield, James Millard and Andrew Dempster (honourable member), for making the last four years so enjoyable at the best of times, and bearable at the worst of times.

Work on DNA transfer was conducted with fellow PhD candidate Craig Woods and early efforts in developing CRISPR-based tools were conducted in collaboration with members of Professor Weihong Jiang's lab at the Shanghai Institute for Biological Sciences.

Outside of academia, I would like to thank my family and friends for their help and encouragement over the last four years. And finally, I would like to thank my fiancé Hymke Moens for her unwavering support, kind words and patience.

Publications

Humphreys, C.M. et al., 2015. Whole genome sequence and manual annotation of *Clostridium autoethanogenum*, an industrially relevant bacterium. *BMC Genomics*, pp.1–10. Available at: http://dx.doi.org/10.1186/s12864-015-2287-5.

Huang, H. et al., 2016. CRISPR/Cas9-Based Efficient Genome Editing in *Clostridium ljungdahlii*, an Autotrophic Gas-Fermenting Bacterium. *ACS Synthetic Biology*, 5(12), pp.1355–1361. Available at: http://pubs.acs.org/doi/abs/10.1021/acssynbio.6b00044.

Abbreviations

%	Percentage
° C	Degrees Celsius
1G	First generation
2G	Second generation
3G	Third generation
4G	Fourth generation
2,3-BD	2,3-Butanediol
5-FOA	5-Flurooratic Acid
ACE	Allele-Coupled Exchange
ATCC	American Type Culture Collection
BLAST	Basic Local Alignment Search Tool
bp	Base pair
C1	Single carbon gases
Cas	CRISPR-associated
CFU	Colony forming unit
CLAU	*Clostridium autoethanogenum*
cm	Centimetre
Cm	Chloramphenicol
CO	Carbon monoxide
CO2	Carbon dioxide
CoA	Co-enzyme A
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
DAM	DNA adenine methyltransferase
DCM	DNA cysteine methyltransferase
DMSO	Dimethyl sulfide
DNA	Deoxyribonucleic acid
DSB	Double Strand Break
dsDNA	Double stranded DNA
ELM	*Eubacterium limosum*
Erm	Erythromycin
EtOH	Ethanol
g	Gram
h	Hour
H_2	Hydrogen
HDR	Homology
HiFi	High Fidelity
kB	Kilobase pair
kDa	Kilodalton
L	Litre
LB	Luria-Bertani medium
LHA	Left Homology Arm
M	Molar

μγ	Microgram
mg	Milligram
min	Minutes
μL	Microlitres
mL	Millilitre
μM	Micromolar
mM	Millimolar
mRNA	Messenger RNA
MCS	Multiple cloning site
NAD(P)+	Nicotinamide adenine dinucleotide (Phosphate), oxidised
NAD(P)H	Nicotinamide adenine dinucleotide (Phosphate), reduced
NCIMB	National Collection of Industrial Food and Marine Bacteria
NHEJ	Non-Homologous End Joining
OD	Optical Density
OH	Overhang
ORF	Open Reading Frame
PAM	Protospacer adjacent motif
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PYG	Peptone Yeast Glucose media
qRT-PCR	Quantitative Real Time PCR
RBS	Ribosome Binding Site
rcf	Relative centrifugal force
RHA	Right Homology Arm
RM	Restriction-Modification
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
SBRC	Synthetic Biology Research Centre (Nottingham)
scr	Screening
SNP	Single Nucleotide Polymorphism
SOE	Splicing by Overlap Extension
ssDNA	Single stranded DNA
TAE	Tris base Acetic acid EDTA buffer
Tm	Thiamphenicol
tRNA	Transfer RNA
V	Volts
w/v	Weight per volume
WT	Wild Type
YTF	Yeast Tryptone Fructose media
Ω	Resistance

Figures

Figure 1.1: UK crude oil production and demand 1998 – 2035 5
Figure 1.2: Global crude oil production 2016 6
Figure 1.3: Crude oil first purchase price 1975 – 2016 7
Figure 1.4: Phylogenetic tree of a range of acetogens 15
Figure 1.5: Transmission electron microscopy of *C. autoethanogenum* 16
Figure 1.6: Wood-Ljungdahl and acetogenesis pathways in acetogens 21
Figure 1.7: Acetate, ethanol and 2,3-butanediol generation in acetogens 27
Figure 1.8: Example of synthetic acetone production pathway 29
Figure 1.9: pMTL80000 vector series 32
Figure 1.10: *E. coli* Dam and Dcm methylation 37
Figure 1.11: Schematic overview of ClosTron gene disruption 40
Figure 1.12: Allelic exchange methodologies 45
Figure 1.13: Overview of CRISPR system classification 52
Figure 1.14: Mechanism of action for Type I CRISPR systems 57
Figure 1.15: CRISPR-Cas9 genome editing in prokaryotes 65
Figure 2.1: SOEing PCR 108
Figure 2.2: GenDB genome annotation pipeline 110
Figure 2.3: Design of pMTL007-EC2-pta 112
Figure 3.1: *C. autoethanogenum* sequence discrepancies with Brown *et al.* (2014) 128
Figure 3.2: *C. autoethanogenum* DSM10061 RM systems 132
Figure 3.3: Comparison of Sexpress and CA434 conjugation frequencies 134
Figure 3.4: BLASTX search results of CLAU_1607 (*cauIII*) 136
Figure 3.5: BLASTX search results of CLAU_2858 (*cauIV)* 137
Figure 3.6: Δ*cauII* colony PCR 139
Figure 3.7: Comparison of wild type and Δ*cauII* conjugation efficiencies 140
Figure 3.8: pMTL-CLAU-del-pyrE plasmid map 142
Figure 3.9: *C. autoethanogenum* Δ*pyrE* strain analysis 143
Figure 3.10: *C. autoethanogenum* RBS library analysis 147
Figure 4.1: Schematic representation of colony PCR screening for knockout detection 160
Figure 4.2: CRISPR-Cas9 knockout generation in *C. ljungdahlii* 161
Figure 4.3: pMTL40000 vector design 163
Figure 4.4: sgRNA module generation via SOEing PCR 166
Figure 4.5: Modification of sgRNA via ssDNA oligo HiFi 168
Figure 4.6: *C. autoethanogenum* Δ*pyrE* colony PCR 169
Figure 4.7: *E. limosum* Δ*cas3* colony PCR 171
Figure 4.8: Sequence and structure comparison between wild type and optimised sgRNA used by Xu *et al* 176
Figure 4.9: StCas9 *G. thermoglucosidius* Δ*ack* colony PCR 177
Figure 4.10: Sequence comparison of *S. pyogenes cas9* (wild type) and pMTL40000 "*trCas9*" 179
Figure 4.11: Relative expression of putative trC9 RBS 181
Figure 4.12: Bookmark complementation of knockouts 183
Figure 4.13: *C. autoethanogenum pta* Bookmark repair colony PCR 186
Figure 4.14: BM4-12 *pyrE* insertion colony PCR 188
Figure 4.15: *C. autoethanogenum* Col3 strain comparison 190
Figure 4.16: Comparison of CAETHG_3768 in wild type and colony 3 strains of *C. autoethanogenum* 192
Figure 4.17: SNP2 modification *Eco*RV diagnostic screen 195

Figure 5.1: *C. autoethanogenum* CRISPR region 205
Figure 5.2: Protospacer 1.27a putative PAM analysis 213
Figure 5.3: Reductive PAM assay of 3' 1.27a putative PAM 216
Figure 5.4: Predicted *C. autoethanogenum* 3' PAM 216
Figure 5.5: Bottom-up PAM screen 218
Figure 5.6: GFP expression in *E. coli* BL21-AI pACYC-CLAU-CRISPR cultures harbouring pET-GFP vectors 221
Figure 5.7: *C. autoethanogenum* Δ*cas3* colony PCR and CRISPR activity comparison 224
Figure 5.8: *C. autoethanogenum* eCas strain development 226
Figure 5.9: PAM and spacer sequences available in the *C. butyricum* thiolase promoter 229
Figure 6.1: Acetate, ethanol and 2,3-butanediol generation in acetogens 250
Figure 6.2: Plasmid map of pMTL-ACE-del-pta 252
Figure 6.3: pMTL43151-CLAU-ack and pMTL43151-CLAU-pta 255
Figure 6.4: *C. autoethanogenum* Δ*ack* colony PCR 256
Figure 6.5: *C. autoethanogenum* Δ*pta* colony PCR 257
Figure 6.6: *C. autoethanogenum pta* repair colony PCR 258
Figure 6.7: CO growth comparison of Δ*pta*, repaired *pta*, and Col3 strain of wild type *C. autoethanogenum* 260
Figure 8.1: Promoter strength in *C. autoethanogenum* 274

Tables

Table 1.1: Comparison of first, second, third and fourth generation biofuels 11
Table 1.2: Overview of restriction-methylation systems in prokaryotes..... 35
Table 2.1: Bacterial strains used in this study..... 71
Table 2.2: Plasmids used in this study 73
Table 2.3: Oligonuclotides used in this study..... 79
Table 2.4: DNA sequences synthetised for this study 91
Table 2.5: Suppliers used in this study 92
Table 2.6: Clostridial growth media 97
Table 2.7: Stock solutions for growth media 98
Table 2.8: Media supplements used in this study 99
Table 2.9: Example of reaction conditions of Q5 Polymerase 107
Table 2.10: Example of reaction conditions of DreamTaq Polymerase..... 107
Table 3.1: Gram-positive replicon characterisation plasmids 122
Table 3.2: CDS discrepancies between with Brown *et al.* (2014) 125
Table 3.3: RBS sequences for study in *C. autoethanogenum* 146
Table 4.1: Examples of use of pMTL40000 usage other bacteria..... 172
Table 4.2: Published PAM/sgRNA sequences for StCas9 genome editing..... 174
Table 4.3: StCas9 PAMs for *G. thermoglucosidius* genome editing 175
Table 4.4: Bookmark seed sequences used in this study..... 184
Table 4.5: SNP/SND modified target loci and restriction enzyme introduction..... 194
Table 5.1: Spacer analysis of *C. autoethanogenum* CRISPR locus 1..... 207
Table 5.2: Spacer analysis of *C. autoethanogenum* CRISPR locus 2..... 208
Table 5.3: Spacer analysis of *C. autoethanogenum* CRISPR locus 3..... 209
Table 5.4: Putative PAM and protospacer sequences for 1.27a PAM analysis..... 213
Table 5.5: Plasmid library for top down *C. autoethanogenum* PAM elucidation 214
Table 5.6: *E. limosum* CRISPR loci..... 231
Table 5.7: Putative protospacers of *E. limosum* ATCC 8486..... 231
Table 5.8: *E. limosum* ATCC 8486 PAM analysis plasmids 232
Table 5.9: Analysis of clostridial CRISPR systems 234
Table 5.10: CRISPR loci of *C. aceticum* and *A. woodii* identified in this study..... 235
Table 5.11: CRISPR loci of industrially relvent clostridia identified in this study..... 237
Table 5.12 CRISPR loci of clinically relvent *Clostrdium* identified in this study 238
Table 5.13: Potential PAMs identified in *Clostridium*..... 239

Contents

Declaration iii
Abstract iv
Acknowledgements v
Publications vi
Abbreviations vii
Figures ix
Tables xi
Contents xii
1. Introduction 2
1.1. Renewable energy: demand, production and limitations of biofuels 4
1.1.1. Finite fossil fuels 4
1.1.2. Energy security and self-sufficiency 5
1.1.3. Carbon concerns 8
1.1.4. Biofuels 10
1.2. Acetogenic bacteria for biofuel generation 13
1.2.1. Acetogen overview 14
1.2.2. *Clostridium autoethanogenum* 16
1.2.3. Substrate utilisation in acetogens 17
1.2.3.1. Organic substrates 17
1.2.4. The Wood-Ljungdahl pathway 18
1.2.4.1. The methyl branch 18
1.2.4.1.1. Formate synthesis 19
1.2.4.1.2. Formyl- and methyl- THF synthesis 21
1.2.4.1. The carbonyl branch and acetyl-CoA generation 22
1.2.5. Industrially relevant products from acetogens 23
1.2.5.1. Acetate synthesis 24
1.2.5.2. Ethanol synthesis 24
1.2.5.3. 2,3-butanediol synthesis 25
1.2.6. Pathway engineering in acetogens 27
1.2.6.1. Optimisation of existing products in acetogens 28
1.2.6.2. Plasmid-based introduction of novel product pathways in acetogens 28
1.2.6.3. Limitations to pathway engineering in acetogens 30
1.3. DNA transfer and genome editing in Clostridia 31
1.3.1. Plasmid vectors 32
1.3.2. DNA transfer 33
1.3.2.1. Conjugation 33
1.3.2.2. Electroporation 33
1.3.3. Barriers to DNA transfer 34
1.3.3.1. Restriction-modification systems 35
1.3.3.2. Avoiding host RM systems 37
1.4. Clostridial genetic engineering and genome editing 38
1.4.1. Conventional genetic engineering in Clostridia 39

1.4.1.1. ClosTron 39
1.4.1.2. Allelic exchange methodologies 41
1.4.1.2.1. ACE-mediated *pyrE* deficient strain generation 42
1.4.1.3. Limitations of conventional clostridial genetic editing strategies 45
1.4.2. Prokaryotic genetic engineering advancements 46
1.4.2.1. Recombineering 46
1.4.2.2. Engineered nucleases for genome editing 48
1.5. CRISPR systems and their applications 49
1.5.1. From genome unknown to genome editing 50
1.5.2. Classification 51
1.5.3. Spacer acquisition 53
1.5.4. Class 1 CRISPR systems 56
1.5.4.1. Class 1 CRISPR loci and RNA processing 56
1.5.4.2. Class 1 effector complexes and target cleavage 56
1.5.4.1. Type IV Cas systems 58
1.5.5. Class 2 CRISPR systems 59
1.5.5.1. Type II CRISPR systems 59
1.5.5.1. Type V CRISPR systems 59
1.5.6. Genome editing with CRISPR systems 60
1.5.6.1. Cas9 genome editing 61
1.5.6.2. Cpf1 genome editing 62
1.5.6.3. Class 1 genome editing 62
1.5.7. Transcription regulation with CRISPR systems 63
1.5.8. Current progress in prokaryotic genome editing 64
1.5.8.1. Teething problems 64
1.5.8.2. Successful genome editing in prokaryotes 66
1.5.9. Considerations for CRISPR-mediated genome editing in acetogens 67
1.6. Project aims 69

2. Methods and materials 71
2.1. Strains and vectors 71
2.2. Primers and oligonucleotides 79
2.3. Chemicals and suppliers 92
2.4. General bio-informatics tools 93
2.4.1. DNA sequence analysis 93
2.4.2. Protein structure analysis 93
2.4.3. Promoter and terminator sequence analysis 93
2.4.1. Restriction-modification system analysis 94
2.4.2. Plasmid design and visualisation 94
2.4.3. Oligonucleotide design 94
2.4.4. Data analysis and presentation 94
2.5. General microbiology techniques 95
2.5.1. Growth conditions and media 95
2.5.1.1. Growth conditions 95
2.5.1.2. Strain storage and revival 95
2.5.1.1. Media composition 96
2.5.2. Preparation of chemically competent *E. coli* cells 100
2.5.3. Transformation of *E. coli* via heat shock 100
2.5.4. Preparation of chemically competent *C. ljungdahlii* 101

2.5.5. Transformation of *C. ljungdahlii* via electroporation 101
2.5.6. Preparation of chemically competent *E. limosum* 102
2.5.7. Phase contrast microscopy 102
2.6. General molecular biology techniques 103
2.6.1. Preparation of genomic DNA from Clostridia 103
2.6.2. Preparation of plasmid DNA from *E. coli* 103
2.6.3. Restriction enzyme digests 103
2.6.4. De-phosphorylation of linear DNA 103
2.6.5. Plasmid DNA ligation 104
2.6.6. Homology-based plasmid assembly 104
2.6.7. Agarose gel electrophoresis 104
2.6.8. DNA extraction from agarose gel 105
2.6.9. Polymerase chain reaction 105
2.6.1. SOEing PCR 106
2.6.2. DNA fragment sequencing 108
2.7. Gram-positive replicon modularisation 109
2.8. Genome annotation 109
2.9. Generation of ClosTron mutagenesis plasmids 111
2.9.1. Target selection 111
2.9.2. Plasmid synthesis 111
2.9.3. Mutant screening 111
2.10. Allele Coupled Exchange and Allelic In-Frame Deletion 112
2.10.1. Development of *pyrE* deficient strains 112
2.10.2. Allelic in-frame deletion 113
2.11. Generation of CRISPR-Cas9 mutants 114
2.11.1. Target selection 114
2.11.2. Generation of new sgRNA seed regions 115
2.11.3. Generation of homology arm cassettes 115
2.11.4. Bookmark design 115
2.12. Spacer identification and PAM analysis 115
2.1. CAT assay for *C. autoethanogenum* expression analysis 116

3. ..... Strain and DNA transfer optimisation 119
3.1. Introduction 119
3.1.1. Genome sequence data 119
3.1.2. DNA transfer and plasmid characterisation 121
3.1.3. *pyrE* deficient strain development 122
3.1.4. Precise gene expression 123
3.2. Manual annotation of *C. autoethanogenum* 124
3.2.1. CLAU_2316-2646 124
3.2.2. Overall genome analysis 127
3.3. Restriction-modification based approach to DNA transfer optimisation 130
3.3.1. *C. autoethanogenum* RM system analysis 131
3.3.2. Type IV restriction analysis 133
3.3.3. Type II RM system analysis 135
3.3.3.1. *In silico* analysis of Type II RM systems 135
3.3.3.1.1. Cau100611I 135
3.3.3.1.2. Cau10061ORFGP (*cauIII*) 136
3.3.3.1.3. Cau10061ORFFP (*cauIV*) 137

3.3.3.2. Type II RM knockout generation and strain analysis 138
3.4. *pyrE* deficient strain generation 141
3.5. Further pCB101and pCACET replicon analysis 144
3.6. Electroporation 144
3.7. RBS and promoter comparison 145
3.7.1. RBS library generation and analysis 145
3.8. Conclusion 149

4. Development of exogenous CRISPR systems for genome editing in clostridia 158
4.1. Introduction 158
4.2. Early iterations 159
4.3. Proof of concept 160
4.3.1. Tool development in *C. ljungdahlii* 161
4.4. Plasmid modularisation and construction 162
4.4.1. Optimisation of vector construction methods 165
4.5. *pyrE* knockout 169
4.7. pMTL40000 in *E. limosum* 170
4.8. pMTL40000 nomenclature generation and use in other organisms 172
4.8.1. Expanding the range of nucleases available for pMTL40000 173
4.8.1.1. StCas9 sgRNA 175
4.9. pMTL40000 nuclease analysis: trCas9 178
4.9.1. *In silico* analysis of trCas9 178
4.9.2. trCas9 nuclease comparison 180
4.10. Bookmark complementation 182
4.10.1. Bookmark 1 185
4.11. Bookmark 2 and 3 187
4.12. Bookmark 4-12 187
4.13. Development of CRISPR based SNP/SND generation 189
4.13.1. SNP/SND locus analysis 191
4.13.1.1. SNPs 191
4.13.1.2. SNDs 192
4.13.2. SNP/SND vector design 193
4.13.3. SNP/SND generation 195
4.14. Conclusion 196

5. Development of endogenous CRISPR systems for genome editing and transcriptional control 203
5.1. Introduction 203
5.2. *In silico* CRISPR analysis 205
5.2.1. *C. autoethanogenum* CRISPR system 205
5.2.2. *C. autoethanogenum* CRISPR loci 206
5.2.3. PAM identification 210
5.2.4. Dispute with Pyne and colleagues 211
5.2.5. Spacer acquisition assay 212
5.2.6. Protospacer 1.27a analysis 212
5.3. Top-down *C. autoethanogenum* PAM elucidation 214
5.4. Bottom-up PAM elucidation 217
5.4.1. Preliminary proof of concept 219
5.4.1.1. Proof of concept conclusions 221

5.5. *C. autoethanogenum* engineered Cas operon (eCas) strain for transcriptomic control 223
5.5.1. Deletion of *cas3* 223
5.5.1. eCas plasmid proof of principle 225
5.6. Analysis of the *E. limosum* endogenous CRISPR system 229
5.7. Analysis of other acetogenic and clostridial CRISPR systems 233
5.7.1. CRISPR systems in acetogens 235
5.7.2. CRISPR systems in other industrially relevant *Clostridium* 236
5.7.3. CRISPR systems in clinically relevant *Clostridium* 238
5.7.4. Protospacer-like sequences identified from clostridial CRISPR spacers sequences 239
5.8. Conclusion 240
5.8.1. Analysis of the *C. autoethanogenum* Type IB CRISPR system 241
5.8.1. Analysis of the Type IB *E. limosum* CRISPR system 243
5.8.2. Development of a robotics-based assay for PAM identification 244
5.8.3. Analysis of other Clostridial CRISPR systems 245

6. Analysis of acetate pathway mutants 248
6.1. Introduction 248
6.1.1. ATP generation in acetogens 248
6.1.2. Phosphotransacetylase and acetate kinase 250
6.2. Strain development 252
6.2.1. ACE and ClosTron mutagenesis 252
6.2.2. pMTL40000-mediated Δ*ack* and Δ*pta* strain generation 254
6.2.2.1. *ack* deletion 256
6.2.2.2. *pta* deletion 257
6.2.2.3. Bookmark mediated repair of *ack* and *pta* mutants 257
6.3. Strain comparison and solvent profile analysis 258
6.3.1. Gas adaptation 258
6.3.1. CO growth experiment 259
6.4. Conclusion 261

7. Conclusion and summary 266
7.1. Addressing the Objectives of the Study 266
7.2. Future work 272

8. Appendix 274
A) Phylogenetic tree construction 274
B) *C. autoethanogenum* promoter analysis 274

9. Bibliography 274

“It was written I should be loyal to the nightmare of my choice”

Joseph J Conrad, Heart of Darkness

CHAPTER ONE

INTRODUCTION

1. Introduction

Due to their ability utilise C1 (single carbon) gases as sole carbon and energy sources, acetogens have garnered significant interest as sustainable production chassis for industrially relevant chemicals. Via the Wood-Ljungdahl pathway, these prokaryotic anaerobes are capable of chemoautotrophic growth through reduction of carbon monoxide (CO), or carbon dioxide (CO_2) in the presence of hydrogen (H_2). This pathway is a conserved feature of all acetogens, and directly feeds into the group's second (and etymologically derived) defining feature, acetate production (Ragsdale & Pierce, 2008; Wood, 1991).

As well as acetate, acetogens are able to use acetyl CoA generated from the Wood-Ljungdahl pathway as a precursor for production of several industrially relevant chemicals, including 2,3-butanediol, butyrate, ethanol, and butanol (Kopke *et al.*, 2010; Cotter, Chinn & Grunden, 2009; Shen, Brown & Wen, 2014). Many of these products are traditionally petrochemicals, or fulfil similar functions to crude oil derivatives, making these organisms ideal for biofuel production.

Due to socio-economic factors, such as oil price volatility, concerns over carbon emissions, and renewable energy, alternatives to the conventional production methods of petrochemicals are readily sought. Acetogens meet these concerns by providing such products from a broad range of renewable C1 sources. Industrial output of C1 gases has already provided a feedstock for commercialisation of acetogenic biofuel production, and modern

gasification technology enables conversion of practically any common waste-stream into C1 gas.

Renewable chemical production, feedstock plasticity and proven industrial application has provided a foundation for biotechnological use of acetogens. Growing commercialisation and continued academic interest has led to improved understanding and improving of their capability for sustainable chemical generation (Liew *et al.*, 2016; Brown *et al.*, 2014; Abrini, Naveau & Nyns, 1994). However, despite current progress, significant improvement in strain performance is required to maximise production potential.

1.1. Renewable energy: demand, production and limitations of biofuels

Interest in renewable alternatives to the conventional energy sources of oil, natural gas and coal (often collectively termed fossil fuels) has grown significantly in recent decades. This is attributed to two main driving factors; the finite nature of fossil fuels, and the general acceptance of their role in climate change (Bentley, 2016). Tangible effects of environmental issues, combined with oil price volatility has led to continued and significant interest in production of renewable energy, particularly biofuels (McCrone *et al.*, 2016). Defined as fuels derived from living matter, these fuels currently complement crude oil usage, and growth of their production has been highlighted as key to the eventual cessation of fossil fuel use (Aro, 2016).

1.1.1. Finite fossil fuels

The limited nature of oil reserves is not a new cause for concern however, oil production is now regarded as being close to its resource-limited maximum. The significance of this state-of-play, often referred to as “peak oil”, is reflected within the oil price volatility seen in recent years (see Figure 1.3), and the predicted terminal decline in crude oil supply in years to come (Bentley, 2016). Taking the United Kingdom (UK) as an example, as illustrated in Figure 1.1, the country’s ability to meet its own demand for oil has failed since 2005. Since that point, the gap between supply and demand has increased, a trend that is predicted to continue in coming decades (Earp, 2015).

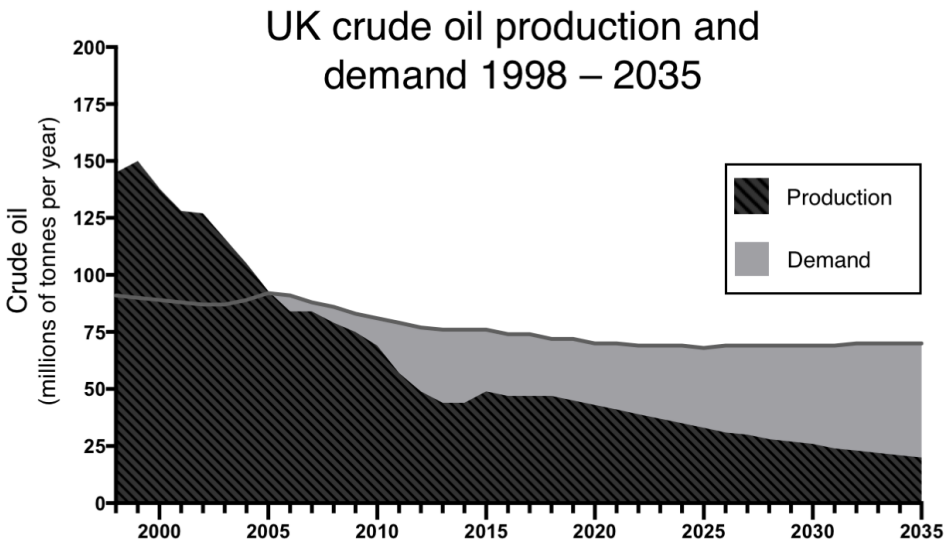


Figure 1.1: UK crude oil production and demand 1998 – 2035
Oil production and demand of the United Kingdom between 1998 and 2035. 1998 to 2015 represents actual production/demand on published by UK Department of Energy & Climate Change (DECC), 2015 onwards represents DECC projected estimates. Based on UKCS Oil and Gas Production Projections Report (Earp, 2015).

Oil plays a crucial role in the global economy; it is the world's largest traded energy resource, and underpins global transport infrastructure, enabling cheap movement of goods and people. Partial substitution of crude oil derived fuels for renewable alternatives is already a reality, and several key factors are driving this trend, as explained in the following section (Miller *et al.*, 2013).

1.1.2. Energy security and self-sufficiency

The gap in supply and demand for crude oil illustrated in Figure 1.1 is not limited to the UK. Many countries lacking the necessary natural resources rely upon trade with better endowed states, as Figure 1.2 illustrates.

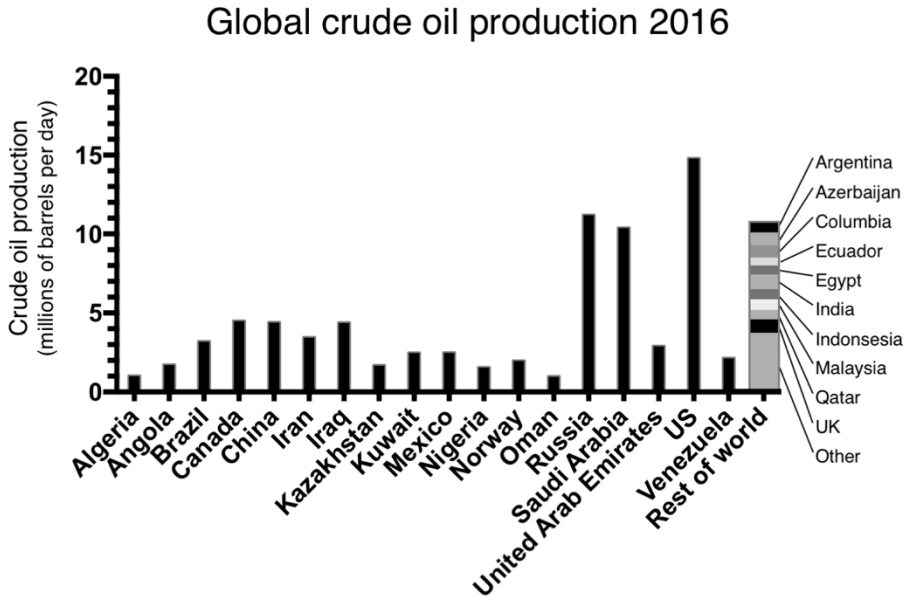


Figure 1.2: Global crude oil production 2016
Crude oil production (millions of barrels per day) of the top 28 crude oil producers in 2016. Main bar graph shows production of; Algeria, Angola, Brazil, Canada, China, Iran, Iraq, Kuwait, Mexico, Nigeria, Norway, Oman, Russia, Saudi Arabia, United Arab Emirates, US (United States), Venezuela, and "rest of world". Rest of world bar is presented in stacked form, comprising; Argentina, Azerbaijan, Columbia, Ecuador, Egypt, India, Indonesia, Malaysia, Qatar, and the UK (United Kingdom). Based on data from US Energy Information Administration (EIA) (EIA, 2017a).

Figure 1.2 shows the top 28 crude oil producers and their respective oil production in 2016. Notably, many of the top oil producers are also infamously politically unstable; even without mentioning internal political issues, recent history has seen numerous conflicts between list members. The 1980's saw the Iran-Iraq war spanning much of the decade, with the Falklands war between the UK and Argentina occurring in 1982. The first Gulf War in the early 1990's was caused by Iraqi invasion and annexation of Kuwait, with Saudi Arabia, the US and the UK intervening on behalf of Kuwait. A fate repeated when a US and UK led coalition invaded Iraq in the 2000's.

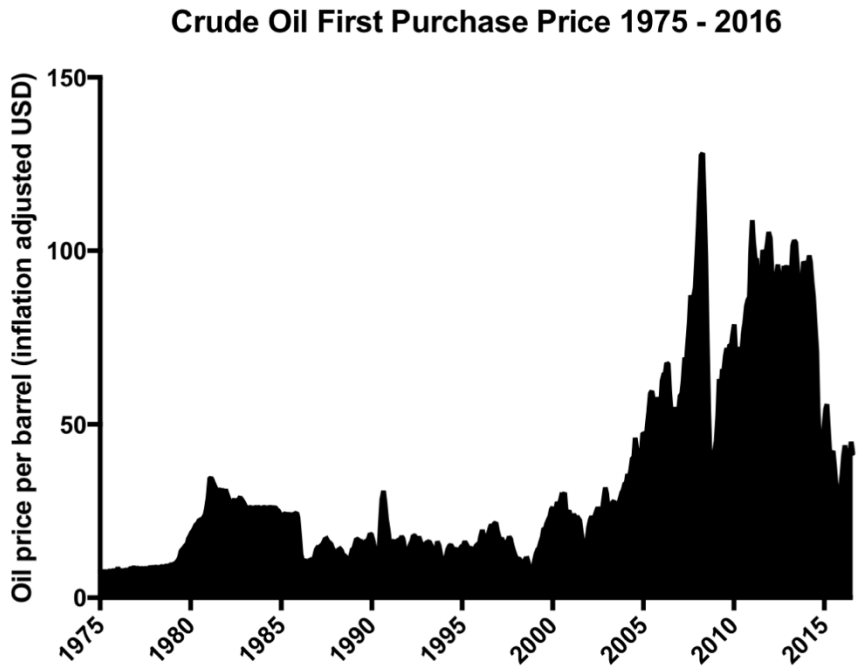


Figure 1.3: Crude oil first purchase price 1975 – 2016
Inflation adjusted first purchase price (USD per barrel) between January 1975 and December 2016. Based on data from US Energy Information Administration (EIA) (EIA, 2017b).

The political instability of oil producing states has been reflected in oil price volatility. Figure 1.3 shows oil price per barrel between 1975 and 2016, with an obvious example of such correlation being the price spike in the early 1990's attributed to political instability in the Middle East and the onset of the Gulf War (Guo & Kliesen, 2005).

From a geopolitical perspective, renewables offer states deficient in oil, coal and natural gas reserves freedom from reliance on foreign markets, buffering them from price volatility. From a transport fuel perspective, their use also requires no such infrastructure overhaul that would be necessary for adoption of electric vehicles. Although there are significant arguments for electric cars, as seen in the rise of premium electric car manufacturer Tesla, widespread adoption of this technology would require huge investment in charging

stations. On a secondary note, current usage of electric cars offers no guarantee of greater renewability, with a significant amount of global electricity being produced by coal-based power plants.

1.1.3. Carbon concerns

Fossil fuels are regarded as drivers of climate change due to their combustion releasing CO_2 into the atmosphere. This release of CO_2 contributes to the "greenhouse gas effect", whereby an increase in atmospheric radiative gas traps radiation in the earth's atmosphere rather than dissipating into space, leading to climate increase. Although combustion of any hydrocarbon fuel, regardless of origin, releases CO_2, fossil fuels are under scrutiny due to their release of carbon that would have otherwise remained in oil fields, coal deposits, or natural gas reservoirs. Conversely, biofuels are regarded as carbon neutral due to being generated from atmospheric CO_2 as a primary carbon source, meaning their generation and subsequent use sees no net increase in greenhouse gas emission (Schiel-Bengelsdorf & Dürre, 2012).

In addition to concerns over CO_2 emissions, fossil fuel combustion has also led to concerns over air pollution. These concerns relate to the release of substances harmful to human health and general biodiversity, and include particulate matter, carbon monoxide and lead. With severe levels of air pollution due to heavy industry and high population, pollution-related health concerns have been growing in China in recent years. This has been reflected in the country's recent legislature, resulting in progressive and far reaching policies including reduction of fossil fuel output (Feng & Liao, 2016).

Beyond China, global interest in biofuel development has driven various initiatives across the socio-political landscape, from crowd sourced local renewable energy projects, to major international summits, such as the 2015 World Climate Agreement, resulting in pledges to reduce carbon output from world leaders. Examples of such government-led action include the British Government's Green Investment Bank (GIB), set up in 2012 with £3.8 billion to fund green energy infrastructure (McCrone *et al.*, 2016). Complementing such grass roots and governmental initiatives, academic and industrial activity has accelerated production capabilities in terms of scalability, available yields and breadth of products available.

With a duel emphasis on the production of renewable energy, and the reduction in CO_2 output, acetogens have been highlighted as at least part of the solution to this global problem. As has been shown on an industrial scale, *Clostridium autoethanogenum* can utilise steel mill off-gas as a carbon and energy source to produce a range of chemicals, including biofuels (Köpke *et al.*, 2011; Liew *et al.*, 2016). This model shows that a waste gas, otherwise rich in the greenhouse gas CO_2, and pollutants including CO and particulate matter, can be utilised for sustainable energy production.

1.1.4. Biofuels

Concerns over the finite nature of fossil fuels, their environmental impact, and their price volatility led to the development and adoption of first generation (1G) biofuels as early as the 1970s. As described in Table 1.1, fuel, namely ethanol and biodiesel, is produced via fermentation or transesterification of traditional food crops. Although sustainable, renewable and carbon neutral, 1G fuels have significant drawbacks, chiefly the fact that using food crops as feedstocks leads to food shortages and drives up food costs. Second generation (2G) biofuels were developed in response to such criticism.

Still derived from plant matter, 2G biofuels utilise non-edible lignocellulosic biomass. Although they offer no direct competition with food, their production does provide significant cause for concern, both in terms of process engineering and ecological and economic impact (Aro, 2016). Efficient utilisation of lignocellulosic biomass has been the subject of much research. Despite some advances, notable drawbacks remain, including fermentation requiring energy intensive pre-treatment of feedstocks, and said feedstocks still impacting on food production (Kovács *et al.*, 2013). Although often derived from agricultural waste such as corn stover, crops such as jatropha are often specifically cultivated for 2G biofuel generation due to their fast growth and other desirable traits. Whilst these crops are inedible and therefore offer no direct competition with food crops, they offer significant indirect competition with food production due to overlap in prerequisite resources such as arable land and water (Fraiture, Giordano & Liao, 2008).

Table 1.1: Comparison of first, second, third and fourth generation biofuels
Summary of first (1G), second (2G), third (3G), and (4G) fourth generation biofuels, describing key features including feedstock and production requirements. Based on description from Ref. (Aro, 2016).

Biofuel	Key features
1G	• Made from conversion of food crops to ethanol or biodiesel via fermentation of transesterification • Requires the use of food crops, arable land and fresh water
2G	• Made from fermentation of non-edible crops, or non-edible components of food crops • Requires the use of arable land and fresh water, and varying degrees of general agricultural infrastructure • Generally, consists of bioethanol production of lignocellulosic biomass
3G	• Biofuels produced via photosynthetic algae without the use of arable land • Refer to specific production of biodiesel, bioethanol, biogas and other algal-based products
4G	• Biofuels produced without the use of arable land and with no destruction of biomass required • Require cheap, readily available, renewable resources

Addressing many of the drawbacks of 1G and 2G biofuels, 3G biofuels are produced via photosynthetic algae. Unlike 1G and 2G fuels, production requires no arable land and has negligible impact on crop production. Despite their ability to synthesise a broad range of products (Deng & Coleman, 1999; Guerrero *et al.*, 2012), engineering limitations associated with industrial scale reactor design and the ability to supply dense cultures with sufficient light energy are cited as bottlenecks to large scale production (Ugwu, Aoyagi & Uchiyama, 2008; Ugoala *et al.*, 2012). Like their 3G counterparts, 4G biofuels also require no arable land, and can utilise cheap, readily available resources. Acetogenic fermentation of C1 gases fall comfortably into this bracket.

1.2. Acetogenic bacteria for biofuel generation

A major benefit of acetogenic fermentation is plasticity of substrate utilisation. Like many anaerobes, acetogens are capable of heterotrophic growth on a variety of sugars. But uniquely, acetogens are also capable of using C1 gases to fulfil their growth requirements via the Wood-Ljungdahl pathway. This broadens the potential range of substrates, not only in terms metabolic diversity, but crucially in terms of feedstock origin.

Already described is the use of steel mill off-gas, an exemplification of a waste gas meeting the requirements to sustain acetogenic bacterial growth. In addition to waste gases, “syngas” (synthesis gas) is a widely available C1 gas consisting of H_2, CO and sometimes CO_2. Syngas can be generated through gasification of a range of waste materials, including biomass, waste plastics, municipal solid waste, and sewage sludge (Kim *et al.*, 2014; Shen, Brown & Wen, 2014; Köpke *et al.*, 2011).

Although an in-depth economic analysis of biofuel production is beyond the remit of this introductory section, the previously described economic, political and environmental issues serve as strong push factors for the development of acetogenic fermentation. Arguments against such development on an economic basis may come from critics highlighting the trend in oil prices dropping in recent years (see Figure 1.3). However, if not viewed as an alternative mode of energy production, acetogenic fermentation can at least be regarded as a powerful carbon mitigation technique. The following section looks to offer an overview of acetogenic bacteria, and describe their common biochemical features that lends them to C1 gas fermentation.

1.2.1. Acetogen overview

The acetogens represent a physiologically defined, yet phylogenetically diverse group of bacteria, found across over 20 genera of eubacteria (Schiel-Bengelsdorf & Dürre, 2012; Drake, Gößner & Daniel, 2008). Members include the thermophilic *Moorella thermoacetica* (previously classified as *Clostridium thermoaceticum*), the α-proteobacteria *Desulfotignum phosphitoxidans,* and the butanol producing *Butyribacterium methylotrophicum*. However, the most well characterised members belong to just two genera; *Eubacterium* and *Clostridium*, including *Eubacterium limosum,* and the closely related *C. autoethanogenum* and *Clostridium ljungdahlii*. Figure 1.4 shows a phylogenetic tree of known acetogens.

The diversity of acetogens is not limited to their phyla, in terms of ecology this group of bacteria is also remarkably diverse. Acetogenic bacteria have been isolated from environments ranging from rice paddy soils (*Sporomusa desulfovibrio*, Rosencrantz et al. 1999), to Mediterranean marine salterns (*Acetohalobium arabaticum*, Ollivier et al. 1994), even to the gastrointestinal tracts of animals, as was the case for *C. ljungdahlii* and *C. autoethanogenum*[1] for example (Breznak, 1994; Mackie & Bryant, 1994; Abrini, Naveau & Nyns, 1994).

Although the name “acetogen” suggests the defining characteristic of this group to be the generation of acetate, the Wood-Ljungdahl pathway is generally regarded as their pivotal feature. This pathway enables metabolism

[1] *C. autoethanogenum* is presumed to be a rabbit gut commensal, however it is worth noting that it was isolated from rabbit faeces rather than the gut itself (Abrini, Naveau & Nyns, 1994)

of CO and CO_2 + H_2 to generate Acetyl- Coenzyme A (Acetyl-CoA), a crucial central metabolite that acts as a precursor to the solvent products associated with industrially relevant acetogens.

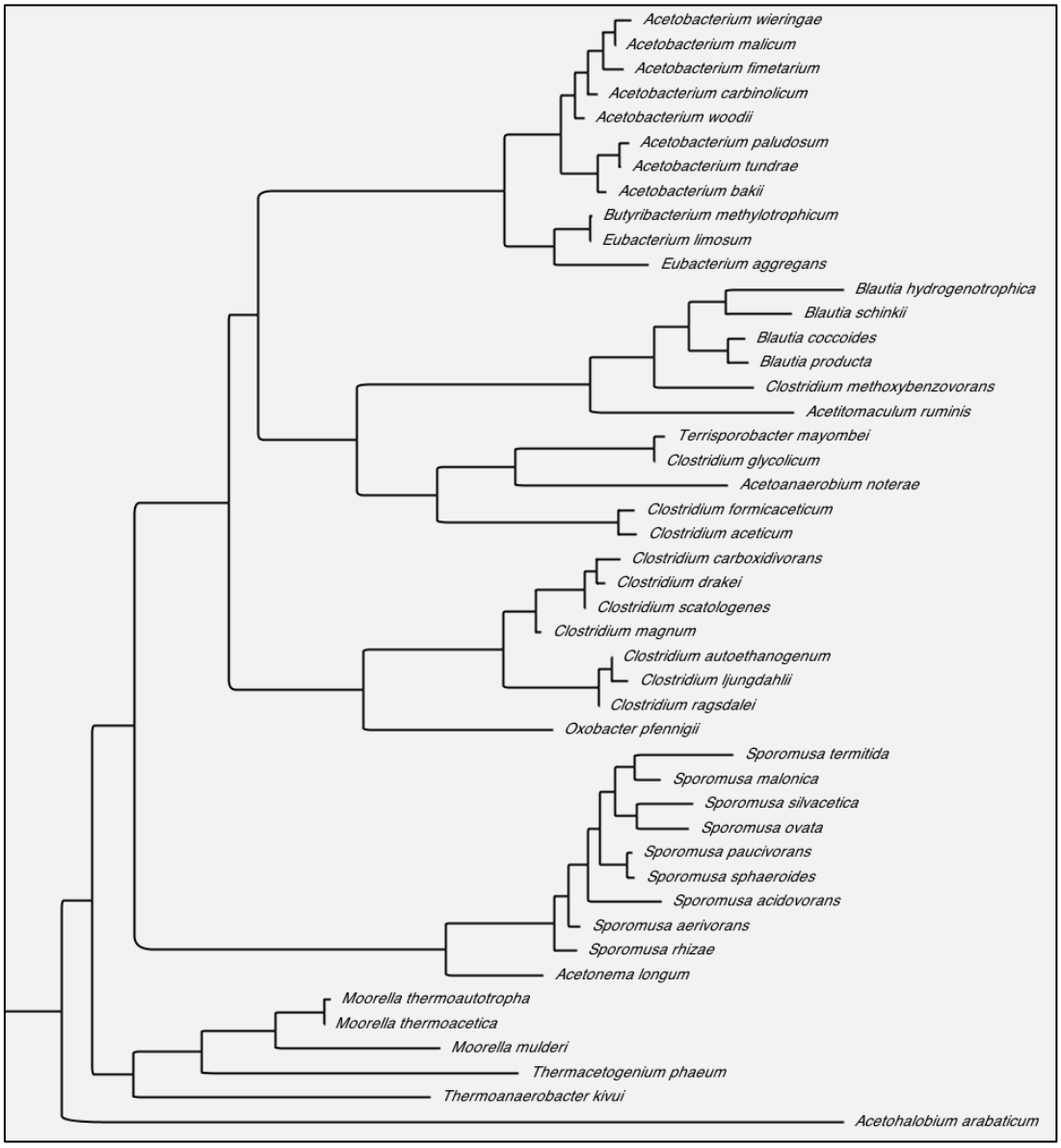


Figure 1.4: Phylogenetic tree of a range of acetogens
Phylogenetic tree based on list of known acetogenic bacteria from Ref. (Schiel-Bengelsdorf & Dürre, 2012), constructed for illustrative purposes as described in Appendix A.

1.2.2. *Clostridium autoethanogenum*

C. autoethanogenum is a Gram-positive, rod-shaped, solventogenic obligate anaerobe, and was the main focus of this study. Unlike many other members of the clostridial genus, the chemoautotrophic potential of the *C. autoethanogenum* metabolism means it requires no sugar-based feedstock to act as an energy or carbon source. Instead, as an acetogen, it is able to solely utilise CO and/or CO_2 + H_2 to produce acetate, 2,3-butanediol and ethanol. This makes *C. autoethanogenum* an ideal candidate for sustainable chemical production, a feature that has not gone un-noticed industrially or academically.

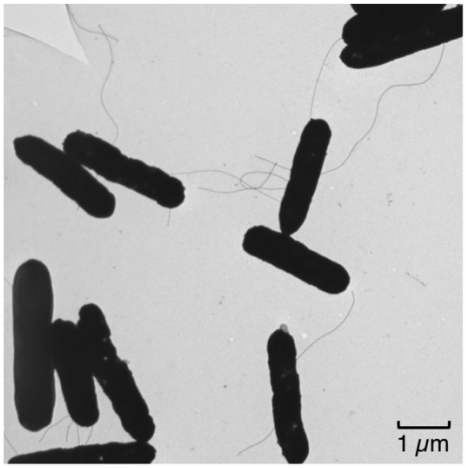


Figure 1.5: Transmission electron microscopy of *C. autoethanogenum*
From unpublished work by Ronja Breitkopf (Breitkopf, Pander & Piatek, 2014).

Although first described by Abrini and coworkers in 1994 (Abrini *et al.,* 1994), this organism has seen little discussion in academic literature until the last decade. Possible explanations for an organism with so much industrial potential to have such little published could be attributed to its highly oxygen sensitive nature, slow growth rate, lack of genomic sequence data, and poorly

established protocols for DNA transfer and genetic engineering. However, due to ongoing efforts, these issues are being addressed.

1.2.3. Substrate utilisation in acetogens

Prior to looking in more depth at how the Wood-Ljungdahl pathway enables chemoautotrophy in acetogens, it is important to take a more holistic view at their carbon fixation and energy generation mechanisms. As well as illustrating the heterotrophic nature of acetogens, this section should also provide context for understanding what makes acetogens unique.

1.2.3.1. Organic substrates

As well as auxotrophic growth on C1 gas, acetogens can generate energy and metabolise carbon through oxidation of a range of organic substrates. These include hexose and pentose sugars, alcohols, methyl groups, and formic acid (Schuchmann & Müller, 2014).

Fermentation of hexoses via glycolysis and pentose via the pentose-phosphate pathway results in the generation of pyruvate. This important metabolic intermediate is either directly used in pathways such as 2,3-butanediol production in *C. autoethanogenum* (Köpke *et al.*, 2011), or can be oxidised to acetyl-CoA via pyruvate dehydrogenase.

1.2.4. The Wood-Ljungdahl pathway

The Wood-Ljungdahl pathway, as illustrated in Figure 1.6, consists of two branches; the methyl (or Eastern) branch and the carbonyl (or Western) branch. Both combined result in the production of acetyl-CoA which serves as a precursor for downstream production of acetate, ethanol and 2,3-butanediol. The methyl branch involves the conversion of CO_2 to methyl-CoFeSP, via several intermediates, which combines with CO from the carbonyl branch to form acetyl-CoA. Through the action of phosphotransacetylase (Pta) and acetate kinase (Ack), the definitive acetogenic product, acetate, is created.

Although acetate is generated in the reductive direction, the bidirectional nature of Pta and Ack enable, when thermodynamically favourable, re-assimilation of acetate into the central metabolism. Examples of such can be seen in methanogens and sulfate reducing bacteria which couple the Wood-Ljungdahl pathway to their respective eponymous exergonic processes, methanogenesis and reduction of sulfate to sulfide. This however is not the case with the clostridial acetogens discussed herein, whereby acetate is one of the main natural products generated through C1 gas fermentation. Further product profiles are discussed in section 1.2.5.

1.2.4.1. The methyl branch

Figure 1.6 illustrates the first step of the methyl branch as being the conversion of carbon monoxide to carbon dioxide, through the activity of a carbon monoxide dehydrogenase. However, if carbon dioxide is present this can be feed directly into the methyl branch, acting as a substrate for formate dehydrogenase mediated generation of formate. From here, the methyl

branch requires a series of tetrahydrofolate (THF)- and cobalamin- dependent reactions that culminate in the generation of methyl-CoFeSP.

First discovered in 1942 in the seminal acetogen, *M. thermoacetica*, carbon monoxide dehydrogenase (CODH) is a nickel dependant enzyme that catalyses the oxidation of CO to CO_2, transferring electrons to a variety of acceptors, most notably ferredoxin (Fontaine *et al.*, 1942). Discovery of CODH and subsequent analysis showed CO could act as a substitute for pyruvate in acetyl-CoA synthesis (Ragsdale & Pierce, 2008), providing the foundation for a credible alternative to glycolysis for carbon fixation. Full understanding of this pathway, including identification of intermediates and their required enzymes, would not be attained until 1985. Central to this complete understanding was the characterisation of tetrahydrofolate (THF)- and cobalamin- dependent enzymes, as discussed below.

1.2.4.1.1. Formate synthesis

Formate acts as a substrate to the first THF dependent reaction in the Wood-Ljungdahl pathway, and it is itself generated through the action of formate dehydrogenase. As mentioned, CODH activity catalyses the conversion of CO to CO_2, however the presence of CO_2 negates the requirement for this reaction, as it can feed directly into the methyl branch. Regardless of origin, CO_2 is reduced to formate, catalysed by the tungsten- and selenocysteine-containing formate dehydrogenase. Due to thermodynamic unfavourability, electrons are provided by NADPH (Schuchmann & Müller, 2014).

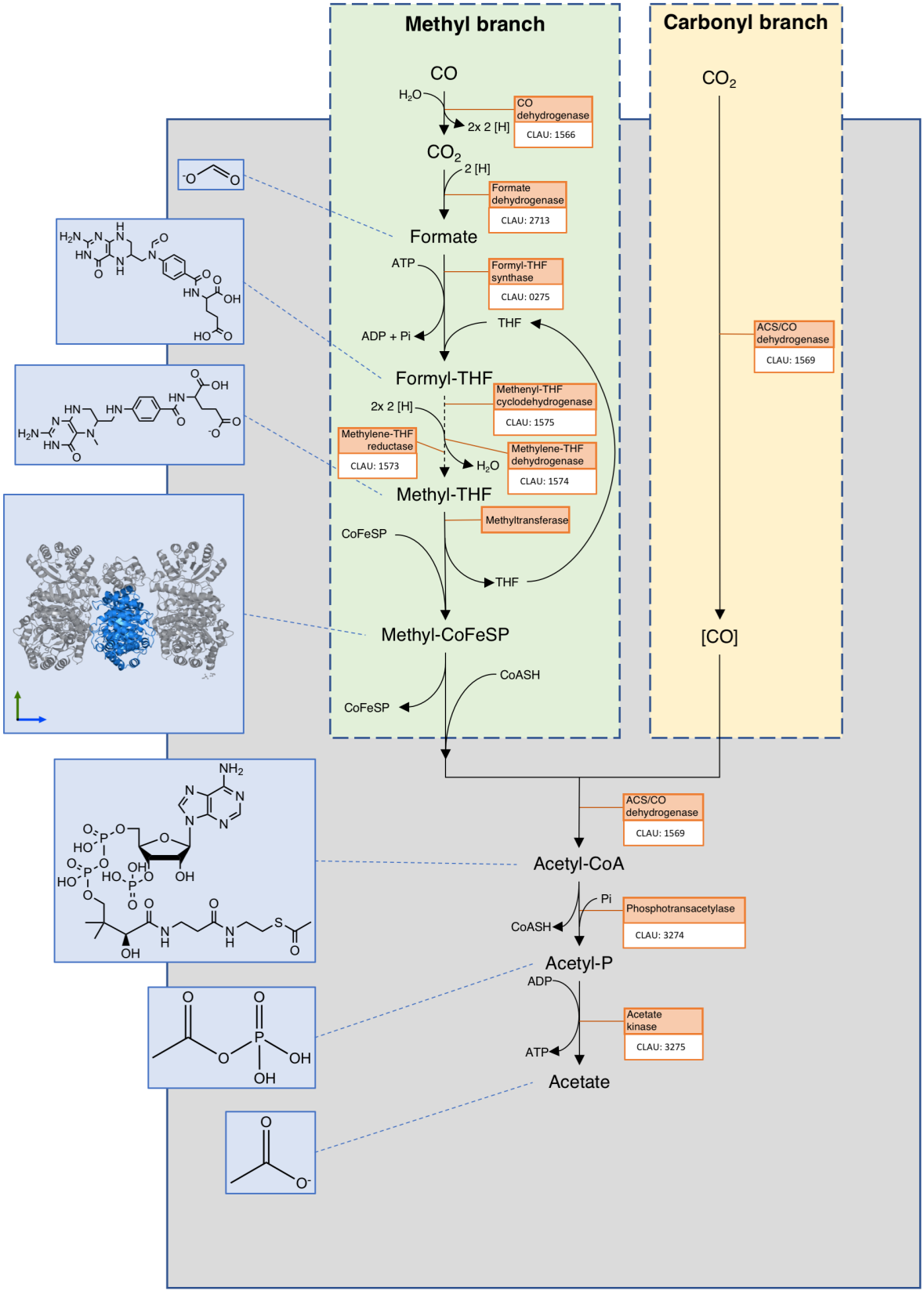

Methyl branch
Carbonyl branch
CO
H2O
2x 2 [H]
CO dehydrogenase
CLAU: 1566
CO2
2 [H]
Formate dehydrogenase
CLAU: 2713
Formate
ATP
Formyl-THF synthase
CLAU: 0275
ADP + Pi
THF
Formyl-THF
2x 2 [H]
Methenyl-THF cyclodehydrogenase
CLAU: 1575
Methylene-THF reductase
CLAU: 1573
Methylene-THF dehydrogenase
CLAU: 1574
H2O
Methyl-THF
Methyltransferase
CoFeSP
THF
Methyl-CoFeSP
CoASH
CoFeSP
CO2
ACS/CO dehydrogenase
CLAU: 1569
[CO]
ACS/CO dehydrogenase
CLAU: 1569
Acetyl-CoA
Pi
CoASH
Phosphotransacetylase
CLAU: 3274
Acetyl-P
ADP
Acetate kinase
CLAU: 3275
ATP
Acetate

Figure 1.6: Wood-Ljungdahl and acetogenesis pathways in acetogens
Reduction of CO to methyl-CoFeSP is shown in the methyl branch of the pathway in a light green box. Reduction of CO_2 to CO displayed in the carbonyl branch of the pathway shown in a light yellow box. Below each branch of the pathway acetogenesis is represented. In methyl branch, THF is tetrahydrofolate. CoASH is Coenzyme-A. Chemical structure of key intermediates displayed in light blue boxes on the left-hand side (with exception of Methyl-CoFeSP (corrinoid iron-sulfur protein) displayed as predicted protein structure) (Goetzl *et al.*, 2011). Enzymes involved in each step are shown in orange boxes, with putative locus tags for their encoding genes shown below in white boxes, based on study by (Schuchmann & Müller, 2014).

1.2.4.1.2. Formyl- and methyl- THF synthesis

During the only ATP dependant step of the Wood-Ljungdahl pathway, formate undergoes condensation with TNF to generate formyl-TNF through the action of formate-THF synthetase. Formyl-THF is in-turn dehydrated and reduced to form methyl-TNF through the activity of several enzymes. As well as being essential for Wood-Ljungdahl mediated production of acetyl-CoA, these two intermediaries are utilised for amino acid and nucleotide synthesis. Generation of methyl-TNF requires the activity of three enzymes; formyl-THF cyclohydrolase, methylene-THF dehydrogenase, and methylene-THF reductase. The cyclohydrolase enables reduction of formyl-TNF to methenyl-THF, which is subsequently reduced via the action of the dehydrogenase to form methylene-THF. The final step towards methyl-TNF is catalysed by methylene-THF reductase, by reducing methylene-THF.

A methyltransferase catalyses the final step of the methyl branch, transferring the methyl group from methyl-THF to corrinoid iron-sulphur protein (CoFeSP), forming methyl-CoFeSP (Schuchmann & Müller, 2014).

1.2.4.1. The carbonyl branch and acetyl-CoA generation

The carbonyl branch of the Wood-Ljungdahl pathway provides CO either directly, or through reduction of CO_2. For the latter, carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS) is required. This bifunctional enzyme then fuses CO with the methyl group from methyl-CoFeSP and CoA to form acetyl-CoA.

1.2.5. Industrially relevant products from acetogens

As

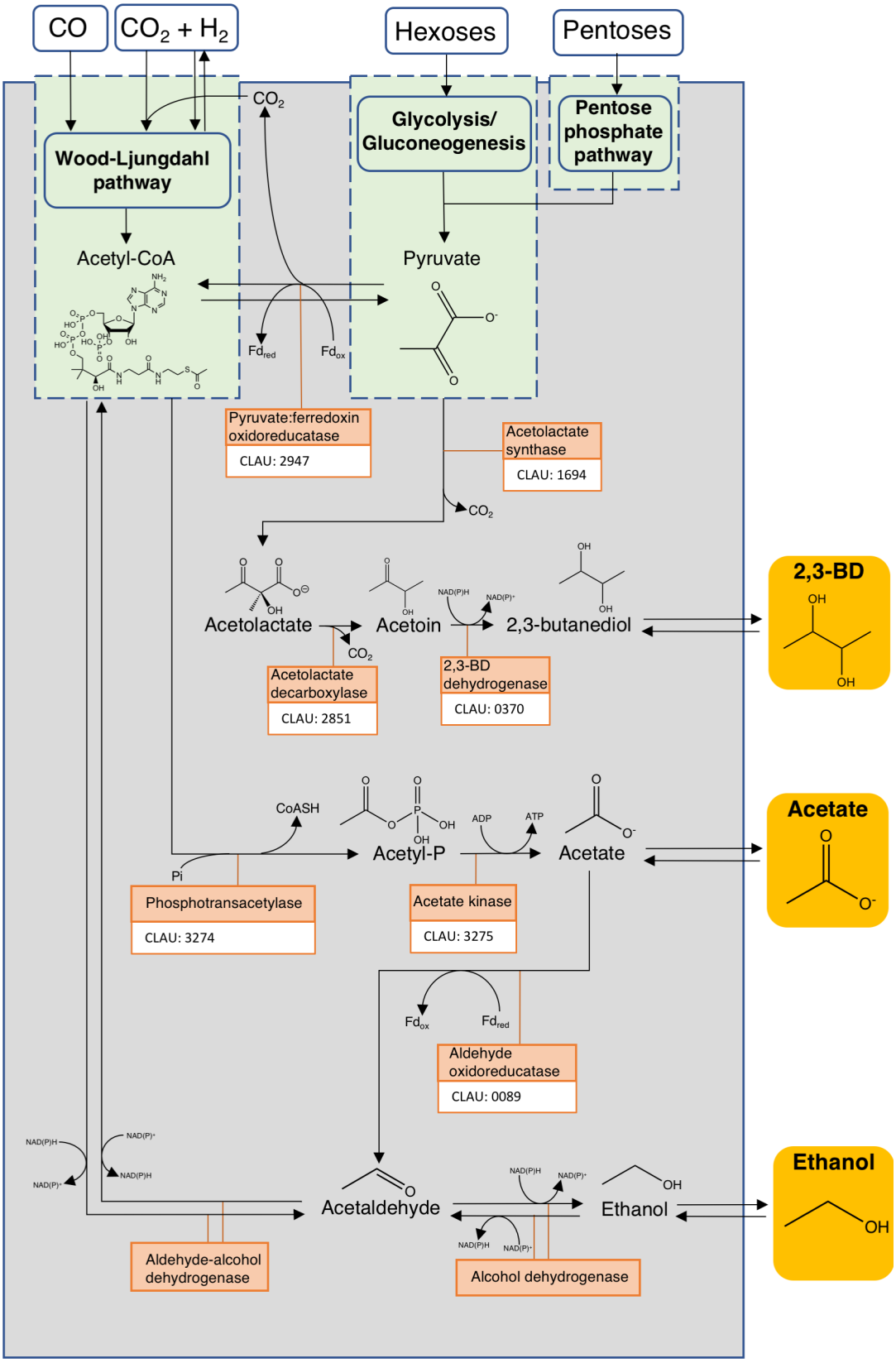


Figure 1.7: illustrates, production of acetate, ethanol and 2,3-butanediol is achieved using acetyl-CoA as a precursor. The heterotrophic nature of acetogenic metabolism is discussed earlier in this section, and results in acetyl-CoA generation either directly from the Wood-Ljungdahl pathway, or from sugar metabolism via pyruvate.

1.2.5.1. Acetate synthesis

Conventionally produced as a petrochemical, acetate is the precursor to many chemicals, including polyvinyl acetates. As the epitomic acetogenic product, unsurprisingly industrial acetogens have been the subject of research into developing more sustainable acetate production.

1.2.5.2. Ethanol synthesis

Ubiquitous with fermentation and almost synonymous with biofuel, ethanol presents the epitome of C1 gas fermentation. As designated by its etymology, *C. autoethanogenum* is capable of the autotrophic generation of ethanol, a trait not unnoticed, or unique among acetogens, as commercial C1 gas fermentation for ethanol production is currently underway using a range of acetogens.

Again, using acetyl-CoA as a precursor, the first step in ethanol biosynthesis requires generation of acetaldehyde using NAD^+ dependant bifunctional acetaldehyde and ethanol dehydrogenases. Additionally, as noted in *C. ljungdahlii* (Liew *et al.*, 2017), acetogens can reduce acetate to acetaldehyde via aldehyde oxidoreductase (AOR) catalysis. Acetaldehyde is subsequently reduced to ethanol via alcohol dehydrogenase activity.

1.2.5.3. 2,3-butanediol synthesis

2,3-butanediol is a commodity chemical which, like acetate, is conventionally derived from crude oil. It is used as a platform chemical for synthesis of a range of chemicals including the solvents kethyl-ethyl-ketone and 1,3-butadiene (Celińska & Grajek, 2009). Production is not limited to acetogens, with several other sugar and citrate fermenting microbes being reported to produce it (Xu *et al.*, 2014; Ng *et al.*, 2012). However, acetogens offer an advantage over microbes in that they can generate 2,3-butanediol from gas fermentation rather than sugar based carbon sources. Acetogenic production of this four-carbon compound was first reported by Köpke *et al*. (2011), with generation observed in *Clostridium ragsdalei*, *C. ljungdahli*, and *C. autoethanogenum*. Unlike acetate and ethanol, 2,3-butanediol synthesis requires pyruvate as a direct precursor. Acetolactate synthase catalyses the generation of the first intermediate in 2,3-butanediol synthesis, acetolactate, by linking two molecules of pyruvate. Found ubiquitiously in nature, this intermediate has an important role in amino acid biosynthesis. For 2,3-butanediol production, acetolactate is split into acetoin and CO_2 by acetolactate synthase. Finally, 2,3-butanediol dehydrogenase catalyses the reduction of acetoin to 2,3-butanediol (Köpke *et al.*, 2011).

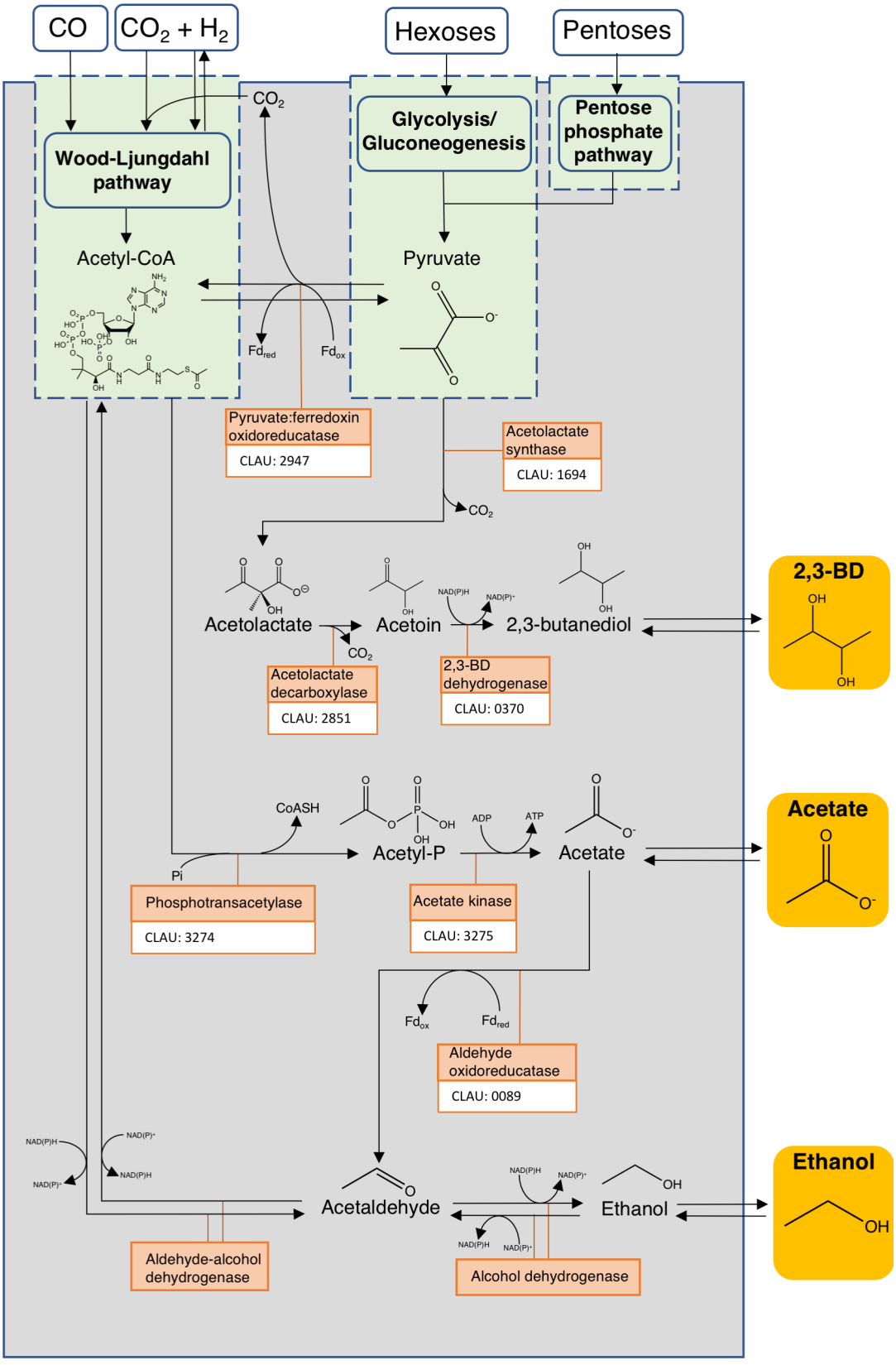

CO
CO2 + H2
Hexoses
Pentoses
CO2
Wood-Ljungdahl pathway
Glycolysis/ Gluconeogenesis
Pentose phosphate pathway
Acetyl-CoA
Pyruvate
Fdred
Fdox
Pyruvate:ferredoxin oxidoreducatase
CLAU: 2947
Acetolactate synthase
CLAU: 1694
CO2
Acetolactate
Acetoin
2,3-butanediol
NAD(P)H
NAD(P)+
CO2
Acetolactate decarboxylase
CLAU: 2851
2,3-BD dehydrogenase
CLAU: 0370
2,3-BD
CoASH
Pi
Acetyl-P
ADP
ATP
Acetate
Phosphotransacetylase
CLAU: 3274
Acetate kinase
CLAU: 3275
Acetate
Fdox
Fdred
Aldehyde oxidoreducatase
CLAU: 0089
Acetaldehyde
Ethanol
Aldehyde-alcohol dehydrogenase
Alcohol dehydrogenase
Ethanol

Figure 1.7: Acetate, ethanol and 2,3-butanediol generation in acetogens White boxes with blue boundaries represent substrates; with CO or $CO_2 + H_2$ feeding into the Wood-Ljungdahl pathway, and hexoses and pentose's feeding into glycolysis/gluconeogenesis and the pentose phosphate pathway respectively. Acetyl CoA is generated through the Wood-Ljungdahl pathway, and through pyruvate:ferredoxin oxidoreductase mediated conversion of pyruvate generated from glycolysis/gluconeogenesis/pentose phosphate metabolism. Products are represented in yellow boxes and enzymes involved in their synthesis pathways are represented in orange boxes. Acetate kinase and phosphotransacetylase govern conversion of acetyl-CoA to acetyl-P, and acetyl-P to acetate respectively. Bifunctional acetaldehyde and ethanol dehydrogenase activity catalyses the conversion of acetyl-CoA to acetaldehyde, which is in-turn converted to ethanol via alcohol dehydrogenase activity. 2,3-butanediol production; pyruvate is converted to acetolactate through acetolactate synthetase activity, which is converted to acetoin via acetlactate decarboxylase. 2,3-butanediol dehydrogenase activity forms 2,3-butanediol from acetoin. Based on figure by Liew *et al*. (2017).

1.2.6. Pathway engineering in acetogens

One conventionally employed approach for solvent optimisation in acetogens has been refinement of anaerobic fermentation conditions. As demonstrated in acetate production in *A. woodii*, modifying parameters such as pH and hydrogen solubility can lead to significant increases in yield of existing products (Weenk, Fitzmaurice & Mossel, 1991). However, through modification of endogenous genes, or introduction of heterologous genes, a metabolic engineering approach can offer further increases in the yields of existing fermentative products, or provide pathways for novel products. To date, several attempts have been made at metabolic engineering in acetogens, with examples across multiple organisms and for optimisation of a wide range of fermentative products.

1.2.6.1. Optimisation of existing products in acetogens

One recent example of pathway optimisation through metabolic engineering can be seen in the increase in ethanol production in *C. autoethanogenum* reported by Liew *et al*. (2017). This approach yielded a 180% increase in autotrophic ethanol production through deletion of *adhE*. Although this step may seem counterintuitive (as shown in Figure 1.7), *adhE* encodes an aldehyde/alcohol dehydrogenase required for production of acetaldehyde, the immediate precursor to ethanol), it provides a good example of a more holistic understanding of acetogenic metabolism.

Again, as shown in Figure 1.7, acetate production is the only solventogenic step to yield ATP, therefore *adhE* deletion redirects carbon flow away from acetaldehyde production and towards acetyl-P production. Acetyl-P is subsequently converted to acetate (producing ATP), which through the action of AOR (aldehyde:ferredoxin reductase) is converted to acetaldehyde. This alternate ethanol pathway had previously been postulated (Kopke *et al.*, 2010; Mock *et al.*, 2015), and deletion of *adhE* confirmed this postulation (Liew *et al.*, 2017).

1.2.6.2. Plasmid-based introduction of novel product pathways in acetogens

C. aceticum offers an early example of pathway engineering for novel product formation. Here, the *C. acetobutylicum* acetone production pathway was introduced into *C. aceticum* on a shuttle vector, resulting in an acetate dependant acetone pathway yielding 9 mg L^{-1} acetone (Figure 1.8) (Lederle, 2011).

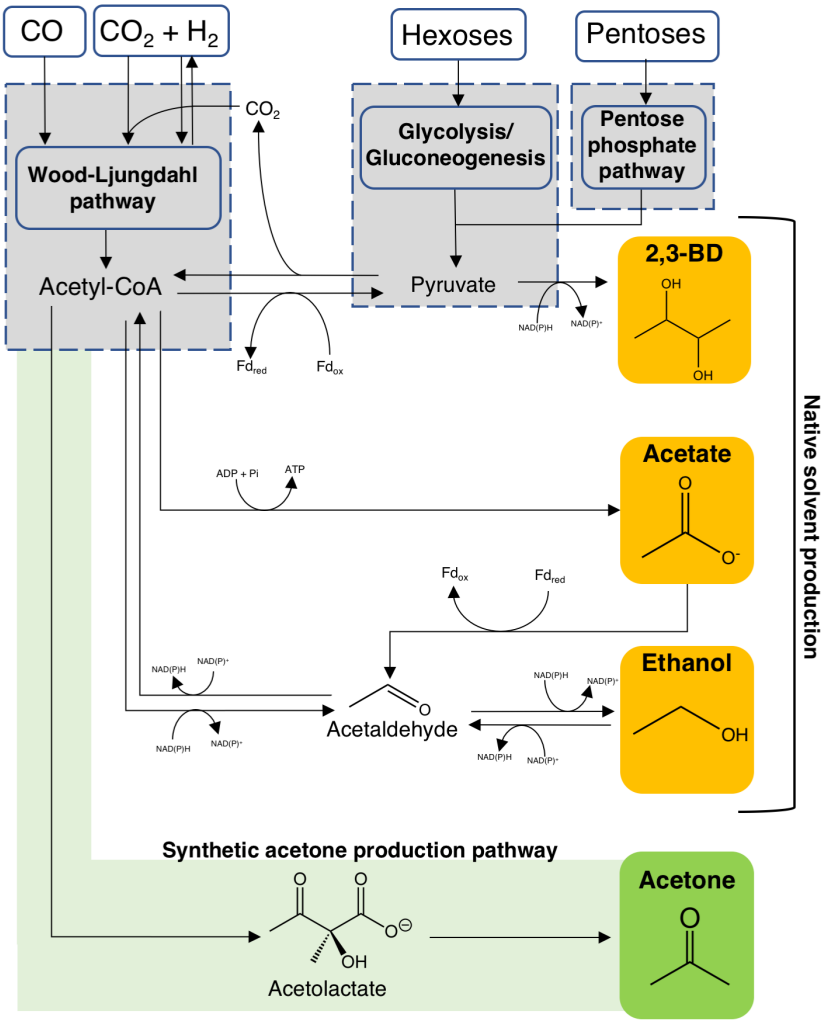


Figure 1.8: Example of synthetic acetone production pathway
As described by Lederle (2011)

Similarly, *C. ljungdahlii* is not naturally capable of butanol production but use of plasmid-borne *C. acetobutylicum* butanol pathway genes has been shown to enable this capability. This approach provided mixed results, with production in exponential growth phase reported to be up to 150 mg L^{-1}, but subsequent measurements showing complete consumption of butanol by *C. ljungdahlii* in later growth stages (Kopke *et al.*, 2010). Although these examples provided proof-of-concept, the nature of plasmid-borne gene expression is limited for industrial use. As noted by Schiel-Bengelsdorf & Durre (2012), this is due to its antibiotic-dependent nature preventing scale up, with genomic integration providing a more industrially feasible approach.

1.2.6.3. Limitations to pathway engineering in acetogens

Research discussed previously by Liew *et al*. (2017) highlights the importance of understanding the metabolic pathways being engineered. As stated, deletion of *adhE* could seem counterintuitive, and as a result prior attempts at increasing ethanol production have centred around over-expression of *adhE* (Peng, Wu & Shao, 2008; Thapa *et al.*, 2015; Yao & Mikkelsen, 2010).

A more holistic understanding of an organism can be gleaned from development of genome-scale model. This approach models the complete network of genes in the organism's metabolism and can be used to identify non-intuitive targets for metabolic engineering. Such a model was published in 2013 for *C. ljungdahlii* (Nagarajan *et al.*, 2013), and PhD student Rupert Norman is currently developing a more in-depth model at SBRC Nottingham.

As well as better understanding of an organism's metabolism, synthetic biology requires efficient genome editing tools. As stated above, expression of genes on plasmids is not an ideal solution for acetogenic fermentation. For production of another non-native product, butyrate, Ueki *et al*. addressed the plasmid issue by incorporating the required genes onto the *C. ljungdahlii* chromosome, using homologous recombination techniques. However, due to insufficient selection pressure, not all the desired gene integrations could be made (Ueki *et al.*, 2014). The limitations of conventional genetic tools available in the genus *Clostridium* act as a significant bottleneck to progress of metabolic engineering. Examples of their development, usage and current limitations are fully detailed in section 1.4.

1.3. DNA transfer and genome editing in Clostridia

Clostridia represent a broad class of Firmicutes, containing a polyphyletic mix of anaerobic bacteria that are generally Gram-positive and spore forming. The most numerous and best studied among these belong to the order Clostridiales, to which many acetogens belong to (see Figure 1.4). Also notable in this order are other biofuel producing and clinically relevant bacteria. An example of the former is *Clostridium acetobutylicum*, synonymous with ABE (Acetone, Butanol and Ethanol) fermentation. Clinically relevant members include the gangrene causing *Clostridium perfringens*, botulism causing *Clostridium botulinum*, and *Clostridium difficile*, a diarrhoea causing opportunistic pathogen associated with hospital acquired infection.

The presence of such well-characterised organisms in the class has led to a significant accumulation of molecular biology tools for their study and development, many of which can be applied to novel or poorly characterised clostridia. Such an approach has recently been laid out by Minton *et al.* (2016) in the paper entitled “A roadmap for gene system development in *Clostridium*”. Here a logical approach to clostridial molecular biology is set out, with analysis of plasmid replicon compatibility, antibiotic resistance, and restriction-methylation systems used to optimise DNA transfer. Once DNA transfer is established, applictation of gene-knockout, complementation and DNA insert techniques is attempted, culminating in forward- and reverse- genetics approaches to synthetic biology.

1.3.1. Plasmid vectors

The pMTL80000 vector series is a modular shuttle plasmid system designed for organism- and application- specific modification. As shown in Figure 1.9, pMTL80000 vector consist of four modules, each flanked by unique eight bp restriction sites for easy exchange. The modules are; a Gram-positive replicon, antibiotic resistance marker, Gram-negative replicon, and an application specific module.

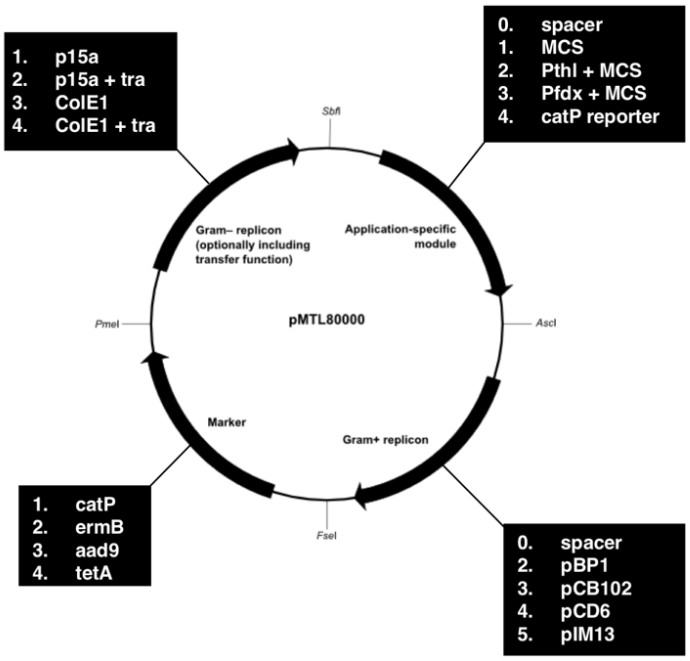


Figure 1.9: pMTL80000 vector series
Modular pMTL80000 vector consisting of Gram-negative replicon (top left) options of p15a and ColE1, each of which with the option of traJ for conjugative transfer. Application specific module (top right) options of spacer, multiple cloning site (MCS), MCS under the control of P_{thl} or P_{fdx} (*C. acetobutylicum* thiolase and *C. sporogenes* ferredoxin promoters, respectively), and a catP reporter gene for analysis of gene expression as described by (Shaw, 1975). Gram positive replicon (bottom right) options; spacer, pBP1, pCB102, pCD6 or pIM13 (from respectively; *C. botulinum*, *C. butyricum*, *C. difficile*, and *Bacillus subtilis*). Antibiotic resistance marker (bottom right) options; *catP* gene from *C. perfringens* encoding chloramphenicol/thiamphenicol resistance, *ermB* and aad9 from *Enterococcus faecalis* encoding erythromycin resistance encoding erythromycin and spectinomycin respectively, and *tetA* from *C. perfringens* plasmid pCW3 encoding tetracycline resistance. Module options are shown in black boxes and numbered according to pMTL80000 nomenclature (Heap *et al.*, 2009).

1.3.2. DNA transfer

Efficient DNA transfer is crucial for genetic and metabolic engineering of any organism. Generally, plasmid transfer is performed either via conjugation or electroporation in *Clostridium*[2].

1.3.2.1. Conjugation

Conjugation involves the direct transfer of genetic material from donor to recipient. In the context of clostridial molecular microbiology, this is conventionally achieved using the CA434 strain of *E. coli* as a donor. First described in *C. acetobutylicum*, this method is commonly used in *C. difficile* and is also effective in *C. autoethanogenum* (Purdy *et al.*, 2002; Williams, Young & Young, 1993; Liew *et al.*, 2016). CA434 is an HB101 strain of *E. coli*, harbouring the R702 plasmid. This IncP plasmid enables mobilisation of non-conjugative plasmids containing a *cis*-acting *oriT* (origin of transfer). In the context of the pMTL80000 vector series, the *oriT* used for conjugative transfer is derived from *E. coli* K12 plasmid RK2 (Ingram, Richmond & Sykes, 1973; Heap *et al.*, 2009).

1.3.2.2. Electroporation

Electroporation involves passing an electric current across the cellular membrane, briefly opening pores and enabling plasmid transfer. Efficient electroporation requires the generation of electrocompetent cells prepared in conditions to optimise transformation efficiency (Dower, Miller & Ragsdale,

[2] The 1980s saw a brief foray into DNA transder via PEG- (polyethylene glycol) mediated generation of Clostridium pasteurianum protoplasts (Minton & Morris, 1983). However, this technique is rarely used in modern day research.

1988). Optimisation of this process involves modulation of a range of parameters, including growth phase of cells and buffer. Other variables in electroporation that can be altered to improve efficiency include the electrical parameters and duration of the pulse delivered to the cells (Minton *et al.*, 2016; Pyne *et al.*, 2013).

1.3.3. Barriers to DNA transfer

The ubiquitous nature of phage and plasmids in microbial environments has driven prokaryotes to evolve mechanisms for protection from potentially harmful foreign invasive DNA elements. Although necessary for molecular biological approaches, and useful for gene acquisition via horizontal gene transfer, DNA transfer in prokaryotes is often detrimental in a native context. As a result, prokaryotes deply two key mechanisms for degradation of foreign DNA; restriction-modification systems, and CRISPR (Clustered Regularly Interspaced Repeats, see section 1.5) systems.

Discovery and characterisation of these two prokaryotic defence measures, separated by several decades, has led to two unforeseen quantum leaps in molecular biology. Firstly, understanding of restriction enzymes led to the birth of recombinant DNA technology. Advances in this field provided the foundation for molecular cloning. Secondly, recent advances in CRISPR-based technology have enabled development of highly efficient genome editing tools.

Although pivotal to the field of molecular biology, these systems, particularly restriction-methylation, have provided significant bottlenecks to DNA transfer in certain prokaryotic species.

1.3.3.1. Restriction-modification systems

Restriction-modification (or restriction-methylation, RM) systems are present in a broad range of prokaryotes and are divided into four types (Types I-IV), as outlined in Table 1.2. Each type consists of restriction endonucleases and methyltransferases, except for Type IV systems which lack methytransferases. Of these systems, Type II are most common.

Table 1.2: Overview of restriction-methylation systems in prokaryotes
Key features of Type I-IV restriction systems described, including mode of action. Based on classification of Type I-III restriction enzymes by Wilson (1991) and Lepikhov *et al.* (2001).

Restriction-methylation system	Key features
Type I	• Complex, multi-subunit combination of endonucleases and methyltransferases • Endonucleases cut at random far from recognition sequence
Type II	• Endonucleases and methyltransferases encoded in two separate genes and have independent activity • Both proteins share the same recognition site and bind competitively • Endonuclease activity results in cleavage at a defined site close to recognition sequence • Commonly used for recombinant DNA applications
Type III	• Consist of endonucleases and methyltransferases that complex and perform duel function • Methyltransferase capable of methylation without forming complex
Type IV	• Consist only of endonucleases • Recognise only modified/methylated DNA

RM systems serve to distinguish native DNA from non-native DNA. Upon the identification of foreign DNA containing a specific recognition site, endonucleases produce a double-stranded cut. In an endogenous context,

this protects from invasive (potentially harmful) DNA elements through nucleolytic degradation of incorrectly methylated DNA. For Type I-III RM systems, endonucleases target specified recognition sites, but are inhibited in the presence of methylation. In contrast, Type IV systems actively cut incorrectly methylated DNA. For these systems, methylated bases that should be non-methylated are degraded, whereas DNA containing no methylation or the correct methylation pattern being protected from degradation.

Type I restriction endonucleases cut at a random site over 1000 bp from their recognition site. As a result, these enzymes have limited applications in molecular biology. In contrast, Type II endonucleases cut at- or within a short distance of- the recognition sequence. This predictability has led to their adoption in molecular biology for *in vitro* cloning techniques (Lepikhov *et al.*, 2001).

Methyltransferases (methylases) add methyl groups to newly synthesised DNA at defined points. For example, the well known *E. coli* methylases Dam (Deoxyadenosine methylase) and Dcm (Deoxycytosine methlylase) act respectively to add methyl groups at adenosine and cytosine residues present in specific DNA motifs (Palmer & Marinus, 1994), as illustrated in Figure 1.10. Different species contain methylases acting on distinct residues within different motifs bases. Although methylation sites can be predicted through algorithms such as NEB REBase, the exact methylation activity of methylases must be confirmed biochemically and therefore, especially for non-model organisms such as acetogens, relatively little information is available.

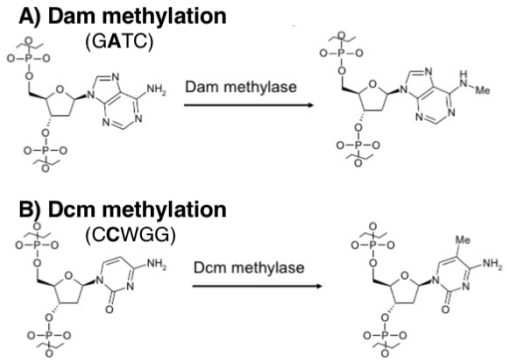


Figure 1.10: *E. coli* Dam and Dcm methylation
A) Methylation of N6 position of adenine within GATC motif via Dam methylase. B) Methylation of C5 position of second cytosine residue in CCWGG motif via Dcm methylase (Palmer & Marinus, 1994).

1.3.3.2. Avoiding host RM systems

In the clostridial field there are two main approaches for avoiding RM systems; mimicking host methylation patterns, and deletion of endogenous restriction endonucleases. Examples of the former include the use of methylation plasmids to methylate shuttle vectors used in *C. acetobutylicum*. Methylation plasmids encoding a *B. subtilis* methyltransferase were harboured in an *E. coli* strain along with a shuttle vector for use in the *C. acetobutylicum* host. Plasmid DNA extraction from *E. coli* results in purified shuttle vector DNA with compatible methylation to the host DNA, protecting it from degradation. This approach was shown to enable transformation via electroporation, something not possible without the methylation vector (Mermelstein & Papoutsakis, 1993). Examples of the latter approach include the deletion of *cpaAIR* in *C. pasteurianum*. This gene encodes a Type II restriction endonuclease, and its deletion enabled unmethylated plasmids to be transformed in, where previously no transformants were observed for wild type (Pyne *et al.*, 2013).

1.4. Clostridial genetic engineering and genome editing

In practice, metabolic engineering is a complex process; gene deletion, gene insertion, codon optimisation, ribosome-binding site selection, promoter selection, and many other factors must be considered. To achieve success, multiple combinations of the above factors must be studied, making the process heavily reliant on efficient methods of molecular biology.

This can be exemplified by the development of an *E. coli* strain capable of producing 1,3-propanediol from D-glucose. *E. coli* can generate 1,3-propanediol using glycerol as a feedstock, however, as a cheaper substrate D-glucose presented itself as a more cost-effective alternative. Despite this pathway being in theory relatively simple to engineer, Genencor reported that in order to generate yields sufficient for industrial production, over 60 genetic modifications were required (Nakamura & Whited, 2003).

Considering *E. coli* is arguably the best-studied microbe, with well-characterised pathways and established genetic tools, this example highlights the need for efficient genome editing tools in acetogens. In contrast to *E. coli*, acetogens are generally slower growing[3], less characterised, and less amenable to genetic manipulation. Therefore, in order to attempt any form of metabolic engineering, efficient tools for making accurate genome modifications are crucial. As discussed in Section 1.2.6, plasmid-based methods of metabolic engineering have significant drawbacks in an acetogenic fermentation context. This means such modifications require

3 For example, *C. autoethanogenum* takes at least 3 days for visible colonies to appear on plates, and a similar amount of time for cell density in liquid cultures to be visible.

permanent modifications to the chromosome, for which two main tools are conventionally used in *Clostridium*; ClosTron, and allelic exchange methodologies.

1.4.1. Conventional genetic engineering in Clostridia

Prior to 2007, targeted mutagenesis in clostridia was limited to basic homologous recombination techniques. Strategies either involved integration of a plasmid backbone for insertional mutagenesis, or introduction of an alternative allele. Both options proved problematic as they were plagued by issues of instability and low efficiency respectively. These issues were partially addressed in 2007 with the development of ClosTron technology, and later with allelic exchange methods.

1.4.1.1. ClosTron

ClosTron mutagenesis relies up targeted insertion of a modified *Lactococcus lactus* group II intron (derived from the *ltrB* gene) into a gene of interest, resulting in gene disruption. This intron contains the *E. faecalis ermB* gene acting as a selectable marker for gene disruption, itself containing a self-splicing group I intron. Due to the presence of the group I intron, the *ermB* gene is disrupted when on the plasmid, and is not able to be expressed unless the group I intron self-excises and the group II intron delivers it into its chromosomal target site. Figure 1.11 illustrates a detailed representation of the ClosTron method of action. Further details can be found in the 2007 paper announcing the technology, and the follow-up paper published three years later (Heap *et al.*, 2010, 2007).

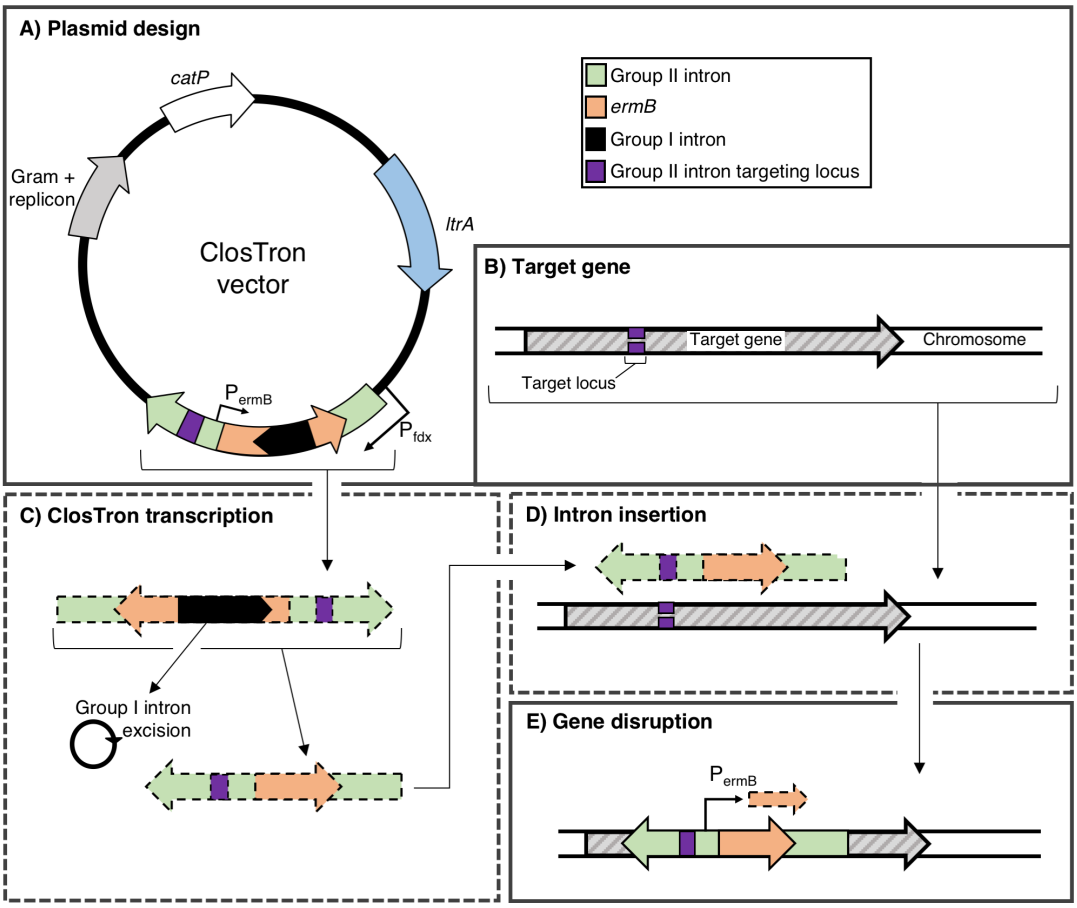


Figure 1.11: Schematic overview of ClosTron gene disruption
A) ClosTron vectors consist of a Gram-negative replicon (not shown) for cloning, and four key components; a deficient Gram-positive replicon, a *catP* gene for plasmid selection, the *Lactococcus lactis ltrA* gene, and the ClosTron Cassette. The ClosTron cassette consists of a Group II intron driven by the *C. sporogenes fdx* promoter (P_{fdx}), within which is retrotransposition-activated marker (consisting of a reverse orientation *ermB* gene and promoter interrupted by a self-splicing Group I intron). This Group I intron can self-splice when transcribed in the forward orientation, meaning transcription of the plasmid-based *ermB* gene from its promoter results in a transcript that remains interrupted by the Group I intron and therefore does not enable antibiotic resistance. B) When transcribed from P_{fdx}, the Group I intron is able to self-excise resulting in a transcript consisting of the Group II intron containing the *ermB* gene and promoter in the antisense orientation. Again transcription of this transcript does not result in antibiotic resistance due to *ermB* being in the wrong orientation. C) As described by Heap *et al.*, the Group II targeting locus can be modified to target any gene of interest, and is achieved through the action of *ltrA*. D) ClosTron insertion results in disruption of the target gene, and the *ermB* promoter is able to drive expression of the uninterrupted *ermB* gene, resulting in erythromycin/clarithromycin resistance. The deficient nature of the Gram-positive replicon enables plasmid loss, and the resulting lack of *ltrA* ensures stability of the Group II intron in the target gene. As described by Heap *et al.* (2007 & 2010).

1.4.1.2. Allelic exchange methodologies

Although proven to be an effective gene disruption technique, ClosTron has significant drawbacks; due to reliance on antibiotic resistance, knockouts are generally limited to single genes[4]. In addition, due to the nature using of intron insertion for disruption, polar affects can occur by affecting expression of downstream genes. Finally, although fast and effective for gene disruption, intron-based methods are ineffective knock-in measures. These issues were at least partially addressed with the next major developments in clostridial genetic engineering; allelic exchange and ACE (Allele Coupled Exchange) (Ng *et al.*, 2013).

Allelic exchange methodologies require a chromosomal gene that enables two-way counter-selection, and a defective Gram-positive replicon. For applications in clostridia, this gene is almost exclusively *pyrE*, encoding orotate phosphoribosyl-transferase (ORPTase), an enzyme essential in pyrimidine biosynthesis. The physiological role of ORPTase is the conversion of orotic acid to orotidine-5'-monophosphate, a precursor for uracil synthesis. However, ORPTase is also able to catalyse the inert 5-fluoro-orotic acid (5-FOA) to the toxic compound 5-fluoro-UMP. Therefore, *pyrE* fulfils the role of a two-way counter-selective gene perfectly; growth on defined media lacking uracil selects for the presence of a functional *pyrE* gene, and growth on media containing 5-FOA counter-selects against it (Minton *et al.*, 2016; Ng *et al.*, 2013).

4 Although ClosTron technology includes FRT (Flippase Recognition Target) sites flanking the *ermB* gene, evidence of their successful excision of the selection marker has not been shown in many organisms

The other essential component for allelic exchange methodologies is a defective Gram-positive replicon. Using plasmids with such replicons slows the population's exponential growth rate, due to plasmid replication being the limiting factor. Here, the plasmid replication rate is lower than the doubling time of the culture, therefore some daughter cells are killed due to lack of plasmid-encoded antibiotic resistance. As allelic-exchange methodologies centre around homologous recombination, a defective replicon enables selection for such an event. If a single cross-over event occurs at one of the homology arms, the plasmid is replicated using genomic replication machinery, relieving the defective replication and increasing growth rate. Therefore, a successful DNA transfer event can be screened via antibiotic resistance, and subsequent single-crossover homologous recombination events via colony size.

1.4.1.2.1. ACE-mediated *pyrE* deficient strain generation

Prior to using allelic exchange methodologies, a *pyrE* deficient strain is created using an ACE vector containing an antibiotic resistance marker, defective Gram-positive replicon, and highly asymmetrical homology arms. First, the longer right homology arm, consisting of the ~1.2 kB region immediately 3' to *pyrE*, directs an initial recombination event. Screening for such an event is achieved through larger colony size on antibiotic selection, with single-crossover integrants benefiting from relief of a growth rate limited by plasmid replicon. These larger colonies are passaged onto FOA containing media, which selects for a double-crossover event. Here, the shorter left homology arm, consisting of ~300 bp of *pyrE* immediately 5' to the last ~200

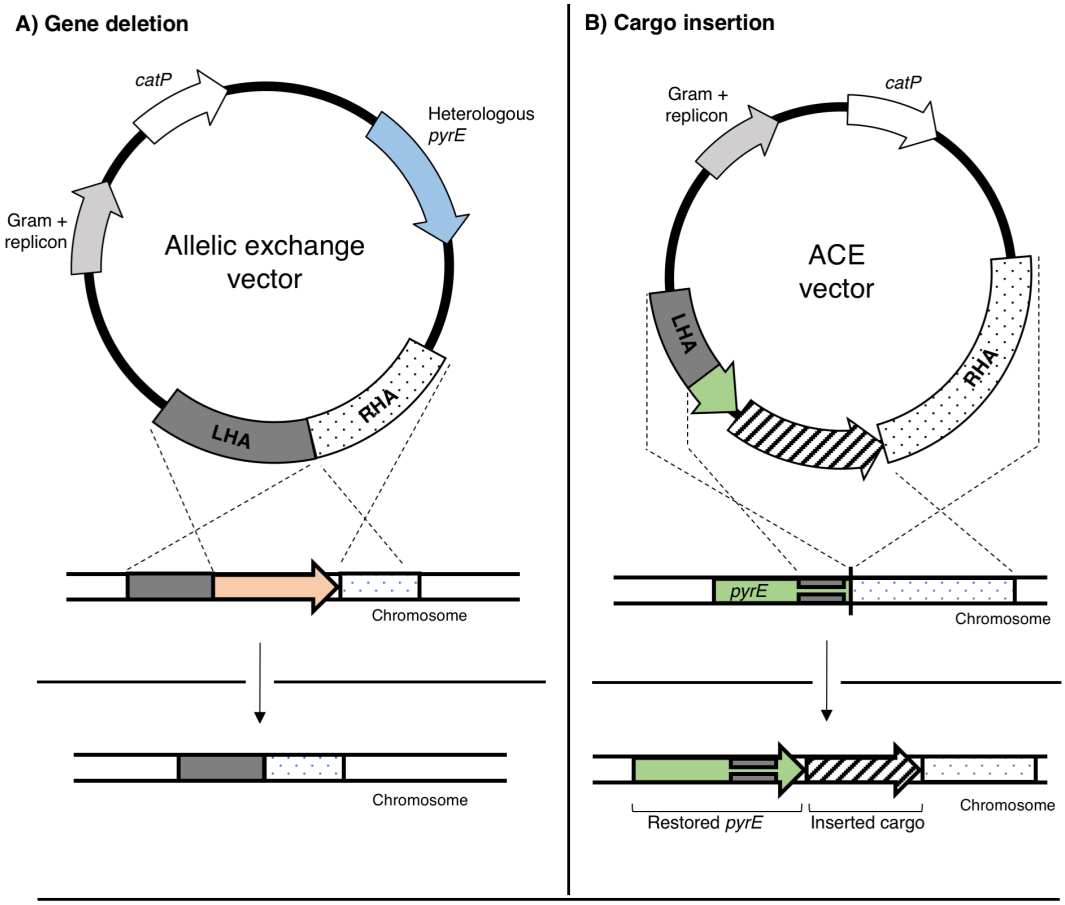
bp of the gene, directs a second recombination event. This leads to removal of the 3' terminal ~200 bp of *pyrE*, removing ORTPase activity and resulting in an FOA resistant phenotype.

As shown in Figure 1.12, once ACE-mediated truncation of *pyrE* is achieved, the *pyrE* deficient strain can be used for both deletion of genes of interest, and insertion of genetic cargo at the *pyrE* locus.

For gene deletions, a vector containing equally sized homology arms, an antibiotic resistance marker, a defective Gram-positive replicon, and a heterologous *pyrE* gene is used. Here antibiotic resistance is used for initial screening of transformants, and again, colony size is used as an indicator for single-crossover events. Once pure single-crossover integrants are confirmed, they are passaged onto FOA-supplemented media. The presence of the heterologous *pyrE* on the vector is toxic under these conditions, meaning this FOA selection forces plasmid excision and therefore a double-crossover event (Minton *et al.*, 2016).

For genetic insertions, complementation of the *pyrE* locus is used to select for double-crossover events. Again, ACE vectors (as described above) as used, but with the desired cargo situated between the smaller left homology arm and larger right homology arm, along with the previously truncated 3' region of *pyrE*. Again, antibiotic resistant colonies are screened for by size, with the larger colonies likely displaying a single-crossover event with the right homology arm. Following identification of single-crossover integrants, defined minimal media lacking uracil is used to select for a double-crossover event. This selection is achieved through the double-crossover event repairing the

pyrE gene and restoring full pyrimidine biosynthesis capabilities. Due to the desired cargo being located between the 3' end of *pyrE* and the right homology arm, its insertion is achieved through this double-crossover event. This approach can be used simply for repair of the *pyrE* gene to restore wild type phenotype, insertion of heterologous cargo, or complementation of gene if the knockout was made in a *pyrE* deficient background.



C) Two-way counterselection

	No uracil supplementation		Uracil supplementation	
	FOA selection	No FOA selection	FOA selection	No FOA selection
pyrE intact				
Δ*pyrE*				

Figure 1.12: Allelic exchange methodologies
A) Allelic exchange vector consisting of; Gram-positive replicon (light grey), *catP* gene (white), heterologous *pyrE* gene, and left homology arm (dark grey) and right homology arm (white with black dots). Homologous recombination between vector homology arms and chromosomal regions results in deletion of the gene of interest (peach). B) ACE (Allele Coupled Exchange) vector consisting of Gram positive replicon, *catP* gene, shorter LHA and longer RHA flanking heterologous genetic cargo (black and white diagonal stripes). Homologous recombination between vector and chromosomal regions results in restoration of the endogenous *pyrE* gene and simultaneous insertion of cargo. C) Illustration of two-way counterselection. Strains with intact *pyrE* are capable of pyramidine biosynthesis and therefore require no supplementation. However, they are also able to degrade uracil analogue FOA to toxic 5-FUMP, meaning they cannot grow in its presence. Conversely, Δ*pyrE* strains have incomplete pyramidine biosynthesis and therefore require uracil supplementation to grow, but can grow in the presence of FOA. Therefore for (A), selection on FOA in Δ*pyrE* strain forces loss of plasmid harbouring an intact *pyrE* gene. For (B), selection for cargo integration is enabled through removal of pyrimidine auxotrophy on minimal media lacking uracil.

1.4.1.3. Limitations of conventional clostridial genetic editing strategies

As previously stated, ClosTron technology is limited in its ability to only generate insertional knockouts. Rather than clean, in-frame deletions, this technique relies upon intron insertion disrupting the gene. However, this creates issues with polar effects, for example, targeting the first gene in an operon may result in a frameshift mutation affecting the downstream genes.

Although allelic exchange methodologies partially address these issues, significant drawbacks remain. Firstly, this technique relies upon generation of *pyrE* deficient which can itself be problematic. FOA resistance is used as a positive selection pressure for this, however prior knowledge of MIC (minimum inhibitory concentration) must be attained. If MIC as high, issues of FOA solubility may pose problems. DMSO is used as a solvent for FOA, but if certain species require high concentrations of FOA, this results in greater DMSO exposure which, at high concentrations can be toxic. Further to this,

the genomic location of *pyrE* is important. In most species of *Clostridium*, it is found at the end of its operon, allowing truncation and insertion of downstream cargo. However, this approach becomes more problematic if *pyrE* is instead location in the middle of the operon.

Beyond issues of FOA MIC and *pyrE* location, a major drawback of this technique it its weak selection pressure. The technique's reliance on screening via colony size can be problematic. For example, if deletion of the gene of interest results in a slower growing phenotype, this may result in no discernible difference in colony size between integrants and non-integrants.

1.4.2. Prokaryotic genetic engineering advancements

Outside of the clostridial genus, several key technological breakthroughs have enabled researchers to develop easier to use and more accurate technique for genetic manipulations. These have been adopted and adapted for use in the genus to varying degrees, as discussed below.

1.4.2.1. Recombineering

Recombineering utilises phage-derived proteins that catalyse integration of single- or double- stranded linear DNA (ssDNA/dsDNA) into chromosomal targets using small homologous flanking regions in the DNA sequence. This technique pioneered in *E. coli* has addressed many of the issues associated with prokaryotic genetic engineering (Sharan *et al.*, 2009).

To date two types of phage-derived systems are available; RecE/T derived from Rac prophage, and Lambda phage derived Red $\alpha/\beta/\gamma$. For applications using dsDNA, a recombineering approach using an exonuclease, such as RecE, is required to degrade the ssDNA into ssDNA. This component of the

system is not required when using ssDNA, however, such an approach limits construct size and complexity. Recombineering is thought to work through homologous recombination of ssDNA to the lagging strand of the replication fork during DNA replication, acting as an Okazaki fragment. Successful integration results in one of the newly synthesised strands containing the desired mutation, however, practically the efficiency of this technique is relatively low (Muyrers *et al.*, 2000; Costantino & Court, 2003; Sharan *et al.*, 2009).

Despite wide success in other prokaryotes (van Kessel & Hatfull, 2007; van Pijkeren & Britton, 2012; Sharan *et al.*, 2009; Mosberg, Lajoie & Church, 2010), recombineering has not been widely reported in clostridia. Previous attempts are thought to have failed because RecT is host dependent and the Gram-positive clostridial genus is too evolutionarily distinct from *E. coli.*

However, Dong *et al*. recently reported successful recombination in *C. acetobutylicum* through use of putative *recT* gene from *C. perfringens*, although this approach is unfortunately limited to ssDNA, due to no RecE homolog being present (Dong *et al.*, 2014).

Despite appearing as a promising leap forward in clostridial engineering, this recombineering approach has not been widely adopted. Issues for such poor uptake could be attributed to limitations on size and complexity of ssDNA fragments, as well as lack of host specificity beyond *C. acetobutylicum*. Recombineering offers much faster generation of modified strains compared to conventional techniques and is not reliant on antibiotic resistance markers. However, this is a double-edged sword as lack of selection results in low

efficiency; 0.1-1% for point mutations, and as low as 10^{-6} for larger changes. Coupling of this technique with CRISPR-Cas9 selection (see section 1.5) has addressed these issues in other organisms (Barrangou & van Pijkeren, 2016), but this does not solve the underlying lack of compatible recombinases for acetogens.

1.4.2.2. Engineered nucleases for genome editing

Researchers involved in the eukaryotic taxa saw a huge leap forward in terms of genetic tools available due to the development and implementation of programmable nucleases for targeting genome editing. First described with Zinc-Finger Nucleases (ZFNs), then TALENS (Transcription Activator-Like Effector Nucleases), and finally and most notably, CRISPR (Gaj, Gersbach & Barbas, 2013; Joung & Sander, 2013; Kim & Kim, 2014). These methods all rely upon targeted cleavage of genomic DNA

ZFNs were first developed by combining the DNA-binding domain of eukaryotic transcription factors (Zinc Finger Proteins, ZFPs) with the nuclease domain of the *Fok*I restriction enzyme. This chimera can show flexible but specific targeting and can be reprogrammed for any target sequence(Sakuma & Woltjen, 2014). Like ZFNs, TALENs also utilize the *FokI* restriction enzyme cleavage domain. Instead of ZFPs, TALENs use DNA-binding domains derived from transcription activator-like effector proteins. TALENs offer greater flexibility of targetable sites compared to ZFNs, for example they are able to target enhancers or miRNA coding sequences, whereas ZFNs are unable to do so (Joung & Sander, 2013).

Although far more effective than previous techniques, both these ZFNs and TALENs have been dwarfed by the advent of a third nuclease-dependent technique, CRISPR-Cas9.

1.5. CRISPR systems and their applications

The CRISPR acronym was coined by Jansen *et al*. in 2002, referring to clustered regularly interspaced palindromic repeats present in many prokaryotes. The fact these arrays often contain phage or plasmid derived sequences in “spacer” regions provided the foundation for the understanding of the role of CRISPR systems in prokaryotic adaptive immune defence. Identification of CRISPR associated (Cas) genes enabled their function and interaction with the CRISPR array to be hypothesised, based upon the activity of known related genes. This understanding has since been implemented to adapt CRISPRs for use as genetic engineering tools.

CRISPR systems are present in 90% of sequenced archaea and at least 40% of sequenced bacteria, including *C. autoethanogenum* and *E. limosum*. These systems function as adaptive defence mechanisms against invasive elements such as bacteriophage and plasmids (Marraffini & Sontheimer, 2008). Current dogma states that Cas proteins recognise phage DNA and insert part of a phage derived specific sequence (known as a protospacer) into the CRISPR array, where it is known as a spacer. If re-exposed to the phage or plasmid, other Cas proteins can complex with the transcribed CRISPR array and use the spacer as a targeting reference for site-specific cleavage, destroying the invasive DNA element (Bhaya, Davison & Barrangou, 2011; Heidrich & Vogel, 2013).

1.5.1. From genome unknown to genome editing

Long before the elucidation of their function, let alone the coining of the "CRISPR" acronym, the advent of whole genome sequencing revealed unexplained repeat regions in the *E. coli* (Ishino *et al.*, 1987). Dismissed as unimportant, the significance of these findings fell undetected until Franciso Mojica, a young Spanish researcher, found similar repeats in the halotrophic archaea *Haloferax mediterranei.*

Sequencing of *H. mediterranei* resulted in the identification of palindromic repeat sequences interspaced with other DNA of uniform length but varying sequence (Mojica, Juez & Rodriguez-Valera, 1993). In an attempt to understand the nature of these repeats, Mojica sought examples in related species and eventually more distant prokaryotes, including that of the *E. coli* repeats described by Ishino and colleagues. Over the course of the next decade Mojica identified examples of CRISPR loci, an acronym he helped coin in 2002, and their associated (Cas; CRISPR associated) genes (Mojica & Montoliu, 2016). Their role in adaptive immune response was elicited in 2003, but it would take most of the remaining decade to understand how (Lander, 2016).

Research by Barrangou and coworkers showed adaptive immunity in *Streptococcus thermophilus* against phage attack was governed by acquisition of spacers into CRISPR loci, from phage derived protospacers (Barrangou *et al.*, 2007). Subsequent research showed the necessity of Cas9 for phage resistance, and noted the presence of its two nucleolytic domains (Barrangou *et al.*, 2007). These findings were built upon by Marraffini and

Sontheimer (2008) who demonstrated the nucleolytic activity of Cas9. The identification of Cas9 as the sole requirement for cleavage, and the identification of RNA elements guiding Cas9 had been achieved by the end of the decade (Jinek *et al.*, 2012), providing the platform for the adaptation of CRISPR-Cas9 for genome editing as described in Chapters 3 and 4.

1.5.2. Classification

To date, the main Cas gene of interest has been *S. pyogenes cas9*, coding for the Cas9 nuclease that has been pivotal in genome editing. However, the Type II system to which this Cas9 belongs is just one of a cast of many types of CRISPR system offering adaptive defence against invasive DNA elements. Over the last decade, much research has gone into characterising and classifying CRISPR systems into different classes, types, and subtypes. The seminal unified classification system was put forward by Makarova *et al.* (2011), dividing systems into three main types (Type I, Type II and Type III). This classification his since been updated, retaining the “type” classification but dividing systems into two classes; Class 1 encompassing Type I, III and IV CRISPR systems, and Class 2 encompassing Type II and V systems (Makarova *et al.*, 2015). Figure 1.13 visualises this updated classification system, illustrating the functional components of each type of CRISPR system.

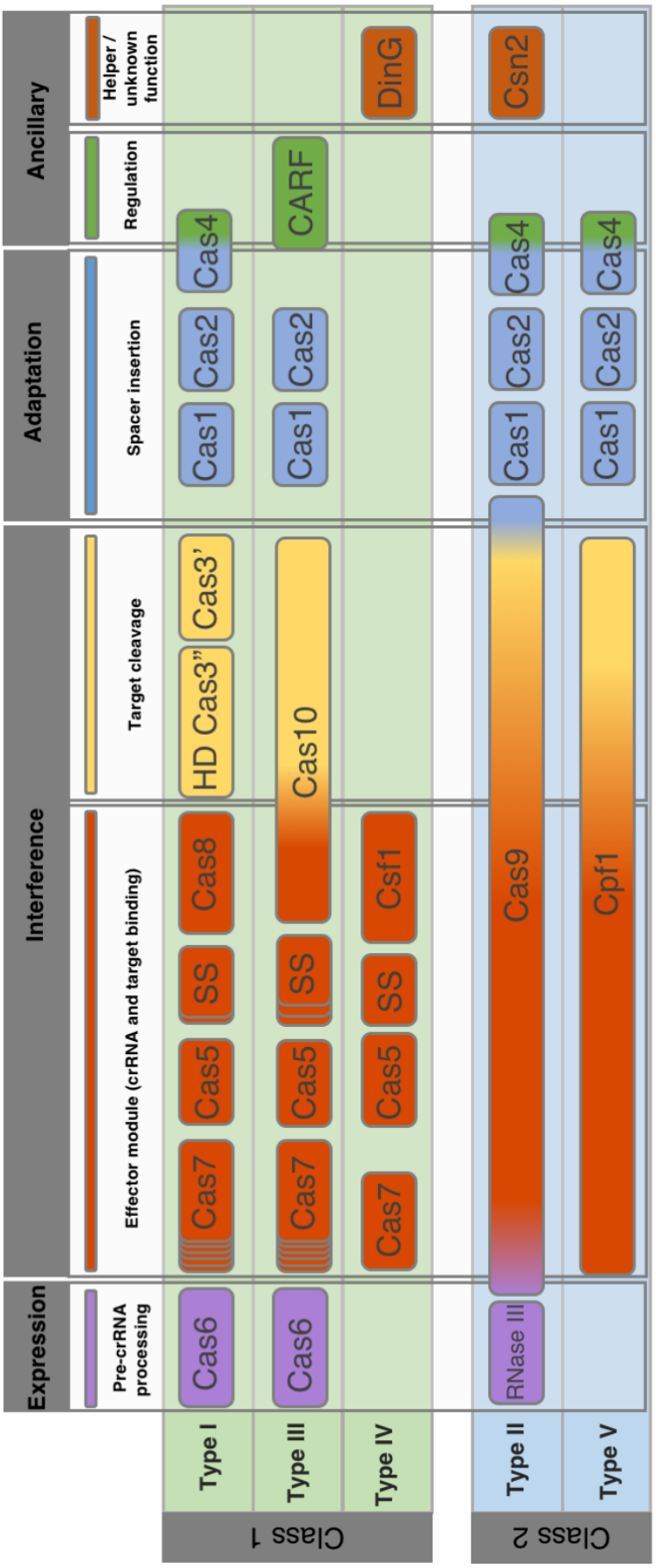


Figure 1.13: Overview of CRISPR system classification
Overview of CRISPR system classifications as described by Makarova *et al.*, figure adapted from Makarova *et al.* (2015).

1.5.3. Spacer acquisition

As adaptive defence mechanisms, CRISPR systems are able to detect foreign DNA, integrate small sequences of it into a reference library in the form of spacers, and use this library to guide targeted degradation. Protospacer adjacent motifs (PAMs) are crucial in this process, firstly in identifying target sites for CRISPR-mediated cleavage, and secondly in determining target specificity. In both classes of CRISPR systems, PAMs provide redundancy for target specificity (Fineran *et al.*, 2014).

Although spacers within the processed CRISPR array transcript specify target sites, these sites are only cleaved if found directly adjacent to a correct PAM sequence. PAM specificity ensures that CRISPR loci themselves are not subject to cleavage by Cas nucleases, as although spacers are identical in sequence to the protospacers from which they are derived, they lack the necessary PAM for nucleolytic activity to occur (Marraffini, 2015). This is illustrated in Figure 1.14, showing the identification of a protospacer, incorporation of its sequence in spacer form into a CRISPR locus, and the transcribed CRISPR array providing targeting for protospacer cleavage.

Although the lack of PAMs adjacent to spacers in CRISPR loci explains why the CRISPR loci are not targeted, this does not explain the ability of CRISPR systems to discern “self” from “non-self”. Several studies have shown CRISPR systems have a strong preference for uptake of non-chromosomal DNA. The mechanism used to achieve avoidance of autoimmunity is not fully understood, however, it has been hypothesised that it could involve a bias towards exposed ends of DNA (Levy *et al.*, 2015).

Studies in *E. coli* have shown that despite PAMs being evenly distributed throughout its genome, chromosomally-derived spacers are rare and show strong bias towards specific regions of the genome[5] (Díez-Villaseñor *et al.*, 2013). These areas, such as those proximal to replication terminus (*Ter)* sites and Chi (Crossover hotspot instigator) sites frequently have exposed ends of dsDNA[6], leading to the hypothesis that such exposed ends are preferential (if not essential) for spacer-acquisition. This hypothesis is consistent with the nature of plasmid and bacteriophage replication and copy number being much higher than that of chromosomal DNA. Therefore, any preference towards exposed ends of dsDNA favours the uptake of foreign DNA (Levy *et al.*, 2015; Yosef, Goren & Qimron, 2012).

The proteins involved in this spacer acquisition process are the highly conserved Cas1 and Cas2. Aside from the hallmark of CRISPR loci, the archetypal feature of almost all CRISPR systems is the presence of *cas1* and *cas2* genes[7]. These proteins are ubiquitously essential, and in some cases the sole requirement for spacer acquisition (Díez-Villaseñor *et al.*, 2013; Datsenko *et al.*, 2012). This is the case in the Type IE system of *E. coli* K12,

5 The incorporation of chromosomal-derived spacers would ordinarily be lethal due to targeted nucleolytic activity towards the chromosome and is, therefore, not observed in native context. However, this study was performed in the BL21-AI strain of *E. coli*, lacking Cas genes, thereby rendering any chromosomal-derived spacers non-lethal (Díez-Villaseñor *et al., 2013)*.

6 Chi sites are regularly occurring short sequences evenly distributed in bacterial genomes involved in double-strand break repair. RecBCD helicase-exonuclease activity degrades DNA from the site of damage until it reaches a Chi sites. Along with Ter, Chi sites are among the most common regions for exposed DNAchi(Levy *et al.*, 2015).

7 With the exception of Type IV CRISPR systems

where the two proteins form a heterohexameric complex. Both proteins have catalytic activity, however, only the activity of Cas1 appears to be essential. The contribution of Cas2 appears to be in providing the crab-like shape of the Cas1-Cas2 complex. Although Cas2 has RNase and DNase activities, the Cas1-Cas2 complex is still able to insert new spacers in CRISPR loci when these catalytic domains are disrupted (Díez-Villaseñor *et al.*, 2013; Hayes *et al.*, 2016).

In contrast to the *E. coli* Type I system, the Type II system of *S. pyogenes* also requires the multifunctional Cas9 for spacer acquisition. As well as being the sole protein required for target interference (compared to the multi-protein Cascade complexes of Class 1 systems), it is thought the PAM recognition capabilities of Cas9 are required to identify new spacers (Heler *et al.*, 2015). The other main protein associated with spacer acquisition is Cas4, found in Type 1-A, I-B, I-C, I-D, and II-B CRISPR systems. Research has shown that Cas4 provides exonuclease activity and is generally thought to be involved in spacer insertion with various hypotheses for other functions (Zhang *et al.*, 2012).

1.5.4. Class 1 CRISPR systems

As illustrated in Figure 1.13, Class 1 systems consist of the structurally similar Type I and Type III, as well as putative Type IV CRISPR systems. Over 90% of all documented CRISPR systems belong to this class, which is reflected in the diverse range of prokaryotes containing such CRISPR systems. The defining feature of these systems is their use of multi-protein effector complexes, in contrast to the single-protein effectors found in Type II systems, such as *S. pyogenes* Cas9 (Makarova *et al.*, 2015).

1.5.4.1. Class 1 CRISPR loci and RNA processing

Class 1 CRISPR systems require a single RNA element for effective targeting. Unlike their Class 2 counterparts that require crRNA and tracrRNA, Class 1 systems only require the former (Wakefield, Rajan & Sontheimer, 2015). Once the CRISPR locus is transcribed to form pre-crRNA, the Cas6 ribonuclease cleaves the repeat sequence. The resulting products of this cleavage contain the spacer region flanked by a 5' handle and 3' stem-loop structure. Cas6 processes pre-CRISPR RNA, resulting in spacers flanked by a 5' handle and 3' stem-loop structure (Richter *et al.*, 2012; Lintner *et al.*, 2011; Li, 2015).

1.5.4.2. Class 1 effector complexes and target cleavage

Class 1 effector complexes consist of several Cas proteins that interact with CRISPR RNA (crRNA) to recognise and bind spacer-encoded targets. Although the sequence similarity of effector complex coding genes is almost undetectable between Type I and Type III systems, structurally they are extremely similar.

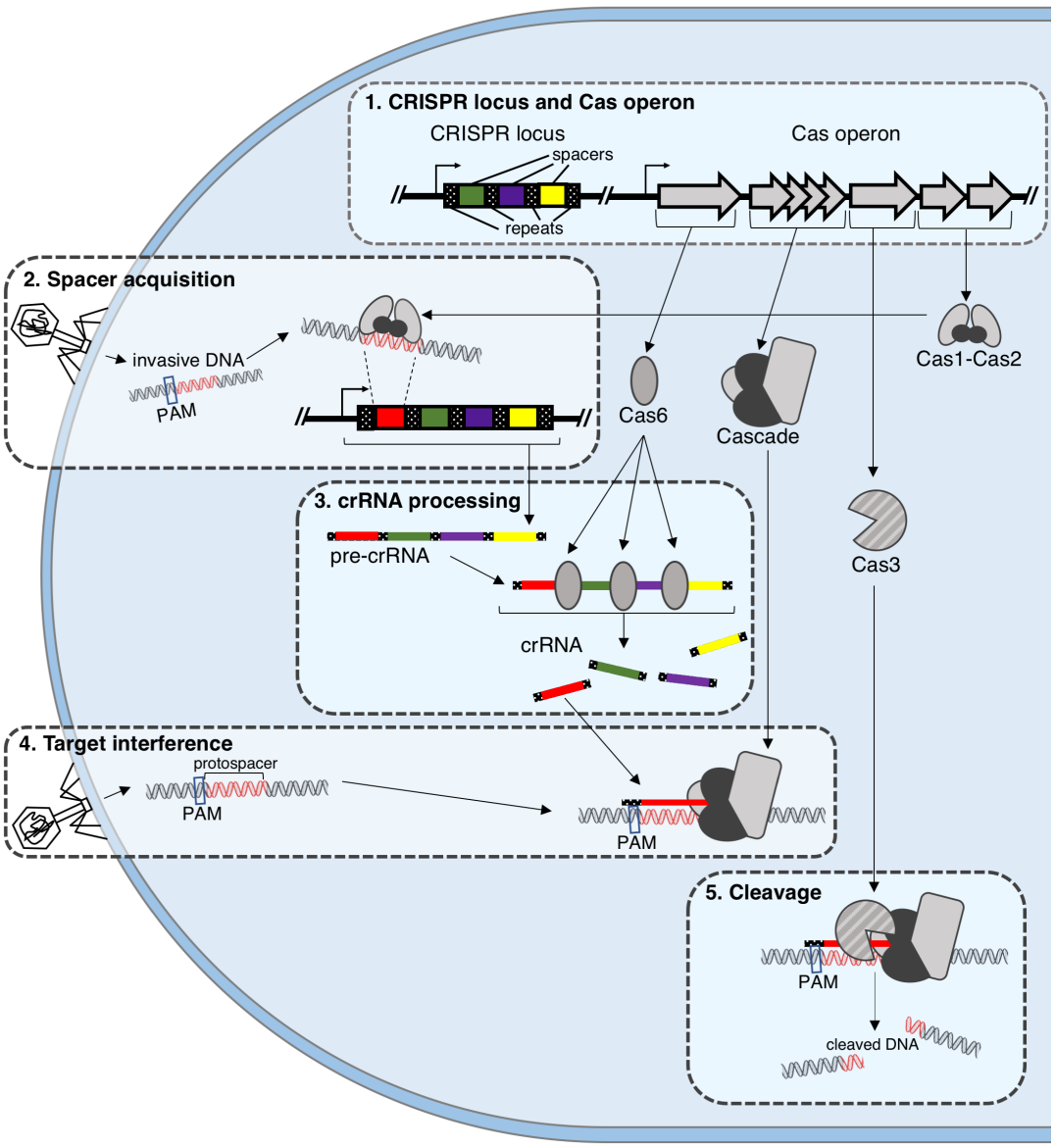


Figure 1.14: Mechanism of action for Type I CRISPR systems
1) CRISPR loci contain repeat regions interspaced with foreign DNA-derived spacers. The Cas operons of Type I systems contains several Cas genes, including the readily conserved *cas1* and *cas2*, their hallmark nuclease encoding *cas3* gene, and genes encoding the "Cascade" complex. 2) Spacer acquisition is governed by a Cas1-Cas2 complex recognising foreign DNA and integrating it into a CRISPR locus in the form of a spacer sequence. 3) Transcription of CRISPR loci results in pre-crRNA (pre-CRISPR RNA), which, through the action of Cas6, is processed into active crRNA. 4) Once processed, crRNA is able to complex with Cascade, guiding it towards its designated target sequence. 5) Interaction of Cascade with the target site enables recruitment of Cas3, resulting in targeted cleave of foreign DNA. PAM (Protospacer Adjacent Motif) is a small sequence found proximal to the protospacer and if required for spacer integration and target interference (see Chapter 5).

Both Type I and Type III Cascade complexes contain a single subunit of Cas5, an RNA binding protein responsible for binding to the 5' handle of crRNA (Makarova *et al.*, 2015). Cas5 interacts with the large subunit of the effector complex (Cas8 in Type I systems, and Cas10 in Type III systems (Cass *et al.*, 2015; Hrle *et al.*, 2013)). The final important conserved component of the Cascade complexes is the multiple Cas7 subunits that function as "rulers"; they bind the spacer region of crRNA and form a chain of monomers, the length of which depends on spacer length (Makarova *et al.*, 2015).

In Type III systems, as well as being the large subunit of the effector system, Cas10 acts as the nucleolytic component through its HD- and cyclase domains. However, Type I systems require recruitment of Cas3 for nucleolytic activity. This recruitment occurs upon the effector complex (often referred to as "Cascade") identifying the spacer encoding protein sequence flanked by the correct PAM. Upon recognition and interaction with the PAM, an R-loop structure is formed that is necessary for Cas3 interaction (Huo *et al.*, 2014; Sinkunas *et al.*, 2011). As well as DNA targeting, ssRNA targeting has been shown in Type III systems through the activity of Cas7 family proteins (Hrle *et al.*, 2013).

1.5.4.1. Type IV Cas systems

Several bacterial genomes have been shown to contain putative Type IV systems, all of which are functionally uncharacterised, lack *cas1* and *cas2*, and appear to also lack CRISPR loci. It has been hypothesised that these systems act as mobile modules (due to often being found on plasmids), that are able to use crRNA from other CRISPR systems (Makarova *et al.*, 2015).

1.5.5. Class 2 CRISPR systems

Unlike their multi-domain Class 1 counterparts, effector modules of Class 2 systems are encoding by a single gene, notably Cas9 and Cpf1 that belong to Type II and Type V systems respectively (Heidrich & Vogel, 2013). Other tentative effectors have been identified for Type V systems (C2c4, C2c8, C2c5, C2c10, C2c9), as well as Cas13 effectors for the newly classified Type VI systems, however these systems lay beyond the remit of this introductory section (Shmakov *et al.*, 2017; Smargon *et al.*, 2017).

1.5.5.1. Type II CRISPR systems

As well as crRNA, Type II systems require a second RNA component, tracrRNA (*trans*-acting CRISPR RNA), for effective targeting. This accessory RNA molecule is partially complementary to crRNA repeat regions, and is required for processing pre-crRNA to form mature crRNA. Once guided to the target site by mature crRNA, Cas9 is able to introduce a double-strand break (DSB) through activity of RuvC and HNH, each cleaving one strand of DNA (Makarova *et al.*, 2011, 2015).

1.5.5.1. Type V CRISPR systems

Unlike Type II systems, Type V systems do not require tracrRNA. Instead crRNA processing is completely governed by Cpf1. This large protein contains a RuvC-like nuclease domain homologous to that of Cas9, but lacks the HNH domain described for Cas9 above. As shown in experiments whereby the HNH domain of Cas9 is inactivated, the presence of only the RuvC domain results in nicking of target DNA (Shmakov *et al.*, 2017).

1.5.6. Genome editing with CRISPR systems

The adaptive nature of CRISPR-Cas systems is key to their application in genetic engineering. As providing a new target for the system only requires the incorporation of a new spacer, it is possible to provide the system with an artificial spacer targeting any gene of interest. Adding to this, CRISPR systems function with multiple spacers, meaning modified systems can offer fast and accurate targeting of many genes simultaneously. The benefits of this technology are clear and its potential has been realised in the successful manipulation of the genomes of many eukaryotic organisms. However, until recently, such success has not been mirrored in the bacterial taxa, despite CRISPR systems being native to prokaryotes.

Modified CRISPR-Cas systems can disrupt target genes through introduction of double strand breaks (DSBs), or single strand nicks. This technology has been widely used in eukaryotes for disruption of target genes, forcing them to use error-prone non-homologues end joining (NHEJ) to repair their chromosome. The error-prone nature of NHEJ results in INDELs (insertional or deletional mutations) and therefore inactivates the target gene (Mladenov & Iliakis, 2011). Generally, the NHEJ approach is not suitable for use in prokaryotes (due to lack of sophisticated DNA repair mechanisms), however as discussed in section 1.5.6, homology direct repair (HDR) techniques can be used. These rely upon homologous recombination, with DSB or nicking to select against unmodified targets.

1.5.6.1. Cas9 genome editing

Examples of Cas9-mediated genome editing are wide ranging and easy to find in almost any taxa. Use of *S. pyogenes* Cas9 (SpCas9) can be seen in large-scale genome modifications in mice (Fujita & Fujii, 2013), insertion of heterologous genes in zebrafish (Auer *et al.*, 2014), and targeted mutations in wheat (Upadhyay *et al.*, 2013). Use of Cas9 has also expanded to development of multiple knockouts simultaneously through multiplex RNA design (Mali *et al.*, 2013b; Guo *et al.*, 2014; Jao, Wente & Chen, 2013). Aside from the *S. pyogenes* Type II CRISPR system, other systems have been adopted, with Cas9 proteins being used from bacteria including *Neisseria meningitides*, *Streptococcus thermophilus*, and *Treponema denticola* (Esvelt *et al.*, 2013).

As well as the "wild type" SpCas9 that offers DSB capabilities, modified "nicking" (nCas9) versions have also been developed. SpCas9 is a 1.4 kDa protein consisting of, amongst other features, two catalytic domains (RuvC and HNH). Each domain cleaves a separate DNA strand. For RuvC, four amino acid residues have been identified as essential (histidine 983, aspartate 986, aspartate 10, and glutamate 762). For HNH, mutation of either histidine 840, or asparagine 863, to an alanine residue has been shown sufficient for inactivation of nucleolytic activity (Nishimasu *et al.*, 2014). The result of mutation of either domain is a Cas9 variant only capable of nickase activity (cutting a single strand). Use of nCas9 to nick just one strand of DNA has applications in eukaryotic organisms where larger genomes mean more likelihood of off-target cutting, and here two sgRNAs are used to direct nCas9

to two separate targets of the gene of interest (Ran *et al.*, 2013a, 2013b; Mali *et al.*, 2013a). Use of both conventional Cas9, and nCas9 variants has been described in the literature for prokaryotes (Xu *et al.*, 2015b; Li *et al.*, 2016).

1.5.6.2. Cpf1 genome editing

Regarded as simpler systems due to lacking the requirement for tracrRNA, Type V systems such as that found in *Franscisella novicida* utilise Cpf1 as a nuclease. The nature of this simpler RNA system lends itself to significantly easier multiplex genome editing, as shown by the editing of four genes simultaneously in mammalian cells (Zetsche *et al.*, 2016).

As well as RNA simplicity, other potential benefits include a more A/T rich PAM and a different cleavage mechanism. The former could lend Cpf1 to open more targets in A/T rich organisms. With regards to the latter, unlike Cas9, Cpf1 does not cut its target site. Instead, it cuts outside of its RNA-specified target. This crucial difference lends itself to more accurate HDR-mediated genome editing in eukaryotes, as it drastically reduced the chance of NHEJ, meaning (as in most prokaryotes), only HDR mutants can survive (Fagerlund, Staals & Fineran, 2015).

1.5.6.3. Class 1 genome editing

Due to their effector complexes requiring multiple genes, use of Class 1 CRISPR systems has limited appeal in eukaryotes. In contrast, due to their presence in prokaryotes, interest in using endogenous systems for genome editing has garnered much interest. Section 1.5.8 further discusses the use of endogenous CRISPR systems in genome editing.

1.5.7. Transcription regulation with CRISPR systems

As seen with the creation of nCas9 for nickase activity, the *cas9* gene can be modified to change the behaviour of the Cas9 protein. In what has proven a success in both eukaryotes and prokaryotes, Cas9 endonuclease activity can be removed altogether, providing a platform for gene regulation. This catalytic inactivation is achieved through introducing mutations in both the HNH and RuvC domains, creating a “dead” Cas9 (dCas9).

Use of dCas9 in transcriptional downregulation (knockdown) has been dubbed CRISPRi (CRISPR interference). Like RNAi (RNA interference), CRISPRi reduces the overall translation of mRNA into protein, but the way this is achieves differs significantly. Whereas RNAi utilises complementary siRNA to complex Argonaut proteins and cleave mRNA, CRISPRi uses dCas9 to physically interfere with RNA polymerase, restricting transcription. Like with conventional CRISPR-Cas9 approaches, guide RNA targets dCas9 to a specified target site (promoter region or 5’ end of a gene), where it complexes but does not cleave. CRISPRi has seen multiple applications since its first published use (Larson *et al.*, 2013), notably in prokaryotes where, due to lack of necessary molecular machinery, RNAi is not possible. Use of Class 2 CRISPR systems for Cascade-based knock down has also been shown in bacteria, as detailed in section 5.1.

1.5.8. Current progress in prokaryotic genome editing

At the start of this project, little ground had been made in adapting CRISPR-Cas systems for use in prokaryotes. Despite CRISPR systems being native to prokaryotes, progress lagged compared to eukaryotes. However, as described below, significant progress has been made within the last few years.

1.5.8.1. Teething problems

Much of the reason for the initial success in CRISPR-based technology in eukaryotes not being replicated in prokaryotes is attributed to NHEJ. The prevalence of NHEJ in eukaryotes allowed for a simplified vector construction and lack of reliance on HDR meant higher efficiencies. Early experiments attempting to repeat the success of Cas9-mediated cleavage in eukaryotes invariably resulted in cell death in prokaryotes as noted by Bikard and colleagues (2012). Here it was observed that the only *E. coli* colonies present on selection plates were those containing mutations in their PAM. Although initially these results may have seemed disappointing, they illustrate clearly the utility of CRISPR-Cas technology in prokaryotes; a powerful counter-selection mechanism not reliant on antibiotics. Full utilisation of this technology has enabled development of tools that can attain fast and efficient genome editing suitable for the iterative nature of synthetic biology.

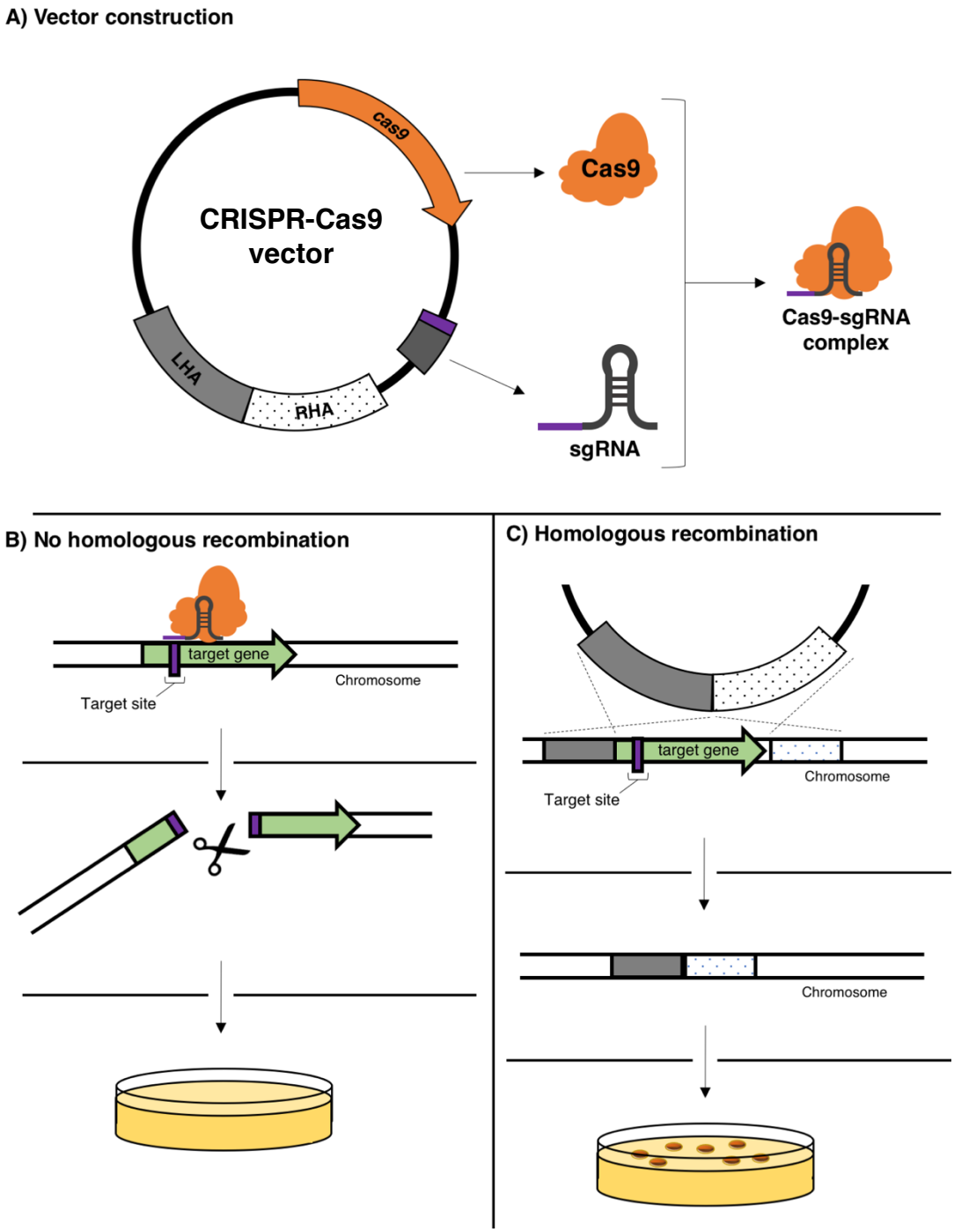


Figure 1.15: CRISPR-Cas9 genome editing in prokaryotes
A) Vectors contain *cas9* (orange), a guide RNA modified for targeting gene of interest (dark grey/purple), and homology arms designed flanking said gene of interest (light grey/white with black spots). B) If no homologous recombination event occurs, Cas9 is guided to the target site within the gene of interest where it introduces a lethal double-strand break (DSB). Screening based on antibiotic selection results in no colonies. C) Homologous recombination occurs between the plasmid homology arms and the chromosome, resulting in excision of the target gene. Screening based on antibiotic selection results in colonies with desired mutation, Cas9 is not lethal as the target has been removed from the genome and therefore no DSB occurs.

1.5.8.2. Successful genome editing in prokaryotes

Jiang *et al*. (2013) designed the first successful exhibition of genetic engineering in prokaryotes using CRISPR-Cas9. Here they demonstrated that using CRISPRs for HDR in bacteria with naturally high recombination frequencies was possible, as well as showing that combining recombineering and CRISPR technologies could enable HDR in bacteria with low recombination frequencies. Instead of combating the fact that DSBs prevent conventional CRISPR-Cas editing by killing cells, Jiang *et al*. incorporated this lethal capability to use it to drive selection of homologous recombination. Figure 1.15 illustrates how CRISPR-Cas9 selection works in for screening homologous recombination events in prokaryotes. Unlike the study by Jiang *et al*. (2013), here homology arms are provided on the same plasmid as Cas9 and the guide RNA.

S. pyogenes is naturally highly recombinogenic, however introduction of oligonucleotides with homology arms to a gene of interest results in relatively low levels of homologous recombination. Jiang *et al*. (2013) demonstrated that by providing CRISPR-Cas mediated targeting of untransformed strains, it is possible to drive the rate of homologous recombination up. Unlike *S. pyogenes*, wild type *E. coli* is not recombinogenic. However, introduction of λ-Red recombineering genes into *E. coli*, such as *E. coli* HME63, enables effective homologous recombination to take place. These two early examples highlight the key bottleneck to use of CRISPR-based genome editing in prokaryotes; recombination frequency.

1.5.9. Considerations for CRISPR-mediated genome editing in acetogens

For synthetic biology approaches, a chassis organism is selected based on desirable characteristics complementing the end goal, in the case of this project, *C. autoethanogenum* was used as a chassis due to its natural solvent profile when grown on C1 gas. Once selected, this chassis is taken forward through iterative steps of genome editing and strain analysis. In order to achieve efficient genome editing with CRISPR-based technology, as stated above, sufficient recombination frequency is required. However, as is the case with *C. autoethanogenum,* the chassis may not have high recombination frequencies. To adapt this technology to non-recombinogenic species, researchers have the choice of several approaches; A) recombineering, B) increasing transformation frequency, or, C), controlling the nucleolytic ability of Cas9.

With regards to option A, as outlined in section 1.4, a recombineering approach offers significant advantages, but is limited by host specificity. Therefore, this approach is restricted to species with compatible recombineering systems developed. Options B and C both contribute not by directly increasing the frequency of recombination, but increasing the likelihood of recombination occurring. Although this difference may seem subtle it is important; if a recombination event has an approximate likelihood of 1 in 10,000, a transformation frequency of 10,000 CFU per transformation would be required to see one (assuming Cas9 cleavage is 100% efficient). One solution is option B; increasing the transformation frequency to surpass

recombination frequency. As discussed in section 1.3.3, there are several approaches to increasing DNA transfer. Assuming every viable cell that is successfully transformed goes on to make a colony, selection for the presence of Cas9 will cause cell death in any un-modified cell. If the likelihood of recombination is 1 in 10,000, and the CFU of the transformation is above this number, in theory at least one mutant should be observed.

Finally, option C represents a more intensive approach, but one that may reap more rewards. This approach itself can either decrease the efficiency of Cas9, or place it under the control of an inducible system. Both approaches allow more opportunity for homologous recombination to occur. Successful implementation of both these techniques are further discussed in section 1.5.6.

1.6. Project aims

The aim of this project was to develop and implement genetic tools for use in acetogens, specifically *C. autoethanogenum*, with an end goal of using these tools to better understand and improve solvent production. These goals can be divided into three major objectives:

1. Optimise DNA transfer methods and modify the *C. autoethanogenum* endogenous restriction-methylation systems to enable sufficient DNA transfer to enable CRISPR-based approaches to genome editing.

2. Using CRISPR-Cas9 technology, design and implement a modular plasmid system for efficient, in-frame knockouts of target genes in *C. autoethanogenum*. Using the modular nature of this system, apply it for genome editing applications in other members of the genus.

3. Analyse the endogenous CRISPR system of *C. autoethanogenum*, and identify its PAM. Once identified, use this PAM to enable modification of the CRISPR system for use in genome editing and knockdown applications.

4. Using tools developed in objectives 2 and 3, create mutants in the *ack* and *pta* genes involved in acetogenesis, as well as other relevant targets. Analyse and compare the mutant strains to wild type to assess the impact on solvent production.

CHAPTER TWO

METHODS AND MATERIALS

2. Methods and materials

2.1. Strains and vectors

Table 2.1: Bacterial strains used in this study

Strain	Description/Genotype	Source/Reference
***Clostridium autoethanogenum* JA1-1 DSM 10061**	Wild type	(Abrini, Naveau & Nyns, 1994)
***Clostridium autoethanogenum* JA1-1 DSM 10061 Col3**	Isolated from steady-state fermenter	(Henstra, 2016)
***Clostridium autoethanogenum* Col3, Δ*caulI* mutant**	Col3 strain with in-frame *caulI* knockout	This study
***Clostridium autoethanogenum* Col3, Δ*ack* mutant**	Col3 strain with in-frame *ack* knockout	This study
***Clostridium autoethanogenum* Col3, Δpta mutant**	Col3 strain with in-frame *pta* knockout	This study
***Clostridium autoethanogenum* Col3, Δ*cas3* mutant**	Col3 strain with in-frame *cas3* knockout	This study
***Clostridium autoethanogenum* Col3, eCas strain**	Col3 strain modified to have CRIPSR locus A and Cas operon under control of strong constitutive promoter, and modified CRISPR locus A containing truncated, synthetic spacer region	This study
***Clostridium autoethanogenum* Col3, eCas strain, Δ*cas3* mutant**	Col3 Δ*cas3* mutant strain modified to have CRIPSR locus A and Cas operon under control of strong constitutive promoter, and modified CRISPR locus A containing truncated, synthetic spacer region	This study
***Clostridium ljungdahlii* ATCC 49587**	Wild type	(Tanner, Miller & Yang, 1993)
***Clostridium ljungdahlii* Δ*pta* mutant**	Wild type *C. ljungdahlii* with in-frame *pta* knockout	This study, (Huang *et al.*, 2016)
***Eubacterium limosum* ATCC 8486**	Wild type	(Genthner, Davis & Bryant, 1981)
***Eubacterium limosum* ATCC 8486 Δ*cas3* mutant**	Wild type *E. limosum* with in-frame *cas3* knockout	This study
Geobacillus thermoglucosidasius	Wild type	(Suzuki *et al.*, 1983)

NCIMB 11955		
***Geobacillus thermoglucosidasius* NCIMB 11955 Δack**	*G. thermoglucosidasius* wild type strain with in-frame *ack* knockout	This study
***Clostridium butyricum* ATCC 19398**	Wild type	(Cummins & Johnson, 1971)
***Clostridium aceticum* ATCC 35044**	Wild type	(Braun, Mayer & Gottschalk, 1981)
***Escherichia coli* NEB 5α**	K12 strain of *E. coli* with phenotype: uA2 Δ(argF-*lacZ*)U169 *phoA glnV*44 Φ 80 Δ(*lacZ*)M15 *gyrA*96 r*ecA1 relA1 endA1* thi-1 *hsdR*17	NEB
***Escherichia coli* CA434**	Conjugative transfer strain of E. coli HB101 (thi-1 hsdS20 (r-B, m-B) supE44 recAB ara-14 leuB5proA2 lacY1 galK rpsL20 (*strR*) xyl-5 mtl-1) harbouring R702 (Tra+, Mob+ conjugative plasmid)	(Purdy *et al.*, 2002)
***Escherichia coli* Sexpress**	Conjugative transfer strain derived from NEB Express (fhuA2 [lon] ompT gal sulA11 R(mcr-73::miniTn10--TetS)2 [dcm] R(zgb-210::Tn10--TetS) endA1 Δ(mcrC-mrr)114::IS10) harbouring R702 (Tra+, Mob+ conjugative plasmid)	(Woods, 2017)
***Escherichia coli* BL21**	B strain of E. coli fhuA2 [lon] ompT gal [dcm] ΔhsdS	NEB

Table 2.2: Plasmids used in this study

Plasmid	Description/Genotype	Source
pMTL81151	pMTL80000 modular vector with pCB101 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	Generated from pMTL83151 as part of this study
pMTL82151	pMTL80000 modular vector with pBP1 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	(Heap *et al.*, 2009)
pMTL83151	pMTL80000 modular vector with pCB102 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	(Heap *et al.*, 2009)
pMTL84151	pMTL80000 modular vector with pCD6 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	(Heap *et al.*, 2009)
pMTL86151	pMTL80000 modular vector with pIM13 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	(Heap *et al.*, 2009)
pMTL87151	pMTL80000 modular vector with p19 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	(Millard, 2017)
pMTL88151	pMTL80000 modular vector with pCACET5 Gram-positive replicon, CatP resistance marker, and ColE1+*tra* Gram-negative replicon.	Generated from pMTL83151 as part of this study
pMTLcas-pta	pMTL83151 backbone containing *S. pyogenes* Cas9/sgRNA driven by divergent promoters (*C. acetobutylicum* P_{thl} and P_{araE} respectfully), with specific sgRNA seed region and homology arms to delete *C. ljungdahlii pta*	This study
pMTL43151-CLAU-pyrE	Modular CRISR vector containing trCas9 (see Chapter 4) driven by *C. acetobutylicum* P_{thl} and sgRNA driven by P_{araE}. Specific sgRNA seed region and homology arms to delete *C. autoethanogenum pyrE*	This study
pMTL43151-CLAU-IIG	pMTL43151-CLAU-pyrE with sgRNA seed and homology arms swapped to target *C. autoethanogenum cauII*	This study
pMTL43151-CLAU-pta	pMTL43151-CLAU-pyrE with sgRNA seed and homology arms swapped to target *C. autoethanogenum pta*, replacing it with Bookmark (BM1) sequence	This study
pMTL43151-CLAU-ack	pMTL43151-CLAU-pyrE with sgRNA seed and homology arms swapped to target *C. autoethanogenum ack*, replacing it with Bookmark (BM1) sequence	This study
pMTL43151-CLAU-cas3	pMTL43151-CLAU-pyrE with sgRNA seed and homology arms swapped to target *C. autoethanogenum cas3*	This study
pMTL43151-CLAU-eCas	pMTL43151-CLAU-pyrE with sgRNA seed and modified to swap *C. autoethanogenum* CRISPR locus 2 and downstream region for a synthetic CRISPR locus and Cas operon promoter	This study

pMTL43151-BM-CLAU-ack-repair	pMTL43151-CLAU-pyrE with sgRNA seed swapped to target BM1sequence at the *ack* locus of the Δ*ack* strain of *C. autoethanogenum*, and homology arms designed to complement the *ack* deletion	This study
pMTL43151-BM-CLAU-pta-repair	pMTL43151-BM-CLAU-ack-repair with same sgRNA seed to target BM1sequence at the *pta* locus of the Δ*pta* strain of *C. autoethanogenum*, and homology arms designed to complement the *ack* deletion	This study
pMTL43151-CLAU-pyrE-BM4-12	pMTL43151-CLAU-pyrE vector with Bookmark 4-12 sequences inserted between homology arms	This study
pMTL43151-CLAUc3-SNP1	pMTL43151-CLAU-pyrE with sgRNA seed and modified to swap *C. autoethanogenum* change the SNP present at the "SNP1" locus of *C. autoethanogenum*, reverting genotype to wild type *C. autoethanogenum*	This study
pMTL43151-CLAUc3-SNP2	pMTL43151-CLAU-pyrE with sgRNA seed and modified to swap *C. autoethanogenum* change the SNP present at the "SNP2" locus of *C. autoethanogenum*, reverting genotype to wild type *C. autoethanogenum*	This study
pMTL43151-CLAUc3-SNP3	pMTL43151-CLAU-pyrE with sgRNA seed and modified to swap *C. autoethanogenum* change the SNP present at the "SNP3" locus of *C. autoethanogenum*, reverting genotype to wild type *C. autoethanogenum*	This study
pMTL43151-St1-Geo-ack	pMTL43151-CLAU-pyrE with *S. pyogenes cas9* and sgRNA swapped for St1Cas9, and St1 sgRNA and homoloy arms for targeting *G. thermoglucosidasius ack*.	This study
pMTL43151-St3-Geo-ack	pMTL43151-CLAU-pyrE with *S. pyogenes cas9* and sgRNA swapped for St3Cas9, and St3 sgRNA and homoloy arms for targeting *G. thermoglucosidasius ack*.	This study
pMTL43151-ELM-cas3	pMTL43151-CLAU-pyrE with sgRNA seed and homology arms swapped to target *E. limosum cas3*	This study
pMTL007C-E2::CLAU-pta	ClosTron vector for targeted intron disruption of *C. autoethanogenum pta*	Synthesised by DNA 2.0, designed as details by Heap *et al*. 2010
pPR1-CLAU-del-pta	Allelic exchange vector designed for in-frame deletion of *C. autoethanogenum pta*	This study
pMTL82254	Vector backbone for assessing *catP* expression, consisting of pMTL82251 backbone, and application specific module containing *catP* driven by *C. acetobutylicum* P_{thl} with native RBS	(Zhang *et al*., 2016)
pMTL82254-RBS1	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS1"	This study
pMTL82254-RBS2	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS2"	This study

pMTL82254-RBS3	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS3"	This study
pMTL82254-RBS4	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS4"	This study
pMTL82254-RBS5	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS5"	This study
pMTL82254-RBS6	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS6"	This study
pMTL82254-RBS7	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS7"	This study
pMTL82254-RBS8	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS8"	This study
pMTL82254-RBS9	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS9"	This study
pMTL82254-RBS10	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS10"	This study
pMTL82254-RBS11	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS11"	This study
pMTL82254-RBS12	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS12"	This study
pMTL82254-RBS13	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS13"	This study
pMTL82254-RBS14	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS14"	This study
pMTL82254-RBS15	pMTL82254 modified to exchange native RBS for *C. autoethanogenum* "RBS15"	This study
pMTL82254-trRBS	pMTL82254 modified to exchange native RBS for the putative truncated section of trCas9	This study
pMTL82254-duelRBS	pMTL82254 modified to include the RBS for the putative truncated section of trCas9 between the native P_{thl} RBS and *catP.*	This study
pMTL86151-1.27-5'	pMTL86151 backbone containing CLAU spacer 1.27 and 5 nt sequence found immediately upstream of putative-protospacer sequence in *C. ljungdahlii*	This study
pMTL86151-1.27-3' (pMTL86151-1.27-ATTAA)	pMTL86151 backbone containing *C. autoethanogenum* spacer 1.27 and 5 nt sequence found immediately downstream of the same sequence in *C. ljungdahlii*	This study

pMTL86151-1.27-5'+3'	pMTL86151 backbone containing *C. autoethanogenum* spacer 1.27 and 5 nt sequences found immediately up/downstream of the same sequence in *C. ljungdahlii*	This study
pMTL86151-1.27-5'-noPAM	pMTL86151 backbone containing *C. autoethanogenum* spacer 1.27	This study
pMTL86151-1.27-TTTAA	pMTL86151-1.27-3' with 3' region modified to "TTTAA"	This study
pMTL86151-1.27-CTTAA	pMTL86151-1.27-3' with 3' region modified to "CTTAA"	This study
pMTL86151-1.27-GTTAA	pMTL86151-1.27-3' with 3' region modified to "GTTAA"	This study
pMTL86151-1.27-AATAA	pMTL86151-1.27-3' with 3' region modified to "AATAA"	This study
pMTL86151-1.27-ACTAA	pMTL86151-1.27-3' with 3' region modified to "ACTAA"	This study
pMTL86151-1.27-AGTAA	pMTL86151-1.27-3' with 3' region modified to "AGTAA"	This study
pMTL86151-1.27-ATAAA	pMTL86151-1.27-3' with 3' region modified to "ATAAA"	This study
pMTL86151-1.27-ATCAA	pMTL86151-1.27-3' with 3' region modified to "ATCAA"	This study
pMTL86151-1.27-ATGAA	pMTL86151-1.27-3' with 3' region modified to "ATGAA"	This study
pMTL86151-1.27-ATTTA	pMTL86151-1.27-3' with 3' region modified to "ATTTA"	This study
pMTL86151-1.27-ATTCA	pMTL86151-1.27-3' with 3' region modified to "ATTCA"	This study
pMTL86151-1.27-ATTGA	pMTL86151-1.27-3' with 3' region modified to "ATTGA"	This study
pMTL86151-1.27-ATTAT	pMTL86151-1.27-3' with 3' region modified to "ATTAT"	This study
pMTL86151-1.27-ATTAC	pMTL86151-1.27-3' with 3' region modified to "ATTAC"	This study
pMTL86151-1.27-ATTAG	pMTL86151-1.27-3' with 3' region modified to "ATTAG"	This study
pACYC-duet	Expression vector with p15A Gram-negative replicon and chloramphenicol resistance marker. Contains two T7 promoter driven multiple cloning sites. Designed for compatibility with pET-duet.	Sigma Aldrich

pET-duet	Expression vector with ColE1 Gram negative replicon and ampicillin resistance marker. Contains two T7 promoter driven multiple cloning sites. Designed for compatibility with pET-duet.	Sigma Aldrich
pACY-CLAU-CRISPR	pACYC-duet vector modified to include the Cas6, Cascade, and Cas3 component of the *C. autoethanogenum* CRISPR system, as well as a synthetic CRISPR locus targeting spacer 1.27	This study
pET-GFP-target	pET-duet vector modified to include GFP, and spacer 1.27 flanked by degenerate 6 nt regions.	This study
ELM-CR2.12-5'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 2.12 and 5 nt sequence found immediately upstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR2.12-3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 2.12 and 3 nt sequence found immediately downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR2.12-5'+3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 2.12 and 5 nt sequence found immediately up/downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR2.12-noPAM	pMTL86151 backbone containing *E. limosum* ATCC 8486 spacer 2.12	This study
ELM-CR3.11-5'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.11 and 5 nt sequence found immediately upstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.11-3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.11 and 3 nt sequence found immediately downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.11-5'+3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.11 and 5 nt sequence found immediately up/downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.11-noPAM	pMTL86151 backbone containing *E. limosum* ATCC 8486 spacer 3.11	This study
ELM-CR3.28-5'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.28 and 5 nt sequence found immediately upstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.28-3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.28 and 3 nt sequence found immediately downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.28-5'+3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.28 and 5 nt sequence found immediately up/downstream of the same sequence in *E. limosum* KIST 618	This study

ELM-CR3.28-noPAM	pMTL86151 backbone containing *E. limosum* ATCC 8486 spacer 3.28	This study
ELM-CR3.29-5'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.29 and 5 nt sequence found immediately upstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.29-3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.29 and 3 nt sequence found immediately downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.29-5'+3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.29 and 5 nt sequence found immediately up/downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.29-noPAM	pMTL86151 backbone containing *E. limosum* ATCC 8486 spacer 3.29	This study
ELM-CR3.30-5'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.30 and 5 nt sequence found immediately upstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.30-3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.30 and 3 nt sequence found immediately downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.30-5'+3'	pMTL87151 backbone containing *E. limosum* ATCC 8486 spacer 3.30 and 5 nt sequence found immediately up/downstream of the same sequence in *E. limosum* KIST 618	This study
ELM-CR3.30-noPAM	pMTL86151 backbone containing *E. limosum* ATCC 8486 spacer 3.30	This study

2.2. Primers and oligonucleotides

Table 2.3: Oligonucleotides used in this study

Oligonucleotide	Sequence	Description
CACET-F	ATATATATGGCGCGCCGTGAAACATTTAATAGTAACGACGTCGTTAC	Forward primer for amplification of *C. aceticum* pCACET plasmid replicon
CACET-R	ATATATATGGCCGGCCGCACGAAAAATAGCTAATGCACTCG	Reverse primer for amplification of *C. aceticum* pCACET plasmid replicon
CB101-F	ATATATATGGCGCGCCCCAGTACTATCACAATTCGCAC	Forward primer for amplification of *C. butyricum* pCB101 plasmid replicon
CB101-R	ATATATATGGCCGGCCAAAGAGAGCAATTAAGCCCTC	Reverse primer for amplification of *C. butyricum* pCB101 plasmid replicon
CatP-R1	CAAGTTTATCGCTCTAATGAAC	Screening primer for sequencing of Gram-positive replicons in pMTL80000 vectors
pCB102-R1	CTGTTATGCCTTTTGACTATC	Screening primer for sequencing the applications specific module in pMTL80000 vectors containing the pCB102 Gram-positive replicon
Cas9F1-PthlOH	AATTTAGGAGGTTAGTTAGAATGGATAAGAAATACTCAATAGGCTTAG	Forward primer for amplification of SpCas9 for cloning into pMTL43151 with overhang for SOEing PCR fusion to *C. acetobutyiicum* P_{thl}
Cas9R1-NotI	ATATATATGCGGCCGCTCAGTCACCTCCTAGCTGACTC	Reverse primer for amplification of *S. pyogenes* Cas9 with *Not*I in overhang for cloning into pMTL43151
PthlF-XbaI	ATATATTCTAGATTTTTAACAAAATATATTGATAAAAATAATAATAGTG	Forward primer for amplification of *C. acetobutylicum* P_{thl} with XbaI in overhang for cloning into pMTL43151
PthlR-Cas9OH	TCTAACTAACCTCCTAAATTTTGATACGGGGTAACAGATAAAC	Reverse primer for amplification of *C. acetobutylicum* P_{thl} with XbaI in overhang for SOEing PCR fusion to SpCas9
Cas9-scr-F1	ATGGATAAGAAATACTCAATAGGCTTAG	Forward sequencing primer for SpCas9
Cas9-scr-F2	GCTTTGTCATTGGGTTTGAC	Forward sequencing primer for SpCas9
Cas9-scr-F3	GTCGATAAAGGTGCTTCAGC	Forward sequencing primer for SpCas9
Cas9-scr-F4	GAACATATTGCAAATTTAGCTGG	Forward sequencing primer for SpCas9

Cas9-scr-F5	CTGACTTCCGAAAAGATTTCC	Forward sequencing primer for SpCas9
Cas9-scr-F6	GAGTTAGAAAACGGTCGTAAACG	Forward sequencing primer for SpCas9
Cas9-scr-R1	GAAACTTAATCATATGCGCTAAGG	Reverse sequencing primer for SpCas9
sgRNA-F-XbaI	ATATATTCTAGATTTATATTTAGTCCCTTGCC	Forward primer for amplification of pMTL43151 sgRNA module with XbaI in overhang
sgRNA-pyrE-sg3-SOE-R	TTTGCCTTGCAAGTTCATAGGTCGACTCTAGAGGATCCC	Reverse internal primer for amplification of sgRNA module with overhang encoding pyrE seed region to allow SOEing PCR
sgRNA-pyrE-sg3-SOE-F	CTATGAACTTGCAAGGCAAAGTTTTAGAGCTAGAAATAGCAAG	Forward internal primer for amplification of sgRNA module with overhang encoding pyrE seed region to allow SOEing PCR
sgRNA-R-MauBI-AscI	ATATATCTCGAGATATATCGCGCGCGATAAAAATAAGAAGCCTGCAAATG	Reverse primer for amplification of pMTL43151 sgRNA module with *Mau*BI *Asc*I in overhang for subsequent homology arm introduction
sgRNA-R-AsiSI-AscI	ATATATCTCGAGATATATCGCGCGCGATAAAAATAAGAAGCCTGCAAATG	Reverse primer for amplification of pMTL43151 sgRNA module with *Asi*SI *Asc*I in overhang for subsequent homology arm introduction
pMTL43151-sgRNA-scr-R	AGTTCCATAGTTTATCCC	Screening primer for Sanger sequencing of pMTL43151 sgRNA region
CLAU-pyrE-LHA-F-MauBI	ATATATATCGCGCGCGTAAATTTAGGTGAAGAACACCATACATATCC	Forward primer for amplification of CLAU *pyrE* LHA with *Mau*BI site in homology arm for cloning into pMTL43151 vector
CLAU-pyrE-LHA-R	ATAATTTAGGAGGACAGTTATGGATAAATAAGTCGAAAAAATCAATGCACGATGCAGAATTGAC	Reverse primer for amplification of CLAU *pyrE* LHA with overhang to enable SOEing PCR
CLAU-pyrE-RHA-F	GTGCATTGATTTTTTCGACTTATTTATCCATAACTGTCCTCCTAAATTATTCCTCTTATTTCAC	Forward primer for amplification of CLAU *pyrE* RHA with overhang to enable SOEing PCR
CLAU-pyrE-RHA-R-AscI	ATATATATGGCGCGCCATATAGTACTGTAATTTTTAAAAAGTTTGAAATAACTAGTTAG	Reverse primer for amplification of CLAU *pyrE* RHA with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-pyrE-scr-F	GGAGTTGACTTACATATGC	Forward screening primer for the CLAU *pyrE* locus
CLAU-pyrE-scr-R	GATGGACCAGTTTTCTTAGG	Reverse screening primer for the CLAU *pyrE* locus
CLAU-IIG-LHA-F-MauBI	ATATATGCGCGCGCGGTGAAAGTATCTTTAATGTTG	Forward primer for amplification of CLAU *cauII* LHA with *Mau*BI site OH

CLAU-IIG-LHA-R	ATAGACTAGCACATTTATTGTGATTTTTATTTTAAGGAGAAAAGTATGG	Reverse primer for amplification of CLAU *caulI* LHA with overhang to enable SOEing PCR
CLAU-IIG-RHA-F	TCTCCTTAAAATAAAAATCACAATAAATGTGCTAGTCTATTTTC	Forward primer for amplification of CLAU *caulI* RHA with overhang to enable SOEing PCR
CLAU-IIG-RHA-R-AscI	ATATATGGCGCGCCAACCTGATATGACTATGAAGG	Reverse primer for amplification of CLAU *caulI* RHA with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-IIG-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACGTAAACAGCTGTGTTGCACAGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting *caulI*
CLAU-IIG-scr-F	TTTTGTAATTCCTGTGGCAGG	Forward screening primer for the CLAU *caulI* locus
CLAU-IIG-scr-R	ACTTGGGGCAGTTATGGATG	Reverse screening primer for the CLAU *caulI* locus
CLAU-pta-LHA-F-AsiSI	ATATATATGCGATCGCATTGGGTAGGCTCTTCTGC	Forward primer for amplification of CLAU *pta* left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-pta-LHA-R-BM1-OH	TCCTCTGAGGAATCCACCCAGCATCATGTTCATTTCCTCCCTTT	Reverse primer for amplification of CLAU *pta* left homology arm with overhang to enable SOEing PCR
CLAU-pta-RHA-F-BM1-OH	ATGCTGGGTGGATTCCTCAGAGGAAAGTAATAACAAAAAGCATAAATGATTC	Forward primer for amplification of CLAU *pta* right homology arm with overhang to enable SOEing PCR
CLAU-pta-RHA-R-AscI	ATATATATGGCGCGCCATCACCGCATCTAGTTCCC	Reverse primer for amplification of CLAU *pta* right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-pta-sg1-HiFi	TATAAATCTTAAGGAGGAGTTTTCGTCGACGTATTTCCAGATCTCCAAGCGTTTTAGAGCTAGAAATAGCAAGTTAAAAT	Oligo for HiFi introduction of sgRNA seed targeting *pta*
CLAU-pta-scr-F	GCAGAAGGATAGGACATACCC	Forward screening primer for the CLAU *pta* locus
CLAU-pta-scr-R	AACTTTCTTCTATGAGGAGGTCC	Reverse screening primer for the CLAU *pta* locus
CLAU-ack-LHA-F-AsiSI	ATATATGCGATCGCTTGTCTTAGCTGAAGGAGAAGAAG	Forward primer for amplification of CLAU *ack* left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-ack-LHA-R-BM1-OH	TCCTCTGAGGAATCCACCCAGCATTTTCATGTTTAATATTCCTCCTAAAAATG	Reverse primer for amplification of CLAU *ack* left homology arm with overhang to enable SOEing PCR
CLAU-ack-RHA-F-BM1-OH	ATGCTGGGTGGATTCCTCAGAGGAAAATAAGACTTGACTTTTAATGCTCATC	Forward primer for amplification of CLAU *ack* right homology arm with overhang to enable SOEing PCR

CLAU-ack-RHA-R-AscI	ATATATATGGCGCGCCGCTTTTTAACTGCTTCTACCAC	Reverse primer for amplification of CLAU *ack* right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-ack-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACGCTCTTGTAGATAAAAAACAGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting *ack*
CLAU-ack-scr-F	AAAAATTTGGAGTAAGGCAAAGG	Forward screening primer for the CLAU *ack* locus
CLAU-ack-scr-R	GCTTAAAATATCTATTTTGTCCTTAGG	Reverse screening primer for the CLAU *ack* locus
CLAU-cas3-LHA-F-AsiSI	ATATATATGCGATCGCGGATAAAAACTTCCACTATAGTATC	Forward primer for amplification of CLAU *cas3* left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-cas3-LHA-R	TTTAGAGGGTAATTTTTATGTAACTGTTAGGGGAGATG	Reverse primer for amplification of CLAU *cas3* left homology arm with overhang to enable SOEing PCR
CLAU-cas3-RHA-F	TACATCTCCCCTAACAGTTACATAAAAATTACCCTCTAAAAAAAGTAC	Forward primer for amplification of CLAU *cas3* right homology arm with overhang to enable SOEing PCR
CLAU-cas3-RHA-R-AscI	ATATATATGGCGCGCCCTTTCTGGAAAAACTGCG	Reverse primer for amplification of CLAU *cas3* right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-cas3-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACGATGACAATAAAATTGACAGGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting *cas3*
CLAU-cas3-scr-F	GGTAAAATGTACTTGAAGTTTTCA	Forward screening primer for the CLAU *cas3* locus
CLAU-cas3-scr-R	AAGGCAGGCTGTGAAAATG	Reverse screening primer for the CLAU *cas3* locus
CLAU-BM1-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACATGCTGGGTGGATTCCTCAGGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting BM1
CLAU-eCas-LHA-F-AsiSI	ATATATATGCGATCGCATTGAATCCATTCGTAATAAGGAC	Forward primer for amplification of CLAU *eCas* left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-eCas-LHA-R	ATTTGATTTATTTCCTCGAGGATTCAACTTCATTTTCATAAACTC	Reverse primer for amplification of CLAU *eCas* left homology arm with overhang to enable HiFi assembly with eCas synRNA
CLAU-eCas-RHA-F	AATGAAGTTGAATCCTCGAGGAAATAAATCAAATAAACATTAAAGTATACATTC	Forward primer for amplification of CLAU *eCas* right homology arm with overhang to enable HiFi assembly with eCas synRNA

CLAU-eCas-RHA-R-AscI	ATATATATGGCGCGCCTTTACAGATTTCATTTATCAAATATTTTCAC	Reverse primer for amplification of CLAU *eCas* right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-eCas-synRNA-F	AAAGTTATTTGAATTTAATGGCTTAGGGCACTTTTTTATTGAGTTTATG	Forward primer for amplification of CLAU-eCas-synRNA
CLAU-eCas-synRNA-R	TCTCTAGAATGTATACTTTAATGTTTATTTG	Reverse primer for amplification of CLAU-eCas-synRNA
CLAU-eCas-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACTGTAATATACAAATGAATGG GTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting *eCas*
CLAU-eCas-scr-F	GGTCGTTGATGGATTAGTAAATG	Forward screening primer for the CLAU *eCas* locus
CLAU-eCas-scr-R	TCCATTAATTTGCTTAGATCTTCC	Reverse screening primer for the CLAU *eCas* locus
CLAU-c3SNP1-LHA-F-AsiSI	ATATATATGCGATCGCTTCAAAGCTGTGTTCACATTC	Forward primer for amplification of CLAU SNP1 left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-c3SNP1-LHA-R	TATGGTGTCGAGACTAAGGAAGGTACCGTTATTTACTTGAATATAAGCATAACATATTTAGCTG	Reverse primer for amplification of CLAU SNP1 left homology arm with overhang to enable SOEing PCR
CLAU-c3SNP1-RHA-F	AAATAACGGTACCTTCCTTAGTCTCGACACCATAATCTCTGGACAATTTCCAGAG	Forward primer for amplification of CLAU SNP1 right homology arm with overhang to enable SOEing PCR
CLAU-c3SNP1-RHA-R-AscI	ATATATATGGCGCGCCGCCTTTTTCCACAAAAAGG	Reverse primer for amplification of CLAU SNP1 right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-c3SNP1-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACTACGGAGTGGAAACAAAAGAGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting SNP1
CLAU-SNP1-scr-F	GCCCATACTTACACAGAAACAG	Forward screening primer for the CLAU SNP1 locus
CLAU-SNP1-scr-R	TGCGCTTACTTCAACTAATGC	Reverse screening primer for the CLAU SNP1 locus
CLAU-c3SNP2-LHA-F-AsiSI	ATATATATGCGATCGCGATGCAAAAAATAATCAATCCG	Forward primer for amplification of CLAU SNP2 left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-c3SNP2-LHA-R	GATATCTTTGTTACCTATAATTGTCAAAACTATACCATTTTTGTCTGTGAGATATAATGAGAAAC	Reverse primer for amplification of CLAU SNP2 left homology arm with overhang to enable SOEing PCR

CLAU-c3SNP2-RHA-F	GGTATAGTTTTGACAATTATAGGTAACAAAGATATCGTAATGGAGCAGGCAAAGG	Forward primer for amplification of CLAU SNP2 right homology arm with overhang to enable SOEing PCR
CLAU-c3SNP2-RHA-R-AscI	ATATATATGGCGCGCCCTTTTCCTGTACCACTTTCACC	Reverse primer for amplification of CLAU SNP2 right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-c3SNP2-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACGGAATTGTATTAACTATCATGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting SNP2
CLAU-SNP2-scr-F	CTTAAAATATTGGACAAAAGACTTCC	Forward screening primer for the CLAU SNP2 locus
CLAU-SNP2-scr-R	CTGTCATTGTGAATTGACTGG	Reverse screening primer for the CLAU SNP2 locus
CLAU-c3SNP3-LHA-F-AsiSI	ATATATATGCGATCGCGAAGAAGATGTTTTTTCACTGGA	Forward primer for amplification of CLAU SNP3 left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
CLAU-c3SNP3-LHA-R	GGTACCTACCCTAGGTATAGTTGTTACTATTCCAATGCTATCTGAATCTTTTATAGCTCTATATGC	Reverse primer for amplification of CLAU SNP3 left homology arm with overhang to enable SOEing PCR
CLAU-c3SNP3-RHA-F	ATTGGAATAGTAACAACTATACCTAGGGTAGGTACCGTGAGAATTGAAAAAGTTGAAAAG	Forward primer for amplification of CLAU SNP3 right homology arm with overhang to enable SOEing PCR
CLAU-c3SNP3-RHA-R-AscI	ATATATATGGCGCGCCCTTGCCATGTACTGAAGTGC	Reverse primer for amplification of CLAU SNP3 right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
CLAU-c3SNP3-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACGTTACTACAATCCCAAGAGTGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting SNP3
CLAU-SNP3-scr-F	AGAAGATGCTGATGTAAGATCC	Forward screening primer for the CLAU SNP3 locus
CLAU-SNP3-scr-R	CAAGAGTTTCTGCTAGTTGAGG	Reverse screening primer for the CLAU SNP3 locus
ELM-cas3-LHA-F-AsiSI	ATATATATGCGATCGCGGATAAAAGGTACCGCTGTAACC	Forward primer for amplification of ELM *cas3* left homology arm with *Asi*SI site in homology arm for cloning into pMTL43151 vector
ELM-cas3-LHA-R	TCCTCTGAGGAATCCACCCAGCATTGAGGAAAATAAAATGAAGGTTAACG	Reverse primer for amplification of ELM *cas3* left homology arm with overhang to enable SOEing PCR
ELM-cas3-RHA-F	ATGCTGGGTGGATTCCTCAGAGGACATAATGTCTCTCCTCCAGACGG	Forward primer for amplification of ELM *cas3* right homology arm with overhang to enable SOEing PCR

ELM-cas3-RHA-R-AscI	ATATATATGGCGCGCCGCCTACCATATTTTCACAAAGG	Reverse primer for amplification of ELM *cas3* right homology arm with *Asc*I site in homology arm for cloning into pMTL43151 vector
ELM-cas3-sg1-HiFi	ATCTTAAGGAGGAGTTTTCGTCGACGCAACCTCAGAATACATGAAGTTTTAGAGCTAGAAATAGCAAGTT	Oligo for HiFi introduction of sgRNA seed targeting ELM *cas3*
ELM-cas3-scr-F	CTTCAGGCAGAAAATACTG	Forward screening primer for the ELM *cas3* locus
ELM-cas3-scr-R	GAAGCCGCTTTAATCGC	Reverse screening primer for the ELM *cas3* locus
CLAU-RBS1-F	TATACATATGAACACCTCCTTTTTGTAGTTTGCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS1
CLAU-RBS2-F	TATACATATGTACTTACACCTCCTTTGTATTACGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS2
CLAU-RBS3-F	TATACATATGATACACCTCCTATATGTGGTTACGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS3
CLAU-RBS4-F	TATACATATGATTTAACACCTCTTTTTAGAACCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS4
CLAU-RBS5-F	TATACATATGTAAACCTCCTTGTATGATAAAACGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS5
CLAU-RBS6-F	TATACATATGAATATATAGGCAACTCCTTTCTCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS6
CLAU-RBS7-F	TATACATATGTTTAAATTACCTCCTTATCATGCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS7
CLAU-RBS8-F	TATACATATGTCTTCTCCTTTGATCCAATTTTCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS8
CLAU-RBS9-F	TATACATATGTTGTATTAATCAACTCCTCATACGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS9
CLAU-RBS10-F	TATACATATGAACCTTCCTTAATCACATCTTACGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS10
CLAU-RBS11-F	TATACATATGTGTCCTCCTAAATTATTCCTCTCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS11
CLAU-RBS12-F	TATACATATGTTTTAACCCTCCCCATTATTTTCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS12
CLAU-RBS13-F	TATACATATGTTCACTCCAATCTCAATACTATCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS13

CLAU-RBS14-F	TATACATATGTTATATCCCGGAATCTGAGCATCGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS14
CLAU-RBS15-F	TATACAATGATTTACAGGAATTGTAAAATATACGGGGTAACAGATAAACC	Forward primer for amplification of *catP* with overhang to introduce RBS5
RBS-catP-R	GAAGTACATCACCGACG	Reverse primer for amplification of *catP* to enable cloning of new RBS regions into pMTL82254
CLAU-spacer-1.27-5'PAM-F	ATATATCTCGAGCTTCAGCACCTAAGTAATCTATAAGTAAAGTATCAACTCTTAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL86151 amplification specific module with overhang to introduce CLAU 1.27 spacer with putative 5' PAM
CLAU-spacer-1.27-3'PAM-F	ATATATCTCGAGGCACCTAAGTAATCTATAAGTAAAGTATCAACTCTTTTAATAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL86151 amplification specific module with overhang to introduce CLAU 1.27 spacer with putative 3' PAM
CLAU-spacer-1.27-5'+3'PAM-F	ATATATCTCGAGCTTCAGCACCTAAGTAATCTATAAGTAAAGTATCAACTCTTTTAATAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL86151 amplification specific module with overhang to introduce CLAU 1.27 spacer with putative 5' and 3' PAM
CLAU-spacer-1.27-noPAM-F	ATATATCTCGAGGCACCTAAGTAATCTATAAGTAAAGTATCAACTCTTAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL86151 amplification specific module with overhang to introduce CLAU 1.27 spacer with no putative PAM
CLAU-spacer-1.27R	ATATATATGGGCGCGCCATAAAAATAAGAAGCC	Reverse primer for amplification of pMTL86151 amplification specific module
1.27-ATTAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27-TTTAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCTTTAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27-CTTAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCCTTAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27-GTTAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCGTTAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27-AATAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCAATAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi

1.27_ACTAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCACTAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_AGTAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCAGTAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATAAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATAAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATCAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATCAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATGAA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATGAAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATTTA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTTAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATTCA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTCAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATTGA-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTGAAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATTAT-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTATAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATTAC-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTACAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
1.27_ATTAG-HiFi-oligo	CAGCTATGACCGCGGCCGCTGTATCATTAGAAGAGTTGATACTTTACTTATAGATTACTTAGGTGCATAAAAATAAGAAGCCTGCATTTGC	Oligonucleotide for introduction of 1.27 spacer with modified 5' PAM region via oligo HiFi
CLAU-endocas-HiFi-F	TTTTGTTTAACTTTAATAAGGAGATATACCATGGAAATAAATCAAATAAACATTAAAG	Forward primer for amplification of Cas6, Cascade and Cas3 coding region of the Cas operon for introduction into pACYC-duet via HiFi assembly

CLAU-endocas-HiFi-R	TTCTGTTCGACTTAAGCATTATGCGGCCGCTTACATAAACAAATTTGCCTCATC	Reverse primer for amplification of Cas6, Cascade and Cas3 coding region of the Cas operon for introduction into pACYC-duet via HiFi assembly
CLAU-endo1.27-F	AAAGTTATTTGAATTTAATGGCTTAGGGCACTTTTTTATTGAGTTTATG	Forward primer for amplification of synthetic CLAU crRNA
CLAU-endo1.27R	AGAATGTATACTTTAATGTTTATTTGATTTATTTCCATATGTATTTCCTCCTAATCTTTTTACA	Reverse primer for amplification of synthetic CLAU crRNA
GFP-F-NdeI	TATAAGAAGGAGATATACATATGACTGCGCTAACCGAG	Forward primer for amplification of GFP with *Nde*I site in overhang for cloning into pET-duet vector
GFP-R-PacI	CAGCAGCCTAGGTTAATTAACTAGACTATCACTGATAAGTATCTAAATCGAC	Reverse primer for amplication of GFP with *Pac*I site in overhang for cloning into pET-duet vector
1.27-degPAM-HiFi-oligo	ACCATCATCACCACAGCCAGGATCCNNNNNNGCACCTAAGTAATCTATAAGTAAAGTATCAACTCTTNNNNNNGCGGCCGCATAATGCTTAAGTCGAA	Oligonucleotide for introduction of 1.27 spacer target region flanked by 6 nt degenerate regions
M13-R	CAGGAAACAGCTATGAC	Sequencing primer
T7-F	TAATACGACTCACTATAGGG	Sequencing primer
ELM-spacer2.08-5'PAM-F	ATATATCTCGAGAAGCCAAGTTGTCCATTTTCAAAGGCATTCAAAACATCCAGAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.08 spacer with putative 5' PAM
ELM-spacer2.08-3'PAM-F	ATATATCTCGAGAAGTTGTCCATTTTCAAAGGCATTCAAAACATCCAGAGACAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.08 spacer with putative 3' PAM
ELM-spacer2.08-5'+3'PAM-F	ATATATCTCGAGAAGCCAAGTTGTCCATTTTCAAAGGCATTCAAAACATCCAGAGACAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.08 spacer with putative 5' and 3' PAM
ELM-spacer2.08-noPAM-F	ATATATCTCGAGAAGTTGTCCATTTTCAAAGGCATTCAAAACATCCAGAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.08 spacer with no putative PAM
ELM-spacer2.12-5'PAM-F	ATATATCTCGAGTGGTCCACGTTCGCCTTGCGGCCCGGTTTCCCCTTTCAGAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.12 spacer with putative 5' PAM

ELM-spacer2.12-3'PAM-F	ATATATCTCGAGCACGTTCGCCTTGCGGCCCGGTTTCCCCTTTCAGCGATAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.12 spacer with putative 3' PAM
ELM-spacer2.12-5'+3'PAM-F	ATATATCTCGAGTGGTCCACGTTCGCCTTGCGGCCCGGTTTCCCCTTTCAGCGATAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.12 spacer with putative 5' and 3' PAM
ELM-spacer2.12-noPAM-F	ATATATCTCGAGCACGTTCGCCTTGCGGCCCGGTTTCCCCTTTCAGAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 2.12 spacer with no putative PAM
ELM-spacer3.11-5'PAM-F	ATATATCTCGAGGCACCGGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.11 spacer with putative 5' PAM
ELM-spacer3.11-3'PAM-F	ATATATCTCGAGGGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCATGAACAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.11 spacer with putative 3' PAM
ELM-spacer3.11-5'+3'PAM-F	ATATATCTCGAGGCACCGGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCATGAACAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.11 spacer with putative 5' and 3' PAM
ELM-spacer3.11-noPAM-F	ATATATCTCGAGGGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.11 spacer with no putative PAM
ELM-spacer3.28-5'PAM-F	ATATATCTCGAGTTTTTTGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.28 spacer with putative 5' PAM
ELM-spacer3.28-3'PAM-F	ATATATCTCGAGTGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGATGACTAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.28 spacer with putative 3' PAM
ELM-spacer3.28-5'+3'PAM-F	ATATATCTCGAGTTTTTTGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGATGACTAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.28 spacer with putative 5' and 3' PAM
ELM-spacer3.28-noPAM-F	ATATATCTCGAGTGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.28 spacer with no putative PAM

ELM-spacer3.29-5'PAM-F	ATATATCTCGAGTTCGTGGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.29 spacer with putative 5' PAM
ELM-spacer3.29-3'PAM-F	ATATATCTCGAGGGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGATGATCAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.29 spacer with putative 3' PAM
ELM-spacer3.29-5'+3'PAM-F	ATATATCTCGAGTTCGTGGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGATGATCAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.29 spacer with putative 5' and 3' PAM
ELM-spacer3.29-noPAM-F	ATATATCTCGAGGGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGAAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.29 spacer with no putative PAM
ELM-spacer3.30-5'PAM-F	ATATATCTCGAGTTCTTGCCGATGAAGCCGGAGCCATGGACAGTTATCCGGTAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.30 spacer with putative 5' PAM
ELM-spacer3.30-3'PAM-F	ATATATCTCGAGGCCGATGAAGCCGGAGCCATGGACAGTTATCCGGTTGAGGAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.30 spacer with putative 3' PAM
ELM-spacer3.30-5'+3'PAM-F	ATATATCTCGAGTTCTTGCCGATGAAGCCGGAGCCATGGACAGTTATCCGGTTGAGGAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.30 spacer with putative 5' and 3' PAM
ELM-spacer3.30-noPAM-F	ATATATCTCGAGGCCGATGAAGCCGGAGCCATGGACAGTTATCCGGTAAGCTTGGCACTGGCCGTC	Forward primer for amplification of pMTL87151 amplification specific module with overhang to introduce ELM 3.30 spacer with no putative PAM
ELM-spacer-R	ATATATATGGGCGCGCCATAAAAATAAGAAGCC	Reverse primer for amplification of pMTL86151 amplification specific module

Table 2.4: DNA sequences synthesised for this study
Details of DNA sequences synthesised. In description *C. autoethanogenum* and *C. acetobutylicum* are referred to as CLAU and CAC respectively.

Synthetic DNA	Sequence	Description
pACYC-CLAU-1.27-synRNA	GCGGCCGCATAATGCTTAAGTCGAACAGAAAGTAATCGTATT GTACACGGCCGCATAATCGAAATTAATACGACTCACTATAGG GGAATTGTGAGCGGATAACAATTCCCCATCTTAGTATATTAG TTAAGTATAAGAAAGGAGATATACATTTCCAACTAATTTGGAG GATAGAGATGTCTTAAGTCATTTGAATTCAATGGATGAGGGT ACTTTTTTTATTGAGTTTGTGAAAAGTGAAAAACGGTTGGAA AATTTTCTGTTGATGCCTTTATTATTAAGGGTTTTAGGACTA TAATGAAAATAAGAAATGGCTATTTTACTGTGGTTGAACCTC AACATAAGATGTATTTAAATGCACCTAAGTAATCTATAAGTA AAGTATCAACTCTTGTTGAACCTCAACATAAGATGTATTTAA ATTGTGTTATTAGGGATAACAGGGTAATTTTTCCGTTAGTTG AACCTCAACATAAGATGTATTTAAATTTAATTAACCTAGGCT GCTGCCACCGCTGA	Synthetic CLAU 1.27 RNA designed to be cloned into pACYC-duet
CLAU-eCas-synRNA	GGGCACTTTTTTATTGAGTTTATGAAAATGAAGTTGAATCTA TGCTAGTTATTAAAACTGTTGAGAATTAATTTTCTCAGCAGT TTTTTATTTTGTACGTAACATTCGTTACAGACCTATACCTTC TATTTTAATATAATGTACTTATAAAATCAATGCTAGTGTATC AAAATTTAGGAGGTTAGTTCTTCCAACTAATTTGGAGGATAG AGATGTCTTAAGTCATTTGAATTCAATGGATGAGGGTACTTT TTTTATTGAGTTTGTGAAAAGTGAAAAACGGTTGGAAAATTT TCTGTTGATGCCTTTATTATTAAGGGTTTTAGGACTATAATG AAAATAAGAAATGGCTATTTTACTGTGGTTGAACCTCAACAT AAGATGTATTTAAATGCACCTAAGTAATCTATAAGTAAAGTA TCAACTCTTGTTGAACCTCAACATAAGATGTATTTAAATTGT GTTATTAGGGATAACAGGGTAATTTTTCCGTTAGTTGAACCT CAACATAAGATGTATTTAAATGTAAAATAAATGTGCCTCAAC TTAGATGTTAAGGCACATTTATTTTATATATGTGTAGTAGCC TGTGAAATAAGTAAGGAAAAAAAAGAAGTAAGTGTTATATAT GATGATTATTTTGTAGATGTAGATAGGATAATAGAATCCATA GAAAATATAGGTTATACAGTTATATAAAAATTACTTTAAAAA TTAATAAAAACATGGTAAAATATAAATCGTATAAAGTTGTGC TCGAGTGTAAAAAGATTAGGAGGAAATACATATGGAAATAAA TCAAATAAACATTAAAGTATACATTCTAGAGA	Synthetic CLAU RNA with *ISce*-I site to allow oligo-HiFi introduction of new spacer regions and CAC P_{fdx} to drive the Cas operon. Designed for pMTL43151-CLAU-eCas
trRBS-genestring-oligo	AGGAAACAGCTATGACCGCGGCCGCTTTTTAACAAAATATAT TGATAAAAATAATAATAGTGGGTATAATTAAGTTGTTAGAGA AAACGTATAAATTAGGGATAAACTATGGAACTTATGAAATAG ATTGAAATGGTTTATCTGTTACCCCATGGATAAGAAATACTC AATAGGCTTAGATATCGGCACAAATAGCGTCGGATGGGCGGT GATCACTGATGATTATAAGGTTCCGTCTAAAAAGTTCAAGGT TCTGGGAAATACAGACCGCCACAGTATCAAAAAAAATCTTAT AGGGGCTCTTTTATTTGACAGTGGAGAGACAGCGGAAGCGAC TCGTCTCAAACGGACAGCTCGTAGAAGGTATACACGTCGGAA GAATCGTATTTGTTATCTACAGGAGATTTTTTCAAATGAGAT GGTATTTGAAAAAATTGATAAAA	trCas9 RBS with overhangs for cloning into pMTL82254
duelRBS-HiFi-oligo	AGGAAACAGCTATGACCGCGGCCGCTTTTTAACAAAATATAT TGATAAAAATAATAATAGTGGGTATAATTAAGTTGTTAGAGA AAACGTATAAATTAGGGATAAACTATGGAACTTATGAAATAG ATTGAAATGGTTTATCTGTTACCCCGTATCAAAATTTAGGAG GTTAGTTAGAATGGATAAGAAATACTCAATAGGCTTAGATAT CGGCACAAATAGCGTCGGATGGGCGGTGATCACTGATGATTA TAAGGTTCCGTCTAAAAAGTTCAAGGTTCTGGGAAATACAGA CCGCCACAGTATCAAAAAAAATCTTATAGGGGCTCTTTTATT TGACAGTGGAGAGACAGCGGAAGCGACTCGTCTCAAACGGAC AGCTCGTAGAAGGTATACACGTCGGAAGAATCGTATTTGTTA TCTACAGGAGATTTTTTCAAATGAGATGGTATTTGAAAAAAT TGATAAAA	trCas9 and CAC P_{thl} native RBS with overhangs for cloning into pMTL82254
BM4-12	GTGCATTGATTTTTTCGACTTATTTAGGGTTGTGGGTTGTAC GGAAGGATTTCTGATATTACTGTCACAGGACCGATACCGTTT ACGAAATAGGTGAAGATCAGGCTATCACTGAGGTCCGGAGCT CCGATAAAAAATGGTATTGATTCTCTTCAAGTAGAGGCCATT GTACTATCATGCTAGAGGATGCAGTCGGCTGTAGAAAGAGGC GACTGCATTTTATTATGTAAGGATCCATAACTGTCCTCCTAA ATTAT	BM4-12 sequences for cloning between pMTL43151-CLAU-pyrE homology arms

2.3. Chemicals and suppliers

Table 2.5: Suppliers used in this study

Supplier	Product(s)
New England Biolabs	Restriction endonucleases, Q5® High-Fidelity DNA polymerase, Antarctic Phosphatase, NEBuilder HiFi mastermix,
Promega	T4 DNA ligase
Oxoid	Compounds used for media preparation
Sigma-Aldrich	Compounds used for media preparation, organic and inorganic chemicals
Thermofisher	O'Gene 1 kB plus ladder, SYBR™ Safe, DreamTaq Polymerasecompounds used for media preparation
TCI Europe	5-Fluoroorotic Acid
MP Biomedicals	Agarose electrophoresis powder
Sartorius	Minisart® Syringe Filters
Pro-lab diagnostics	Microbank™ -80 °C cell storage tubes

2.4. General bio-informatics tools

2.4.1. DNA sequence analysis

Genomic DNA sequence data was visualised primarily using Benchling, with Artemis being used for this function initially. Basic local alignment searches were performed using the NCBI BLASTn algorithm, with BLASTx used for protein sequence analysis. Sequence data alignments were performed primarily using the Benchling Sequence Alignment tool, with EMBOSS Needle being used for this function previously.

2.4.2. Protein structure analysis

Functional protein domains were predicted using Pfam versions 27.0-31.0, a service from the Wellcome Trust Sanger Institute using *Pfamseq*, a database based on UniProt. This tool was used for comparative structural analysis of proteins.

2.4.3. Promoter and terminator sequence analysis

Putative promoter sequence were identified using BPROM (Bacterial PROMoter prediction programme), an algorithm provided by Softberry[8]. Terminators sequences and predicted secondary structures were identified and analysing using ARNold[9], an algorithm provided by the Institute of Genetics and Microbiology (University Paris-Sud).

[8] available from: http://linux1.softberry.com/cgi-bin/programs/gfindb/bprom.pl
[9] available from: http://rna.igmors.u-psud.fr/toolbox/arnold/

2.4.1. Restriction-modification system analysis

Whole genome sequence data was submitted to REBASE (Restriction Enzyme dataBASE) for restriction-modification system analysis[10]. This web-based platform uses published and unpublished data on cleavage sites, isochizomers, methylation sensitivity to predict DNA methyltransferases, homing endonuclease, nicking endonucleases, specificity subunits, control proteins and helicase domain proteins.

2.4.2. Plasmid design and visualisation

Plasmid design, *in-silico* construction and analysis was performed primarily using Benchling[11], with these functions being provided by UGENE software by UniProt prior to use of Benchling[12].

2.4.3. Oligonucleotide design

Oligonucleotide design for PCR reactions was routinely performed manually from DNA sequences visualised as described above. Analysis of oligonucleotides for use as primers was primarily performed on Benchling, including predicted secondary structure formation, primer dimerization and DNA melting temperatures. Prior to, and occasionally in parallel with use of Benchling, this analysis was performed using NEB Tm calculator[13].

2.4.4. Data analysis and presentation

GraphPad Prism was used for statistical analysis and result presentation[14].

---

[10] available from: http://rebase.neb.com/rebase/rebase.html
[11] available from: https://benchling.com/
[12] available from: http://ugene.net/
[13] available from: http://tmcalculator.neb.com/
[14] available from: https://www.graphpad.com/scientific-software/prism/

2.5. General microbiology techniques

2.5.1. Growth conditions and media

2.5.1.1. Growth conditions

E. coli strains were cultured aerobically at 37 °C unless otherwise stated. All *E. coli* cultures were grown on LB media (both liquid and solid plates), with appropriate antibiotic selection as stated in table 8. For liquid culture, *E. coli* were grown for 16 hours at 250 rpm unless otherwise stated.

Clostridial strains were cultured at 37 °C under anaerobic conditions. DNA transfer, stock handling and day-to-day microbiology work was performed in a Don Whitley Scientific MG1000 workstation, under an atmosphere of 80% Nitrogen, 10% Carbon Dioxide, and 10% Hydrogen. All solid media was reduced inside the workstation for a minimum of 6 hours, with all liquid media reduced for at least 16 hours. Liquid growth inside the workstation was performed statically.

2.5.1.2. Strain storage and revival

Plasmid harbouring *E. coli* strains were stored at -80 °C using Microbank™tubes provided by Pro-lab diagnostics according to the manufacturer's instructions. Clostiridial strains were grown to an OD_{600} of at least 1, and stored in 20% glycerol containing media at -80 °C.

2.5.1.1. Media composition

For this study, all media were prepared with components as outlined in Table 2.6. For solid media, 1.5% (w/v) No. 1 Bacteriological Agar (Oxoid) was added to liquid broth prior to autoclave sterilisation. This sterilisation was performed at 121 °C, 100 kPa for 20 minutes unless stated otherwise.

Table 2.6 includes stock solution in components, the details of which are provided in Table 2.7. For application specific purposes, various supplements were included in media where stated. All media supplements were stored and prepared as per manufacturer instructions, and filter-sterilised through a 0.2 μm membrane prior to use. Supplement concentrations were used as stated in Table 2.8, and stocks were stored for no more than four weeks prior to use.

Table 2.6: Clostridial growth media
Media composition for culturing bacteria. YTF (Yeast, Tryptone, Fructose) media, CBM (Clostridial Basal Media), PYG (Peptone, Yeast, Glucose) media, PET-C MES (ATCC medium 1754), LB (Lysogeny Broth), SOC (Super Optimal Broth). Media pH and organisms used are shown below media name. Dotted lines indicate liquid constituents for media.

Medium	Constituents	Quantity added ($g\ L^{-1}$)
YTF pH to 5.8 w/ HCl *C. autoethanogenum* *C. ljungdahlii*	Yeast Extract	10
	Tryptone	16
	Fructose	10
	NaCl	0.2
	Acid trace element solution	1 mL
	Base trace element solution	1 mL
	Vitamin stock solution	1 mL
CBM pH to 5.8 w/ HCl *C. acetobutylicum*	Glucose	10
	$MgSO_4.7H_2O$	0.2
	$MnSO_4.4H_2O$	0.01
	$FeSO_4.7H_2O$	0.01
	Casein hydrolysase	4
	p-aminobenzoic acid	1 mg
	Biotin	2 μg
	Thiamine HCl	1 mg
PYG pH to 7.2 w/ NaOH *E. limosum*	Tryptone	5
	Peptone	5
	Yeast extract	10
	Beef extract	5
	Glucose	5
	H_2HPO_4	2
	$NaHCO_3$	0.4
	L-cysteine	0.5
	Trace mineral solution	2 mL
	Tween 20	0.2 mL
	Wolfe's vitamin solution	1 mL
	Haemin solution	10 mL
PET-C MES pH to 5.7 w/ NaOH *C. autoethanogenum*	MES	20
	Macronutrients solution	20 mL
	Trace metal solution	10 mL
	Wolfe's vitamin solution	1 mL
LB *E. coli*	Tryptone	10
	Sodium Chloride	10
	Yeast extract	5
SOC *E. coli*	Yeast extract	5
	Tryptone	20
	NaCl	0.6
	KCl	0.2
	$MgSO_4$	2.4

Table 2.7: Stock solutions for growth media
Composition of stock solutions used in making media detailed in Table 2.6. Stock concentrations and the media in which they are used are indicated in brackets and italics respectively. Units of constituents per litre are shown in the quantity added column. Dotted lines indicate addition of liquid components.

Stock solution	Constituents	Quantity added (L^{-1})
Acid trace element solution (1000X) *YTF*	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
	ZnCl2	78 mg
	$CuSO_4.5H_2O$	50 mg
	$AlK(SO_4)_2.12H_2O$	50 mg
	HCl	50 mM
Base trace element solution (1000X) *YTF*	Na_2SeO_3	58 mg
	Na_2WO_4	53 mg
	$Na_2MbO_4.2H_2O$	52 mg
	NaOH	10 mM
Wolfe's vitamin solution (1000X) *YTF, PYG, PET-C*	Biotin	2 mg
	Folic acid	2 mg
	Pyridoxine hydrochloride	10 mg
	Thiamine HCl	5 mg
	Riboflavin	5 mg
	Nicotinic acid	5 mg
	Calcium D-pantothenate	5 mg
	Vitamin B_{12}	0.1 mg
	p-Aminobenzoic acid	5 mg
	Thioctic acid	5 mg
Trace mineral solution (500X) *PYG*	NH_4Cl	50 g
	NaCl	40 g
	KCl	5 g
	KH_2PO_4	5 g
	$MgSO_4.7H_2O$	10 g
	$CaCl_2.2H_2O$	1 g
Haemin solution (100X) *PYG*	Haemin	0.5 g
	NaOH (1M)	10 mL
Macronutrient solution (100X) *PET-C MES*	Nitriolotriacetic acid	2 g
	NH_4Cl	10 g
	NaCl	80 g
	KCl	10 g
	KH_2PO_4	20 g
	$MgSO_4.7H_2O$	20 g
	$CaCl_2.2H_2O$	4 g
	$Fe(SO_4)_2(NH_4)_4.6H_2O$	4.8 g

Trace metal solution (100X) *PET-C MES*	Nitriolotriacetic acid	2 g
	$MnSO_4.H_2O$	1 g
	$Fe(SO_4)_2(NH_4)_4.6H_2O$	0.8 g
	$CoCl_2.6H_2O$	0.2 g
	$ZnSO_4.7H_2O$	0.2 mg
	$CuCl_2.2H_2O$	0.02 g
	$NaMoO_4.2H_2O$	0.02 g
	Na_2SeO_3	0.02 g
	$NiCl_2.6H_2O$	0.02 g
	$Na_2WO_4.2H_2O$	0.02 g

Table 2.8: Media supplements used in this study
List of media supplements added to media after sterilisation. Information on stock concentration, solvent and storage is provided, as well as working concentration for use in both *E. coli* and clostridial species used in this study.

Supplement	**Stock solution**			**Working concentration**	
	Concentration (mg mL^{-1})	Solvent used	Storage temperature (°C)	***E. coli*** (μg mL^{-1})	**Clostridia** (μg mL^{-1})
Chloramphenicol (Cm)	25	100% EtOH	-20	25	N/A
Thiamphenicol (Tm)	15	50% EtOH	4	N/A	15
Erythromycin (Erm)	50	100% EtOH	4	500	10
Kanamycin (Kan)	50	H_2O	-20	50	N/A
D-Cycloserine (Cs)	50	H_2O	4	N/A	250
5-Fluororotic Acid (5-FOA)	150	DMSO	N/A (used fresh)	N/A	4000
Uracil	1	H_2O	N/A (used fresh)	N/A	20

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

Chemically competent cells were prepared as follows for all strains of *E. coli* used in this study. 10 mL LB starter cultures were inoculated and grown overnight. 500 mL conical shake flasks were then used to culture the final growth culture, with 300 mL LB inoculated with 5% (v/v) starter culture. OD_{600} was monitored until cultures reached 0.4 – 0.6, at which point cultures were divided into 6 x 50 mL Falcon tubes and left on ice for 10 minutes. Following this, Falcon tubes were centrifuged at 4000 rcf, 4 °C for 10 minutes. The supernatant was discarded and pellets resuspended in 10 mL 100 mM $CaCl_2$ (20% (v/v) glycerol), with cell suspensions combined into 2 x 50 mL Falcon tubes (2 x 30 mL aliquots). Cell suspensions were chilled on ice for 15 minutes, and then centrifugation was repeated with pellets being resuspended in a final volume of 2 mL 100 mM $CaCl_2$ (20% (v/v) glycerol). After 1-2 hours on ice, cell suspensions were divided into 1.5 mL Eppendorf tubes in 100 μL aliquots which were stored at -80 °C.

2.5.3. Transformation of *E. coli* via heat shock

All chemically competent *E. coli* were subject to the same transformation conditions. Competent aliquots were removed from storage at -80 °C and left on ice for 20 minutes. 1-5 μL of 10-200 μg mL^{-1} plasmid DNA was subsequently introduced to the thawed competent cells, and left on ice for another 30 minutes. Heat shock was then conducted for exactly 30 seconds at 42 °C, followed by a 5 minute incubation on ice. Following this, 950 μL of SOC media was added per transformation, and the resulting recovery culture

was incubated at 37 °C, 250 rpm for 1 hour. Typically, 75 μL of recovery was plated onto selective media, with the remaining culture centrifuged at 4000 rcf for 1 minute. Supernatant was poured away, and the pellet resuspended in residual supernatant and plated onto a second petri dish of selective media.

2.5.4. Preparation of chemically competent *C. ljungdahlii*

10 mL YTF starter cultures of *C. ljungdahlii* were inoculated until ~OD_{600} 1-2 was reached. Following this 2 x 150 mL YTF in Erlenmeyer flasks were inoculated using 10% inoculum of starter culture and grown until OD_{600} 0.2-0.3. Upon reaching appropriate OD_{600} cultures were divided into 3 x 50 mL Falcons and centrifuged at 4000 rcf for 5 minutes at 4 °C. Supernatants were discarded and pellets resuspended and pooled into a total 50 mL SMP 5.8 (1 mM $NaPO_4$, 1 mM $MgCl_2$, 270 mM sucrose, pH 5.8). The above described wash step was repeated with another 50 mL SMP 5.8, and centrifugation performed again. Supernatant was discarded and pellet resuspended in a final volume of 2 mL SMP 5.8 (10% DMSO). The cell suspension was divided into 2 mL screw cap tubes in 200 μL aliquots that were stored at -80 °C.

2.5.5. Transformation of *C. ljungdahlii* via electroporation

Cells were removed from -80 °C and immediately transported on ice to the anaerobic workstation. Cells were immediately added to 2 mm electroporation cuvettes (pre-chilled) and introduced to 10 μL plasmid DNA (~100 ng mL^{-1}), and left on ice for 10 minutes. Following this, the cells were electroporated inside the workstation at 1000V, 50 μF, 200 Ohm (ca. 5 ms). Immediately following electroporation, 1 mL YTF was added and the cell suspension

transferred to 15 mL Falcon tube for overnight recovery. After 20 hours, recovery cultures were centrifuged anaerobically at 4000 rcf for 1 minute and resuspended in 100 μL PBS which was spread onto selective media.

2.5.6. Preparation of chemically competent *E. limosum*

10 mL PYG *E. limosum* starter cultures were grown overnight, until an OD_{600} of 0.1-1 was reached. Starter cultures were used to inoculate (5% (v/v)) 300 mL PYG broth. At OD_{600} 0.3, the liquid culture was divided between 6 x 50 mL Falcon tubes which were centrifuged at 7000 rcf, 4 °C for 10 minutes. Supernatants were discarded and SMP 5.8 (1 mM $NaPO_4$, 1 mM $MgCl_2$, 270 mM sucrose, pH 5.8) was used to combine and resuspend pellets into 2 x 25 mL cell suspensions in 50 mL Falcon tubes. Centrifugation was repeated with pellets resuspended and combined in a total of 10 mL SMP 5.8. Centrifugation was repeated and pellets resuspended in a final 1 mL SMP 5.8 (10% (v/v) DMSO). The cell suspension was divided into 2 mL screw cap tubes in 200 μL aliquots that were stored at -80 °C

2.5.7. Phase contrast microscopy

Throughout this study, phase contrast microscopy was used to ensure culture purity based on phenotype. 2 μL culture samples were loaded onto microscope slides and observed.

2.6. General molecular biology techniques

2.6.1. Preparation of genomic DNA from Clostridia

For general use, genomic DNA was extracted using the GenElute™ Bacterial Genomic DNA Kit from Sigma-Aldrich (NA2100).

For next-generation sequencing, phenol-chloroform extraction was performed as described by Humphreys *et al*. (2015).

2.6.2. Preparation of plasmid DNA from *E. coli*

For plasmid DNA extraction, overnight *E. coli* cultures were set up with appropriate selection and the Monarch® Plasmid Miniprep Kit from NEB was used as per the manufacturer's instructions.

2.6.3. Restriction enzyme digests

All restriction enzyme digests, unless stated otherwise, were performed using NEB restriction endonucleases in CutSmart® buffer. For 20 μL reactions, 2 μL of 10X CutSmart® buffer was used, along with 0.5 μL of each restriction enzyme, 1 μg DNA (1-10 μL), and dH_2O to raise volume to 25 μL. Digests were performed for 1 hour at 37 °C.

2.6.4. De-phosphorylation of linear DNA

Dephosphorylation of linear DNA was performed using Antarctic Phosphatase (AP) from NEB. For a 20 μL reaction, 2 μL of 10X Antarctic Phosphatase was used, with 1 μL Antarctic Phosphate and 1-10 μg plasmid DNA, dH_2O was used to bring total reaction volume to 20 μL.

2.6.5. Plasmid DNA ligation

Ligation of digested, AP treated plasmid DNA with digested PCR product inserts was performed using T4 Ligase from Promega. Vector to insert DNA was used at a 1:3 ratio, with a 20 μL reaction consisting of 1 μL T4 ligase, 2 μL T4 ligase buffer, vector DNA, insert DNA, and dH_2O bringing the final reaction volume to 20 μL.

2.6.6. Homology-based plasmid assembly

For homology-based plasmid assembly, NEBuilder® HiFi DNA Assembly Master Mix from NEB was used. This method was used for assembly of linearised vector backbone with PCR fragments, as well as linearised vector and oligonucleotides to bring in small (<50 bp) changes.

Vector to insert ratios were generally 1:2, with a 20 μL reaction consisting of 10 μL 2 X Master Mix, vector and insert DNA, and dH_2O to bring final reaction volume to 20 μL. Reactions were incubated at 50 °C for 1 hour and immediately transformed into chemically competent *E. coli*.

2.6.7. Agarose gel electrophoresis

Agarose gel electrophoresis was performed using gels made up of 0.7-1.0% (w/v) agarose supplied by MP Biomedicals, in TAE buffer (40 mM Tris, 1mM EDTA, 0.1% Glacial Acetic Acid). Gels were prepared by adding agarose powder to 1X TAE buffer, heating in microwave until dissolved, cooling to 55 °C. Upon casting gels, SYBR™ Safe DNA Gel Stain from ThermoFisher was used as per the manufacturer's instructions and added prior to pouring. NEB Loading Dye (6X) was added to DNA samples, which were loaded onto

the gel and separated via gel electrophoresis at 70-120 V for 20-60 minutes in TAE buffer.

2.6.8. DNA extraction from agarose gel

After running, gels were visualised on a blue light box, with any bands of interest excised using a scalpel. Extraction of DNA from the agarose gel was performed using Monarch® DNA Gel Extraction Kits from NEB, as per the manufacturer's instructions.

2.6.9. Polymerase chain reaction

Primers synthesised by Sigma-Aldrich, and designed as described in section 2.4.5. Upon receipt, dry oligonucleotides were resuspended in dH_2O as per the manufacturer's instructions to 100 mM to make stock solutions. 10 mM working primer stocks were made via a 1/10 dilution in dH_2O.

For molecular cloning applications, Q5® High-Fidelity DNA Polymerase 2X Master Mix was used. For 25 μL reactions, 12.5 μL Master Mix was used along with 1.25 μL of each primer, 0.1-1.0 μL of template DNA (10-200 μg), dH_2O was used to bring final volume to 25 μL. An example of the reaction conditions of Q5® polymerase can be seen in Table 2.9.

For colony PCR applications, DreamTaq Green PCR Master Mix was used. For this application, 1 μL loops were used to take up a small amount of cell matter from a colony of interest and resuspend in 20 μL of dH_2O. 5 μL of this cell suspension was used in a PCR reaction, along with 12.5 μL of Master Mix, 1.25 μL of each primer, and dH_2O to a total volume of 25 μL.

PCR conditions illustrated in Tables 2.9 and 2.10

2.6.1. SOEing PCR

Splicing by Overlap Extension PCR (SOE/SOEing PCR), refers to a technique described by Horton *et al*. for splicing together two disparate DNA sequences to form a novel recombinant product. This technique was used to create homology cassettes for knockout clusters used in homology-based methods for genetic engineering.

As illustrated in Figure 2.1, this technique can be used to amplify homology arms from either side of a gene of interest, to generate a homology cassette for knocking out said gene. Primer design for SOEing PCR was performed as described in section 2.4.5, with reaction conditions as described in section 2.6.9 for Q5® Polymerase.

Table 2.9: Example of reaction conditions of Q5 Polymerase
Annealing temperature varies depending on primer design, with elongation lasting 30 second per kB.

		Step	Temperature (°C)	Duration (min:s)
	1	Initial denaturation	98	0:30
30 CYCLES	2	Denaturation	98	0:10
	3	Annealing	52-68	0:30
	4	Elongation	72	1:00
	5	Final elongation	72	2:00
	6	Hold	15	∞

Table 2.10: Example of reaction conditions of DreamTaq Polymerase
Annealing temperature varies depending on primer design, with elongation lasting 1 minute per kB.

		Step	Temperature (°C)	Duration (min:s)
	1	Initial denaturation	95	2:00
30 CYCLES	2	Denaturation	95	0:30
	3	Annealing	52-68	0:30
	4	Elongation	72	2:00
	5	Final elongation	72	5:00
	6	Hold	15	∞

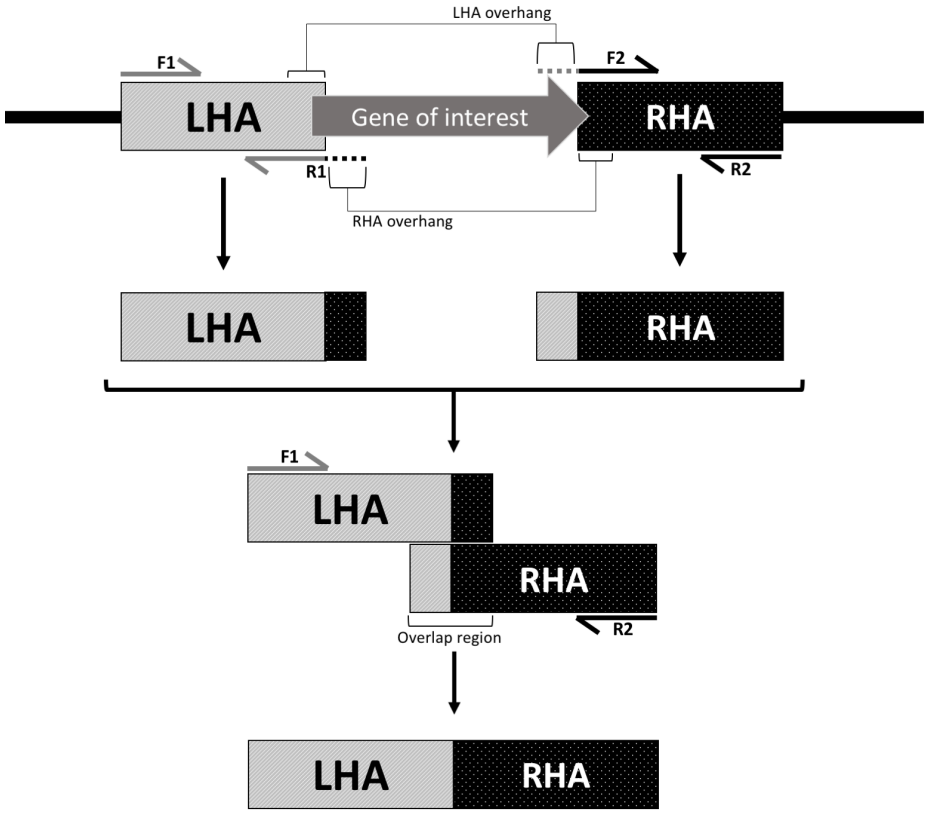


Figure 2.1: SOEing PCR
The left- and right- homology arms (LHA/RHA) of a gene of interest are amplified with two primer pairs; F1 and R1 for the LHA, and F2 and R2 for the RHA. The internal primers (R1 and F2) both contain overhangs derived from the other homology arm to which they are designed to amplify. R1 binds to the LHA anti-sense stand and on its 5' end contains a 20 nt sequence from the 5' end of the RHA. Equally, F2 binds to the start of the RHA, and contains an overhang encoding the last 20 nt of the 3' end of the LHA. Amplification of the homology arms results in LHA containing the 20 nt overhang encoded by R1 at its 3' end, and RHA containing the 20 nt overhang encoded by F2 at its 5' end. The resulting complementary 40 nucleotide region is able to anneal, meaning when both the LHA and RHA PCR products are used as templates for a subsequent PCR with F1 and R2 primers, a recombinant product is created.

2.6.2. DNA fragment sequencing

DNA fragment sequencing was performed by Eurofins (Germany) using primers designed as described in section 2.4.3, which were provided along with sample DNA. Each plasmid used in this study was confirmed via DNA sequencing.

2.7. Gram-positive replicon modularisation

To generate pCACET and pCB101, *C. aceticum* and *C. butyricum* genomic DNA (respectively) were used as templates for PCR amplification. This was achieved using CACETF/CACETR primers for pCACET, and CB101F/CB101R for pCB101. These primers introduced AscI and FseI sites at the 5' and 3' (respectively), which were used to introduce the new replicons into an AscI/FseI digested pMTL83151 vector, thus replacing the existing pCB102 Gram-positive replicon with pCACET and pCB101 respectively.

2.8. Genome annotation

Genome annotation was performed on DNA sequence data produced through an Illumina MiSeq desktop sequencer using a paired-end approach with read lengths of 250 bp. The DNA was extracted from wild type *C. autoethanogenum* JA1-1 DSM 10061 grown on YTF, by postdoctoral researcher Christopher Humphreys, using the DNA extraction method described in section 2.6.1.

The resulting DNA sequence data was analysed using CLC Genomics Workbench, using DSM10061 as a reference sequence (Brown *et al.*, 2015), by Christopher Humphreys. Resulting sequence discrepancies with the reference sequence confirmed via Sanger sequencing of PCR products generated from that locus by Christopher Humphreys.

GenDB version 2.4 was used for automatic sequence annotation, using Prodigal version 2.6.0 for coding region prediction, tRNAScane-SE version 1.21 for tRNA prediction, and RNAMMER version 1.2 for rRNA prediction.

Predicted coding regions were then manually confirmed via the pipeline outlined in Figure 2.2, with amendments added using GenDB V2.4.

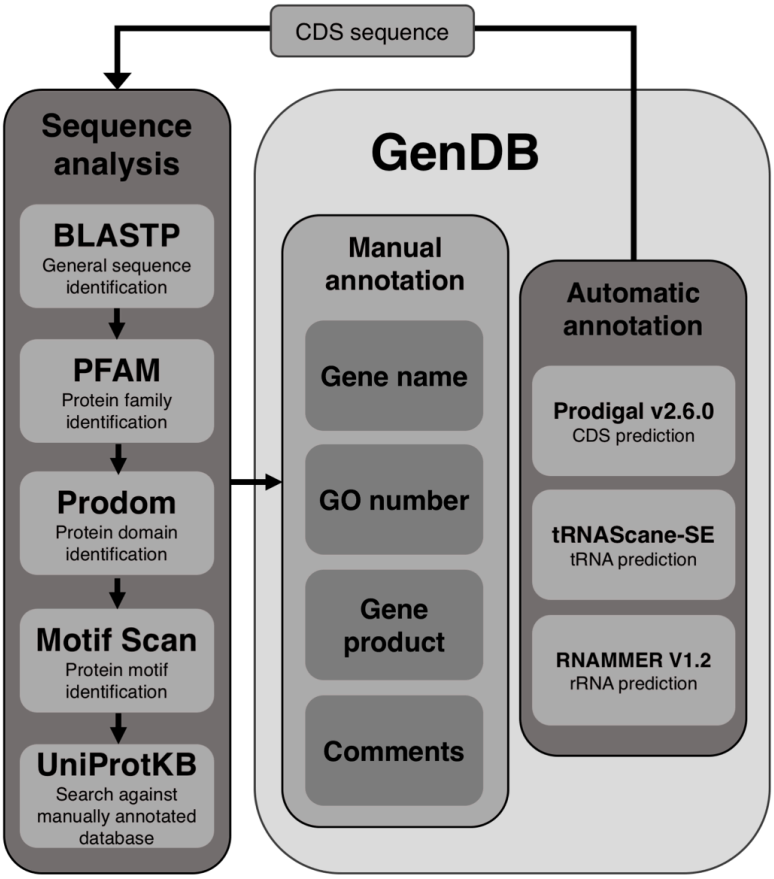


Figure 2.2: GenDB genome annotation pipeline
Amino acid sequences from CDS regions called by Prodigal v26.0 as part of the GenDB automated annotated pipeline were used as query sequences for the following programs: 1) BLASTP: Service provided by NCBI for general sequence identity based on query amino acid similarity to protein sequences in non-redundant protein sequence library. 2)PFAM: EMBL-EBI service that allows query search against database of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). 3) ProDom: InterPro database consisting of protein domain families, searched against by query sequences to identify putative domains present in CDS. 4) Motif Scan: Identifies all known protein motifs in amino acid query sequence, using a range of libraries, including PROSITE, ExPASy, Pfam and InterPro. 5)UniProtKB: Manually annotated database of curated protein sequences with known function, used to verify confidence in annotations based on sequence similarity alone
Individual analysis of each gene enables renaming of subject gene in the event of high sequence identity with many closely related genes of UniProt/BLASTP search results. GO (Gene Ontology) and EC (Enzyme Commission) numbers were also assigned based on high sequence identity, with gene product being listed if information readily available. If manual database search resulted in no consensus protein product, CDS were labelled “hypothetical protein” with added detail with regards to function of conserved domains present and protein family membership.

2.9. Generation of ClosTron mutagenesis plasmids

2.9.1. Target selection

Genes identified for ClosTron mutagenesis were analysed for intron site insertion using the Perutka algorithm (freely available at www.clostron.com). Based on the algorithm score and manual analysis of the site of insertion, the highest score within predicted functional domain was selected.

2.9.2. Plasmid synthesis

Plasmid synthesis of the pMTL007C-E2 plasmid vectors used for mutagenesis was conducted by DNA2.0 (USA). Details shown in Figure 2.3.

2.9.3. Mutant screening

The pMTL007C-E2 vectors were transformed into *E. coli* CA434, which was used as a conjugative donor for *C. autoethanogenum*. Initial transconjugant selection was performed through the acquisition of thiamphenicol resistance using YTF media supplemented in D-cycloserine and thiamphenicol. Individual colonies were streaked to purity using a 1 μL loop, and the resultant clonal cells were harvested with a 10 μL loop, resuspended in 1 mL PBS, and serially diluted to 10^{-1}, 10^{-2}, and 10^{-3}. 100 μL of each 3 of these dilutions, along with the neat suspension were spread onto solid YTF media supplemented with D-cycloserine and clarithromycin to select for insertion of the Group II intron.

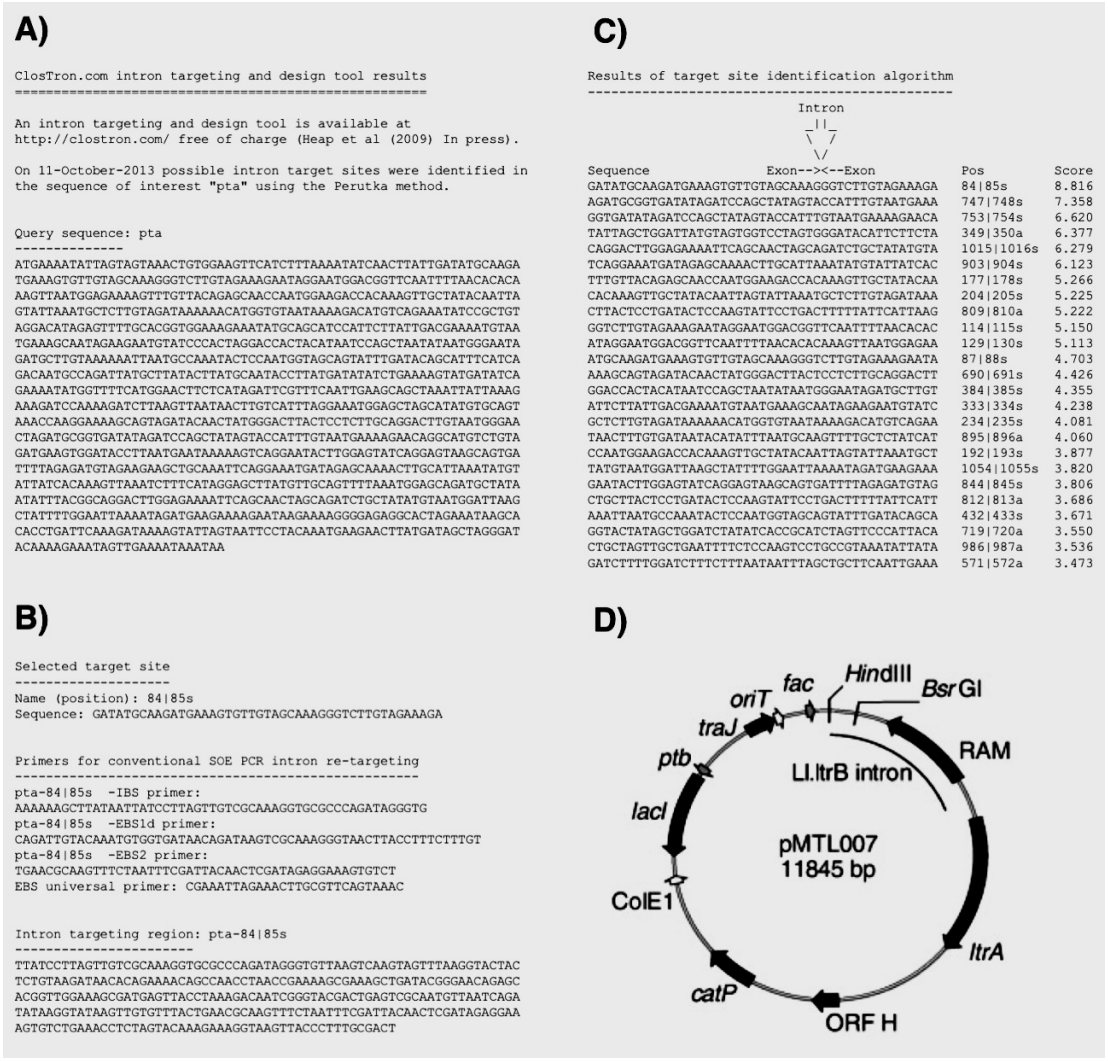


Figure 2.3: Design of pMTL007-EC2-pta
A) Query sequence of the *C. autoethanogenum pta* gene used for Perutka algorithm analysis. B) Selected target site sequence for ClosTron mutagenesis (for intron insertion between position 84/85s), SOE PCR for intron retargeting, and targeting region. C) All potential target site sequences from the *pta* gene with positions within the gene and Perutka scores. D) Plasmid map of pMTL007 vector, adapted from Heap *et al*. (2007).

2.10. Allele Coupled Exchange and Allelic In-Frame Deletion

2.10.1. Development of *pyrE* deficient strains

The plasmid pMTL-CLAU-del-pyrE was conjugated into *C. autoethanogenum* using *E. coli* CA434. YTF plates supplemented with D-cycloserine and thiamphenicol were used to select for pMTL-CLAU-del-pyrE transconjugants. Of the resulting trasnconjugants, larger colonies were identified as possible

single-crossover integrants and struck onto YTF plates supplemented with D-cycloserine and thiamphenicol. To screen for such single crossover integrants, colony PCR was performed of the resulting colonies of the re-streak using primers CLAU-pyrE-internal-F and CLAU-ACE-pyrE-scr-R. Clonal cells of confirmed single-crossover integrants were then harvested using a 10 μL loop, resuspended in 1 mL PBS, and serially diluted to 10^{-1}, 10^{-2}, and 10^{-3}. 100 μL of each 3 of these dilutions, along with the neat suspension were spread onto solid YTF media supplemented with FOA and uracil.

2.10.2. Allelic in-frame deletion

All allelic in-frame deletion was attempted in the *C. autoethanogenum* Δ*pyrE* RB strain developed by Ronja Breitkopf. Plasmids were developed using the pMTL-CLAU-del-pyrE plasmid as a backbone. Using *Nhe*I and *Asc*I restriction sites, the *pyrE* knockout cassette was replaced with knockout cassettes generated via SOEing PCR. Once sequence confirmed, sequenced plasmids were transformed into *E. coli* CA434 and conjugated into *C. autoethanogenum* Δ*pyrE* RB.

Transconjugant selection was achieved using YTF supplemented with D-cycloserine and thiamphenicol. Selection for single-crossover events was performed by selecting faster growing colonies, with these re-struck onto YTF supplemented with D-cycloserine and thiamphenicol. For confirmation of single-crossover mutants, screening primers were used to compare PCR product size to that of a larger wild type locus. Following confirmation of single-crossover integrants, a 10 μL loop was used to harvest clonal cells and

resuspend them in 1 mL PBS. Following resuspension, serial dilution to 10^{-3} was performed which 100 μL of resuspension from each of the dilution series spread onto YTF supplemented with FOA and uracil to force plasmid loss. Of the resulting colonies, screening PCR was repeated to assess if a double-crossover event had occurred at the locus of the single-crossover integrants, or whether the locus had reverted to wild type.

2.11. Generation of CRISPR-Cas9 mutants

Generation of the pMTL40000 vector series is detailed in Chapter 3, which offers details on restriction sites and more detailed information on technique development.

2.11.1. Target selection

In silico selection of sgRNA 20 nt seed region was performed initially for *Clostridium ljungdahlii* manually, with potential targets cross referenced with the entire genome via nucleotide BLAST. Any seed regions with >75% sequence identity elsewhere in the genome were disregarded, with the remaining candidates chosen from based on lack of A/T rich regions.

For all other seed region selection, the Benchling CRISPR design guide tool. Tool parameters were set to “single guide”, with a guide seed length of 20 nt using a 3’ NGG PAM. All potential targets within the gene of interest were then ranked based on off-target (sequence identity elsewhere in the genome) and on-target (Cas9 interaction) algorithm scores. The highest scoring seed regions for each target were used for targeting.

2.11.2. Generation of new sgRNA seed regions

Initial seed region generation was achieved through SOEing PCR, with the sgRNA promoter and sgRNA handle (and terminator) amplified using primers encoding the new seed region. For subsequent and the majority of sgRNA generation, NEB ssDNA oligo HiFi assembly was used.

2.11.3. Generation of homology arm cassettes

Homology arm cassettes were designed with equal sized (750-1000 bp) homology arms either side of the gene of interest. With regards to plasmid nomenclature, the term left homology arm (LHA) was used for the 5' region of the gene of interest, and right homology arm (RHA) for the 3' region. In order to minimise polar effects of gene deletion, homology arms were designed to maintain start and stop codons.

2.11.4. Bookmark design

Bookmark sequences were selected as follows. BM1; synthetic design showing no sequence identity to clostidial genus. BM2/3; successful guide sequences used for gene deletion in *C. difficile* (unpublished). BM4-12; successful guide sequences for gene deletion in *S. pneumoniae* and *B. subtilis*. Confirmation of compatibility with clostridial genus was performed using BLAST optimised for short sequence search against the genus, with all sequences confirmed to have <70% sequence identity to any species within the genus.

2.12. Spacer identification and PAM analysis

For analysis of endogenous CRISPR systems, the CRISPRfinder web based tool from Université Paris-Sud 11 (http://crispr.i2bc.paris-saclay.fr/Server/)

was used to identify CRISPR loci and Cas genes via upload of genome sequence data.

The sequence output function of CRISPRfinder was used to export potential spacer sequences from putative CRISPR loci. Manual *in silico* analysis of each spacer was used to cross-reference the NCBI database (limited to prokaryote and bacteriophage sequences), optimised for somewhat similar sequences via BLASTn. Putative protospacers were selected using a sequence identity threshold of >80%. Putative protospacers were noted for organism of origin and relevance of any associated genome annotative (e.g. phage related protein), and their putative PAMs (5' and 3' 5 nt flanking regions) were recorded. Putative PAMs were assigned an arbitrary score out of 5, with higher sequence identity between the spacer and putative protospacer at the side of the PAM being assigned a higher score. Putative PAMs were then listed in order of descending score to identity conserved domains and higher scoring sequences taken forward for analysis.

Putative PAM analysis was performed through cloning of spacer regions with and without 5' and 3' putative PAMs into pMTL80000 vectors. CFU formation after introduction of these vectors was assessed in order to infer CRISPR-mediated interference.

2.1. CAT assay for *C. autoethanogenum* expression analysis

For RBS analysis, sequences of interest were cloned into the pMTL82254 expression vector, replacing the *C. acetobutylicum* P_{thl} RBS. RBS activity was measured through CatP expression, which was measured via CAT (Chloramphenicol acetyltransferase) assay, as described by Smale (2010).

Following conjugation, colonies were picked in triplicate and used to inoculate liquid YTF cultures supplemented with clarithromycin. Optical density was analysed, and cultures harvested at OD_{600} 0.8-1.0, with volume normalised to 1 mL OD_{600} 1.0. Samples were centrifuged at 16000 rcf for 2 minutes, with supernatant removed via aspiration and cell pellets stored for 7 days at -20 °C. Protein extraction of cell pellets was performed using 250 μL BugBuster® 1X master mix supplemented with Calbiochem Protease Inhibitor Cocktail Set I (2% (v/v)) (both supplied by Merck). Here, pellets were resuspended via vortex and incubated at 37 °C, 200 rpm for 30 minutes. Following incubation, lysate samples were centrifuged at 16000 rcf for 2 minutes, with 200 μL supernatant removed via aspiration and stored at -20 °C for subsequent analysis.

Lysate samples were prepared for CAT assay by adding 2 μL of each lysate sample was added to 98 μL 100mM Tris buffer (pH 7.8) in 96 wells plates. Tecan M1000 injecting plate reader was used to conduct the CAT assay, with 100 μL CAT reaction mastermix (100mM Tris (pH 7.8), 0.167 mM acetyl CoA sodium salt, 0.0833 mM DTNB 0.005% Chloramphenicol (w/v)) added at a rate of 100 μL s^{-1}. Following 2 s orbital shaking (2 mm, 654 rpm), absorbance at 415 nm was measured at intervals of 9 seconds. Expression was quantified in terms of units CatP per mg total protein using rate of reaction calculated from maximum linear rate ((ΔA_{412nm}/min Test -ΔA_{412nm}/min Blank (2) (df)), and normalised for sample protein concentration with Pierce™ BCA Protein assay performed as described by Smith and colleagues (1985).

CHAPTER THREE

C. AUTOETHANOGENUM STRAIN AND DNA TRANSFER OPTIMISATION

3. Strain and DNA transfer optimisation

3.1. Introduction

As an acetogen, the potential of *C. autoethanogenum* as a biofuel production chassis has been outlined in Chapter 1. The primary aim of this study was to realise this potential through metabolic engineering, in order to optimise solvent production capability. In their much-cited review, Fisher and colleagues (2014) comment on metabolic engineering strategies for microbial cell factories (MCFs). Here the use of microbes for complex catalysis of useful end products is described, and the selection criteria for potential chassis discussed. This discussion includes the need for minimisation of metabolic adjustment, with the presence of the Wood-Ljungdahl pathway and its utilisation of C1 gas cited as a key advantage for acetogens over their sugar fermenting counterparts(Fisher *et al.*, 2014). Aside from confirming the potential for acetogens as production chassis already described in Chapter 1, this review also highlights the importance of the following key factors; genome sequence data, efficient DNA transfer, establishment of plasmid-based genomic tools, and controlled gene expression. This chapter looks to address each of these issues in order to fulfil the promise of *C. autoethanogenum* as a chassis.

3.1.1. Genome sequence data

To ensure efficient metabolic engineering and enable accurate genetic engineering, a fully and precisely annotated genome is required. Understanding of which genes and metabolic pathways are present in an

organism is crucial, both with a view to introducing new metabolic pathways, and modifying existing ones.

The first step towards in-depth understanding of the genetics of *C. autoethanogenum* came with the draft genome sequence published by Bruno-Barcena and colleagues (2013). This study used Roche 454 GS FLX Titanium and Ion Torrent PMG techniques to generate a collection of contigs (NCBI GCA_000427255.1), with a total genome size of 4.5 Mbp and overall G+C content of 37.5%. Comparison of the sequence data to that of *C. ljungdahlii* confirmed their phylogenetic similarity, with the authors stating that based on their analysis, the two could in fact be regarded as different strains of the same species.

Following the Bruno-Barcena genome announcement, a study by Brown and colleagues (2014) published the first finished genome sequence (NCBI GCA_000484505.1). This sequence was generated through Pacific Biosciences single molecule DNA sequencing technology and resulted in a smaller reported genome size of 4.4 Mbp. This sequence data underwent a completely automated annotation pipeline, without any stated manual annotation. Despite significant advancement of automated annotation techniques, the lack of manual annotation in this context leads to higher potential for miss-annotation errors being carried through unchecked. Other current advantages of manual annotation also include standardisation of features, such as enzyme names. For example, a genome could contain multiple ORFs with the same predicted function, however if the genes have

high sequence identity to orthologues from different organisms with different annotations, their shared function could go unnoticed.

3.1.2. DNA transfer and plasmid characterisation

Genome sequence data provides a foundation for the understanding of an organism's metabolism, upon which attempts at metabolic engineering can be built. However, efficient DNA transfer is first required to introduce plasmids harbouring the necessary genetic components to do this. Without well characterised plasmids and efficient DNA transfer, metabolic engineering is not possible, therefore development of such methods is crucial.

Many plasmids and methods of transforming or conjugating them in clostridial species have been published (Heap *et al.*, 2009; Jang *et al.*, 2014; Heap *et al.*, 2007). For *C. difficile*, *C. sporogenes*, and *C. botulinum*, the CA434 strain of *E. coli* is routinely used as a conjugal donor (Purdy *et al.*, 2002). This consists of the HB101 strain of *E. coli* (commonly used for recombinant cloning) harbouring the R702 conjugative plasmid (Hedges & Jacob, 1974). Shuttle vectors are transformed into CA434 which is then used to conjugate them into the host strain of interest. Other members of the genus have had successful electroporation techniques developed for DNA transfer (Pyne *et al.*, 2013; Mermelstein *et al.*, 1994), however prior to this study neither conjugation or electroporation had been reported in *C. autoethanogenum*.

As described in section 1.3.1, the pMTL80000 series is modular with four main components; a Gram-positive replicon, an antibiotic resistance marker, a Gram-negative replicon, and an application specific module. As reported in various studies, the range of Gram-positive replicons in the pMTL80000 series

work in a wide variety of organisms. However, these replicons vary in terms of DNA transfer efficiency, copy number and stability from species to species. Therefore, it is crucial to identify the characteristics of each replicon within a given species to determine which to use for a given applications. An example of such research can be seen in the Heap *et al*. (2013) study, describing the varying Gram-positive replicon stabilities of the pMTL series in different clostridia. Table 3.1 displays the plasmids used for Gram-positive replicon analysis in this study, including replicons within the pMTL80000 vector series, as well as new replicons for this study (p19, pCB101, and pCACET).

Table 3.1: Gram-positive replicon characterisation plasmids
Replicons for analysis in *C. autoethanogenum*. The pMTL80000 vector series was used for existing Gram-positive replicons, along with additional Gram-positive replicons new to this study (pCB101 and pCACET).

Modular plasmid	Gram-positive Replicon	Replicon origin
pMTL81151	pCB101	*Clostridium butyricum*
pMTL83151	pCB102	*Clostridium butyricum*
pMTL83a151	pCB102FS	*Clostridium butyricum*
pMTL84151	pCD6	*Clostridium difficile*
pMTL86151	pIP404	*Clostridium perfringens*
pMTL87151	p19	*Clostridium carboxydivorans*
pMTL88151	pCACET	*Clostridium aceticum*

3.1.3. *pyrE* deficient strain development

As described by Ng *et al*., at the time of their 2013 paper, and at the start of this study, allelic exchange using *pyrE*-based counter-selection was the most efficient method of in-frame gene deletion in the genus *Clostridium*. Chapter 1 describes this technique in detail, but pivotal to this technique is the inactivation of the *pyrE*, allowing the use of a plasmid-borne heterologous *pyrE* gene.

3.1.4. Precise gene expression

A final consideration for metabolic engineering is the relative expression of genes being studied. As with replicons, characteristics of promoters and ribosome binding sites (RBSs) vary from organism to organism.

Selection of appropriate promoters and RBSs for gene expression is crucial, and forms a cornerstone of metabolic engineering. One example from within the genus *Clostridium* includes overexpression of *pfkA* and *pykA* genes (encoding phosphofructokinase and pyruvate kinase, respectively) to increase intracellular concentrations of ATP and NADH in *C. acetobutylicum*. This was achieved through use of the promoter from the highly expressed *ptb* (phosphate butyrate transferase), and resulted in increased butanol and ethanol yields (Ventura, Hu & Jahng, 2013).

A commonly used method for assessing relative expression strength is the use of a CAT assay, as described in Chapter 2. The *catP* gene is a common component on plasmid vectors due to it encoding chloramphenicol acetyltransferase, enabling resistance to chloramphenicol. This is achieved through attachment of an acetyl group from acetyl-CoA to chloramphenicol, preventing it from achieving ribosomal interference. By having *catP* under the control of a promoter/RBS, its expression can be quantified through cell lysate acetylation activity. This reaction results in an absorbance shift that can be measured, and comparison of different promoters and RBS sequences offers accurate insight into their relative expression levels.

3.2. Manual annotation of *C. autoethanogenum*

Manual genome annotation was conducted as a collaborative effort, with the initial putative CDS regions divided equally among 12 researchers. These manual annotations, along with SNP analysis conducted by Dr Chris Humphreys resulted in the publication Humphreys and colleagues (2015).

3.2.1. CLAU_2316-2646

The genes specifically annotated within this study were CLAU_2316-2646, and of these 330 genes, 64 were disputed in some way with the annotation presented by Brown and colleagues (2014). Details of such discrepancies are outlined in Table 3.2. In most cases, this involved re-designation of predicted gene products. One example of which being CLAU_2378, originally designated as a RNA methyltransferase. Analysis, however, shows high sequence identity to pyruvate kinase encoding genes, as identified by KEGG, TIGRFAM and Pfam database searches.

The rest of the disputes within the range of genes were either changing more specific annotations to a less specific “hypothetical protein” annotation, or vice versa. For the former, 7 such instances were identified. One example being CLAU_2547, originally annotated as an amino acid permease, this gene shows no sequence identity to any such amino acid sequence. For the latter, more detail was provided for gene products in 15 instances. An example being CLAU_2557, previously identified as a hypothetical protein, this gene has high sequence identity to asparagine synthase genes annotated elsewhere. Also within this range, one hypothetical protein, CLAU_2529 was identified that was not previously reported (Brown *et al.*, 2014).

Table 3.2: CDS discrepancies between with Brown *et al*. (2014)
Displayed are all the discrepancies found in genes CLAU_2316-2646 in this study, with the locus tag from this study included (CLAU), Brown *et al*. locus tag (CAETHG), revised gene product proposed in this study, and original gene product proposed by Brown *et al*. (2014).

Locus tag		Revised gene product	Original gene product
CLAU_2325	CAETHG_2385	Phosphoribulokinase/uridine kinase family protein	ATPase AAA
CLAU_2336	CAETHG_2397	NAD synthase	potassium ABC transporter ATPase
CLAU_2341	CAETHG_2403	Histidine kinase-like ATPase domain-containing protein	anti-sigma regulatory factor
CLAU_2345	CAETHG_2407	Hypothetical protein	SNARE associated protein
CLAU_2351	CAETHG_2413	Hypothetical protein	cell wall hydrolase
CLAU_2353	CAETHG_2415	Holo-(acyl-carrier-protein) synthase	4'-phosphopantetheinyl transferase
CLAU_2357	CAETHG_2419	Transcriptional modulator of MazE/toxin MazF	growth inhibitor PemK
CLAU_2367	CAETHG_2429	Cell cycle protein	membrane protein
CLAU_2368	CAETHG_2430	Peptidoglycan glycosyltransferase	membrane protein
CLAU_2370	CAETHG_2432	UvrABC system protein C	excinuclease ABC subunit C
CLAU_2372	CAETHG_2434	UPF0042 nucleotide-binding protein yhbJ	glmZ(sRNA)-inactivating NTPase
CLAU_2378	CAETHG_2441	Pyruvate kinase	RNA methyltransferase, TrmA family
CLAU_2381	CAETHG_2444	Hypothetical protein	type III effector Hrp-dependent outer protein
CLAU_2385	CAETHG_2448	Hypothetical protein	hydroxyacid dehydrogenase
CLAU_2386	CAETHG_2449	Major facilitator superfamily MFS_1	membrane protein
CLAU_2390	CAETHG_2453	Peptidase dimerization domain protein	succinyl-diaminopimelate desuccinylase
CLAU_2400	CAETHG_2464	GGDEF domain-containing protein	diguanylate cyclase
CLAU_2409	CAETHG_2473	Methyltransferase type 11	hypothetical protein
CLAU_2415	CAETHG_2479	Aspartate ammonia-lyase	Fumarate hydratase class II
CLAU_2416	CAETHG_2480	2-hydroxycarboxylate transporter family protein	Citrate carrier protein
CLAU_2418	CAETHG_2482	HpcH/HpaI aldolase/citrate lyase family protein	Citryl-CoA lyase
CLAU_2439	CAETHG_2503	Putative serine protein kinase PrkA	MarR family transcriptional regulator
CLAU_2441	CAETHG_2505	Drug resistance transporter EmrB/QacA subfamily	major facilitator transporter
CLAU_2443	CAETHG_2507	Efflux transporter RND family MFP subunit	hemolysin D
CLAU_2459	CAETHG_2524	Dolichyl-phosphate-mannose-protein mannosyltransferase family	hypothetical protein
CLAU_2462	CAETHG_2527	Glycosyl transferase family 2	hypothetical protein
CLAU_2464	CAETHG_2529	ABC-2 family transporter protein	hypothetical protein
CLAU_2482	CAETHG_2547	Hypothetical protein	amino acid permease
CLAU_2486	CAETHG_2551	PP-loop domain protein	restriction endonuclease EcoEI subunit S
CLAU_2498	CAETHG_2563	LysM domain containing cysteine-rich secretory protein family	Sporulation uncharacterized protein YkwD
CLAU_2501	CAETHG_2566	DNA-binding domain protein excisionase family	molybdate-binding protein

CLAU_2503	CAETHG_2568	Nitrogen regulatory protein P-II	nitrogen fixation nifHD region glnB 1
CLAU_2505	CAETHG_2570	Nitrogenase molybdenum-iron protein subunit alpha	nitrogenase molybdenum-iron protein beta chain
CLAU_2518	CAETHG_2583	Outer membrane efflux protein	hypothetical protein
CLAU_2519	CAETHG_2584	Outer membrane efflux protein	hypothetical protein
CLAU_2520	CAETHG_2585	ResB family protein	hypothetical protein
CLAU_2521	CAETHG_2586	KWG Leptospira repeat protein	hypothetical protein
CLAU_2533	CAETHG_2598	Exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase	multidrug MFS transporter
CLAU_2534	CAETHG_2599	Glycosyl transferase group 1	hypothetical protein
CLAU_2542	CAETHG_2609	DegT/DnrJ/EryC1/StrS aminotransferase	Glutamine--scyllo-inositol transaminase
CLAU_2552	CAETHG_2622	Exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase	multidrug MFS transporter
CLAU_2554	CAETHG_2625	Glycosyl transferase group 1	O-antigen polymerase
CLAU_2557	CAETHG_2628	Asparagine synthase	hypothetical protein
CLAU_2559	CAETHG_2630	Polysaccharide biosynthesis protein	flagellin modification protein FlmA
CLAU_2560	CAETHG_2631	DegT/DnrJ/EryC1/StrS aminotransferase family	UDP-4-keto-6-deoxy-N-acetylglucosamine 4-aminotransferase
CLAU_2562	CAETHG_2633	NeuB family protein with SAF domain	pseudaminic acid synthase
CLAU_2563	CAETHG_2634	Glycosyl transferase 28 domain-containing protein	pseudaminic acid biosynthesis-associated protein PseG
CLAU_2572	CAETHG_2644	Type IV pilin N-term methylation site GFxxxE family protein	hypothetical protein
CLAU_2579	CAETHG_2652	Type IV pilin N-term methylation site GFxxxE family protein	hypothetical protein
CLAU_2581	CAETHG_2654	MgtC/SapB transporter	methyltransferase
CLAU_2589	CAETHG_2662	Hypothetical protein	von Willebrand factor type A
CLAU_2596	CAETHG_2670	NUDIX hydrolase	hypothetical protein
CLAU_2599	CAETHG_2673	Hypothetical protein	GTP1/OBG protein
CLAU_2604	CAETHG_2678	FecCD transport family protein	corrinoid ABC transporter permease
CLAU_2607	CAETHG_2681	Ig domain protein	hypothetical protein
CLAU_2611	CAETHG_2685	Hypothetical protein DUF2442	GNAT family acetyltransferase
CLAU_2613	CAETHG_2688	FAD-linked oxidase domain protein	D-lactate dehydrogenase (cytochrome)
CLAU_2617	CAETHG_2692	HlyD family secretion protein	transporter
CLAU_2618	CAETHG_2693	HlyD family secretion protein	hemolysin D
CLAU_2624	CAETHG_2699	Transcriptional Coactivator p15	hypothetical protein
CLAU_2625	CAETHG_2700	Peptidoglycan glycosyltransferase	penicillin-binding protein
CLAU_2628	CAETHG_2703	Peptidase family S41 protein	protease
CLAU_2629	CAETHG_2704	Hypothetical protein DUF45	peptidase
CLAU_2630	CAETHG_2705	Putative manganese efflux pump MntP	membrane protein

3.2.2. Overall genome analysis

As detailed by Humphreys and colleagues (2015), Illumina sequencing of *C. autoethanogenum* DSM10061 in this study resulted a 4352627 bp closed genome, with 3969 annotated CDSs. As Figure 3.1 illustrates, of the annotated CDS regions, 8 were not previously identified by Brown and colleagues (2014), and 482 previously unnamed hypothetical protein coding genes were reclassified with an assigned specific function, or conserved domain. A total of 132 CDS regions identified by Brown and colleagues (2014) were reassigned a less specific designation, and 73 were disputed outright.

Inspection of all genes called by Brown and colleagues (2014) showed various instances of DNA sequences being designated as CDS regions, despite the presence of premature stop-codons midway through their sequence. The Humphreys and colleagues paper, towards which this study contributed, highlighted these, along with other instances pseudogenes being annotated as functional.

In terms of nucleotide sequence, many of the apparent inaccuracies mentioned above correlated with loci of apparent single nucleotide insertions. In total, there were 242 instances of apparent single nucleotide insertions in the genome reported by Brown and colleagues (2014) compared to the genome sequence presented in Humphreys and colleagues (2015), with no large INDELS detected. The distribution of these single nucleotide discrepancies is illustrated in Figure 3.1B. Analysis of these discrepancies was conducted, comparing them to the closely related *C. ljungdahlii* genome

(Kopke *et al.*, 2010), and a draft *C. autoethanogenum* genome sequence previously published (Bruno-barcena, Chinn & Grunden, 2013).

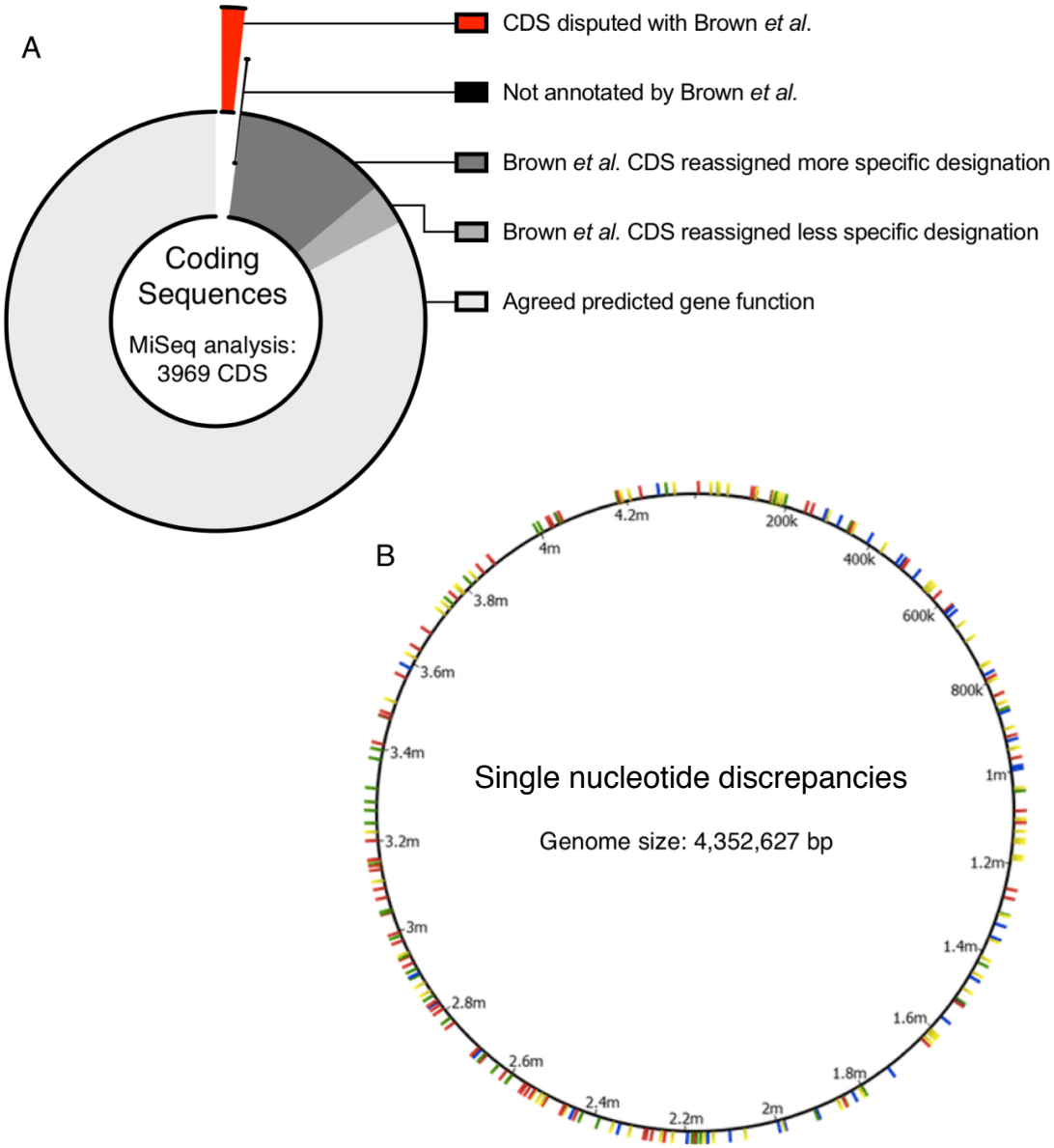


Figure 3.1: *C. autoethanogenum* sequence discrepancies with Brown *et al.* (2014)
A) CDS analysis of MiSeq sequencing data from this study showed a total of 3969 coding sequences, compared to the 4042 annotated by Brown *et al.* (2014) at the time of analysis. The 73 coding regions named by Brown *et al.* (2014) but disputed in this study displayed cut out in black. The remainder of the graph represents the 3969 coding sequences identified in this study, including 8 sequences not annotated by Brown *et al.* (2014), 482 reassigned a more specific function than that offered by Brown *et al.* (2014), and 73 assigned a less specific function than offered by Brown *et al.* (2014).
B) Distribution of 243 insertion sites across the 4352627 bp genome detected by Illumina sequencing in this study, compared to that of Brown *et al.* (2014). Adapted from published figure by Humphreys *et al.* (2015).

Analysis of the loci of the 242 single-nucleotide discrepancies showed they all appeared in both the Bruno-Barcena draft genome and the *C. ljungdahlii* genome. Such concurrence suggested errors in the sequencing published by Brown and colleagues (2014) rather than this study. As shown in Figure 3.1B, the location of these discrepancies appears to be evenly distributed around the genome, and closer analysis revealed the disputed nucleotides to almost exclusively appear in homopolymer regions.

Sanger sequencing performed by Christopher Humphreys was used to confirm the sequence of 10% of the homopolymer discrepancies, confirming their presence in all instances. Using the Sanger verified genome sequence, misidentified frame-shift mutations within the Brown and colleagues (2014) genome were re-designated. Analysis of these regions resulted in this study's more accurate calling of CDS regions, without the previously omitted nucleotides resulting in mistaken calling of stop codons and therefore inaccurate annotation of truncated or frame-shifted genes.

3.3. Restriction-modification based approach to DNA transfer optimisation

Early in this study, DNA transfer in *C. autoethanogenum* was established through conjugation, utilising the CA434 strain of *E. coli* (HB101 harbouring the R702 conjugative plasmid). Attempts were made to optimise this method with regards to ratio of donor to recipient, mating time, and growth stage of donor/recipient. However, the frequency of DNA transfer showed such variation between replicates, no useful conclusions could be drawn. The apparent stochastic nature of this conjugation was reflected in multiple attempts at conjugative transfer failing, despite success in near identical conditions.

Problems with DNA transfer resulted in a consideration of alternative acetogen chassis[15], whilst simultaneously attempting to develop more reproducible methods in *C. autoethanogenum*. The two main hypotheses for host interference with DNA transfer were; 1) RM system interference of incorrectly methylated plasmid DNA, and 2) host CRISPR system interference. The role of CRISPR systems is discussed in Chapter 4, with the role of RM systems discussed herein.

15 This is reflected in this study containing CRISPR tool development work on the acetogens *C. ljungdahlii* and *E. limosum*, as well as the solventogenic *C. acetobutylicum*.

3.3.1. *C. autoethanogenum* RM system analysis

PacBio sequencing data (complete with methylation data) of the *C. autoethanogenum* genome was submitted to NEB REBASE for RM system analysis. Figure 3.2 displays a schematic diagram modified from the results of this analysis, with the 10 genes of interest highlighted. These genes consist of four putative constituents of Type II RM systems, as well as six putative constituents of Type IV RM systems.

In order to determine host RM interference, two approaches were taken, as discussed in this chapter. Firstly, to assess Type IV RM system interference, the Sexpress strain of *E. coli* was developed by PhD student Craig Woods. This was achieved through conjugation of the R702 plasmid into NEB Express. This resulting *E. coli* strain has a Dam^{+}/Dcm^{-} methylation pattern, meaning at the N^{6} position of the adenine in the motif GATC, a methyl group is added. Theoretically, this would provide protection from putative *C. autoethanogenum* Type IV RM systems if they are active. Secondly, restriction endonucleases of putative Type II RM systems were analysed to assess their role in inhibition of DNA transfer.

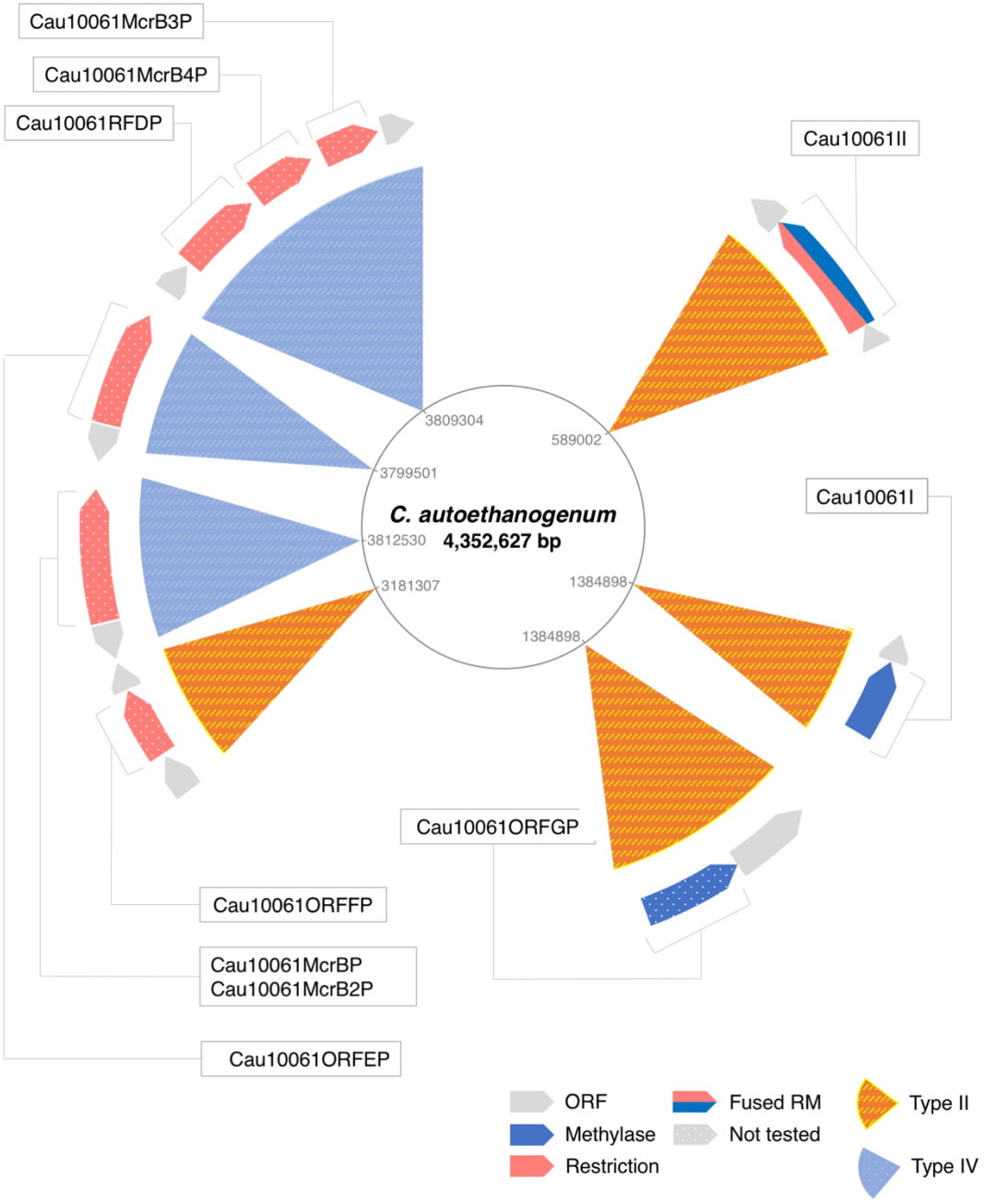


Figure 3.2: *C. autoethanogenum* DSM10061 RM systems
MiSeq data was provided to NEB REBASE (Restriction Enzyme Database) for analysis, resulting in the identification of four potential Type II, and six potential Type IV RM systems. The genomic location of 5' end of gene(s) in question are annotated on the circular map, with regions of interest exploded with either orange (Type II), or blue (Type IV) wedges. Details of these loci are shown with generic ORFs annotated in grey, predicted restriction endonucleases in red, and predicted methylases in blue. White dots are included in the endonucleases or methylases that are reported as untested by NEB, whereas solid colour represent those that have been tested. Based on figure available from REBASE (Roberts *et al.*, 2015)

3.3.2. Type IV restriction analysis

To assess the role played by Type IV restriction systems as barriers to conjugation, a large dataset was required for statistical significance. This was achieved through conjugation of a library plasmids containing the range Gram-positive replicons detailed in Table 3.1. The aim of this experiment was simultaneous comparison of transformation frequency for each of the replicons, along with direct comparison of CA434 and Sexpress strains of *E. coli* as conjugative donors.

The Gram-positive replicons generated in this study (pCB101 and pCACET) were constructed as follows. First, DNA extraction was performed on both *C. aceticum* for pCACET, and *C. butyricum* for pCB101. Following this, using the respective DNA extractions as templates, oligonucleotide primers CB101-F and CB101-R were used to amplify the replicon of pCB101, and CACET-F and CACET-R for pCACET. Primers were designed with overhang-encoded *Asc*I and *Fse*I sites at the 5' and 3' ends respectively, allowing cloning of the replicons into pMTL83151 linearised with the *Asc*I and *Fse*I. Replacement of the original pCB102 replicon in pMTL83151 with the new pCACET and pCACET replicons was confirmed via Sanger sequencing, using screening primer CatP-R1.

As shown in Figure 3.3 in all instances Sexpress provided significantly higher conjugation frequencies compared to CA434. Using CA434 as a donor resulted in 18 - 54 CFU mL^{-1} for the top performing replicons (pCB102, pCB102FS and p19), with these each of these seeing efficiency increases of >1000%. pCB101 and pCD6 each showed modest increases in CFU mL^{-1},

although both were still under 10 CFU mL^{-1}, whilst pCACET failed to produce any detectable colonies using CA434 as a donor strain. The largest efficiency gain was seen with pIP404, with >7400% increase to >2400 CFU mL^{-1}.

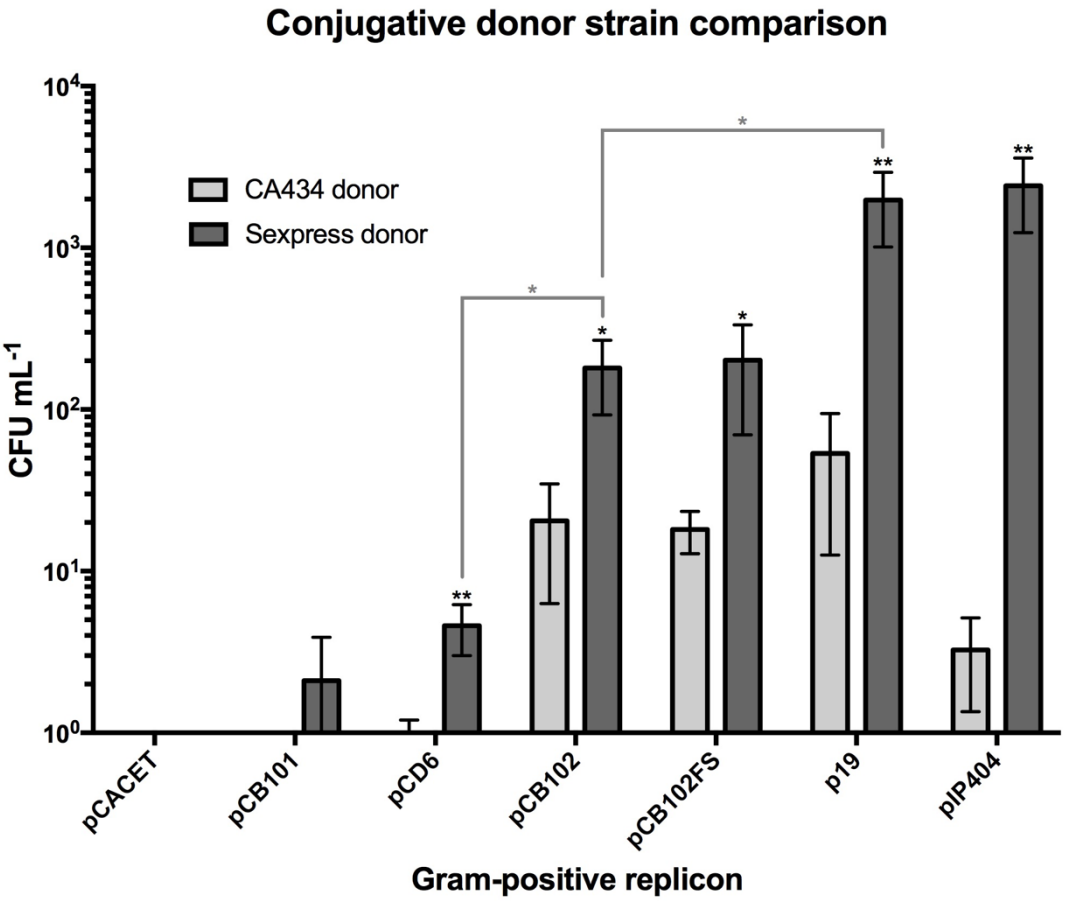


Figure 3.3: Comparison of Sexpress and CA434 conjugation frequencies
CFU mL^{-1} of each Gram-positive replicon (pCACET, pCB101, pCD6, pCB102, pCB102FS, p19, and pIP404) displayed for both CA434 (light grey bars) and Sexpress (dark grey bars) conjugative donor strains of *E. coli*. Standard deviation represented by black bars, with P value for significance of the difference observed between donors for each given replicon indicated as black astrisks above the Sexpress error bar. P value for significance between replicons represented by grey astrisks between replicon data sets. CFU mL^{-1} calculated as number of colony forming units per mL of recipient culture, whereby 200 μL of conjugative donor was used. P values calculated by unpaired T-test, $P \leq 0.05$ represented by one asterisk and $P \leq 0.01$ represented by two astrisks.

3.3.3. Type II RM system analysis

In order to further assess the role of RM systems in conjugation, the Type II systems identified via NEB REBASE analysis were assessed for their contribution to interference of DNA transfer. REBASE analysis identified 4 potential systems, as displayed in Figure 3.2, these are annotated as Cau10061II, Cau10061I, Cau10061RFGP and Cau10061RFFP.

Initially Cau10061II was identified as a potential target due to it containing the only putative endonuclease identified by REBASE. This Type IIG endonuclease (*cauII*), is believed to have bifunctional RM activity, resulting in the performing of both methylation, and restriction activities. Subsequent manual *in silico* analysis of the remaining putative Type II systems resulting in the identification of two further potential endonucleases.

3.3.3.1. *In silico* analysis of Type II RM systems

Aside from *cauII*, no putative endonucleases were identified via REBASE analysis, with only putative methylases identified for the remaining Type II systems. Due to Type II systems generally containing methytranferases and endonucleases in close proximity (Roberts *et al.*, 2015), manual *in silico* analysis was performed on the genes flanking the putative methylases identified by REBASE.

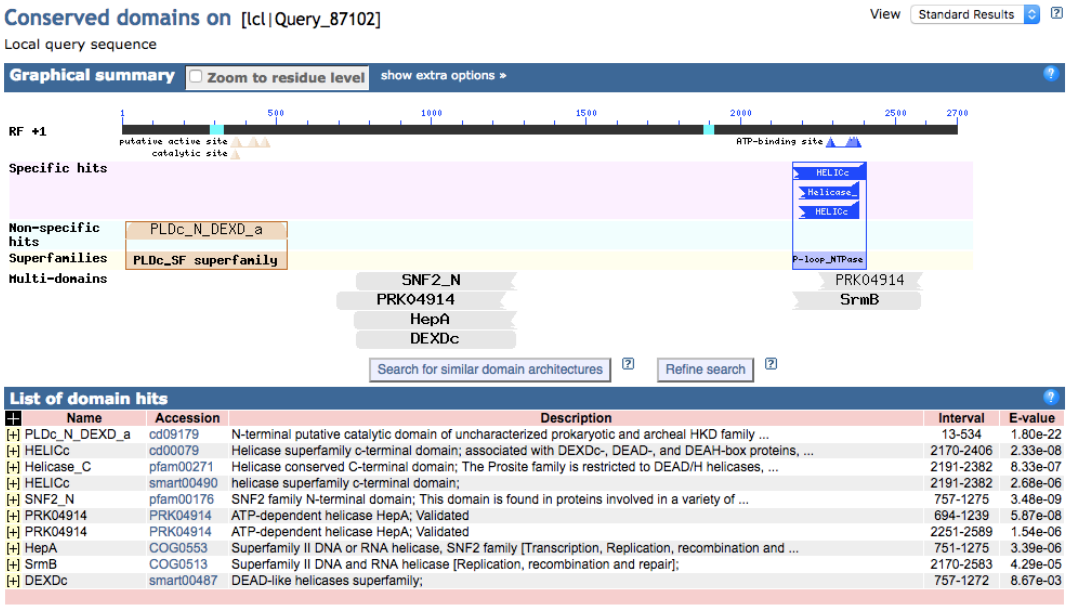
3.3.3.1.1. Cau100611I

The Cau10061I putative methyltransferase, CLAU_1258, was analysed and showed significant similarity to Type II restriction endonucleases, however, no endonucleic domain could be identified. The proximal genes to CLAU_1258 also showed no domains associated with endonuclease activity and therefore

no candidate genes from this putative RM system were selected for knockout analysis.

3.3.3.1.2. Cau10061ORFGP (*cauIII*)

REBASE analysis identified CLAU_1608 (anti-sense) as a putative methytransferase. Analysis of the proximal downstream gene CLAU_1607 identified it as a potential endonuclease, with the presence of a DExD domain, as illustrated in Figure 3.4. The domain is associated with prokaryotic HKD family nucleases, which are fused to DEAD/DEAH box helicase domain, a feature also found in CLAU1607. Due to the presence of such domains, CLAU_1607 was designated *cauIII*.



Name	Accession	Description	Interval	E-value
[+] PLDc_N_DEXD_a	cd09179	N-terminal putative catalytic domain of uncharacterized prokaryotic and archeal HKD family ...	13-534	1.80e-22
[+] HELICc	cd00079	Helicase superfamily c-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins, ...	2170-2406	2.33e-08
[+] Helicase_C	pfam00271	Helicase conserved C-terminal domain; The Prosite family is restricted to DEAD/H helicases, ...	2191-2382	8.33e-07
[+] HELICc	smart00490	helicase superfamily c-terminal domain;	2191-2382	2.68e-06
[+] SNF2_N	pfam00176	SNF2 family N-terminal domain; This domain is found in proteins involved in a variety of ...	757-1275	3.48e-09
[+] PRK04914	PRK04914	ATP-dependent helicase HepA; Validated	694-1239	5.87e-08
[+] PRK04914	PRK04914	ATP-dependent helicase HepA; Validated	2251-2589	1.54e-06
[+] HepA	COG0553	Superfamily II DNA or RNA helicase, SNF2 family [Transcription, Replication, recombination and ...	751-1275	3.39e-06
[+] SrmB	COG0513	Superfamily II DNA and RNA helicase [Replication, recombination and repair];	2170-2583	4.29e-05
[+] DEXDc	smart00487	DEAD-like helicases superfamily;	757-1272	8.67e-03

Figure 3.4: BLASTX search results of CLAU_1607 (*cauIII*)
BLASTX results showing high sequence identity of *cauIII* to HKD family nucleases, specifically the PLDx N DEXD domain and helicase domains consistent in such families.

3.3.3.1.3. Cau10061ORFFP (*cauIV*)

REBASE analysis identified CLAU_2858 as the potential methyltransferase of the putative Cau10061ORFFP RM system. *In silico* analysis of flanking genes provided no potential endonucleic domains, however, the CLAU_2858 gene itself shows sequence identity to the *Dpn*II superfamily domain. Again, the presence of this domain led to the designation of CLAU_2858 as *cauIV* and knockout analysis was attempted. Figure 3.5 illustrates the sequence identity and location of the *Dpn*II superfamily domain in *cauIV*.

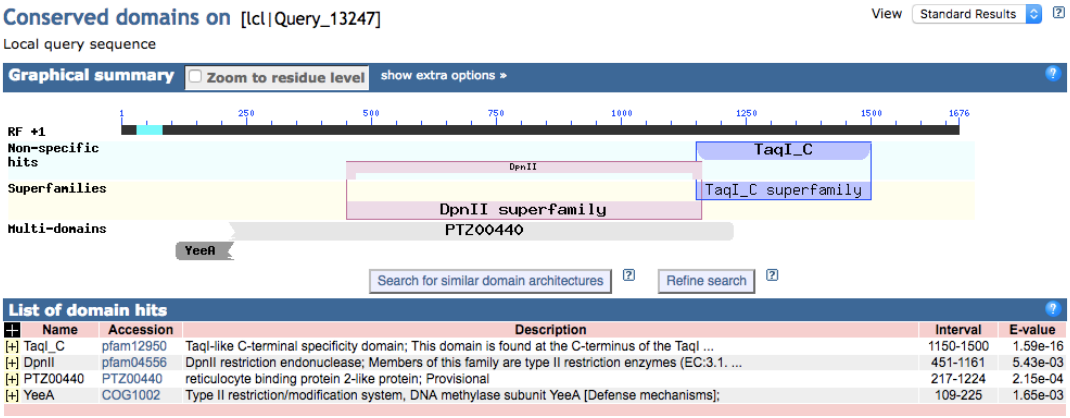


Figure 3.5: BLASTX search results of CLAU_2858 (*cauIV*)
BLASTX results showing high sequence identity of *cauIV* to *Dpn*II superfamily restriction endonuclease domain.

3.3.3.2. Type II RM knockout generation and strain analysis

In this study, the *cauII* endonuclease identified by REBASE analysis was deleted to assess its role in DNA transfer interference. No knockouts were attempted for either of the putative endonucleases (*cauIII* and *casIV*). Knockout generation of *cauII* was performed using a CRISPR-Cas9 approach, the development and further details of which are discussed extensively in Chapter 4.

To generate the pMTL43151-CLAU-IIG vector used for this knockout, the cloning protocol outlined in Chapter 2 was used. The homology cassette was generated via SOEing PCR using primers CLAU-IIG-LHA-F-MauBI, IIG-LHA-R, IIG-RHA-F and IIG-RHA-R-AscI, and cloned into the pMTL43151-CLAU-pyrE vector using *Mau*BI and *Asc*I restriction sites. Once homology arm insertion was sequence confirmed using pCB102-R, the oligonucleotide CLAU-IIG-sg1-HiFi was used to complete the vector via the NEB HiFi ssDNA method outlined in Chapter 2. Once confirmed via Sanger sequencing using pMTL43151-CLAU-sgRNA-scr-R, the finalised vector was transformed into the conjugative door strain *E. coli* Sexpress.

Following conjugation with pMTL43151-CLAU-IIG, *cauII* deletion was confirmed via colony PCR (as shown in Figure 3.6) and Sanger sequencing. Following sequence confirmation, plasmid loss was performed and the strain stored at -80 °C.

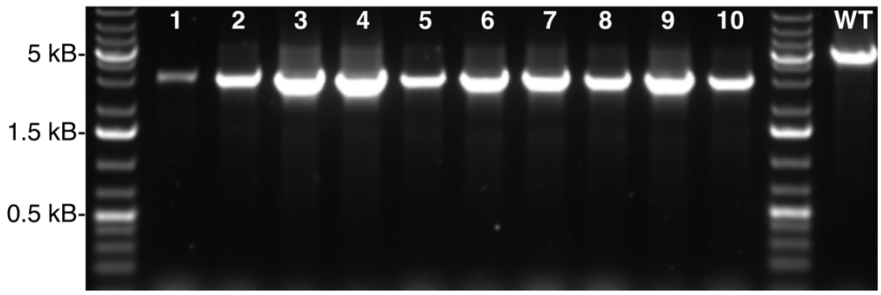


Figure 3.6: Δ*caulI* colony PCR
Lanes 1-10 represent the 10 colonies screened from YTF plates supplemented with D-cycloserine and thiamphenicol, with "WT" lane representing wild type *C. autoethanogenum*. Each of the colonies show and sizes consistent with deletion of the *caulI* gene (~3 kB), with the wild type band showing the predicted 5 kB size of the original *caulI* locus. Oligonucleotide primers for this PCR were "IIG-scr-F" and "IIG-scr-R". See Figure 4.1 for schematic explanation of colony PCR.

To assess the impact of *caulI* deletion, conjugation frequency of the *caulI* deficient *C. autoethanogenum* strain (ΔIIG) was compared to its wild type progenitor strain, performed in technical and biological triplicate with a range of Gram-positive replicons. For each of which, a statistically significant increase in conjugation efficiency was confirmed. This is illustrated in Figure 3.7.

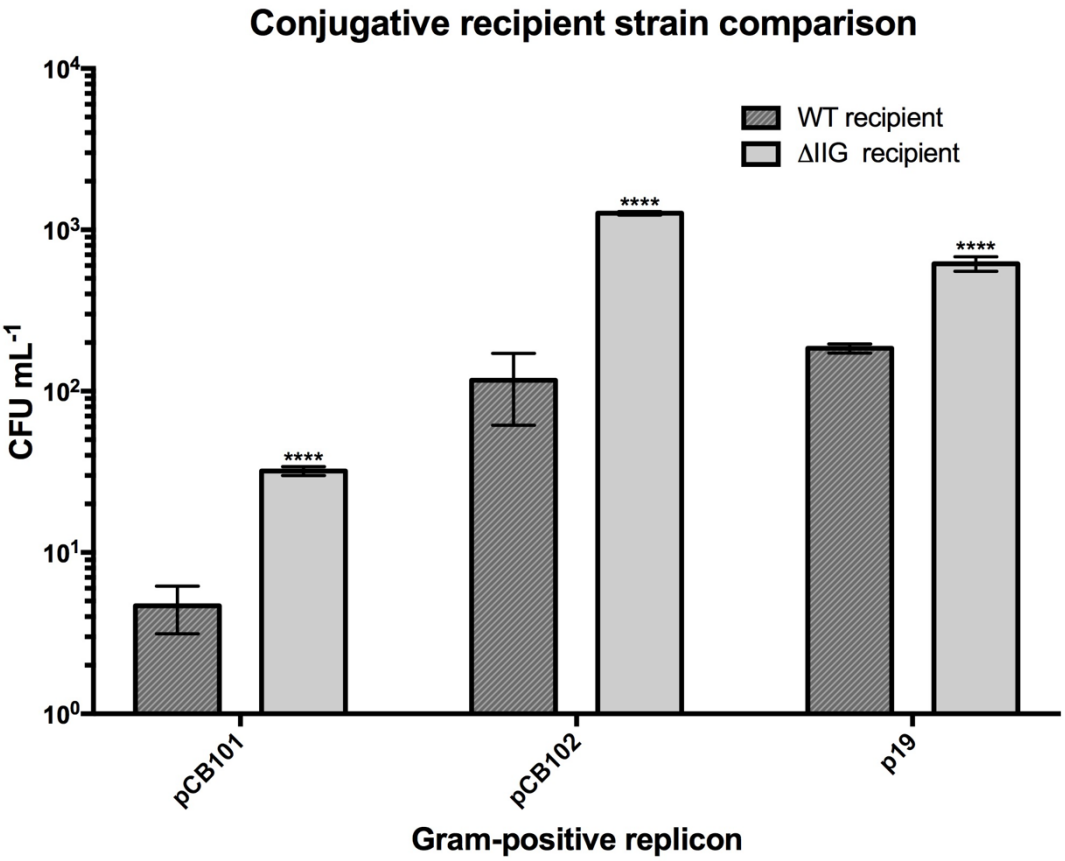


Figure 3.7: Comparison of wild type and Δ*caulI* conjugation efficiencies
CFU mL^{-1} of each Gram-positive replicon (pCB101, pCB102, p19) displayed for both WT (dark grey bars) and ΔIIG (light grey bars) recipient strains of *C. autoethanogenum*. Standard deviation represented by black bars, with P value for significance of the difference observed between recipients for each given replicon indicated as black asterisks above the ΔIIG error bar.
CFU mL^{-1} calculated as number of colony forming units per mL of recipient culture, whereby 200 μL of conjugative donor was used. P values calculated by unpaired T-test, $P \leq 0.0001$ represented by four asterisks.

3.4. *pyrE* deficient strain generation

As described by Minton *et al*. (2016), generation of a *pyrE* deficient strain is intrinsic to the roadmap of clostridial strain development for synthetic biology. This is due to its involvement in both allelic exchange and ACE methodologies. Although these methods have been to some extent eclipsed by CRISPR, at the beginning of this study (and for other researchers working on *C. autoethanogenum* prior to rollout of the pMTL40000 vector series), development of a Δ*pyrE* strain was essential.

The *pyrE* deletion vector, pMTL-CLAU-del-pyrE, was developed using the plasmid pMTL-AMH101 (Liew *et al.,* 2017) as a backbone. The backbone consisted of a pCD6 Gram-positive replicon, CatP module, ColE1 + *tra* Gram-negative replicon module and a *C. ljungdahlii pyrE* truncation cassette. The latter was using *Sac*II and *Asc*I restriction sites with a *C. autoethanogenum pyrE* knockout cassette generated via SOEing PCR with primers CLAU-del-pyrE-LHA-F, CLAU-del-pyrE-LHA-R, CLAU-del-pyrE-RHA-F and CLAU-del-pyrE-RHA-R. Figure 3.8 illustrates the pMTL-CLAU-del-pyrE vector, showing the homology arms for this method are asymmetric, with the LHA consisting of ~0.5 kB and the RHA ~1.2 kB. Once confirmed via Sanger sequencing, the completed vector was transformed into *E. coli* CA434, which was used as a conjugative donor for conjugation into wild type *C. autoethanogenum*.

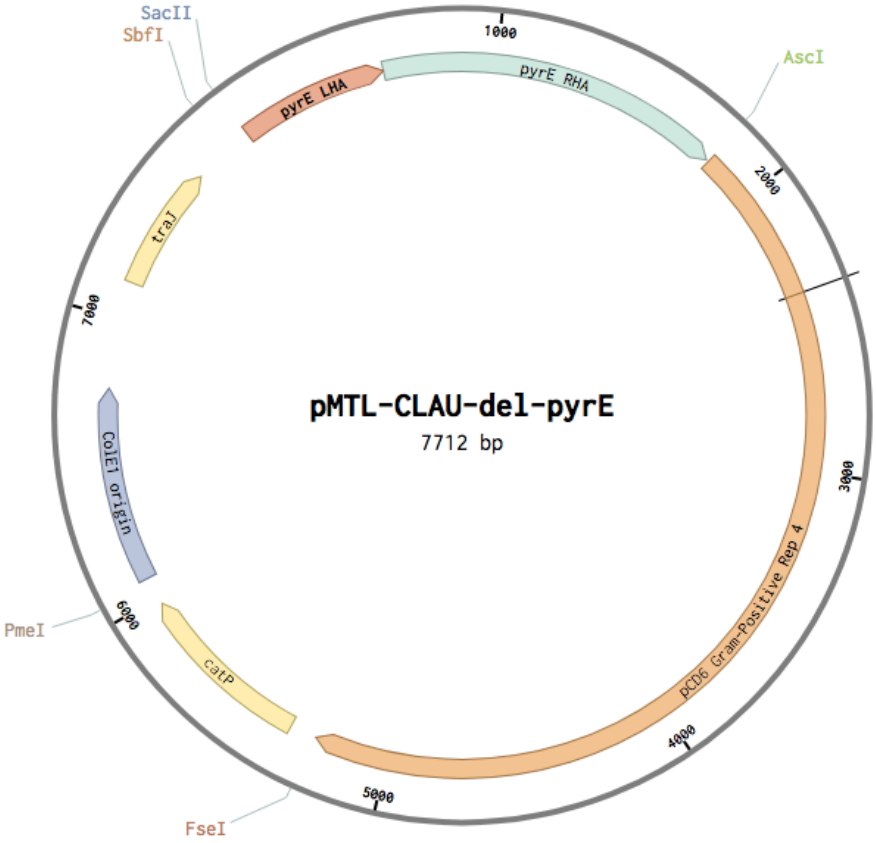


Figure 3.8: pMTL-CLAU-del-pyrE plasmid map
Map displays the 7712 bp plasmid consisting of a pCD6 Gram-positive replicon, chloramphenicol resistance encoding *catP*, ColE1 Gram-negative replicon with *traJ* to enable conjugative transfer, shorter (~0.5 kB) *C. autoethanogenum pyrE* left homology arm (LHA), and longer (~1.2 kB) *pyrE* right homology arm (RHA). As a pMTL80000 vector, modular components are separated by *Fse*I, *Pme*I, *Sbf*I and *Asc*I restriction sites, with knockout cassette flanked by *Sac*II and *Asc*I.

Following conjugation, potential *pyrE* knockouts were screened for as described in Chapter 3, using primers CLAU-pyrE-internal-F and CLAU-ACE-pyrE-scr-R for colony PCR, resulting in one sequence confirmed Δ*pyrE* mutant (*C. autoethanogenum* Δ*pyrE* PR strain). In order to mitigate against SNP accumulation in the Δ*pyrE* strain used by the research group, several researchers independently generated Δ*pyrE* strains. These strains were then sent for next-generation sequencing and analysed by Dr Chris Humphreys. A

Figure 3.9A shows the sequence confirmation of the 3' truncation of the *pyrE* gene in strain *C. autoethanogenum* Δ*pyrE* PR strain, however, as shown in Figure 3.9, this strain accumulated a 15 nt insertion at position 2570061-2570062 of the reference genome and a SNP at position 1743960.

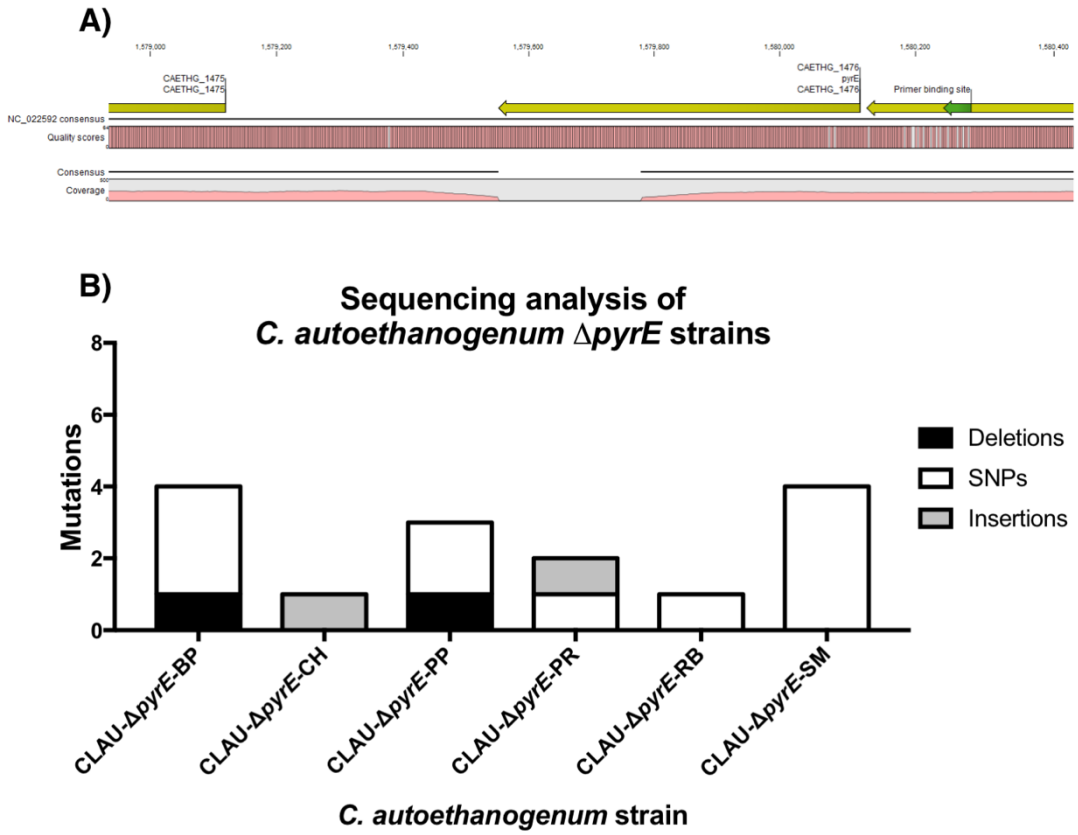


Figure 3.9: *C. autoethanogenum* Δ*pyrE* strain analysis
A) Screenshot of PacBio sequencing data showing truncation of the 3' end of *pyrE* in *C. autoethanogenum* Δ*pyrE* PR strain. Genomic location is indicated on the top axis, with gene loci and primer binding site identified in yellow and green arrows (respectively) below. Consensus sequence and sequence coverage can be seen at the bottom of the image. B) Comparison of mutations in *C. autoethanogenum* strains developed by researchers as follows; Bart Pander (Δ*pyrE* BP), Dr Chris Humphreys (Δ*pyrE* CH), Pawel Piatek (Δ*pyrE* PP), Peter Rowe (Δ*pyrE* PR), Ronja Brietkopf (Δ*pyrE* RB) and Dr Samantha McClean (Δ*pyrE* SM). Cumulative number of mutations present in each strain is shown by total bar size, and type of mutations present are represented by colours within the bars. Black represents deletions, white represents SNPs, and grey represents insertions.

Comparison of all Δ*pyrE* strains generated identified *C. autoethanogenum* Δ*pyrE* RB (generated by Ronja Breitkopf) as the best candidate strain for subsequent work, due to it containing the least deleterious mutations. Allelic exchange attempts to generate *ack* and *pta* mutants (see Chapter 6) used this strain.

3.5. Further pCB101and pCACET replicon analysis

No futher analysis of the new replicons modularised in this study was conducted, due to pCB101 showing low conjugation efficiency, and pCACET failing to show any transconjugants.

3.6. Electroporation

DNA transfer via electroporation offers some benefits over conjugation. For many clostridial species, electocompetent cells can be prepared and stored at -80 °C prior to use. This enables immediate availability of DNA transfer, rather than the pre-culture growing at the correct density that is required for conjugation. Electroporation also requires less space in anaerobic workstations, due to its lack of requirement for mating plates compared to conjugation. For these reasons, avenues for electroporation of *C. autoethanogenum* were explored.

During attempts to optimise conjugation, the *C. ljungdahlii* electroporation protocol described in Chapter 2 was used to prepare electrocompetent cells for *C. autoethanogenum*. These cells were transformed (as described for *C. ljungdahlii* in Chapter 2) with a cocktail of pMTL80000 plasmids, but yielded no observable transformants, despite several attempts.

Attempts at electroporation with this protocol were repeated using the ΔIIG strain described in section 3.3, but again yielded no success. Using a different RM based approach, PhD student Craig Woods attempted electroporation using plasmids purified from an *E. coli* strain mimicking the methylation pattern of *C. autoethanogenum,* however, this again showed no success.

3.7. RBS and promoter comparison

For metabolic engineering, a library of promoter and RBS sequences of known expression strength is invaluable for controlled, tuneable expression of genes of interest. A collaborative effort was put in place to develop such a library, with the RBS library developed and characterised in this study, and the promoter library characterised in *C. autoethanogenum* by PhD candidate Bart Pander.

3.7.1. RBS library generation and analysis

A random selection of RBS sequences from upstream regions of *C. autoethanogenum* genes were selected and analysed against the known rRNA sequence. A range of 15 RBS sequences with ranging similarities to the rRNA sequence were selected, as displayed in Table 3.3, which and numbered according to descending sequence identity.

The sequences were introduced between *C. acetobutylicum* P_{thl} and the *catP* reporter gene, through amplification of the promoter region of the plasmid using the oligonucleotide primers CLAU-RBS-F (1-15) and CLAU-RBS-R, with each reverse primer introducing its associated RBS sequence. Once amplified, the PCR product was digested using *Sac*II and *Nde*I, and ligated into the pMTL82254 plasmid backbone (also digested with the same

restriction endonucleases). Replacement of the existing RBS sequence in the parent vector with the new oligonucleotide encoded RBS sequences was confirmed via Sanger sequencing using CatP-R1, and the sequence verified vectors were transformed into *E. coli* CA434 for conjugation into *C. autoethanogenum*.

Table 3.3: RBS sequences for study in *C. autoethanogenum*
RBS sequences for analysis in *C. autoethanogenum*, start codons shaded in grey. RBS identifiers are provided, along with the *C. autoethanogenum* gene with which they are associated, with "Cac thl" derived from the *C. acetobutylicum* P_{thl} RBS present of the pMTL82254 vector.

RBS	Associated gene	Sequence
1	CLAU 2296	CAAACTACAAAAAGGAGGTGTTCATATG
2	*rplO*	TAATACAAAGGAGGTGTAAGTACATATG
3	*rplK*	TAACCACATATAGGAGGTGTATCATATG
4	*rhaB*	GTTCTAAAAAGAGGTGTTAAATCATATG
5	CLAU 2414	TTTTATCATACAAGGAGGTTTACATATG
6	CLAU 3014	AGAAAGGAGTTGCCTATATATTCATATG
7	CLAU 0011	CATGATAAGGAGGTAATTTAAACATATG
8	*fabH*	AAAATTGGATCAAAGGAGAAGACATATG
9	CLAU 1812	TATGAGGAGTTGATTAATACAACATATG
10	CLAU 2282	TAAGATGTGATTAAGGAAGGTTCATATG
11	*pyrE*	AGAGGAATAATTTAGGAGGACACATATG
12	CLAU 1740	AAAATAATGGGGAGGGTTAAAACATATG
13	*pth*	ATAGTATTGAGATTGGAGTGAACATATG
14	CLAU 0410	ATGCTCAGATTCCGGGATATAACATATG
15	CLAU 0818	TATATTTTACAATTCCTGTAAATCATTG
Cac thl	pMTL82254	ATCAAAATTTAGGAGGTTAGTTCCATTG

Once transformed into *C. autoethanogenum*, the plasmid harbouring strains of *C. autoethanogenum* were designated names corresponding to their plasmid (RBS1-15) and grown in selective liquid YTF media to an OD_{600} of 0.8. At this point, 1 mL of culture was centrifuged at 10,000 rcf for one minute and the supernatant removed and stored at -20 °C. Supernatant samples were prepared for the CatP assay as described in section 2.1, with the assay performed as stated. Relative expression values for each RBS are displayed in Figure 3.10.

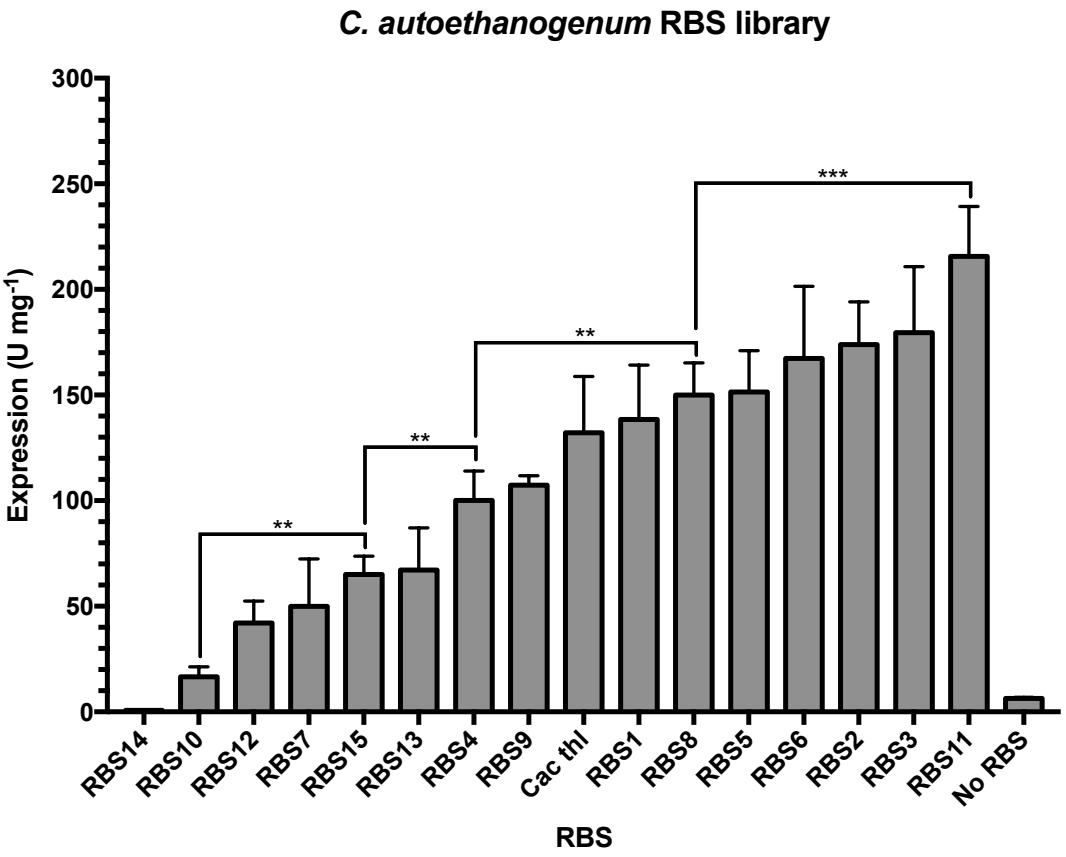


Figure 3.10: *C. autoethanogenum* RBS library analysis
Expression of CatP from vector library consisting of pMTL82254 backbone with C. acetobutylicum Pthl driving catP and the original Pthl RBS substituted for the RBS sequences detailed in Table 3.3. Expression is given as CatP units per mg of total protein. Statistical significance of ≤ 0.01 represented by two asterisks and ≤ 0.001 represented by three asterisks.

The library of RBS sequences provided a range of CatP expression values, which are displayed in Figure 3.10 in order of expression. No expression was seen from CLAU 0410 derived RBS14. The remainder of the library providing five statistically significant tiers of expression, with the following RBS sequences listed in order of strength; RBS 10 (CLAU 2282), RBS 15 (CLAU 0818), RBS 4 (*rhaB*), RBS 8 (*fabH*), RBS 11 (*pyrE*). Data from promoter comparison generated by Bart Pander can be seen in Appendix B.

3.8. Conclusion

The first section of this chapter described the manual annotation of the *C. autoethanogenum* genome. Here, marked improvements were shown in the detail of annotation available compared to the previously published genome (Brown *et al*., 2014), with a handful of previously unknown CDS regions identified and more detailed likely gene functions described for many genes.

As well as improvement of gene annotation, the sequence data also highlighted key single-nucleotide sequence discrepancies with the genome described by Brown and colleagues (2014). Here it was shown that the length of homopolymer regions differed from Illumina-generated sequence data generated in this study. These differences were analysed by comparison to a previously published unclosed genome of *C. autoethanogenum* (Bruno-barcena, Chinn & Grunden, 2013), the published genome sequence of *C. ljungdahlii* (Kopke *et al*., 2010), and Sanger sequencing of the specific loci in question. It was concluded that as these sources corroborated the sequence data generated as described in this study, suggesting the algorithm used in the study of Brown and colleagues (2014) to call homopolymer regions was mis-calibrated. This conclusion was reached due to the even distribution of these regions throughout the genome, and the fact that of the 10% randomly analysed, all confirmed by Sanger sequencing.

Establishment of this more accurately annotated genome provided the foundation of much of the work described in this study. This genome was used as reference for all genome editing targets, as well as for design of

oligonucleotides for amplification from the chromosome. Further to this, *in silico* analysis of the endogenous CRISPR system described in Chapter 5 was conducted using this more accurate sequence.

Although the improvement in genome annotation provided a solid platform for future work in *C. autoethanogenum*, the issue of DNA transfer problems described in section 3.3 provided a far bigger problem. Inconsistent and often failed DNA transfer via the *E. coli* CA434 conjugal donor strain resulted in severely hampered progress for the project. As described in Chapter 5, the knockout of acetogenesis genes was an early objective of this project. However, transfer of genetic tools into *C. autoethanogenum* to achieve this goal proved difficult. Before progress could be made in any meaningful metabolic engineering, an RM system based approach for solving DNA transfer was attempted.

Modification of plasmid methylation has been seen in previously developed techniques for the clostridial genus. Notably, use of the pAN2 plasmid for methylation of shuttle vectors for use in *C. acetobutylicum*. This technique relies upon a *B. subtilis* phage-derived methyltransferase present on the pAN2 plasmid, transformation of *E. coli* harbouring pAN2 with the shuttle vector means that when plasmid DNA is extracted from this duel-plasmid harbouring strain, the shuttle vector is methylated in a manner that protects it from *Cac*8241 DNA restriction activity. Although similar, the methylation approach used for Sexpress as part of this study differs in two vital ways; firstly it is provided in the context of a conjugative donor strain of *E. coli*, and secondly it designed to protect from a Type IV RM system, rather than Type II.

As described by Purdy and colleagues (2002), the CA434 strain of *E. coli* has been used as a conjugative donor across in the clostridial genus, including *C. difficile*, *C. sporogenes* and *C. botulinum*. This strain was initially used for DNA transfer in *C. autoethanogenum* due to difficulties establishing a functional electroporation protocol.

The pAN2 vector protects from Type II restriction endonucleases by addition of methyl groups to bases present in the recognition site of said endonucleases. This methylation interferes with endonuclease activity, and therefore prevents cleavage of the shuttle vectors. However, the Type IV systems for which Sexpress provides protection against in *C. autoethanogenum* differ in that they recognise incorrectly methylated DNA. This difference means that use of a methyltransferase could in fact hinder DNA transfer, as methylation of a previously un-methylated base within a Type IV restriction endonuclease's cleavage site could mark it for destruction.

The Sexpress strain of *E. coli* was developed by Craig Woods by transferring the R702 plasmid to NEB Express. Unlike HB101 which has a Dam^{+}/Dcm^{+} methylation pattern, NEB Express has a Dam^{+}/Dcm^{-} methylation pattern. This difference means that Sexpress, unlike CA434, does not methylate at the C5 position of the second cytosine in the sequences CCAGG and CCTGG. Results from this study conclusively show that this change in donor improves conjugation frequencies for all Gram-positive replicons tested (baring pCACET, which failed to result in any transconjugants). As well as improving conjugation frequencies for pCD6, pCB102, pCB102FS, p19 and pIP404, the

use of Sexpress showed that pCB101 was a viable replicon in *C. autoethanogenum*, whereas no transconjugants were observed when using CA434. These results show the Dam^+/Dcm^- methylation pattern of Sexpress, compared to the Dam^+/Dcm^+ pattern of CA434 improves conjugation efficiency significantly. As a result, all conjugations performed after this study were undertaken using Sexpress as a donor.

The improvement in conjugation efficiency seen in the use of *E. coli* Sexpress enabled the development and implementation of CRISPR-based tools described in Chapter 4. Prior to use of Sexpress, CRISPR-Cas9 plasmids described in this study failed to conjugate. This is most likely due to the size of Cas9 vectors (>10 kB), combined with the potentially deleterious effects of Cas9 (further discussed in Chapter 4). Both these factors are likely to have a negative impact on conjugation frequency, however, the use of Sexpress enabled the conjugation of such plasmids.

Further to Type IV RM analysis, this study looked to improve conjugation efficiencies through Type II RM engineering. Using the pMTL40000 vector system described in Chapter 4, the gene encoding the *Cau*II restriction endonuclease was targeted for deletion. In a similar strategy, the deletion of genes coding for putative Type II restriction endonucleases has previously been performed in *C. acetobutylicum*. Again, this example comes from the *Cac*8241I RM system, but rather than providing protection from restriction via pAN2, the organism's ability to restrict DNA was removed through ClosTron mutagenesis (Dong *et al.*, 2010). As detailed in Chapter 1, this method relies upon disruption of the gene through an intron containing the *ermB* gene.

Expression of the *ermB* gene is only possible if the intron has inserted into the genome, and therefore erythromycin resistance is used as positive selection marker. Although this research shows that the *Cac*8241I gene does appear to be responsible for DNA transfer interference, it also means that ClosTron cannot be used for other in the *Cac*8241I deficient strain due to lack of selection pressure. The result of this study, therefore, shows that deletion of the endonuclease improved transformation frequency, but results in a strain unable to be used for subsequent knockout analysis using ClosTron. In contrast, use of the pMTL40000 system for clean, in-frame deletion of *Cau*II provided the same increase in DNA transfer frequency, but crucially offered no barrier to subsequent genome editing.

Further differences to the Dong *et al*. (2010) study should also be noted in that *Cau*II is putative Type IIG restriction endonuclease, meaning that is theoretically provides methyltransferase and restriction endonuclease activity. Due to the known impact of widespread changes in methylation patterns to the genome, growth experiments of the ΔIIG strain were conducted. This study showed no further of use of the ΔIIG strain, due to its full analysis being completed after other knockouts were made. However, it may well prove useful in the future for conjugation of larger vectors, or those proving problematic for conjugation.

This study has shown that greater understanding of host RM systems, and subsequent modification of donor methylation profile, or recipient endonuclease activity both provide improved conjugation frequency. Similar approaches are currently being attempted in members of the genus that have

thus far seen a lack of DNA transfer, or low DNA transfer, proving inhibitive to synthetic biology approaches.

Further to enabling CRISPR-Cas9 genome editing, Sexpress also enabled the successful conjugation of the previously unused pCB101 Gram-positive replicon from *C. butyricum*, something that was not possible using CA434. Analysis of pCB101, along with the existing range of Gram-positive replicons provided interesting insight into plasmid conjugation frequencies, stability and copy number. This information formed the foundation for development of the pMTL40000 vectors used in *C. autoethanogenum*, as described in Chapter 4. For these, the pCB102 Gram-positive replicon was used due to its high conjugation frequency. Although the pIP404 and p19 replicons both showed higher conjugation frequencies, both showed significant trade-offs. Initially pIP404 was used for pMTL43151 work in *C. autoethanogenum*, however this proved problematic due to colony size being extremely small, regardless of length of incubation. Plasmid copy number analysis provided some potential insight for this, with the low copy number perhaps limiting colony size. Regardless of cause, this trait meant pIP404 was unsuitable for this application, and insight gained from the conjugation frequency experiment provided the selection of the next best candidate.

The p19 Gram-positive replicon has been extensively characterised by PhD candidate James Millard in many clostridia. Although this replicon showed high conjugation frequencies, data from other organisms suggested it would be unsuitable due to high stability in all organisms it was tested. Without an unstable replicon, achievement of plasmid loss can prove difficult. As such,

use of a stable replicon for genome editing purposes would make it hard to make subsequent genome modifications. With specific regards to acetogens, the time penalty associated with the use of a stable replicon would be amplified by the slow growth rates described in Chapter 1. As a result, the unstable pCB102 was used for subsequent pMTL40000 vectors used in *C. autoethanogenum*, and went on to be consistently used for all knockouts made in this study using the pMTL40000 system.

With regards to other Gram-positive replicons, this study found limited use for replicons other than pCB102, however their characterisation should provide a platform for future researchers to select the most appropriate for their intended study. For example, the low copy number and high stability of p19 would provide a good candidate for plasmid-based expression of genes in gas fermentation.

On theme with foundational work providing little further use, development of the *pyrE* deficient strain of *C. autoethanogenum* found no further use in this study. Although at the time of making this mutant, it was viewed as integral to generation of clean in-frame deletion based on *pyrE* counter selection (see Chapter 1), development of the pMTL40000 system meant this was not the case. Despite this, other researchers on the GASCHEM project used the strain created by PhD candidate Ronja Breitkopf to generate knockouts and for expression of heterologous genes from the *pyrE* locus.

As described by Fisher and colleagues (2014), fine-tuned control of gene expression is essential in synthetic biology. Prior to this study, no readily available information on RBS or promoter activity was available. This study

has provided much need data for the former, and research by Bart Pander for that latter. Here, RBS 10 (derived from the upstream region of CLAU 2282) provided the lowest detectable expression, and RBS 11 (derived from the upstream region of *C. autoethanogenum pyrE*) provided the highest expression. With regards to promoters, a range of promoters were identified with varying levels of expression. With regards to this study, these data provided the required information to provide appropriate promoters for work discussed in Chapters 4 and 5.

In conclusion, the foundational work described in this chapter has provided a platform for improved research in *C. autoethanogenum*. The collaborative generation of multiple *pyrE* strains resulted in the identification of one with no obvious deleterious mutants. Although this strain was not used in this study, it was crucial for further work in the GASCHEM project from which this study belongs. The reason this study did not use the *pyrE* based genetic engineering techniques described in Chapter 1 is because of work developing the pMTL40000 system described in Chapter 4, development of which was significantly aided by characterisation of Gram-positive replicons described in this chapter. Further to this, the improved DNA transfer frequencies observed through use of Sexpress enabled successful conjugation of the pMTL40000 vectors, and use of this conjugal donor along with the ΔIIG strain will allow high DNA transfer capabilities in future projects.

CHAPTER FOUR

DEVELOPMENT OF EXOGENOUS CRISPR SYSTEMS FOR GENOME EDITING IN CLOSTRIDIA

4. Development of exogenous CRISPR systems for genome editing in clostridia

4.1. Introduction

As discussed extensively in Chapter 1, CRISPR-based technologies offer marked improvements over the conventional technologies available in clostridia at the beginning of this study. Initially, genes involved in the acetate pathway in *C. autoethanogenum* (*pta* and *ack*) were targeted for disruption with both ClosTron and ACE, however, as described further in Chapter 6, neither technology was sufficient in generating either mutant strain. As a result, the development of a CRISPR-Cas9 system for use in acetogens became a new focus of research.

A handful of successful applications of CRISPR-Cas9 technology in the genus *Clostridium* have been published to date, but at the beginning of this study no such work had been shown. In 2015 exemplification of this technology was shown in *C. beijerinckii* (Wang *et al.*, 2015). Following this, success was seen in *C. acetobutylicum, C. ljungdahlii* (as part of this study), and *C. pasteurianum* (Li *et al.*, 2016; Pyne *et al.*, 2016b). As a result of this study, success has been achieved in *C. autoethanogenum*, with the resulting pMTL40000 vector series being used across many organisms within the genus. Using a different system, Nagaraju *et al*. have published research showing CRISPR-Cas9 knockouts in *C. autoethanogenum*, however this method appears to be less efficient and prone gene truncation rather than generation of clean in-frame deletions (Nagaraju *et al.*, 2016).

4.2. Early iterations

Initial attempts at CRISPR-Cas9 cloning relied upon a two-plasmid system, with “pCRISPR” harbouring a codon-optimised Cas9 and “pTarget” containing the sgRNA and homology arms required for target knockout. This system was designed to enable the generation of a strain harbouring the non-specific pCRISPR plasmid, with introduction of the target-specifying pTarget later. This approach was attempted due to issues with low conjugation efficiencies described in Chapter 3. It was hypothesised that the large size of CRISPR-Cas9 plasmids combined with Cas9 nuclease activity would be too inhibitory for practical use.

Unfortunately, attempts at knocking out *pyrE* as a proof of concept resulted in no success. Introduction of pTarget_pyrE to the pCRISPR strain of *C. autoethanogenum* yielded no colonies. This suggested either successful nuclease activity was occurring without sufficient homologous recombination events, or that conjugation issues were still inhibitory. Since moving away from this two-plasmid system, a similar approach has been published for genome editing in *C. acetobutylicum* (Wasels *et al.*, 2017).

4.3. Proof of concept

After initial work on developing a two-plasmid system, a collaborative effort with the Shanghai Institute of Biosciences (SIBS) resulted in development of a functional system in *C. ljungdahlii* which was subsequently modularised for use across the *Clostridium* genus (Huang *et al.*, 2016). This work utilised a system designed for use in *E. coli* by SIBS, and resulted in a single-plasmid system developed using the pMTL80000 vector series. Using pMTL83151 as a backbone, Cas9, sgRNA, and a homology cassette were inserted in the application specific module. This modularised plasmid series, pMTL40000, was used for all knockouts in this study unless stated otherwise. Knockouts generated in this study using the pMTL40000 system are illustrated with gel electrophoresis images of colony PCRs and confirmed via Sanger sequencing. The diagram in Figure 4.1 illustrates how such gel images can be interpreted.

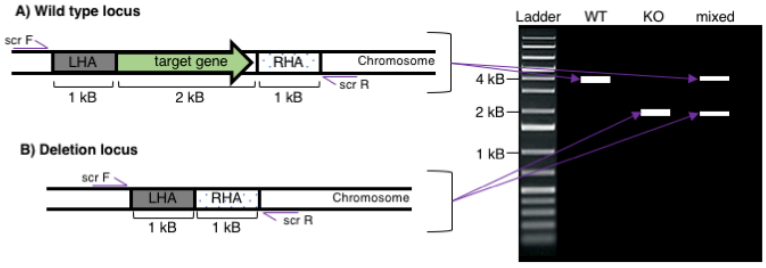


Figure 4.1: Schematic representation of colony PCR screening for knockout detection
A) Wild type locus consisting of a 2 kB target gene flanked by 1 kB homology arms (LHA and RHA) is amplified by screening primers “scr F” (forward) and “scr R” (reverse) that bind outside of the homology arms. As primer binding sites are unique to the chromosomal locus of interest, and not elsewhere (i.e. knockout vector), a PCR band of 4 kB represents an intact, wild type locus.
B) Deletion locus lacking gene of interest (due to successful knockout) will still enable binding of screening primers described above, but amplification will only result in a 2 kB band, due to the deletion of the 2 kB gene. Mixed colonies (where genome editing has not been 100% efficient) will contain sub-populations with knockouts and result in a smaller 2 kB band, and some wild type sub-populations resulting in 4 kB bands.

4.3.1. Tool development in *C. ljungdahlii*

Due to issues with DNA transfer described in Chapter 2, the closely related (>98% sequence identical) *C. ljungdahlii* was used as a model species for development of CRISPR technology in acetogens.

Promoter strength in *C. ljungdahlii* was assessed via a β-galactosidase assay performed by Dr He Huang (SIBS), identifying P_{thl} and P_{araE} as the strongest of the candidates tested. These were used for expression of *cas9* and the sgRNA respectively. Subsequently, the pMTLcas-pta vector was constructed as illustrated in Figure 4.2A, which resulted in knockout generation illustrated in Figure 4.2B.

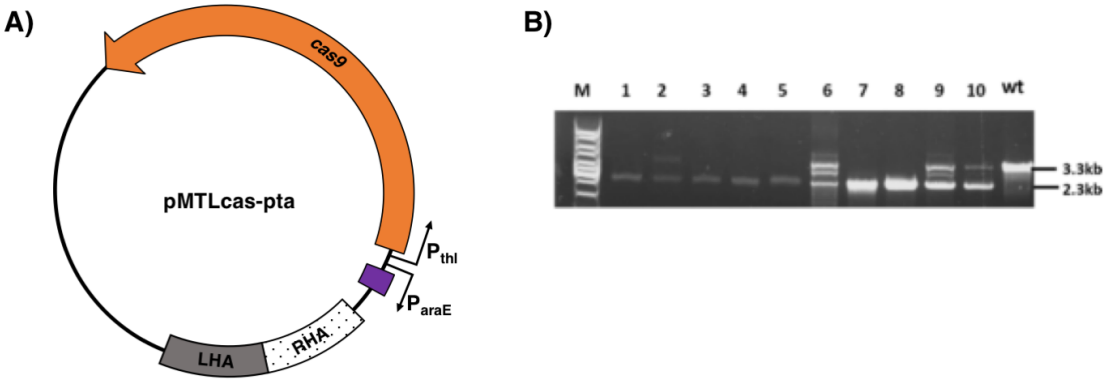


Figure 4.2: CRISPR-Cas9 knockout generation in *C. ljungdahlii*
A) Vector construction of plasmid used to create *pta* knockout in *C. ljungdahlii*, showing *cas9* (orange) and sgRNA (purple) being driven by the divergent P_{thl} and P_{araE} promoters (respectively) derived from *C. acetobutylicum*. The homology cassette downstream of the sgRNA consists of the left homology arm (LHA) and right homology arm (RHA) derived from the immediate up- and down-stream regions of *pta*. All the plasmid components labelled are found in the application specific module of the pMTL83151 backbone. B) Colony PCR results of *C. ljungdahlii* pMTLcas-pta transformants. Lanes 1-10 represent the 10 individual colonies yielded from transformation, with colonies 1, 2, 3, 4, 5, 7 and 8 showing a 2.3 kB band indicating gene deletion. Colonies 2, 6, 9 and 10 show both a 2.3 kB and a 3.3 kB band, indicating a mixed population of *pta* deficient and wild type genotypes. The final lane, "wt", shown a strong 3.3 kB band present in the wild type control. See Figure 4.1 for schematic explanation of colony PCR.

4.4. Plasmid modularisation and construction

Following proof of concept in *C. ljungdahlii*, the pMTLcas-pta vector was modified to allow for easier downstream modification via either classical, or assembly based cloning. As shown in Figure 4.3, the CRISPR-specific components were divided into the following sub-modules; A) nuclease, B) guide RNA, and C) homology cassette.

Submodules A and B are flanked with unique 6-8 nt restriction sites; *Sbf*I and *Xba*I for submodule A, and *Xba*I, and *Asi*SI for submodule B. The latter also contains a *Sal*I site between the guide RNA promoter and the seed region of the guide RNA, enabling oligo-based HiFi assembly (see section 4.4.1). Due to submodule C containing relatively long (>1500 bp) sequences, the 8 nt *Asi*SI and *Asc*I restriction sites were selected to avoid zemblanitous occurrence of recognition sequences within homology arms.

Earlier iterations of this system used a *Mau*BI restriction site in place of *Asi*SI, but as detailed in section 4.4.1, these sites were swapped. Generation of the first pMTL40000 vector, pMTL43151-CLAU-pyrE-sg3 was achieved through the use of a pMTL8151 backbone, Cas9 and sgRNA amplified from pMTLcas-pta, and a homology arm cassette generated through SOEing PCR of the *pyrE* LHA and RHA amplified from genomic DNA.

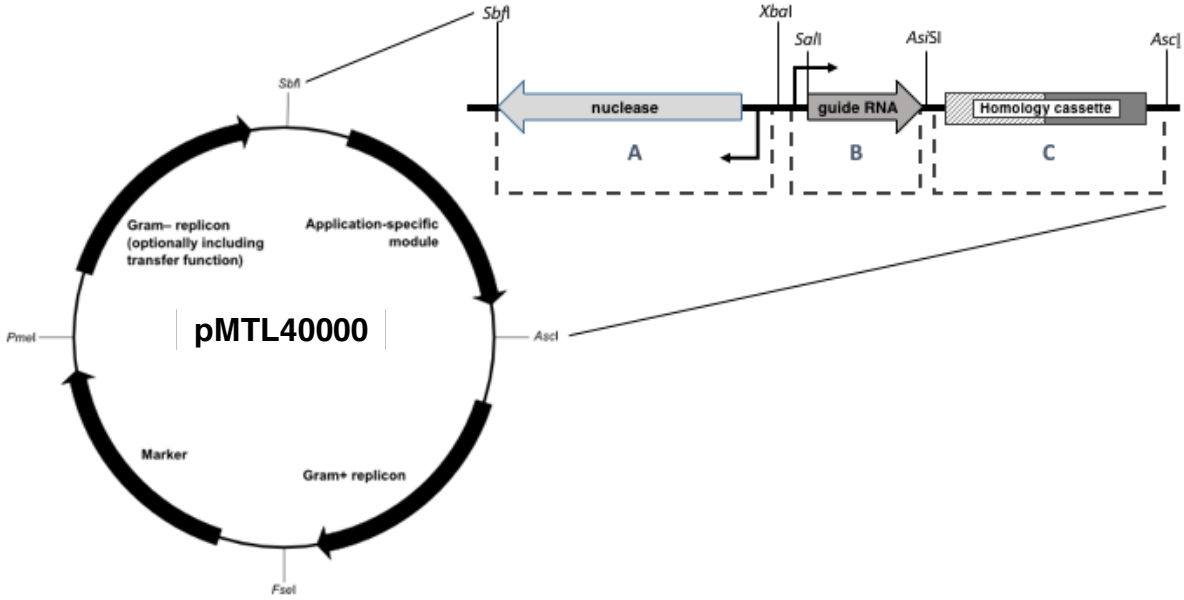


Figure 4.3: pMTL40000 vector design
Modular vector design of pMTL40000, showing three CRISPR-specific sub-modules within the application specific module of the pMTL80000 vector series. Module A consists of a nuclease and promoter (reverse orientation) and is flanked by *Sbf*I and *Xba*I. Module B consists of the RNA component(s) of the nuclease in module A, along with a promoter. This is flanked by *Xba*I and *Asi*SI, and contains an internal *Sal*I site between the promoter and guide RNA. Module C consists of the homology cassette derived from the left- and right- homology arms of the gene of interest, and is flanked by *Asc*I and *Asi*SI restriction sites.

First, *cas9* and P_{thl} from pMTL-cas-pta were amplified using Pthl-F-XbaI and Cas9R-NotI, with the resulting PCR product digested with *Xba*I and *Not*I and cloned into a pMTL83151 backbone linearised with the same enzymes. Following Sanger sequence confirmation with CatP-R1, *pyrE* targeting sgRNA cassette was generated and inserted described below. Using pMTL-cas-pta as a template, P_{araE} was amplified with sgRNA-F-XbaI and sgRNA-pyrE-sg3-SOE-R, with the latter primer containing the 20 nt *pyrE* targeting seed region as an overhang. Simultaneously, the sgRNA handle and its associated terminator were amplified using sgRNA-pyrE-sg3-SOE-F and sgRNA-R-MauBI-AscI primers, with the former containing the 20 nt *pyrE* targeting seed

region as an overhang. Using the flanking sgRNA-F-XbaI and sgRNA-F-MauBI-AscI primers, these two PCR products were spliced via SOEing PCR using the overlapping seed region. The resulting spliced PCR product was digested with *Xba*I and *Asc*I, and cloned in the Cas9 containing pMTL83151 vector described above (also digested with *Xba*I and *Asc*I). Again, CatP-R1 was used for Sanger sequence confirmation of this intermediate vector, prior to insertion of the *pyrE* homology arm cassette, resulting in the final pMTL43151-CLAU-pyrE vector.

For homology cassette generation, primer pairs CLAU-pyrE-LHA-F-MauBI + CLAU-pyrE-LHA-R, and CLAU-pyrE-RHA-F + CLAU-pyrE-RHA-R-AscI were used to generate the *pyrE* LHA and RHA respectively. Following this, the flanking primers CLAU-pyrE-LHA-F-MauBI and CLAU-pyrE-RHA-R-AscI were used to splice the homology arms, with the subsequent PCR product digested with *Mau*BI and *Asc*I. The digested homology arm cassette was cloned in a *Mau*BI and *Asc*I linearised backbone of the intermediate *cas9* and sgRNA containing pMTL83151 vector described above, to generate the final pMTL43151-CLAU-pyrE vector. Final vector sequence confirmation was achieved using CatP-R1 and sgRNA-F-XbaI primers.

The final pMTL40000 vector design can be seen in Figure 4.3. Here a generic vector is displayed, illustrating the modularity of vector design. For pMTL43151-CLAU-pyrE, the pCB102 Gram-positive replicon was used, along with the CatP selective marker, and ColE1 + *tra* Gram-negative replicon module. With regards to the CRISPR-specific modules, as described above, the nuclease module (A) consists of P_{thl} driving *cas9*, the guide RNA module

(B) consists of P_{araE} driven sgRNA containing the *pyrE* targeting seed region, and the homology cassette consists of 1 kB regions immediately up- and down-stream of the *pyrE* gene.

4.4.1. Optimisation of vector construction methods

Due to the inherent secondary structure of sgRNA, affordable DNA synthesis such as GeneArt® Strings from ThermoFisher is an option for insertion of new seed regions. Attempts at improving vector construction was therefore limited to laboratory-based techniques. For early iterations of the pMTL40000 vector series, as described in previously in section 4.4, a SOEing PCR-based approach was used to generate the entire guide RNA submodule, which was inserted into the vector backbone using *Xba*I and *Mau*BI[16] restriction sites. However, this approach proved time consuming and problematic due to small PCR product size being hard to distinguish from primer dimerization, as illustrated in Figure 4.4. Due to such problems, ssDNA oligo methods for vector construction were explored using NEB HiFi.

16 *Mau*BI was selected initially as an 8 bp restriction site, however, due to similarity to *Asc*I restriction sequence (CG^CGCGCG and GG^CGCGCC, respectively), the homology cassette of the pMTL40000 (submodule C) would often integrate in the incorrect orientation. This was amended by the substitution of *Mau*BI with *Asi*SI (GCGAT^CGC).

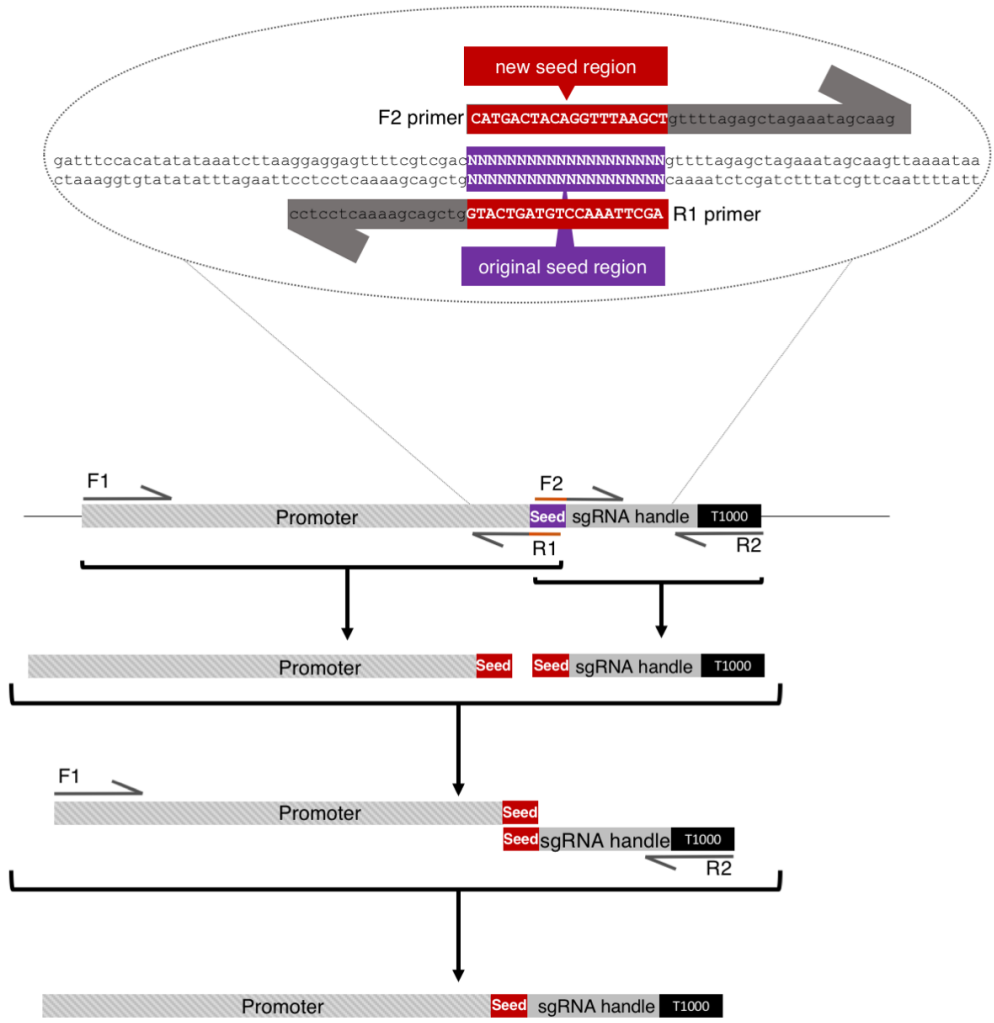


Figure 4.4: sgRNA module generation via SOEing PCR
To generate the sgRNA module via SOEing PCR, primers were designed for amplification of the sgRNA promoter (F1 and R1), and for the sgRNA handle and terminator (F2 and R2). The resulting products contained overhangs encoding the new sgRNA seed region, illustrated in red, which were used for SOEing PCR using flanking primers (F1 and R2). The final PCR product was then cloned into pMTL4000 using *Xba*I and *Mau*BI sites encoded by flanking primers.

Oligo HiFi was performed through *Sal*I linearisation of an existing pMTL40000 vector, which was dephosphorylated to prevent re-circularisation. As illustrated in Figure 4.5, ssDNA oligos were designed with 25 nt regions of overlap, which proved sufficient for assembly of plasmids with the desired mutations. To assess the efficiency of this assembly-based method, four new sgRNA seed regions were designed, and used to construct vectors as described in Chapter 2.

Cloning of the pMTL40000 vectors using ssDNA oligo HiFi was attempted initially using 5 different seed regions, each yielding 6-15 colonies per transformation. For each, 6 clones were Sanger sequenced to analyse the method efficiency. For those tested, an average of >50% efficiency was obtained, meaning over half had the desired 20 nt seed region modification without SNPs occurring in the seed region, or 25 nt flanking regions encoded by the oligo. SNPs were observed in these regions in ~40% of clones sequenced, with the remaining samples being made up of unmodified (re-circularised) backbone.

As a result of successful oligo HiFi, this method was taken forward for all subsequent sgRNA modification. For all vectors constructed in this study, SOEing PCR and conventional restriction-ligation based cloning was used for insertion of homology arms.

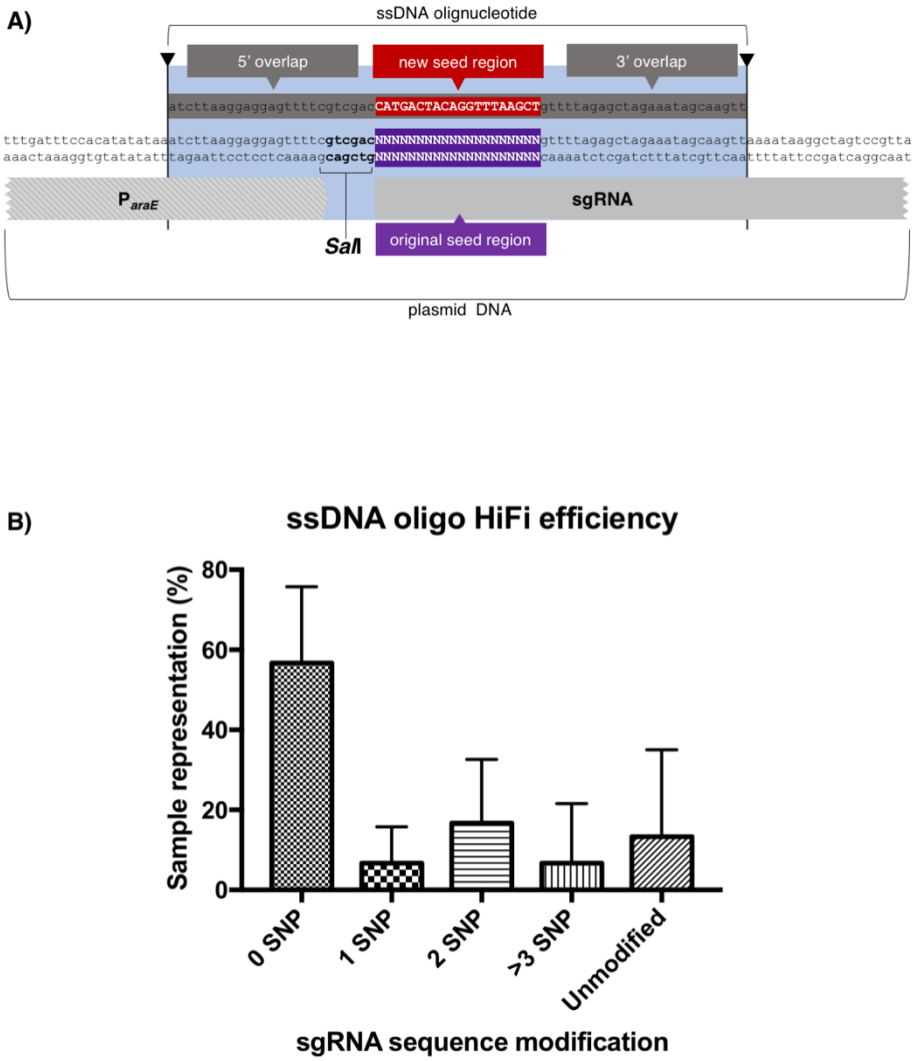


Figure 4.5: Modification of sgRNA via ssDNA oligo HiFi
A) Example of ssDNA oligo design for introduction of new sgRNA with new 20 nt seed region located between 25 nt flanking regions homologous to the 3' end of P_{araE} and 5' end of the sgRNA handle.
B) Efficiency of ssDNA oligo HiFi assembly. 100% efficiency was observed in >50% of samples sequenced, with single SNPs observed in seed or overlap in ~7% of samples, 2 SNPs in ~17% of samples, and >3 SNPs in ~7% of samples. The remaining colonies analysed were verified via sequencing to be re-circularised backbone vector, with no modification attained.

4.5. *pyrE* knockout

Having optimised DNA transfer (as described in Chapter 3), the pMTL40000 vector series was used to generate pMTL43151-CLAU-pyrE, a vector designed to enable clean, in-frame deletion of *pyrE*. Conjugation yielded 15 colonies, all of which were screened via colony PCR, resulting in identification of 14 knockout mutants, with the remaining colony (Figure 4.6, lane B2) displaying wild type genotype. Following colony PCR, plasmid loss was performed and the knockouts were confirmed via Sanger sequencing.

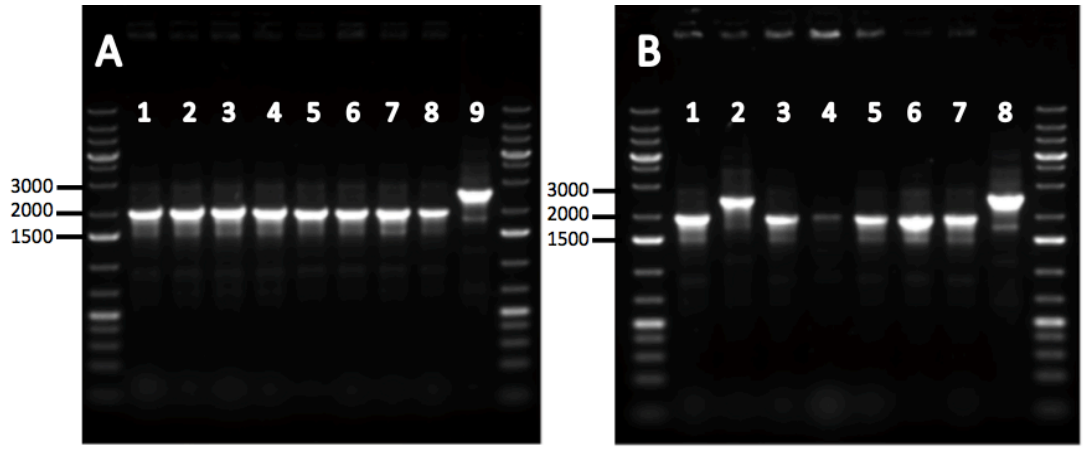


Figure 4.6: *C. autoethanogenum* Δ*pyrE* colony PCR
Gel image A displays 8 of the 15 pMTL43151-CLAU-pyrE harbouring *C. autoethanogenum* colonies screened via colony PCR (A1-8), with the remaining 7 colonies displayed in gel image B (B1-7). Lanes A9 and B8 represent a wild type control from the conjugation recipient culture. Knockout mutants in A1-8, B1, and B3-B7 all display ~2 kB bands consistent with deletion of the pyrE gene, whereas the wild type controls (A9 and B8) and B2 display a wild type band size of 3 kB. See Figure 4.1 for schematic explanation of colony PCR.

4.7. pMTL40000 in *E. limosum*

As discussed in Chapter 3, problems with DNA transfer in *C. autoethanogenum* led to the search for a possible alternative chassis. The relatively uncharacterised *E. limosum* was chosen as a likely candidate, due to preliminary research by PhD candidate Robert Mansfield, showing production of ethanol and butyrate when grown on C1 gas.

For proof of principle in *E. limosum*, *cas3* was chosen as target, due to its role in the endogenous CRISPR system (see Chapter 5). Following replicon characterisation by PhD candidate Robert Mansfield, pCB102 was again identified as being sufficient. As a result, the pMTL43151-CLAU-pyrE vector was used as a basis for generation of pMTL43151-ELM-cas3, with the sgRNA seed region swapped via HiFi assembly using oligonucleotide ELM-cas3-sg1-HiFi as described in section 4.4.1. Following Sanger sequence confirmation of this intermediate vector with sgRNA-F-XbaI, the homology cassette was generated using primers ELM-cas3-LHA-F-AsiSI, ELM-cas3-LHA-R, ELM-cas3-RHA-F, and ELM-cas3-RHA-R-AscI, and inserted using *AsiSI* and *AscI* as described in section 4.4.

Following Sanger sequence confirmation, the pMTL43151-ELM-cas3 vector was transformed into *E. limosum*. Screening of the resulting five thiamphenicol-resistant colonies on the initial selection plate resulted in the identification of two colonies displaying bands consistent with *cas3* deletion, two colonies displaying bands consistent with a wild type *cas3* locus but with faint knockout bands, and one colony displaying just a wild type band (Figure 4.7). Following a single passage on selective media, the progeny of all five

colonies displayed band sizes consistent with *cas3* deletion, confirmation of which was achieved via Sanger sequences. Colony PCR and sequencing was performed using ELM-cas3-scr-F and ELM-cas3-scr-R.

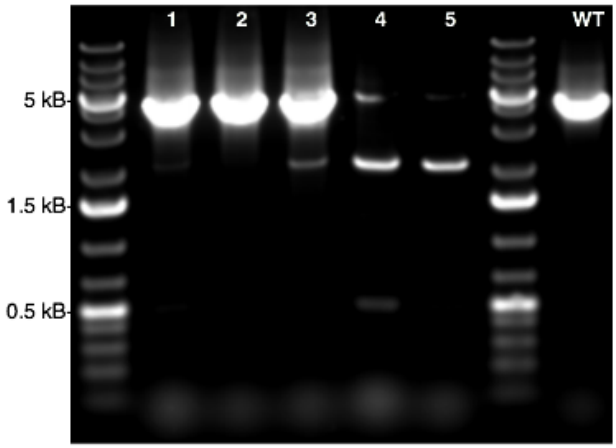


Figure 4.7: *E. limosum* Δ*cas3* colony PCR
Lanes 1-5 represent colony PCR products generated with ELM-cas3-scr-F and ELM-cas3-scr-R from 5 thiamphenicol resistant colonies present after transformation of *E. limosum* with pMTL43151-ELM-cas3. Lane "WT" represents the wild type control, displaying a band size of ~4 kB. Lanes 4 and 5 display the predicted ~2 kB band size consistent with deletion of *cas3*. See Figure 4.1 for schematic explanation of colony PCR.

4.8. pMTL40000 nomenclature generation and use in other organisms

As described, the pMTL40000 system was designed to offer full compatibility to the existing pMTL80000 vector series, enabling easy adaptation for use in other organisms. Once proof of concept for knockout of *pyrE* in *C. autoethanogenum* was attained, the system was tested in a range of other organisms, including the health-related *C. difficile* and *C. botulinum*. Table 4.1 shows the current organisms the pMTL40000 system has been used in.

Table 4.1: Examples of use of pMTL40000 usage other bacteria
Experimental design considerations as stated in this study were used for all initial knockouts in that stated organisms, with researchers then going on to make further knockouts independently. Knockouts generated in *C. botulinum* were performed by PhD student Daphne Groethius, *C. difficile* knockouts were generated by PhD student Patrick Ingle, and

Organism	Gene(s) targeted	pMTL40000 modifications
C. autoethanogenum ATCC	*pyrE*, *cas3*, *pta*, *ack*	N/A
C. botulinum ATCC 3502	boNT/A1	N/A
C. sporogenes NCTC 10696	*helA, cls*	N/A
E. limosum	*cas3, pyrE*	N/A
C. difficile 630 *C. difficile DH1916*	*ermB1*, *ermB2, hsdR1*	P_{tcdB} used to drive sgRNA
C. pasteurianum DSM 525	*ack, pta*	P_{J23119} used to drive sgRNA
C. saccharoperbutylacetonicum N1-4	*cas1*	P_{J23119} used to drive sgRNA

4.8.1. Expanding the range of nucleases available for pMTL40000

To offer a fully comprehensive modular system, exemplification of the interchangeability of individual sub-modules of pMTL40000 was achieved in the thermophile *Geobacillus thermoglucanosidans*[17]. This organism was selected for testing of the *Streptococcus thermophilus* St1Cas9 and St3Cas9 nucleases, due to its thermophilic nature matching that of the origin of the nucleases.

S. thermophilus contains two active CRISPR systems, both of which have been adapted for genome editing purposes in a handful of examples (Muller *et al.*, 2016; Karvelis *et al.*, 2013; Muller *et al.*, 2016; Xu *et al.*, 2015a; Kleinstiver *et al.*, 2015). As outlined by Selle *et al.*, both systems are Type IIA (Selle, Klaenhammer & Barrangou, 2015). In order to modify CRISPR systems for genome editing, the system's PAM must be identified, as well as the RNA components required for targeting. Table 4.2 summarises published PAMs and sgRNA sequences used for St1Cas9 and St3Cas9 genome editing applications, from which a consensus was drawn for use in *G. thermoglucosidius*.

17 All vector design, including sgRNA sequence, PAM selection and target selection was performed in the study, with vector construction and DNA transfer performed by PhD candidate Matthew Lau.

Table 4.2: Published PAM/sgRNA sequences for StCas9 genome editing
Examples of St1Cas9 and St3Cas9 uses in academic literature were used to compile PAM and sgRNA usage. For each instance, the PAM and sgRNA sequence were derived from the research paper (and supplementary materials) stated in the table.

	Ref.	Application	PAM used	sgRNA sequence
St1	(Muller *et al.*, 2016)	Mammalian	5'-NNAGAAW-3'	GTTTTTGTACTCTCAAGATTCA ATAATCTTGCAGAAGCTACAAA GATAAGGCTTCATGCCGAAATC AACACCCTGTCATTTTATGGCA GGGTGTTTT
	(Kleinstiver *et al.*, 2015)	Mammalian	5'-NNAGAA-3'	GTTTTTGTACTCTCAAGATTTA AGTAACTGTACAACGAAACTTA CACAGTTACTTAAATCTTGCAG AAGCTACAAAGATAAGGCTTCA TGCCGAAATCAACACCCTGTCA TTTTATGGCAGGGTGTTTTT
	(Steinert *et al.*, 2015)	Arabidopsis	5'-NNAGAA-3'	GTTTTTGTACTCTCAAGATTTA AGTAACTGTACAACGAAACTTA CACAGTTACTTAAATCTTGCAG AAGCTACAAAGATAAGGCTTCA TGCCGAAATCAACACCCTGTCA TTTTATGGCAGGGTGTTTTTTT
	(Esvelt *et al.*, 2013)	Mammalian	5'-NNAGAAW-3'	GTTTTTGTACTCTCAAGATTTA AGTAACTGTACAACGAAACTTA CACAGTTACTTAAATCTTGCAG AAGCTACAAAGATAAGGCTTCA TGCCGAAATCAACACCCTGTCA TTTTATGGCAGGGTGTTTTTTT
St3	(Muller *et al.*, 2016)	Mammalian	5'-NGGNG-3'	GTTTTAGAGCTGTGTTGTTTGT TAAAACAACACAGCGAGTTAAA ATAAGGCTTAGTCCGTACTCAA CTTGAAAAGGTGGCACCGATTC GGTGTTTTT
	(Xu *et al.*, 2015a) "Opti sgRNA"	Yeast	5'-NGGNG-3'	GTTTTAGAGCTGTAGGATCCAA CAGCAGTTAAAATAAGGCTTAG TCCGTACTCAACTTGAAAAGGT GGCACCGATTCGGTGTTTTTTT TT
	(Xu *et al.*, 2015a) "WT sgRNA"	Yeast	5'-NGGNG-3'	GTTTTAGAGCTGTGTTGTTTGC GAATGGTTCCAAAACGAAAAAC CATTCGAAACAACACAGCGAGT TAAAATAAGGCTTAGTCCGTAC TCACTTGAAAAGGTGGCACCGA TTCGGTGTTTTTTTT

As shown in Table 4.2, papers published showing use of *S. thermophilus* for genome editing identify the PAM for St1Cas9 and St3Cas9 to be 5'-NNAGAA(W)-3' and 5'-NGGNG-3' respectively. For St1Cas9, there seems to be some debate for the necessity of the W (A/T) at the 7th position, whilst Xu and colleagues (2015a) similarly report that the G at the 5th position for the St3Cas9 PAM is not essential. With these considerations in mind, a conservative consensus PAM sequence was used, including the debated nucleotides, as shown in Table 4.3.

Table 4.3: StCas9 PAMs for *G. thermoglucosidius* genome editing

Nuclease	Consensus PAM
St1Cas9	5'-NNAGAW-3'
St3Cas9	5'-NGGNG-3'

4.8.1.1. StCas9 sgRNA

For genome editing applications, the two RNA components of Type II CRISPR systems are usually fused in a synthetic guide RNA (sgRNA). As shown in Table 4.2, sgRNA usage varies between studies. This is due to usage of either fused complete tracrRNA and crRNA, or a synthetic version, streamlined to contain the minimum features required for targeting. This streamlining is illustrated in Figure 4.8, showing the comparison of "wild type" guide RNA to an optimised sequence.

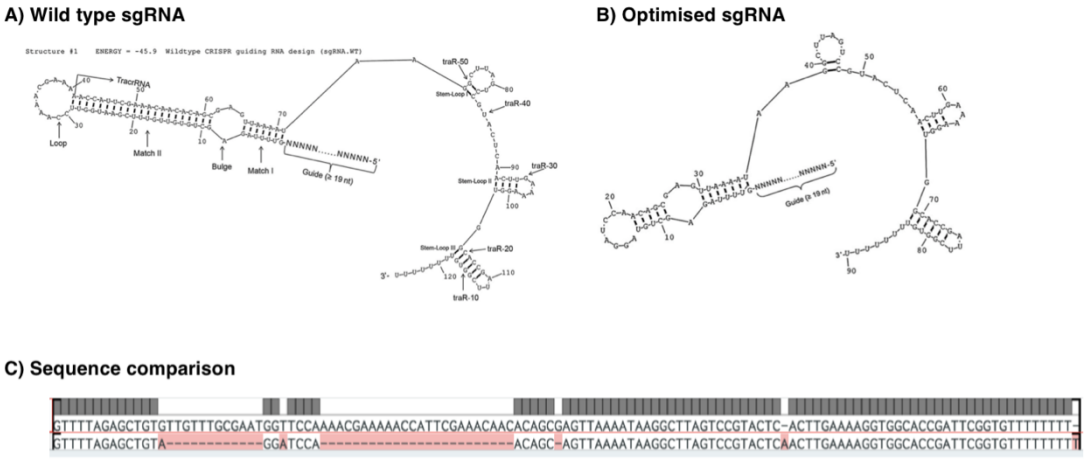


Figure 4.8: Sequence and structure comparison between wild type and optimised sgRNA used by Xu *et al*.
A) Shows the structure on "wild type" sgRNA, consisting of the entire sequences of tracrRNA and crRNA fused into a single transcript. B) Shows the optimised sgRNA design by Xu *et al*., which is significantly shorter due to the removal of non-essential sequence sections. C) Shows the comparison of wild type (top) and optimised (bottom) sgRNA sequences, with removed sequence highlighted in red.

Muller and colleagues (2016) detailed optimised versions of sgRNA for both St1Cas9 and St3Cas9, neither of which showed reduced nuclease activity. For this reason, the Mullet and colleagues (2016) sequences detailed in Table 4.3 were used for genome editing in *G. thermoglucosidias*. Target selection was achieved as described for pMTL40000 in Chapter 2, with sgRNA seed regions designed against several genes, for both Sp1Cas9 and Sp3Cas9. As illustrated in Figure 4.9, knockouts were successfully generated using both nucleases with the chose PAM sequence, showcasing the use of thermophilic nucleases within the pMTL40000 vector series. In this instance, the deletion of *ack* is showcased, generated with pMTL43151-St1-GB-ack and pMTL43151-St3-GB-ack vectors containing the St1Cas9 and St3Cas9 nucleases respectively, and their respective guide RNAs.

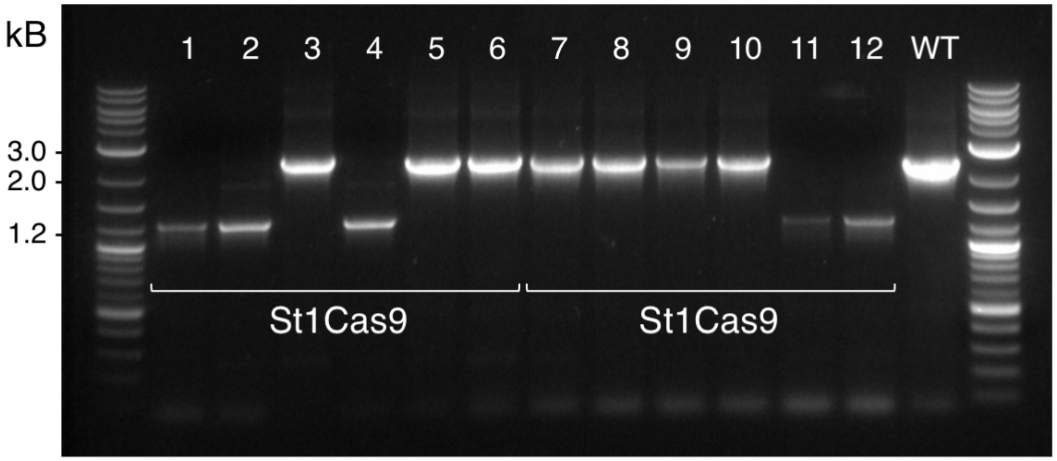


Figure 4.9: StCas9 *G. thermoglucosidius* Δ*ack* colony PCR
Lane "WT" shows the PCR product generated through amplification of wild type gDNA with *ack* screening primers, generating a 2.4 kB product. Lanes 1-6 show 6 colonies screened from initial selection plates after transformation of *G. thermoglucosidius* with pMTL40000-GB-St1-ack and pMTL40000-GB-St3-ack. Lanes 1,2 and 4 display 1.2 kB product sizes, indicating deletion of *ack*, with lanes 3, 5 and 6 showing the 2.4 kB size associated with a wild type *ack* locus. Lanes 7-12 show mutants screened from initial selection plates after transformation, with lanes 7-10 wild type (2.4 kB) product sizes, and lanes 11/12 displaying 1.2 kB bands indicating deletion of *ack*. Following colony PCR, knockout mutants displaying the correct locus size were confirmed via sequencing. See Figure 4.1 for schematic explanation of colony PCR.

4.9. pMTL40000 nuclease analysis: trCas9

Having been successfully used in a range of organisms, routine sequencing of a pMTL40000 vector resulted in the identification of a single nucleotide insertion at the extreme 5' end of the gene. Due to an abnormally long Sanger sequencing read, this insertion was detected and deemed too deleterious to enable generation of a functional Cas9. However, analysis of the progenitor of this plasmid identified this insertion again, and further analysis revealed the presence of the insertion from the creation of the pMTL40000 vector series.

4.9.1. *In silico* analysis of trCas9

In silico analysis suggests that a novel Cas9 variant (termed trCas9), is most likely being produced through translational initiation at a new start codon directly preceded by an RBS. As illustrated in Figure 4.10, the adenine insertion occurs in a poly-adenine region ~130 nucleotides into the gene, resulting in a frameshift-introduced stop codon truncating the product from the original start site to 53 nucleotides. The trCas9 product is proposed to be transcribed from a novel start codon 268 nucleotides downstream of the original ATG start codon, the proposed RBS for which is illustrated in Figure 4.10D.

The expressed trCas9 gene is 268 nucleotides smaller than the wild type SpCas9 gene, and is comprised of 1279 amino acids compared to 1368. Due to this truncation, the RuvCI nucleolytic domain is disrupted. As shown extensively in the literature, and exemplified by the commonly used D10A Cas9 variant, disruption of this domain results in removal of one of Cas9's nucleolytic domains and instead results in nickase activity. Therefore, trCas9

is assumed to act as a nickase, resulting in single strand nicks rather than double strand breaks.

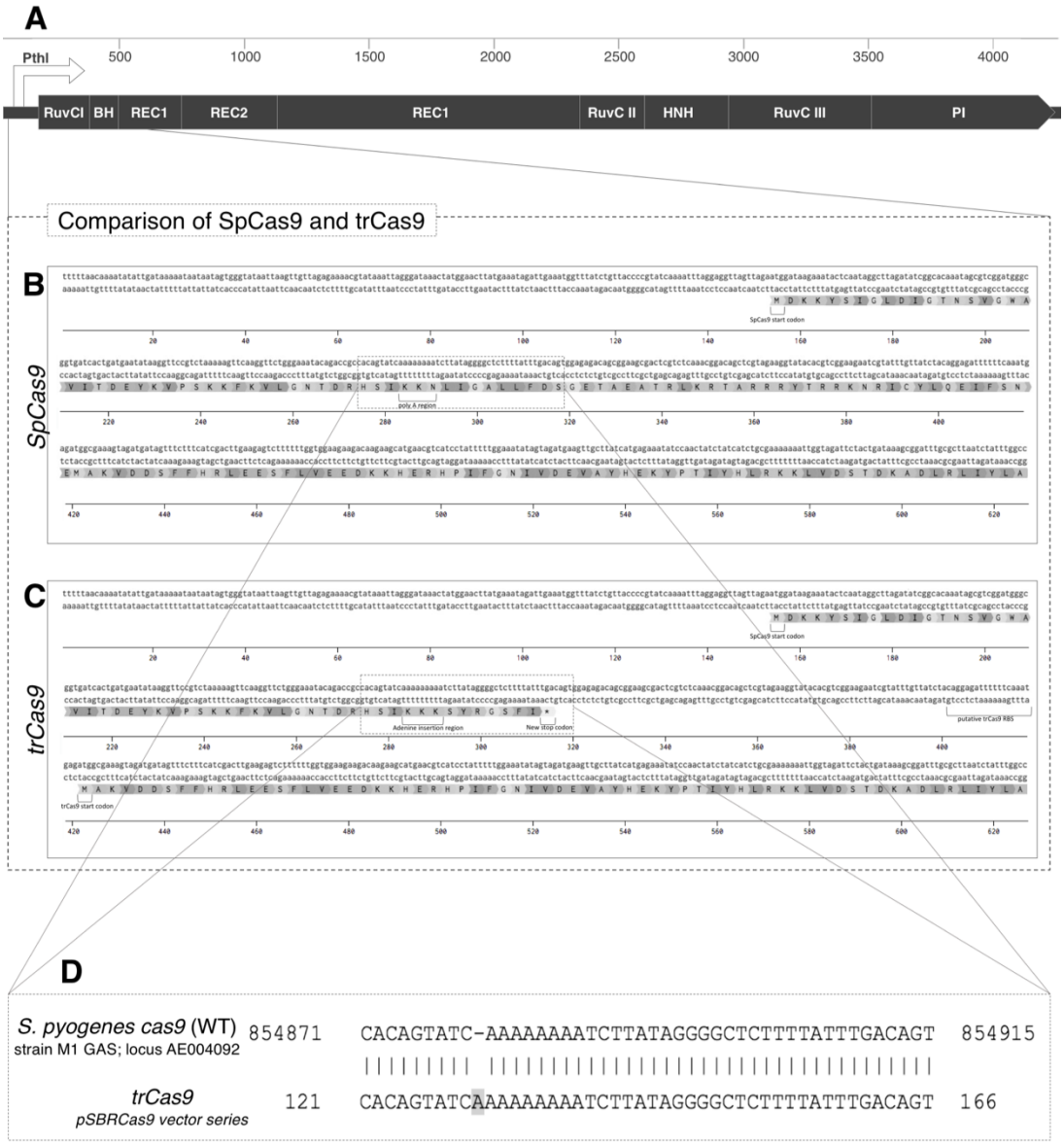


Figure 4.10: Sequence comparison of *S. pyogenes cas9* (wild type) and pMTL40000 "*trCas9*"
A) Overview of functional domains within SpCas9. B) Nucleotide sequence of 5' end of SpCas9, preceded by the *C. acetobutylicum* thiolase promoter. The SpCas9 methionine start codon is highlighted, with codon sequence of the remainder of the gene present. C) Nucleotide sequence of trCas9. Nucleotide sequence shown to be identical to SpCas9, with the exception of an adenine insertion in the poly-adenine region highlighted. Truncation of the initial translated product is shown by the frameshift introduced stop codon, with amino acid sequence of the predicted protein shown below. The putative RBS for trCas9 is highlighted, along with the methionine start codon. D) Nucleotide sequence comparison of SpCas9 and trCas9 at site of insertion.

4.9.2. trCas9 nuclease comparison

Once identified, numerous attempts to compare trCas9 activity to that of wild type SpCas9 and its D10A nickase variant (nCas9) were made. However, despite several cloning strategies no successful SpCas9 or nCas9 variants could be made. Although direct comparison of nucleolytic activity in *C. autoethanogenum* or other organisms could not be attained, the efficiency of the pMTL40000 system appears to offer comparable efficiency to other published systems. Lack of success with the two-plasmid system described in section 4.2 could also be attributed to increased lethality of the codon optimised, double nuclease, compared to the nicking activity of trCas9.

It was hypothesised that the expression of trCas9 was less lethal in *E. coli* and therefore more amenable to molecular cloning than SpCas9 or nCas9. Although vector construction issues hindered direct assessment of this hypothesis, relative levels of expression of the different nucleases under the control *C. acetobutylicum* P_{thl} was assessed. However, as shown Figure 4.11 the putative trCas9 RBS showed no expression.

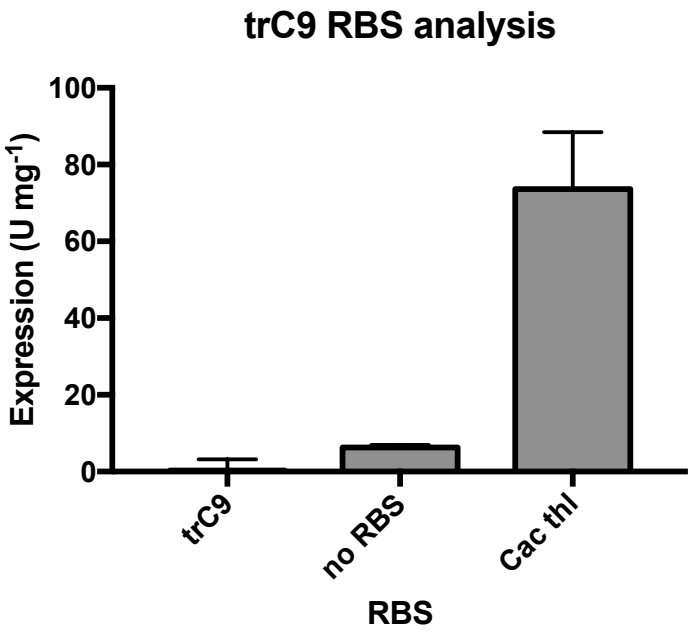


Figure 4.11: Relative expression of putative trC9 RBS
Expression of CatP from expression vector pMTL82254 with C. acetobutylicum Pthl (Cac thl), and a modified pMTL82254 vector with the Pthl RBS region replaced with the putative "trCas9" RBS (trC9). Expression is given as CatP units per mg of total protein.

4.10. Bookmark complementation

Conventional gene complementation within the genus *Clostridium* relies upon either expression of the deleted gene from a plasmid vector, or chromosomal integration at the *pyrE* locus. These methods fail to re-create the exact expression profile of the gene of interest, and therefore often fail to show a full wild type phenotype. To ensure any phenotype observed through CRISPR-Cas9 mediated genome editing was due to targeted deletion, rather than distal SNP accumulation, a novel complementation method, 'Bookmark', was designed.

As illustrated in Figure 4.12, Bookmark consists of a 20 nt protospacer sequence, a PAM, and a single extra nucleotide. This 24 nt sequence is used as a placeholder at the locus of the deleted gene, between the start and stop codon, and allows targeted for insertion of further cargo, such as the original gene. The 24 nt length of the sequence mitigates against polar effects due to being in-frame and therefore maintaining the original gene's functional stop codon.

Initially Bookmark was envisioned as a single sequence to be used across the *Clostridium* genus, however, due to the prospect of multiple knockouts being made in a single strain, other Bookmark sequences were designed. Table 4.4 shows the origin of each Bookmark seed sequence, with BM1 being a synthetic design, and all other seeds derived from examples of successful prokaryotic CRISPR-Cas9 genome editing.

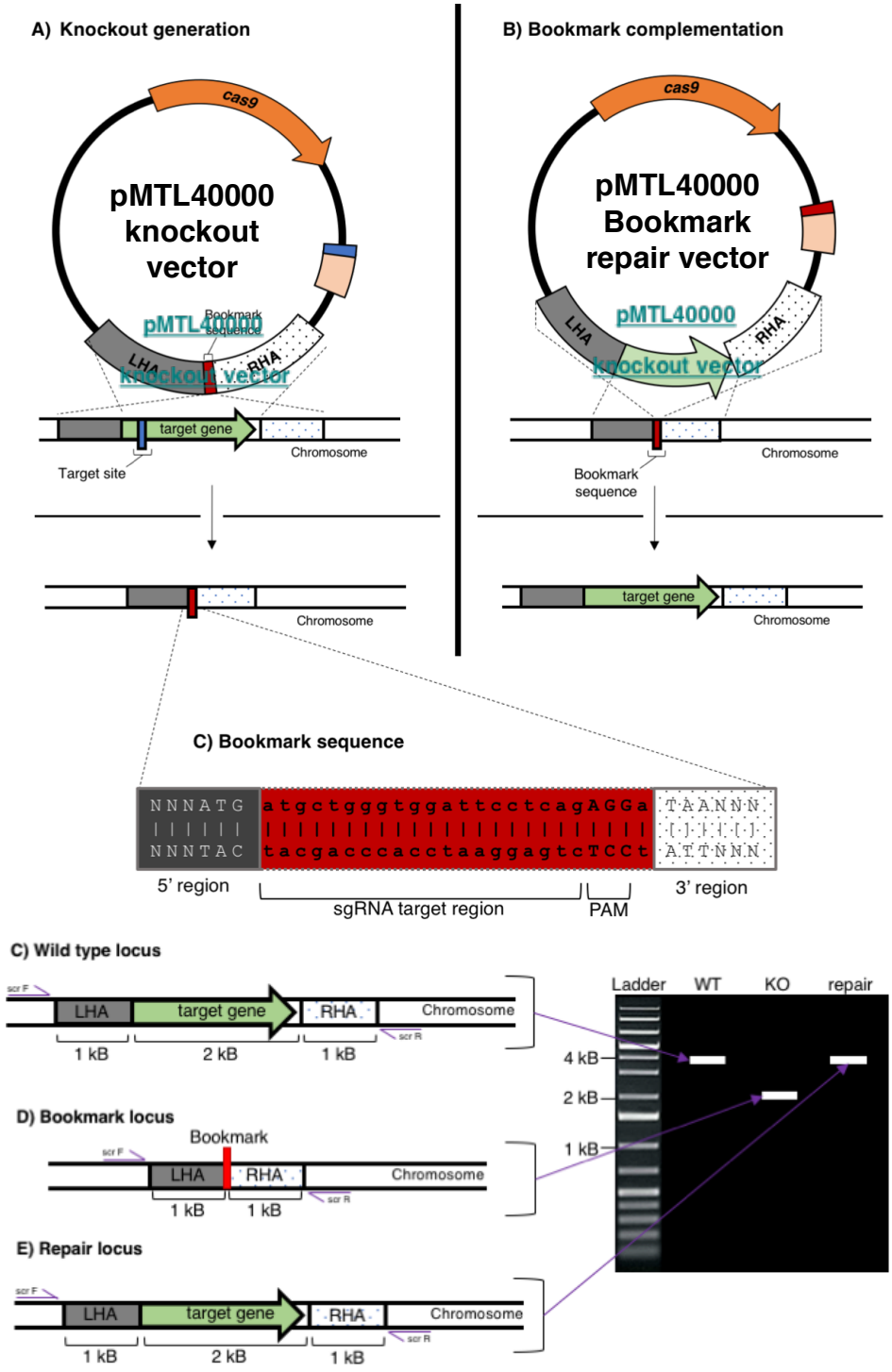


Figure 4.12: Bookmark complementation of knockouts
A) pMTL40000 vector used to generate a clean, in-frame deletion of the target gene shown in green. The seed region within the sgRNA on the vector (blue and pink) direct the plasmid encoded Cas9 protein to cleave the target site (shown in blue) within the gene of interest. Cas9 cleavage selects for a homologous recombination event deleting the gene, and replacing it with the bookmark sequence between the left- and right- homology arms (LHA and RHA). B) Targeting of the Bookmark sequence. The pMTL40000 Bookmark repair vector contains an sgRNA seed region, as shown in red, that targets the Bookmark sequence that was inserted as described above. This targeting selects for a homologous recombination event, again using the same LHA

and RHA, this time inserting the original gene as cargo between the homology arms. Cas9-mediated cleavage ensures only modified cells survive. C) Bookmark sequence analysis. The Bookmark sequence is located directly between the LHA encoded start codon (ATG) on the 5' side, and RHA encoded stop codon (TAA) on the 3' side. The sequence itself consists of a 20 nt target region not encoded elsewhere on the genome, a 5'-NGG-3' PAM (AGG) to enable targeting, and an extra adenine nucleotide to ensure the sequence in 24 nucleotides in length, and therefore does not cause frameshift-related polar affects.
(C), (D) and (E) represent wild type, bookmark deletion, and repair loci, with their respective screening colony PCR bands represented on a schematic gel image. See Figure 4.1 for schematic explanation of colony PCR.

Table 4.4: Bookmark seed sequences used in this study
Details of Bookmarks used in this study, including organism of origin and 20 nucleotide seed sequence. BM1 was generated in this study, with BM2 and BM3 derived from successful pMTL40000-mediated genome editing in *Clostridium difficile* by PhD student Patrick Ingle. BM4-12 were derived from examples of successful prokaryotic genome editing in academic literature (Jiang *et al.*, 2013; Altenbuchner, 2016; Oh & van Pijkeren, 2014).

Bookmark	Seed sequence	Details
BM1	ATGCTGGGTGGATTCCTCAG	Synthetic, designed for minimum off-target and maximum on-target activity
BM2	AACATCAAATCAAACGTGCC	From successful targeting in *C. difficile*
BM3	TTGCAAGAACAGTTTAACCC	From successful targeting in *C. difficile*
BM4	AGGGTTGTGGGTTGTACGGA	*S. pneumoniae bgaA*
BM5	ATTTCTGATATTACTGTCAC	*S. pneumoniae srtA*
BM6	ACCGATACCGTTTACGAAAT	*S. pneumoniae ermB*
BM7	TGAAGATCAGGCTATCACTG	*B. subtilis amyE*
BM8	TCCGGAGCTCCGATAAAAAA	*B. subtilis pulC*
BM9	TATTGATTCTCTTCAAGTAG	*B. subtilis trpC*
BM10	CCATTGTACTATCATGCTAG	*L. reuteri strA*
BM11	ATGCAGTCGGCTGTAGAAAG	*L. reuteri sdp6* 1
BM12	CGACTGCATTTTATTATGTA	*L. reuteri sdp6* 2

4.10.1. Bookmark 1

Proof of concept for Bookmark 1 was tested using the *pta* and *ack* knockouts discussed in detail in Chapter 6. The initial knockouts were made using pMTL43151-CLAU-pta-BM1 and pMTL43151-CLAU-ack-BM1 respectively. Generation of the sgRNA for each was conducted as described in section 4.4.1. Homology cassettes were constructed using SOEing PCR as described in Chapter 6, with the internal overhang sequence being comprised of "BM1". BM1, as described in Table 4.4, was a synthetically designed target region, designed for maximal on-target and minimal on-target activity across the *Clostridium* genus.

Once *pta* and *ack* mutants were generated (see Chapter 6), complementation was attempted. For the repair of *pta*, pMTL43151-CLAU-pta-repair was generated. CLAU-pta-LHA-F-AsiSI and CLAU-pta-RHA-R-AscI were used to generate the homology cassette from *C. autoethanogenum* gDNA, which was cloned into pMTL43151-CLAU-pyrE using *AsiSI* and *Asc*I, as described in section 4.4. Once homology arms were sequence confirmed, the CLAU-BM-HiFi oligo was used to introduce a new sgRNA seed region targeting the BM1 sequence present in place of the *pta* gene in the ∆*pta* strain (as described in section 4.4). Once sequence confirmed, this vector was transformed into Sexpress *E. coli* and conjugated into *C. autoethanogenum* ∆*pta*, yielding 32 colonies. 12 of these colonies were passaged onto selective media, and subsequent colony PCR showed them all to be repaired strains (Figure 4.13), with some showing possible sub-populations of knockout mutants.

Subsequently plasmid loss was performed prior to Sanger sequence verification and storage at -80 °C.

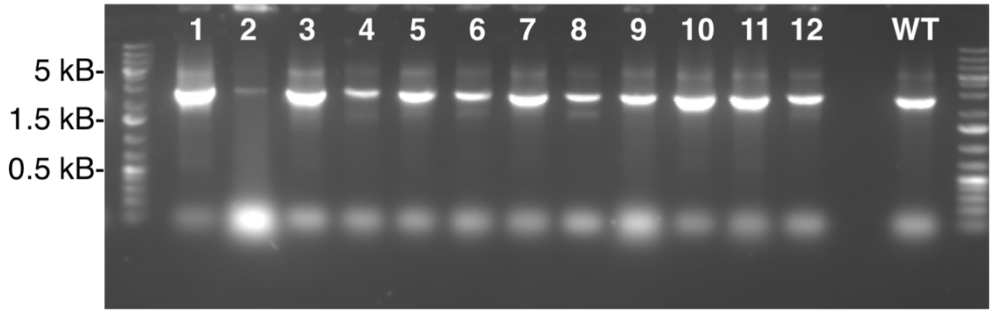


Figure 4.13: *C. autoethanogenum pta* Bookmark repair colony PCR
Colony PCR of *pta* repair strains generated through conjugation of pMTL43151-CLAU-BM-pta-repair into Δ*pta* strain of *C. autoethanogenum*. As detailed in Figure 4.12, all colonies have returned to the same 2.5 kB size as wild type and therefore been successfully been repaired.

To repair the deletion of *ack*, the pMTL43151-CLAU-BM-ack-repair was generated. The homology cassette was generated through the use of CLAU-ack-LHA-F and CLAU-ack-RHA-R to amplify genomic *C. autoethanogenum* gDNA. The homology cassette was cloned inserted into a pMTL43151-CLAU-BM-pta-repair backbone, replacing the *pta* repair cassette using *Asc*I and *Asi*SI restriction sites. Due to issues with the Δ*ack* strain's growth rate and limited cell density (see Chapter 6), generation of an *ack* repair strain was hampered. Multiple attempts to conjugate the repair vector resulted in no thiamphenicol resistant colonies.

4.11. Bookmark 2 and 3

Bookmark 2 and 3 sequences were based on seed regions used by PhD candidate Patrick Ingle for successful gene deletions in *C. difficile* 630. These were designed to be introduced for knockouts of *cauIII* and *cauIV* restriction endonucleases, however, due to time constraints these vectors were not completed. Therefore, the use of these bookmark sequences was never tested.

4.12. Bookmark 4-12

To analyse the potential of the Bookmark 4-12 sequences, a DNA sequence containing all 9 sequences interspaced with 5'-NGG-3' was synthesised by DNA 2.0. Using *pyrE* homology arms described in section 4.5, NEB HiFi assembly was used to splice the synthetic DNA sequence to the *pyrE* LHA and RHA, which was subsequently used as a DNA template for PCR amplification with CLAU-pyrE-LHA-F-AsiSI and CLAU-pyrE-RHA-R-AscI. Using pMTL43151-CLAU-ack as a backbone, the BM4-12 *pyrE* homology cassette was cloned using *Asi*SI and *Asc*I restriction site as detailed in section 4.4. Following sequence confirmation, the CLAU-pyrE-sg3-HiFi oligonucleotide was used for ssDNA HiFi insertion, as detailed in section 4.4.1.

After final sequence confirmation, pMTL43151-CLAU-pyrE-BM4-12 was conjugated into *C. autoethanogenum* (wild type). This resulted in several colonies, 12 of which were screened via colony PCR using CLAU-pyrE-scr-F and CLAU-pyrE-scr-R, as shown in Figure 4.14.

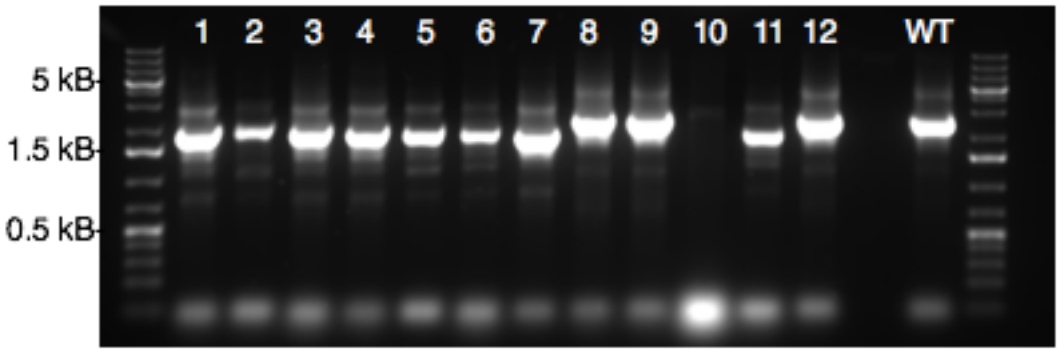


Figure 4.14: BM4-12 *pyrE* insertion colony PCR
Lanes 1-12 represent colony PCR products of thiamphenicol resistant colonies present on pMTL43151-CLAU-pyrE-BM4-12 *C. autoethanogenum* selection plates. Lane "WT" represents the wild type *pyrE* locus size (~2.5 kB), with the expected modified locus (*pyrE* replaced with BM4-12 sequences) size of ~2 kB observed in lanes 1, 3-7 and 11. See Figure 4.1 for schematic explanation of colony PCR.

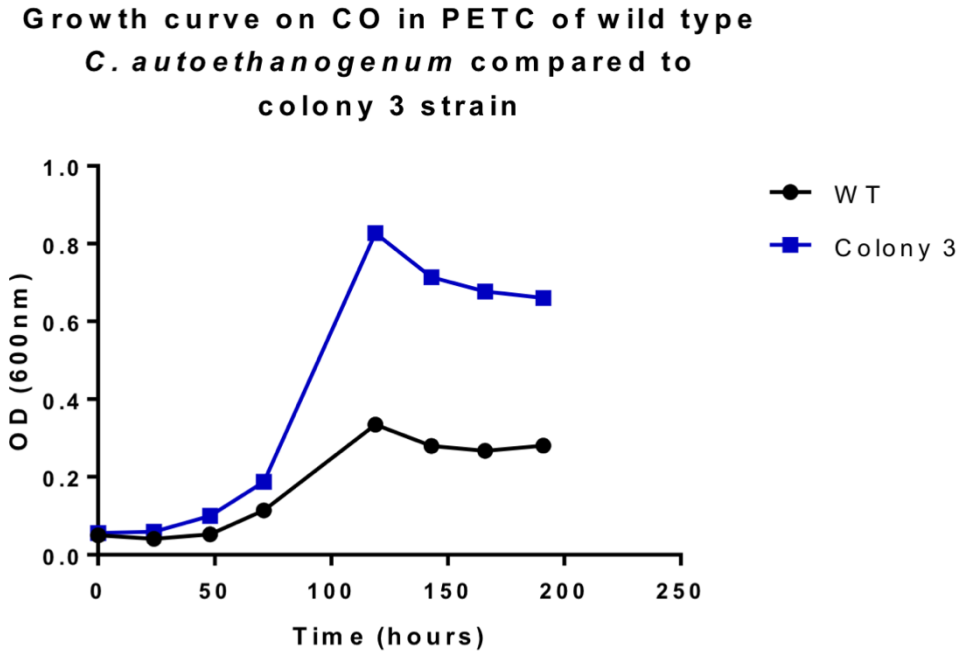
Colony PCR identified several mutants which were confirmed via Sanger sequencing with CLAU-pyrE-scr-F to contain the BM4-12 sequences. Unfortunately, due to time constraints targeting of these sequences was not achieved.

4.13. Development of CRISPR based SNP/SND generation

The impact of SNPs, rather than solely large genomic mutations on phenotypic variance has garnered much attention in recent years across the biological sphere. A prominent example within the clostridial species *C. difficile*. Here, two research groups disputed the relative importance of different toxin coding genes within the *C. difficile* 630 strain. Although in name (and under superficial genomic analysis) both groups used the same strain, deeper sequence analysis revealed SNPs that could explain the different results.

In a similar example from the GASCHEM project, of which this study is a part, continual fermentation of wild type *C. autoethanogenum* resulted in the generation of the fermenter adapted strain "Col3". Analysis of steady state cultures after >100 days fermentation resulted in the identification of 3 strains of interest (Col1, Col3 and Col3). This interest stemmed from improved performance, either in terms of fermenter stability or growth rate. Growth comparison between each strain (data not shown) identified Col3 as the best performing of the progeny strain in terms of growth rate. Illustration of this improved growth rate compared to wild type is illustrated in Figure 4.15.

Genetic analysis performed by Dr. Anne Henstra identified that each of the progeny strains differed from the progenitor strain in terms of several SNPs or SNDs. Analysis of Col3 is displayed in Figure 4.15, showing the location of 3 SNPs and 2 SNDs not present in the wild type strain. Due to time constraints, the fitness contribution of each SNP was planned to be analysed in isolation, with further analysis of SNP combinations to be performed in the future.



Position	Name	Progenitor	Col1	Col2	Col3	Type
1649394	N/A	+	+	+	+	SNP
2570271	N/A	+	+	+	+	SNP
3469169	N/A	+	+	+	+	SNP
3469310	N/A	+	+	+	+	SNP
4317806	N/A	+	+	+	+	SNP
4318262	N/A	+	+	+	+	SNP
479136	N/A	-	+	+	-	SNP
1271252	N/A	-	+	+	-	SNP
2606221	N/A	-	+	+	-	SNP
483403	**SNP1**	**-**	**-**	**-**	**+**	**SNP**
611671	**SNP2**	**-**	**-**	**-**	**+**	**SNP**
2605689	**SNP3**	**-**	**-**	**-**	**+**	**SNP**
3480533	**SND1**	**-**	**-**	**-**	**+**	**SND**
4053355	**SND2**	**-**	**-**	**-**	**+**	**SND**

Figure 4.15: *C. autoethanogenum* Col3 strain comparison
A) Growth curve comparing wild type and Col3 strains of *C. autoethanogenum* grown on CO in PET-C media. B) Genomic comparison of progenitor strain, Col1, Col2, and Col3, to wild type. All strains share the first 6 SNPs listed, suggesting the progenitor had these SNPs and passed them on to the progeny strains (Col1, Col2, and Col3). Col1 and Col2 share 3 SNPs not present in the progenitor strain, and Col3 has 3 unique SNPs and 2 unique SNDs.

4.13.1. SNP/SND locus analysis

The following section details the location of each SNP/SND being studied using co-ordinates from Humphreys and colleagues (2015). Using the annotations of this genome generated in the study (see Chapter 3), analysis of the function of each of the mutations is assessed.

4.13.1.1. SNPs

Located at position 483403 Humphreys and colleagues (2015), SNP1 constitutes an Asp175Tyr mutation in the transcription regulator Crp/Fnr family gene CAETHG_0443. This family of proteins generally act as positive transcription factors via DNA-binding activity (Körner *et al.*, 2003).

Located at position 611671 Humphreys and colleagues (2015), SNP2 constitutes an Asp86Asn mutation in another transcriptional regulator. As described in the annotation in this study, the CAETHG_0554 gene is a Fis family regulator. Examples of this regulatory family include in the two-component signal transduction system of *C. perfringens* VirR, regulating virulence factors (Nikolskaya & Galperin, 2002).

SNP3 is found in a putative signal transduction protein, containing putative CBS (cystathionine β-synthase) and DRTGG domains. These domains are known to be involved in protein regulation, as described in *Clostridium perfringens* (Tuominen *et al.*, 2010). This SNP causes a Leu46Ile mutation in CAETHG_2437.

4.13.1.2. SNDs

SND1 was identified to be in the putative promoter region for CAETHG_3226, a putative positive regulator of sigma E. The locus of this deletion is between the predicted -10 box, and the RBS, 94 nt upstream of the start codon.

SND2 was identified to be upstream of the putative promoter region for CLAU_3768. Annotation performed as part of this study, resulting in the Humphreys and colleagues paper, identified this gene as a hypothetical protein with no further details. Further analysis of the locus found no conserved functional domains, as shown in Figure 4.16.

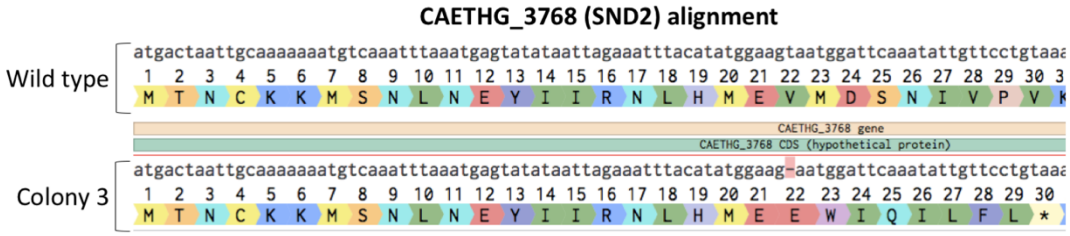


Figure 4.16: Comparison of CAETHG_3768 in wild type and colony 3 strains of *C. autoethanogenum*
The top section shows the nucleotide and amino acid sequence of wild type CAETHG-3768, with the corresponding sequences of colony 3 displayed below. Deletion of thymine residue at CAETHG location 4053355 in the colony 3 strain results in a frame shift mutation affecting the amino acid sequence from the 22nd residue. As a consequence of this frameshift mutation, a stop codon is introduced at residue 30.

4.13.2. SNP/SND vector design

Using the pMTL40000 system previously discussed, vectors pMTL43151-SNP1, pMTL43151-SNP2, pMTL43151-SNP3, pMTL43151-SND1, and pMTL43151-SND2 were designed. As illustrated below, modifications were designed to be introduced either in the sgRNA-designated target site, or in proximal regions to the target site in the homology arms. In both instances, a number of silent mutations were introduced in the 20 nt target region. This tactic was employed to ensure counter-selection of non-edited cells, but enable edited cells to escape Cas9-mediated cleavage. In addition, silent mutations were also designed to insert or remove restriction site in order to enable a restriction digest-based screen of colony PCR products. Both SNDs detected in Col3 appear in putative promoter regions, for this reason target sites downstream of the gene were selected in order to avoid alteration of the promoter activity through other nucleotide modifications.

Table 4.5: SNP/SND modified target loci and restriction enzyme introduction
Details of each modification designed for introduction of SNPs that alter coding sequence, or SNDs in promoter sequences. Target sites are modified with silent mutations and restriction sites inserted or removed as detailed in table.

SNP1 (843403)

Suitable target site 14 nucleotides downstream of SNP, BanI site introduced between SNP and target site (TGTACC to GGTACC)

	Y	**G**	**V**	**E**	**T**	**K**	**E**
Original	TAC	GGA	GTG	GAA	ACA	AAA	GA
Modified	TAT	GGT	GTC	GAG	ACT	AAG	GA

SNP2 (611671)

Target site 4 nucleotides upstream of SNP, single nucleotide silent mutation 10 nucleotides downstream of SNP to introduce EcoRV site (GATATT to GATATC)

	G	**I**	**V**	**L**	**T**	**I**	**I**
Original	GGA	ATT	GTA	TTA	ACT	ATC	AT
Modified	GGT	ATA	GTT	TTG	ACA	ATT	AT

SNP3 (2605689)

Target site 9 nucleotides downstream of SNP, double silent mutation 7 nucleotides downstream of target region to introduce KpnI site (GGAACG to GGTACC)

	V	**T**	**T**	**I**	**P**	**R**	**V**
Original	GTT	ACT	ACA	ATC	CCA	AGA	GT
Modified	GTA	ACA	ACT	ATA	CCT	AGG	GT

SND1 (3480533)

Target site 214 nucleotides downstream of deletion, modification of target site removes BsaWI site and introduces DdeI site.

	V	**D**	**A**	**K**	**N**	**P**	**V**
Original	GTA	GAT	GCA	AAA	AAT	CCG	GT
Modified	GTT	GAA	GCT	AAG	AAC	CCC	GT

SND2 (4053355)

Target site 111 nucleotides downstream of deletion, modification of target site introduces MfeI restriction site.

	Y	**A**	**I**	**D**	**T**	**L**	**N**
Original	TAT	GCT	ATA	GAT	ACA	TTA	AA
Modified	TAC	GCA	ATT	GAC	ACT	TTG	AA

4.13.3. SNP/SND generation

Unfortunately, repeated attempts at screening SNP changes in *C. autoethanogenum* failed to result in the identification of any SNP changes for any of the loci. Figure 4.17 illustrates this for the SNP2 locus, whereby despite multiple attempts and passages, pMTL43151-CLAU-SNP2 transconjugants showed no change in restriction profile. As shown in Table 4.5, if the desired targeted mutations had been made, a second EcoRV site should have should have been inserted, altering the restriction profile. However, as displayed in Figure 4.17, no change is restriction profile is observed between and of the transconjugants and the wild type control.

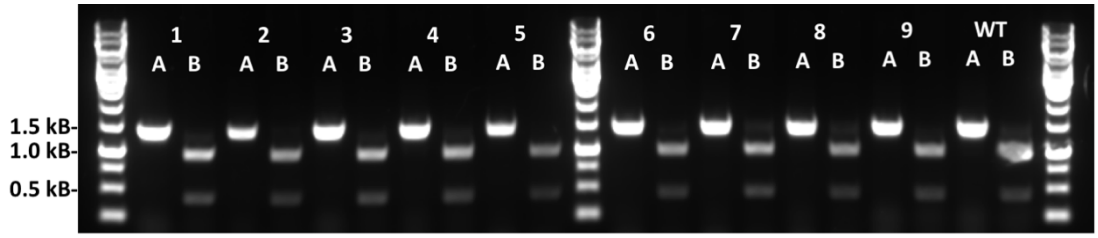


Figure 4.17: SNP2 modification *Eco*RV diagnostic screen
Lanes 1 to 9 represent PCR products generated with CLAU-SNP2-scr-F and CLAU-SNP2-scr-R from 9 *C. autoethanogenum* Tm^R colonies conjugated with pMTL43151-CLAU-SNP2. Lane "WT" shows the PCR product of the SNP2 locus with wild type *C. autoethanogenum* culture used as a template. "A" lanes represent undigested PCR products, whilst "B" lanes represent *Eco*RV digested PCR products.

4.14. Conclusion

At the beginning of this study, no published research had been conducted on the use of CRISPR-Cas9 based systems in the genus *Clostridium*. Unfortunately, due to DNA transfer issues, progress developing the pMTL40000 system was hindered. However, once DNA transfer was improved as described in Chapter 3, successful gene deletion using this system was achieved in *C. autoethanogenum*. Following proof of principal, this system was used in several other members of the genus, with its modularity proving useful for efficient modification to suit species-specific requirements.

Over the course of this study, several papers have been published reporting the application of CRISPR-Cas9 systems in clostridial species more amenable to DNA transfer (Xu *et al.*, 2015b; Pyne *et al.*, 2016a; Wasels *et al.*, 2017). Further to this, the use of a CRISPR-Cas9 system in *C. autoethanogenum* has also recently been reported, however, this paper reports much lower knockout efficiencies and of unintended gene truncations (rather than in-frame deletions) (Nagaraju *et al.*, 2016).

Collaborative efforts with the Shanghai Institute of Biological Sciences resulted in the design a functional CRISPR-Cas9 vector for use in clostridia, pMTL-cas-pta. Following the generation of the *pta* mutant described in this study, the mutant strain was characterised by He Huang and the findings published (Huang *et al.*, 2016). Although use of the pMTL-cas-pta vector showed efficient genome editing, no transconjugants were observed when

preliminary attempts at assessing its function in *C. autoethanogenum* were made.

In order to provide an easier platform for vector modification for use in *C. autoethanogenum* and beyond, the pMTL40000 vector series was designed. Although no major changes were designed between pMTL-cas-pta and pMTL43151, transconjugants were observed for the latter for *pyrE* mutant generation. After this proof of principle, the vector series was employed to make a range of other knockouts in various members of the genus. Retrospective analysis of the resulting pMTL40000 vector series showed all vectors used to generate knockouts contained what should be a deleterious frameshift mutations. Despite the elongation of a poly-thymine at the 5' end of the *S. pyogenes* Cas9 encoding *cas9* gene, efficient genome editing across the genome was still observed.

A logical explanation for the frame-shifted Cas9 still functioning is that the frame shift results in truncated version of Cas9 (trCas9), still able to function as a nickase. This would result in disruption of the HNH nucleolytic domain (as has been shown in the essential requirement of the D10 residue), with transcriptional read-through enabling transcription of a truncated but non-frame-shifted gene still containing a functional RuvC nucleolytic domain.

Consistent with the trCas9 hypothesis, use of nickase versions of Cas9 (nCas9) have shown more efficient genome editing in prokaryotes, including *C. beijerinckii*. The lack of the DSB activity of a fully functional Cas9 could crucially decrease the lethality, allowing more opportunity for recombination events to occur. This could be a possible explanation for the much higher

efficiencies observed using the pMTL40000 vector series in *C. autoethanogenum* (and other members of the genus), to the lower efficiencies report by Nagaruju *et al*. (2016) in *C. autoethanogenum*. Further to this, as described in Chapter 6, the use of pMTL43151-CLAU-pta resulted in the generation of *pta* knockouts, whereas pMTL-cas-pta showed a lack of observable transconjugants.

In order to compare trCas9 to the original spCas9, multiple attempts to clone SpCas9 into the pMTL43151 vector series was made. Unfortunately, despite these efforts, none of the intended vectors could be isolated. As a consequence, it did not prove possible to obtain a direct, in-context comparison of the two nucleases. As a result, the hypothesis of a truncated Cas9 was tested through analysis of CatP expression with the putative truncated region inserted immediately downstream of the thiolase RBS sequence. However, this putative RBS region failed to show any expression of CatP when inserted into the pMTL82254 reporter vector in place of the *C. acetobutylicum* P_{thl} RBS. This leaves the question of how effective genome editing can be achieved using pMTL40000 vectors, despite *cas9* containing a mutation that should be entirely deleterious.

Despite the uncertainty of how *trCas9* is able to provide nuclease activity, modularisation of the pMTL40000 vector series provided easy adoption in other members of the genus. With equally, high frequencies were observed in *E. limosum*, *C. difficile*, *C. botulinum*, *C. sporogenes, C. pasteurianum* and *C. saccharoperbutylacetonicum.* Additionally, this modularity enabled further expansion of the range of nucleases available to use. The thermophilic

St1Cas9 and St3Cas9 from *S. thermophilus* both showed high efficiency for use in *G. thermoglucosidians*. Combining the existing module of the pMTL80000 with a range of nucleases allows the pMTL40000 series to be used across a wide range of prokaryotes.

Although the work conducted in the study was unable to be published prior to other CRISPR-Cas9 research, the Bookmark technique developed in this study provides potential for a genuine advancement in the field. As shown with *C. autoethanogenum pta*, replacement of a gene of interest with a 24 nucleotide sequence provides the opportunity to complement the knockout. Unfortunately, due to time constraints other sequences could not be tested, however, 9 potential sequences were identified and inserted in at the *C. autoethanogenum pyrE* locus for future assessment. The identification of alternative Bookmark sequences would provide the ability to make multiple mutants and return and complement them at any time. This technique offers significant improvements over the existing gold standard ACE method. Whereas ACE complementation involves inserting a deleted gene at the *pyrE* locus, Bookmark inserts the gene at its original locus. This locus-specific restoration allows exact restoration of gene expression and therefore greater confidence in knockout characterisation. Although increasing lower costs of whole genome sequencing may make gene complementation less necessary, the inevitable accumulation of SNPs whilst generating mutants makes a case for the relevance of this new gold standard method of complementation.

Unlike Bookmark, attempts to develop the range of CRISPR-Cas9 applications to SNP correction failed in this study. Despite intensive planning,

this approach failed to introduce any detectable genomic changes in the three loci attempted. For each of these attempts, multiple colonies were seen on initial selection plates, but neither upon initial screening, or screening of multiple passages were any modifications detected. Explanations for this failure could be due to the following: insufficient selection pressure, deleterious effects of modifications, high off-target activity, or low on-target efficiency.

With regards to selection pressure, the pMTL40000 vector series used for SNP modifications was the same used for other knockouts. This means expression of sgRNA and Cas9 should provide the same high efficiencies observed with the knockouts (assuming high on- and off-target efficiency scores). However, rather than the deletion of the target locus, the SNP modification method discussed in this chapter modified ~7 of the 21 nucleotides. Assuming the behaviour of heterologous Cas9 systems to be similar across different host organisms, this would suggest that the modification of at least 35% of the nucleotides in the target region, including modifications at the 5' end.

With regards to deleterious effects of modifications, the likelihood of this being inhibitory to targeting is low. Although uncharacterised, none of these genes appear to be essential. Additionally, the modifications introduced were all silent mutations, meaning at most gene expression could be slightly altered. Also unlikely to contribute towards the failure of this method is high off-target activity. Not only were the sgRNA sequences analysed for their off-target activity, showing a lack of similar sequences in the *C. autoethanogenum*

genome, but there was a high number of colonies observed on selection plates. High off-target activity would likely result in a lower number of colonies observed due to increased lethality, and this was not the case. Discounting the first three factors discussed, the remaining explanation for SNP modification failure is low on-target activity. As the SNPs analysed in this study were predefined, unlike with the deletion of genes with a wide selection of potential target sites, targets were limited to those proximal to the SNP. The choice of sub-optimal target sequences is the most likely explanation of the failure for each of instances.

As well as failing to show any detectable modifications in the three instances attempted, this method proved intensive in planning stages. For each SNP, proximal target sites were identified and analysed. Following this, a manual search for proximal partial or complete 6-8 nt restriction enzyme recognition sites was performed to identify sites to modify. Finally, silent mutations in the target site and selected restriction site were designed, and oligonucleotides to implement these changes performed. Although unsuccessful, it is likely selection of alternate target sites may have provided more success. However, this only exemplifies the intensive nature of the design process.

Within the context of non-essential genes, a more fruitful approach would be a two-step technique using the Bookmark technique previously described. Here, the gene of interest could be deleted (if non-essential), or partially truncated, providing an opportunity to further characterise the gene, and Bookmark used to introduce the desired modification.

CHAPTER FIVE

DEVELOPMENT OF ENDOGENOUS CRISPR SYSTEMS FOR GENOME EDITING AND TRANSCRIPTIONAL CONTROL

5. Development of endogenous CRISPR systems for genome editing and transcriptional control

5.1. Introduction

As outlined in Chapter 1 and Chapter 3, the use of exogenous Type II CRISPR systems for genome editing and transcriptomic control has been exemplified in a wide range of organisms. However, the use of endogenous systems in bacteria have shown comparatively less exposure, despite several advantages over their foreign counterparts.

Use of Cas9 based tools requires establishment of sufficient DNA transfer methods, due to vector size. In contrast, utilisation of endogenous CRISPR systems enables genome-based Cas proteins to be repurposed for genome editing or transcriptomic control, meaning that smaller plasmids with lower DNA transfer requirements can be used (Pyne *et al.*, 2016a).

The term CRISPRi is used to describe CRISPR based interference of transcription, which is generally achieved using dCas9, a nucleolytically inactive version of Cas9. Here, the sgRNA still guides dCas9 to the target site, namely the promoter or 5' end of the gene of interest, where it physically interferes with RNA polymerase, reducing the rate of transcription (Larson *et al.*, 2013). This approach has also been exemplified in the Type IE CRISPR system of *E. coli*, using the Cascade complex for interference in strains deficient in the Cas3 nuclease normally responsible for CRISPR nucleolytic activity (Rath *et al.*, 2015). With regards to *C. autoethanogenum*, the use of its endogenous system for transcription control is particularly applicable as it negates the need for either genomic integration of dCas9, or continual plasmid selection during fermentation. Previous *in silico* analysis of the

C. autoethanogenum CRISPR system has briefly been reported in two separate studies (Pyne *et al.*, 2016a; Brown *et al.*, 2014), however no attempt to modify this system for use as a tool has been attempted.

The main barrier to overcome in order to harness the potential of endogenous CRISPR systems is identification of the PAM of the CRISPR system. As described in Chapter 1, the PAM is the Protospacer Adjacent Motif, a short sequence found next to protospacers that is necessary for targeting (Kleinstiver *et al.*, 2015; Hayes *et al.*, 2016; Leenay *et al.*, 2016). For any intact CRISPR system, identification of any spacer encoded sequence adjacent to a correct PAM will result in nucleolytic activity. For this reason, PAMs are not encoded within CRISPR loci, as this would result in lethal self-targeting. Further to this, nowhere else on the genome explicitly outlines the sequence of a PAM, therefore uncharacterised CRISPR systems must first have their PAMs identified prior to their adaptation as tools. Two methods can be employed for PAM identification. In one approach, *in silico* identification of protospacers is undertaken from spacer sequence identity and subsequent analysis of PAMs. In the other strategy, *in situ* analysis of the CRISPR system within its natural host is performed either by assessing spacer acquisition when exposed to phage or plasmid sequences, or assessing which sequences adjacent to known spacers illicit CRISPR interference (Leenay *et al.*, 2016).

5.2. *In silico* CRISPR analysis

In silico analysis was performed on the genomes of *C. autoethanogenum* and *E. limosum* to determine the type of CRISPR system they both contained, and to analyse their CRISPR loci for potential information on PAM usage.

5.2.1. *C. autoethanogenum* CRISPR system

Through GenDB annotation described in section 3.2, the *C. autoethanogenum* CRISPR operon was analysed. This identified the presence of 8 Cas genes and three proximal CRISPR loci. This analysis supports the previously published Brown and colleagues (2014) work which commented on the presence of the CRISPR system in *C. autoethanogenum*, and partially confirmed its activity through transcriptomic analysis showing expression of the operon. As identified in this reserch and illustrated in Figure 5.1, the operon consists of 8 Cas genes (*cas6*, *cas8b*, *cas7*, *cas5*, *cas3*, *cas4*, *cas1*, and *cas2*), and has a terminator following the last gene, *cas2*.

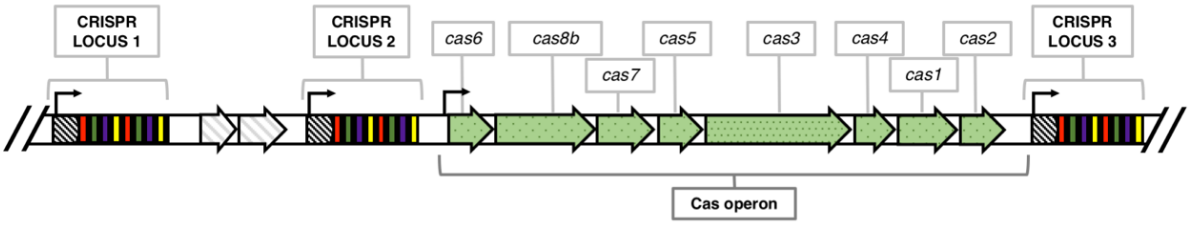


Figure 5.1: *C. autoethanogenum* CRISPR region
The *C. autoethanogenum* Cas operon is found on the reverse strand and consists of *cas6*, *cas8b*, *cas7*, *cas5*, *cas3*, *cas4*, *cas1*, and *cas2*. These genes are represented in the Humphreys *et al.* genome annotation generated as part of this study by the locus tags CAETHG_1394-9. The Cas operon has two proximal upstream CRISPR loci and one proximal downstream loci.

As detailed in Chapter 1, the presence of *cas3* makes this a Type I CRISPR system (Makarova *et al*., 2011). In the most recent classification of CRISPR systems, Makarova *et al*. (2015) propose a series of CRISPR system sub-types based on Cas operon architecture. Based on this classification, the *C. autoethanogenum* system is a Type I-B system, with the same operon composition to the example of *C. kluyveri* used by Makarova and colleagues (2015). This classification corroborates *in silico* analysis by Pyne and colleagues (2016a) on the *C. autoethanogenum* CRISPR system which also identified it as Type I-B, along with *C. pasteurianum*. Analysis in this study also identified several putative Type I-B systems in *Clostridium* species, as described in section 5.7.

5.2.2. *C. autoethanogenum* CRISPR loci

Although Brown and colleagues (2014) offered some analysis of the *C. autoethanogenum* endogenous CRISPR system, this was limited to identification of CRISPR loci and transcriptomic analysis of the Cas operon. Building on this, Pyne and colleauges (2016a) performed some limited *in silico* analysis of the CRISPR loci of *C. autoethanogenum*, identifying a single putative protospacer sequence also independently identified in this study.

To further analyse the CRISPR loci of *C. autoethanogenum*, CRISPRfinder was used to identify possible loci in the Humphreys and colleagues (2015) sequence generated in this study. This analysis corroborated the 3 loci identified in previous studies, which were termed CRISPR locus 1, 2, and 3. Using an approach detailed in Chapter 2, the spacers of these loci were analysed in attempt to find their putative protospacer sequences, and

therefore identify the PAM of the *C. autoethanogenum* Type I-B CRISPR system.

Table 5.1: Spacer analysis of *C. autoethanogenum* CRISPR locus 1
In silico analysis of CRISPR locus 1. Protospacer ID is signified by CAETHG prefix, followed by Spacer co-ordinate, with letters to differentiate between alternative hits of the same spacer sequence against putative protospacer sequence. Organism and organism ID of protospacer as signified along with strand of protospacer. The putative protospacer sequences are signified along with putative 5' and 3' PAM sequences, and similarity information. Sequence identity is signified, with 5' and 3' weightings for PAM confidence. Notes column details significant information of putative protospacer origin.

Protospacer ID	Protospacer origin			Putative Protospacer and PAM sequences			Similarity and scoring			Notes
	Organism of origin	Organism ID	Strand	5' PAM	Putative Protospacer	3' PAM	Identity	5' Score	3' Score	
CAETHG_1.06.a	Clostridium scatologenes strain ATCC 25775	gb|CP009933.1|	Plus/Minus	CAGAA	TTTATATCTATGCCTTTTAACAAATTAAATCCACC	CTCCT	72%	0	1	N/A
CAETHG_1.06.b	Clostridium pasteurianum DSM 525	gb|CP009268.1|	Plus/Plus	AAAAT	TTACATTGCAAAAAATTTAACAATTTAAATACCTC	TATAA	56%	0	0	ABC-type polysaccharide
CAETHG_1.07.a	Clostridium botulinum strain NCTC 8550	gb|CP010521.1|	Plus/Minus	ATTCT	TATAACTTTTTTATCCTTTAATTCTTCTATAGCTTG	CTTCA	64%	0	1	ABC transporter
CAETHG_1.07.b	Clostridium botulinum E3 str. Alaska E43	gb|CP001078.1|	Plus/Minus	GCTTA	TTGTTCTTATAGATCCTTTAGTCTTGGGTGTGCTAA	ATTCA	61%	0	0	Phage integrase family protein
CAETHG_1.07.c	Clostridium sp. M2/40	emb|HG917868.1|	Plus/Minus	TCAAA	TTTCACATTTCTTCTTGTTAATTCTTCTATAGCTTC	TTCTC	53%	0	2	N/A
CAETHG_1.07.d	Clostridium ljungdahlii DSM 13528	gb|CP001666.1|	Plus/Plus	GTCAT	CTTCTCAGCTGTTTCCTTTAATTCTTCTATAGTCAT	AACTG	53%	0	0	N/A
CAETHG_1.08.a	Bacillus thuringiensis plasmid pBMB0232	gb|CP005939.1|	Plus/Minus	TCACG	CCTCATATTCTTTTACTAATTTATAAAAAGAAGAA	CGTTT	75%	2	0	N/A
CAETHG_1.08.b	Bacillus thuringiensis HD-789 plasmid pBTHD789-3	gb|CP003766.1|	Plus/Minus	GTTTT	CTTCATATTCTTTCAATATTTTATAAAAAGTATTT	TTCTT	83%	1	2	N/A
CAETHG_1.08.c	Bacillus thuringiensis HD-789	gb|CP003763.1|	Plus/Minus	GTTCT	CTTCATATTCTTTTAGTATTTTATAAAAAGTATTT	TTCTT	83%	0	2	N/A
CAETHG_1.08.d	Bacillus thuringiensis HD1002	gb|CP009351.1|	Plus/Plus	GTTCT	CTTCATATTCTTTTAGTATTTTATAAAAAGTATTT	TTCTT	83%	1	2	Hypothetical protein
CAETHG_1.08.e	Bacillus thuringiensis HD1002 plasmid 4	gb|CP009350.1|	Plus/Plus	GTTTT	CTTCATATTCTTTCAATATTTTATAAAAAGTATTT	TTCTT	83%	1	2	N/A
CAETHG_1.08.f	Bacillus cereus strain AH820 plasmid pPER272	gb|DQ889677.1|	Plus/Minus	GTTCT	CTTCATATTCTTTTAGTATTTTATAAAAAGTATTT	TTCTT	83%	1	2	N/A
CAETHG_1.08.g	Clostridium pasteurianum NRRL B-598	gb|CP011966.1|	Plus/Plus	ATTAG	ATTTAGTTACCTTTAATAATTTATAATATGTATTT	TTCAA	64%	0	2	DNA binding protein
CAETHG_1.08.h	Clostridium carboxidivorans P7	gb|CP011803.1|	Plus/Minus	AAACT	TATAATATTATTTTAATAATTTATAAAATGAAGAC	CTATT	69%	0	0	N/A
CAETHG_1.10.a	Clostridium pasteurianum BC1	gb|CP003261.1|	Plus/Minus	TAGAG	AATATATCTATATGCTTAATACATGTAGTATCTACTT	TTAAA	89%	2	0	phage portal family protein
CAETHG_1.11.a	Clostridium beijerinckii strain NCIMB 14988	gb|CP010086.1|	Plus/Plus	TTGCT	TTTCTTGCTTATTAGCAATTTTATAAGCCTTGTTTC	AAGAG	67%	0	0	N/A
CAETHG_1.16.a	Clostridium beijerinckii strain NCIMB 14988	gb|CP010086.1|	Plus/Minus	AGTCA	AAGAATGTAAATTGGACTTTATTTTCATTTTCTTTT	TCAAC	69%	0	2	N/A
CAETHG_1.16.b	Clostridium phage phiMMP04	gb|JX145342.1|	Plus/Plus	CTGTA	ATATCTCTATAATCAACTTTATTTTCATTTTCTTCT	AAGTC	56%	0	1	N/A
CAETHG_1.23.a	Geobacillus thermoglucosidasius strain DSM 2542	gb|CP012712.1|	Plus/Plus	TTAAT	TTTTTGGGAATTACCCTTTAAAATGGGAATAGGTA	TTGTT	69%	1	0	N/A
CAETHG_1.24.a	Clostridium saccharobutylicum DSM 13864	gb|CP006721.1|	Plus/Minus	GTGTC	TGCCTTATGTTATGAGGGATATTAATTGCTGGGATAG	TGAAG	75%	0	2	N/A
CAETHG_1.24.b	Clostridium botulinum DNA, complete genome, strain: 111	dbj|AP014696.1|	Plus/Minus	ATGTT	TACCTTACTTTATGAGAGATATTAATTGTTGGGATAG	TAAAG	61%	0	2	N/A
CAETHG_1.24.c	Clostridium perfringens str. 13 DNA	gb|CP000312.1|	Plus/Minus	TTAAT	TTGTAGATATTAATAGAGATATTATTTAACTCACTCA	TAAAA	61%	0	0	Spacer
CAETHG_1.27.a	Clostridium ljungdahlii DSM 13528	gb|CP001666.1|	Plus/Plus	CTTCA	GCACCTAAGTAATCTATAAGTAAAGTATCAACTCTT	TTAAT	100%	4	4	Phage related protein
CAETHG_1.27.b	Clostridium beijerinckii strain NCIMB 14988	gb|CP010086.1|	Plus/Minus	CTTCA	GCACCTAAATAATCTATAAATAATGTATCAACTCTC	TTTTC	89%	2	1	Phage-like protein
CAETHG_1.27.c	Clostridium scatologenes strain ATCC 25775	gb|CP009933.1|	Plus/Minus	CCTCA	GCGCCTAAGTAATCTATAAGTAAGGTATCAATCCTT	TTAAT	81%	3	1	Phage-like protein
CAETHG_1.27.d	Clostridium cellulovorans 743B	gb|CP002160.1|	Plus/Minus	CGGTC	GCTCCAAGATAGTCTATAAGTAAGGTATCAACTCTT	TTAAT	69%	1	3	N/A
CAETHG_1.27.e	Clostridium pasteurianum NRRL B-598	gb|CP011966.1|	Plus/Plus	TAGAA	CTACATAAGTAACAC-CGTTTAAAGTATCAACTCTT	GTAGT	44%	0	1	Leucine tRNA-ligase

Table 5.2: Spacer analysis of *C. autoethanogenum* CRISPR locus 2
In silico analysis of CRISPR locus 2. Protospacer ID is signified by CAETHG prefix, followed by Spacer co-ordinate, with letters to differentiate between alternative hits of the same spacer sequence against putative protospacer sequence. Organism and organism ID of protospacer as signified along with strand of protospacer. The putative protospacer sequences are signified along with putative 5' and 3' PAM sequences, and similarity information. Sequence identity is signified, with 5' and 3' weightings for PAM confidence. Notes column details significant information of putative protospacer origin.

Protospacer ID	Protospacer origin			Putative Protospacer and PAM sequences			Similarity and scoring			Notes
	Organism of origin	Organism ID	Strand	5' PAM	Putative Protospacer	3' PAM	Identity	5' Score	3' Score	
CAETHG_2.02.a	Bacillus thuringiensis MC28 plasmid pMC429	gb|CP003694.1|	Plus/Minus	CCTGG	AATTGTACAATTTGATTTGTTTCTTTGTATAAGCGA	TTTGC	64%	1	0	N/A
CAETHG_2.02.b	Puniceispirillum phage HMO-2011	gb|GU557055.1|	Plus/Plus	AGGAG	TAATACCACTATCTATTTGTTTCTTTATATATACAT	TTTTA	56%	0	1	N/A
CAETHG_2.02.c	Clostridium perfringens strain FORC_003	gb|CP009557.1|	Plus/Plus	TAAGA	AAACCTTGATATTTATCACGTTCTTTATTTATAGCT	TCTGT	36%	0	1	Phage tail length tape measure protein
CAETHG_2.02.d	Clostridium beijerinckii ATCC 35702	gb|CP006777.1|	Plus/Minus	GATAA	ATGCCATCTCTTATATCAAGTTCTTTATTTATACCT	GTCTC	39%	0	0	TP901 family phage tail tape measure protein
CAETHG_2.03.a	Clostridium botulinum DNA	dbj|AP014696.1|	Plus/Plus	GAAAA	AATAGTTTCGTAAATAATAAAATTACT-CTATTATT	TATTT	69%	0	1	N/A
CAETHG_2.03.b	Clostridium botulinum CDC_1436 plasmid pCBG,	gb|CP006909.1|	Plus/Minus	ACATA	TTGCTATTATTAATTAATAAAATTACTTAATTGGAA	ATCCA	47%	0	0	N/A
CAETHG_2.03.c	Clostridium pasteurianum NRRL B-598	gb|CP011966.1|	Plus/Plus	ATATT	ATTCCATAAATATTTAATAAAATTACTGATATTATT	AATGT	67%	0	1	N/A
CAETHG_2.03.d	Clostridium beijerinckii strain NCIMB 14988	gb|CP010086.1|	Plus/Minus	ATATT	ATTCCATAAATATTTAATAAAATTACTGATATTATT	AATGT	67%	0	1	N/A
CAETHG_2.03.e	Clostridium beijerinckii ATCC 35702	gb|CP006777.1|	Plus/Minus	ATTGT	TCCATTTTATTAATTACCATAAGTATTGCTTATCTT	ACTTT	39%	0	0	Phage related protein
CAETHG_2.03.f	Clostridium clariflavum DSM 19732	gb|CP003065.1|	Plus/Minus	CAACT	ATAAATAAAATAATTAATATCATTAATGCTATTATA	ATTTT	36%	0	1	N/A
CAETHG_2.05.a	Clostridium ljungdahlii DSM 13528	gb|CP001666.1|	Plus/Plus	TTATA	ACTTCGCTAAATAGTGTTTTAACGAAGTTATAAAAA	GTCAC	75%	2	1	N/A
CAETHG_2.05.b	Clostridium sporogenes strain DSM 795	gb|CP011663.1|	Plus/Plus	ATAAC	AGAAAAATAAATTATGTTTTAAGGAAGTTGAATAAT	ATGAA	64%	0	0	N/A
CAETHG_2.05.c	Clostridium sporogenes strain NCIMB 10696	gb|CP011663.1|	Plus/Plus	ATAAC	AGAAAAATAAATTATGTTTTAAGGAAGTTGAATAAT	ATGAA	61%	0	0	Phage replisome organiser
CAETHG_2.07.a	Clostridium sporogenes strain DSM 795	gb|CP011663.1|	Plus/Plus	AACTA	AGCTAATGACTAATATAATATTAATAAGTAAGTTTA	AATAA	53%	0	0	Phage-like protein
CAETHG_2.07.b	Clostridium sporogenes strain NCIMB 10696	gb|CP009225.1|	Plus/Minus	AACTA	AGCTAATGACTAATATAATATTAATAAGTAAGTTTA	AATAA	53%	0	0	Phage-like protein
CAETHG_2.07.c	Clostridium sporogenes strain NCIMB 10696	gb|CP009225.1|	Plus/Minus	CAAAT	GAAAAATGACCAATATAATATGTATTACAACATTGC	TAAAA	31%	1	0	Phage holin protein
CAETHG_2.07.d	Clostridium botulinum Prevot_594 genome	gb|CP006902.1|	Plus/Minus	CAAAT	GAAAAATGACCAATATAATATGTATTACAACATTGC	TAAAA	31%	1	0	Phage related protein
CAETHG_2.07.e	Clostridium perfringens strain FORC_003	gb|CP009557.1|	Plus/Plus	CAAAT	GAAAAATGACCAATATAATATGTATTACAACATTGC	TAAAA	31%	1	0	N/A
CAETHG_2.07.f	Clostridium perfringens strain FORC_003	gb|CP009557.1|	Plus/Minus	TCCAA	CATTAAGTAACCATATAATATTATCTACTAATCCAG	TAAGT	39%	0	0	Phage terminase
CAETHG_2.07.g	Clostridium perfringens strain FORC_003	gb|CP009557.1|	Plus/Plus	TATTA	TAGAATTTTTTAACTTAATAATATTATCTTTATTTT	TATCT	36%	0	0	Phage infection protein
CAETHG_2.07.h	Peptoclostridium difficile strain Z31	gb|CP013196.1|	Plus/Minus	ATTTT	CCAAAATGACTAATATAATAATAAATGTACCCTTTC	GCACA	56%	0	0	N/A
CAETHG_2.07.i	Clostridium pasteurianum DSM 525	gb|CP009268.1|	Plus/Minus	TATAG	TTTAATTATATAATAGAATATTATCTCATCCATTAT	AGGAT	0%	0	0	N/A
CAETHG_2.07.j	Clostridium carboxidivorans P7	gb|CP011803.1|	Plus/Plus	TAGCA	AAGATAATGTTCAAGCAGTATTATCTTATCCATTAT	ATGAT	50%	0	1	N/A
CAETHG_2.07.k	Clostridium carboxidivorans P7	gb|CP011803.1|	Plus/Minus	ATAGT	AAAATATGACTAATATAATTGTATATAACATATACG	AATAG	50%	0	1	N/A
CAETHG_2.07.l	Clostridium butyricum strain JKY6D1 chromosome 1	gb|CP013352.1|	Plus/Minus	AATAA	TTAAAATGACTAATACAATATTAGCTATTATAACAA	TTACA	61%	0	0	N/A
CAETHG_2.09.a	Clostridium ljungdahlii DSM 13528	gb|CP001666.1|	Plus/Minus	CAAAT	AGTTTGGAGTAATAAATTATAGTTAATGTTATAATA	TAAAG	56%	1	0	N/A
CAETHG_2.09.b	Bacillus thuringiensis strain XL6 plasmid	gb|CP013001.1|	Plus/Minus	TCACA	TAATTGGAGTAATAAATTAGCTACAGAGACCTTAGC	ATTAT	44%	0	0	N/A

Table 5.3: Spacer analysis of *C. autoethanogenum* CRISPR locus 3
In silico analysis of CRISPR locus 1. Protospacer ID is signified by CAETHG prefix, followed by Spacer co-ordinate, with letters to differentiate between alternative hits of the same spacer sequence against putative protospacer sequence. Organism and organism ID of protospacer as signified along with strand of protospacer. The putative protospacer sequences are signified along with putative 5' and 3' PAM sequences, and similarity information. Sequence identity is signified, with 5' and 3' weightings for PAM confidence. Notes column details significant information of putative protospacer origin.

Protospacer ID	Protospacer origin			Putative Protospacer and PAM sequences			Similarity and scoring			Notes
	Organism of origin	Organism ID	Strand	5' PAM	Putative Protospacer	3' PAM	Identity	5' Score	3' Score	
CAETHG_3.05.a	Clostridium perfringens strain CP1 plasmid pCpb2-CP1	gb|JQ655732.1|	Plus/Plus	TTAGT	CAATACAGATACCAATCTTAGGAATTTACTAATGTT	TGGTA	47%	0	0	N/A
CAETHG_3.05.b	Clostridium perfringens plasmid pJIR3844	gb|JN689217.1|	Plus/Plus	TTAGT	CAATACAGATACCAATCTTAGGAATTTACTAATGTT	TGGTA	47%	0	0	N/A
CAETHG_3.12.a	Clostridium botulinum Prevot_594	gb|CP006902.1|	Plus/Minus	TACTG	TAAAAAATACAGATAAAGAAATAGATACTGTAGCA	GTATC	67%	2	0	N/A
CAETHG_3.12.b	Clostridium perfringens strain FORC_003	gb|CP009557.1|	Plus/Minus	TTCAC	AAAACAATACAGATATTGAAATAAAAGAAATCTGC	GAAAT	67%	0	0	N/A
CAETHG_3.13.c	Staphylococcus epidermidis strain SP1 plasmid pSP01	gb|KR230047.1|	Plus/Plus	GATTG	GTAGTCACTTAGATAAAGAAATCGAAGATTTAAAC	CAACA	61%	0	0	N/A
CAETHG_3.05.a	Clostridium perfringens strain CP1 plasmid pCpb2-CP1	gb|JQ655732.1|	Plus/Plus	TTAGT	CAATACAGATACCAATCTTAGGAATTTACTAATGTT	TGGTA	47%	0	0	N/A
CAETHG_3.05.b	Clostridium perfringens plasmid pJIR3844	gb|JN689217.1|	Plus/Plus	TTAGT	CAATACAGATACCAATCTTAGGAATTTACTAATGTT	TGGTA	47%	0	0	N/A
CAETHG_3.12.a	Clostridium botulinum Prevot_594	gb|CP006902.1|	Plus/Minus	TACTG	TAAAAAATACAGATAAAGAAATAGATACTGTAGCA	GTATC	67%	2	0	N/A
CAETHG_3.12.b	Clostridium perfringens strain FORC_003	gb|CP009557.1|	Plus/Minus	TTCAC	AAAACAATACAGATATTGAAATAAAAGAAATCTGC	GAAAT	67%	0	0	N/A

Analysis of each of the CRISPR loci resulted in the identification of the putative protospacers listed in Tables 5.1, 5.2 and 5.3. Putative PAM sequences were scored based on the sequence identity of the protospacer to the spacer. Unfortunately, the number of spacers with significant similarity to any putative protospacers was too low to draw any significant conclusion regarding a conversed sequence. However, this analysis did identify spacer 1.27 as having 100% sequence identity to a putative protospacer in *C. ljungdahlii* (protospacer 1.27a).

As also identified by Pyne and colleagues (2016a), protospacer 1.27a is present in a phage-related protein. Aside from the absence of a CRISPR system, *C. ljungdahlii* is extremely similar (>98% sequence identity) to *C. autoethanogenum*. This key distinction theoretically leads to a second difference; the increased presence of phage-related sequences in the *C. ljungdahlii*.

5.2.3. PAM identification

C. autoethanogenum and *C. ljungdahlii* are extremely similar organisms, displaying 98% sequence identity. One key difference between them is the presence of the Type I-B CRISPR system in *C. autoethanogenum*. As could be expected for a bacterium lacking a CRISPR system (compared to an almost identical species with one), *C. ljungdahlii* contains more phage-related proteins, including protospacer 1.27a.

With the similarities of these two *Clostridium* species in mind, and the known narrow host range of phages, it is conceivable to imagine that spacer 1.27 and protospacer 1.27a are either derived from the same, or closely related phage. Continuing this hypothetical scenario, lack of a CRISPR system would enable phage integration in *C. ljungdahlii*, but conversely, the presence of a CRISPR system in *C. autoethanogenum* would prevent this. The former would result in the presence of the putative 1.27a protospacer sequence in a phage-related gene on the *C. ljungdahlii* chromosome, and the latter would result in the integration of a spacer in a *C. autoethanogenum* CRISPR locus (spacer 1.27). Based on this assumption, the immediate 5' and 3' regions to protospacer 1.27a may provide information on the PAM of the Type I-B CRISPR system. If, like the putative protospacer sequence, the flanking regions of 1.27a are also sequence identical to the phage that resulted in the integration of spacer 1.27, in their presence, any CRISPR locus encoded sequence should be cleaved. As PAMs are known to occur on either the 5', or 3' side of protospacers, analysis of each putative PAM in isolated would in theory identify which side the PAM of the *C. autoethanogenum* system lies.

5.2.4. Dispute with Pyne and colleagues

In research published by Pyne and colleagues (2016a) offering in depth analysis of the *C. pasteurianum* Type I-B CRISPR system, limited *in silico* analysis of the *C. autoethanogenum* CRISPR system was also provided. Here, spacer 1.27 was also identified as being derived from *C. ljungdahlii*, along with proposal of a 5' PAM (5'-PAM-protospacer-3') sequence, 5'-ATTAA-3'.

Pyne colleagues (2016a) claim the Type I CRISPR system always have 5' PAMs. However, this is disputed both by the literature, and more in depth *in silico* analysis performed by Luo and colleagues (2015).

Analysis of the CRISPR locus 1 provides only two putative promoters, both of which occur on the negative strand of the Humphreys and colleagues (2015) genome (published as part of this study, see Chapter 3). Figure 5.1 illustrates the Cas operon in *C. autoethanogenum*, and also shows all 3 CRISPR loci are in the same antisense orientation (displayed as sense for easier comprehension). In this regard, spacer 1.27 has the sequence displayed in Table 5.1, meaning the putative 5' and 3' PAMs from protospacer 1.27a are 5'-CTTCA-3', and 5'-TTAAT-3' respectively. This is contrary to Pyne and colleagues' (2016a) identification of the reverse-complement of this sequence, which identifies 5'-AATTA-3' as the putative 5' protospacer. Although this difference is to some extent trivial, with a 3' PAM of 5'-TTAAT-3' and 5' PAM of 5'-ATTAA-3' being essentially the same, it is important to note the following: This study independently identified the same PAM as Pyne

and colleagues (2016a) but for reasons explains above regards it as a 5'-spacer-PAM-3' sequence of 5'-TTAAT-3'.

5.2.5. Spacer acquisition assay

In an attempt to gain insight into *C. autoethanogenum* PAM usage, a spacer acquisition assay was performed. Here, a cocktail of conjugative donor plasmids was used to conjugate *C. autoethanogenum* repeatedly without selection pressure. Unfortunately, after over twenty iterations of this process, colony PCR yielded no observable increase in size in any of the 3 CRISPR loci, suggesting no spacer uptake.

5.2.6. Protospacer 1.27a analysis

To assess the viability of the 1.27a protospacer hypothesis, an experiment was designed to assess CRISPR activity upon the protospacer region. Here, the 5 nt sequences immediately flanking 1.27a in the *C. ljungdahlii* phage related protein were cloned separately along with the protospacer into pMTL86151. To achieve this, the application specific module of pMTL86151 was amplified using a separate forward primer for each of the plasmid variations (CLAU-spacer-1.27-5'PAM-F, CLAU-spacer-1.27-3'PAM-F, and CLAU-spacer-1.27-noPAM-F), and the same reverse primer (CLAU-spacer-1.27-R). The overhang of each forward primer contained the protospacer sequence and relevant PAM(s), as described in Table 5.4 as well as a *Pst*I restriction site. The universal reverse primer overhang included an *Asc*I restriction site, enabling the resulting PCR product to be digested with these enzymes and inserted into a pMTL86151 vector linearised with the same enzymes via classic molecular cloning.

This resulted in a putative 5' PAM plasmid (pMTL86151-5'PAM), putative 3' PAM plasmid (pMTL86151-3'PAM), a plasmid without either putative PAM (pMTL86151-noPAM) and a pMTL86151 backbone control vector. In theory, analysis of conjugation frequency of each vector would enable identification of which encoded a functional PAM. As illustrated in Figure 5.2, the vector containing the 1.27a protospacer sequence with its immediate 5' putative PAM showed a 50% reduction in CFU formation compared to its 3' counterpart and the pMTL86151 control vector. This suggests an active CRIPSR system, whereby spacer 1.27 encodes targeting of its corresponding target sequence, in the presence of 5' sequence of protospacer 1.27a from a *C. ljungdahlii* phage related protein.

Table 5.4: Putative PAM and protospacer sequences for 1.27a PAM
Sequences represented 5'-3' with putative PAMs and protospacer 1.27 detailed. PAMs of interest are highlighted in bold and underlined, with italicised PAMs representing unmodified plasmid sequences.

Plasmid	5' PAM	Protospacer Sequence	3' PAM
pMTL86151-5' PAM	**CTTCA**	GCACCTAAGTAATCTATAAGTAAAGTATCAACTCTT	*AAGCT*
pMTL86151-3' PAM	*TCGAG*	GCACCTAAGTAATCTATAAGTAAAGTATCAACTCTT	**TTAAT**
pMTL86151-no PAM	*TCGAG*	GCACCTAAGTAATCTATAAGTAAAGTATCAACTCTT	*AAGCT*
pMTL86151	N/A	N/A	N/A

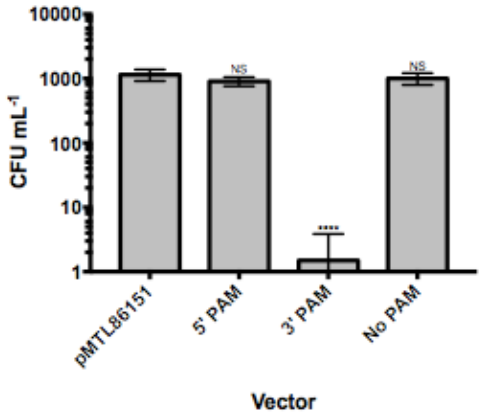


Figure 5.2: Protospacer 1.27a putative PAM analysis
Conjugation efficiencies of 1.27A vectors outlined in Table 5.4 in wild type *C. autoethanogenum*.

5.3. Top-down *C. autoethanogenum* PAM elucidation

Due to the identification of protospacer 1.27a in *C. ljungdahlii*, a fully degenerate approach to PAM identification was not required. Testing each possible perturbation of a 5 nt sequence would result in >1000 potential plasmids, far beyond the capabilities of manual efforts. Instead, the verification of the 3' putative 5'-TTAAT-3' PAM being sufficient for CRISPR targeting of a spacer 1.27 presented the opportunity to refine a minimal PAM sequence through a top-down approach.

In this approach, each of the 5 nucleotides were varied to their 3 possible alternatives, whilst maintaining the remaining 4 nucleotides of the sequence. To generate this library, pMTL86151 was digested with *Pst*I and *Asc*I, and NEB HiFi oligo assembly performed using the oligonucleotides outlined in Table 5.5.

Table 5.5: Plasmid library for top down *C. autoethanogenum* PAM elucidation
Plasmids used to study the 3' PAM of the *C. autoethanogenum* Type IB CRISPR system, and the oligonucleotides used to generate them.

Plasmid	HiFi Oligonucleotide
pMTL86151-1.27-ATTAA	1.27-ATTAA-HiFi-oligo
pMTL86151-1.27-TTTAA	1.27-TTTAA-HiFi-oligo
pMTL86151-1.27-CTTAA	1.27-CTTAA-HiFi-oligo
pMTL86151-1.27-GTTAA	1.27-GTTAA-HiFi-oligo
pMTL86151-1.27-AATAA	1.27-AATAA-HiFi-oligo
pMTL86151-1.27-ACTAA	1.27_ACTAA-HiFi-oligo
pMTL86151-1.27-AGTAA	1.27_AGTAA-HiFi-oligo
pMTL86151-1.27-ATAAA	1.27_ATAAA-HiFi-oligo
pMTL86151-1.27-ATCAA	1.27_ATCAA-HiFi-oligo
pMTL86151-1.27-ATGAA	1.27_ATGAA-HiFi-oligo
pMTL86151-1.27-ATTTA	1.27_ATTTA-HiFi-oligo
pMTL86151-1.27-ATTCA	1.27_ATTCA-HiFi-oligo
pMTL86151-1.27-ATTGA	1.27_ATTGA-HiFi-oligo
pMTL86151-1.27-ATTAT	1.27_ATTAT-HiFi-oligo
pMTL86151-1.27-ATTAC	1.27_ATTAC-HiFi-oligo
pMTL86151-1.27-ATTAG	1.27_ATTAG-HiFi-oligo

As illustrated in Figure 5.3, modification of each of the first two residues showed significant increases in CFU mL^{-1} for all nucleotides other than thymine. For the third and fourth residues, modification from adenine to thymine resulted in significant increases in CFU mL^{-1}, with both showing moderate (but not statistically relevant) increases when modified to cytosine and guanine. For the fifth residue, no significant increase was observed for any modification.

Through inverting the data illustrated in Figure 5.3, a more refined PAM prediction can be attained, with residues resulting in lower CFU mL^{-1} being represented to higher degree. As shown in Figure 5.4, based on this data, a conservative 3' PAM prediction would be 5'-TTAA-3'. A less conservative predicted PAM of 5'-TTVV-3' is possible, however, without further investigation, not statistically certain.

C. autoethanogenum reductive PAM assay

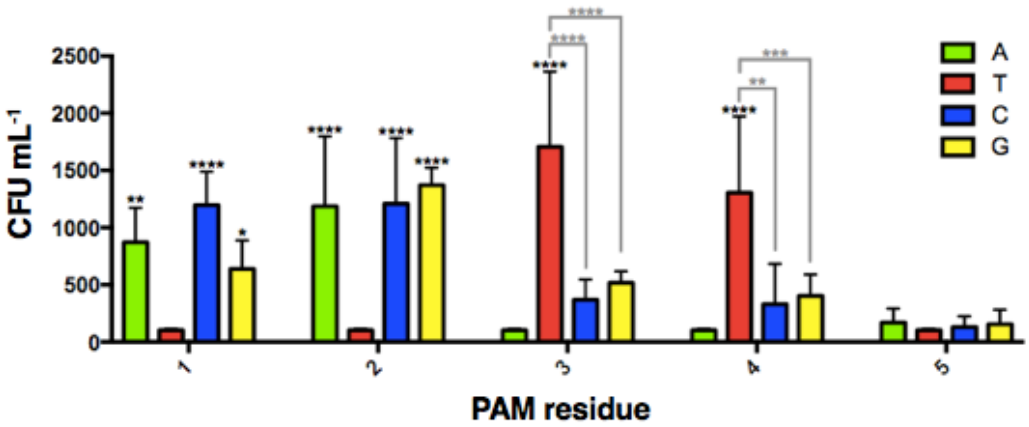


Figure 5.3: Reductive PAM assay of 3' 1.27a putative PAM
This figure shows CFU mL^{-1} of variants of the 5'-TTAAT-3' putative 3' PAM from protospacer 1.27 identified as described in Figure 5.2. A series of pMTL86151 plasmids were constructed with each possible variation of the PAM, with the CFU mL^{-1} for each displayed as shown above. Each nucleotide residue is labelled 1-5 (5'-3'), with the following colours representing each nucleotide (from left to right); green (adenine), red (thymine), blue (cytosine), yellow (guanine). The control vector (pMTL86151-1.27-TTAAT) CFU mL^{-1} is represented within each residue; in residue 1, 2 and 5 as thymine, and in residues 3 and 4 as adenine. Statistical significance between the modified versions of this control plasmid and the original vector calculated via two-way ANOVA ($F(12, 60)=12.48$), and show in black above error bars. P values for statistical significance between cytosine/guanine and thymine for residues 3 and 4 represent by grey bars and asterisks $P<0.05$ represented by single asterisk, $P\leq 0.01$ represented by two astericees, $P\leq 0.001$ represented by three astericees, and $P \leq 0.0001$ represented by four astericees.

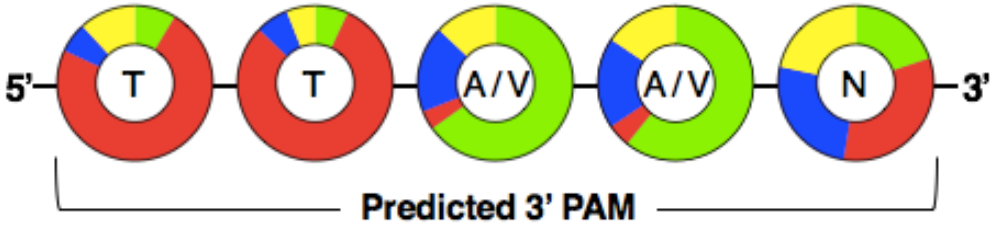


Figure 5.4: Predicted *C. autoethanogenum* 3' PAM
Under the assumption that functional PAM sequences enable CRISPR-mediated reduction of conjugation, this figure displays the likely required residues for the 3' PAM for the *C. autoethanogenum* Type IB system. CFU mL^{-1} for variants of each residue are inverted and displayed relative to each other in parts of a whole graphs. Using the same colour coding described for Figure 5.3, the predicted PAM sequence can be visualised as shown. Resulting in the conservative predicted reduced PAM of 5'-TTAA-3', with a less conservative prediction of 5'-TTVV-3'.

5.4. Bottom-up PAM elucidation

Analysis of reductive PAM assay appears to identify the 3' PAM sequence as being 5'-TTAA-3'. However, data on the third and fourth residues was not conclusive in identifying the minimal essential nucleotide sequence. Despite significant differences observed between modified residues, the stochastic nature of conjugation as a method of DNA transfer limits the accuracy of this method of screening. Such inaccuracies could be attributed to variable amounts of plasmid transfer during conjugation due to various factors affecting the conjugative donor.

To further understand the nature of the *C. autoethanogenum* PAM, a robotics screen was designed to elucidate the exact sequence, or sequences, able to enable CRISPR activity. Using the high-throughput capabilities of a robotics setup, the limitations of manual microbiology described in section 5.3 can be overcome. With regards to PAM screening, this requires the generation of a degenerate library two orders of magnitude bigger than that described in section 5.3, allowing full coverage of all potential PAM sequences.

For this study, unfortunately, no anaerobic robotics setup was available. Therefore, a bottom-up approach for PAM elucidation was designed, with the CRISPR expressed in *E. coli* and a robotics manifold used to analyse its interference with a library of plasmids containing the 1.27 spacer with degenerate 6 nt sequences either side. The experimental design is illustrated in Figure 5.5.

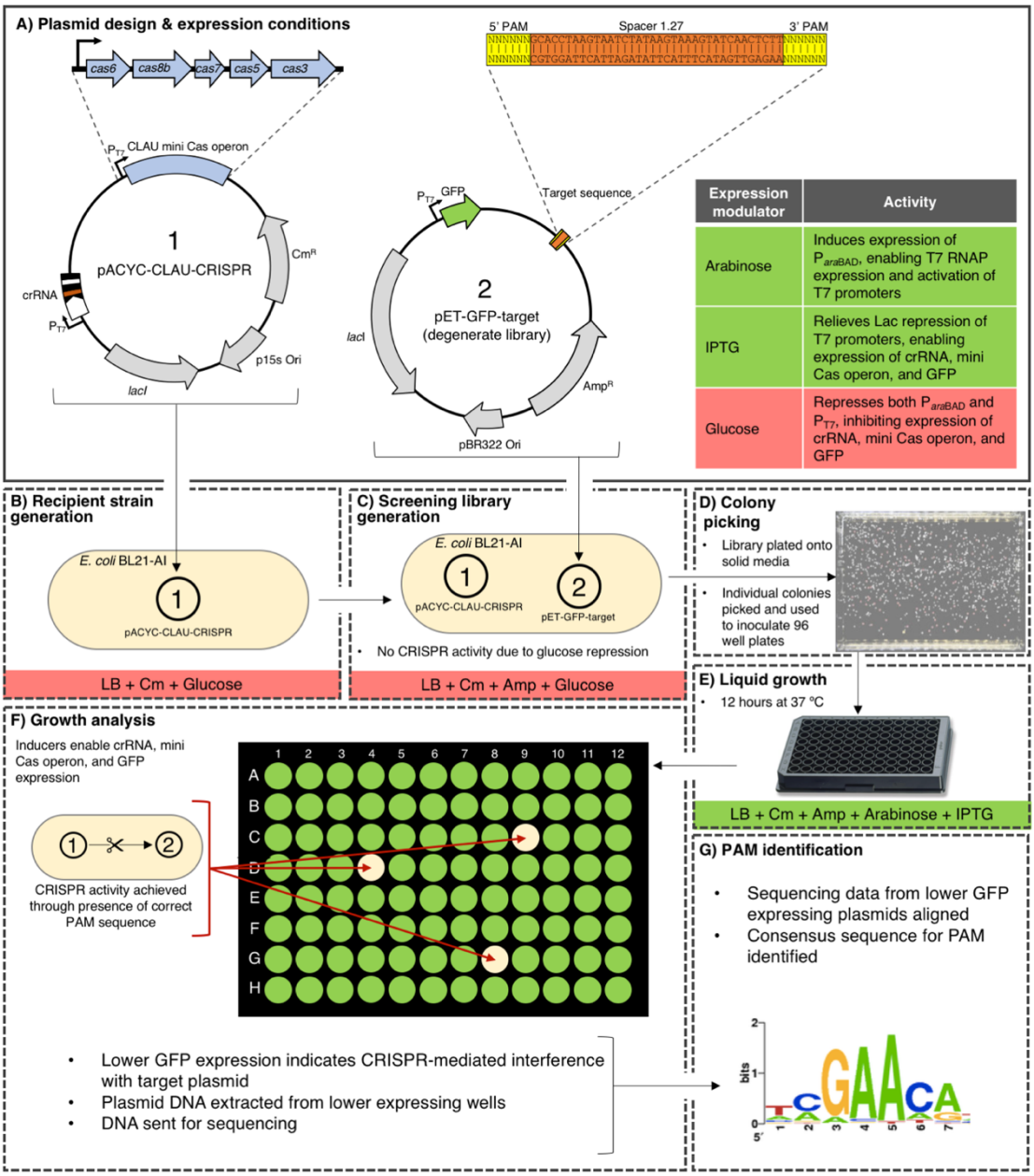


Figure 5.5: Bottom-up PAM screen

Flow through of robotics based PAM screen. A) Vector composition of pACYC-CLAU-CRISPR (designed to target spacer 1.27 with CRISPR components under inducible T7), and pET-GFP-target (containing protospacer 1,27 flanked by 6 nt degenerate sequences), and table of expression modulator activity. pACYC-CLAU-CRISPR consists of a pACYC-duet backbone. B) *E. coli* BL21-AI harbouring pACYC-CLAU-CRISPR capable of inducible CRISPR activity when a protospacer 1.27 present proximal to a correct PAM sequence. C) pET-GFP-target library transformed in to BL@10AI pACYC-CLAU-CRISPR with glucose repressing the CRISPR and GFP expression D) Library of colonies, each theoretically containing a unique PAM sequence inoculate liquid media containing with no upregulation of CRISPR and GFP. F) Cultures displaying lower fluorescence are presumed to have pET-GFP-target copy number reduced (lowering GFP expression), due to CRISPR-activity. G) Plasmids showing reduced GFP expression are sequenced and their putative PAM regions aligned to gain a consensus sequence for the *C. autoethanogenum* PAM.

As shown in Figure 5.5, this approach consists of two plasmids; pACYC-CLAU-CRISPR and pET-GFP-target. The former consists of the necessary Cas genes for CRISPR activity (*cas6*, *cas8b*, *cas7*, *cas5*, and *cas3)* and a synthetic CRISPR locus, using a pACYC-duet backbone. The Cas genes were amplified from *C. autoethanogenum* gDNA using primers CLAU-endocas-HiFi-F and CLAU-endocas-HiFi-R, and inserted into *Nco*I and *Pac*I digested pACYC-duet MCS1 via NEB HiFi assembly. Subsequently, the synthetic crRNA containing spacer 1.27 was amplified from a DNA 2.0 GeneStrings fragment using CLAU-endo1.27-F and CLAU-endo1.27-R primers. This PCR product was inserted via NEB HiFi assembly into the intermediate plasmid described above (digested with *Not*I and *Pac*I). Sequence confirmation of the final vector was confirmed via Sanger sequencing with M13-R and T7-F.

For the pET-GFP-target plasmid, the *Aequorea victoria* green fluorescence protein gene was amplified using GFP-F-NdeI and GFP-R-PacI primers from a plasmid supplied by PhD candidate Matthew Lau, and inserted via classic molecular cloning into pET-duet using *Nde*I and *Pac*I. Following Sanger sequencing confirmation with M13-R, the intermediate vector was digested with *Bam*HI and *Asc*I and the oligonucleotide 1.27-degPAM-HiFi-oligo was used to create the degenerate library via NEB ssDNA HiFi assembly.

5.4.1. Preliminary proof of concept

To gain proof of concept for this bottom-up PAM screen, a trial run of the robotics flowthrough was tested using a series of control vectors, in collaboration with Dr Gareth Little. The BL21-AI strain containing pACYC-

CLAU-CRISPR vector was transformed independently with the following three vectors; pET-GFP, pET-GFP-TTAAT, and pET-GFP-target (degenerate PAM library). Following transformation, a series of colonies were picked and grown in liquid media, with their fluorescence and OD_{600} measured after 12 hours. Dividing fluorescence by OD_{600} provides a metric for CRISPR activity on the target plasmid

In Figure 5.6, clonal cultures inoculated from single colonies are grouped into bins defined by their range of fluorescence/OD_{600}. Plotting of the frequency of cultures within these bins shows distinct distribution patterns for three plasmids used for the study.

For the pET-GFP-TTAAT positive control, the entire library has a low fluorescence: OD_{600} ratio compared to the other two vectors. With almost all cultures measured displaying a ratio of $<5 \times 10^7$.In contrast, the pET-GFP negative control shows almost all cultures to have a ratio of $>5 \times 10^7$, showing practically no overlap with the fluorescence: OD_{600} distribution for the positive control.

For the pET-GFP-target vector library, a slightly lower average fluorescence:OD_{600} ratio was observed compared to the negative control. Again the majority of cultures displayed a fluorescence:OD_{600} ratio above 5×10^7. However, unlike the negative control, there is a significant distribution of cultures below this threshold.

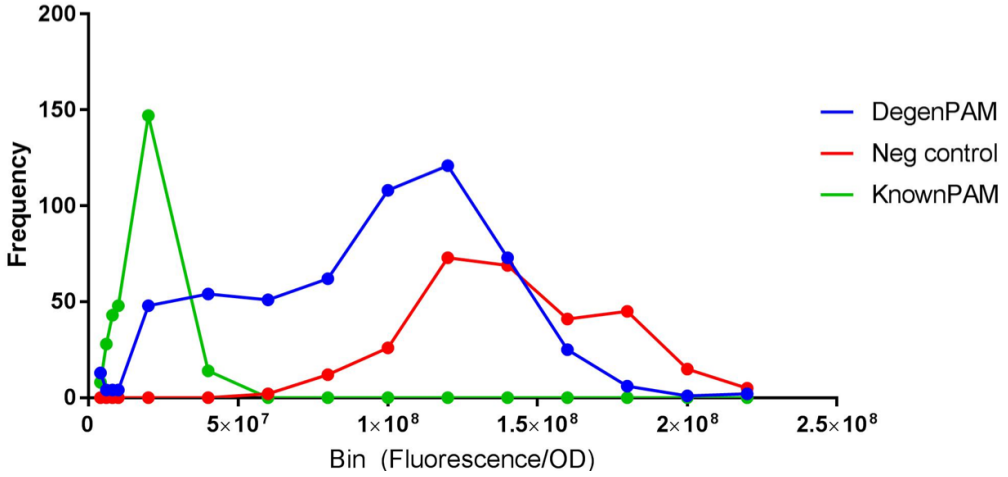


Figure 5.6: GFP expression in *E. coli* BL21-AI pACYC-CLAU-CRISPR cultures harbouring pET-GFP vectors
This histogram represents distribution frequency of pET-GFP, pET-GFP-TTAAT, and pET-GFP-target fluorescence:OD_{600} ratios. Cultures were inoculated from individual colonies resulting from transformations of the pET vector into *E. coli* BL21-AI pACYC-CLAU-CRISPR strain, with fluorescence and OD_{600} measured after 12 hours. Cultures were grouped into bins based on defined ranges of fluorescence:OD_{600} ratio, and the frequency of cultures in said bins plotted.
The blue line displays cultures from the degenerate pET-GFP-target vector library, containing the 1.27 target region flanked by degenerate 6 nt PAM sequences. The red line represents the negative control, cultures harbouring pET-GFP vector lacking the 1.27 target region. The green line represents the positive control, cultures harbouring the pET-GFP-TTAAT vector, containing the 1.27 target region, and the 3' PAM 5'-TTAAT-3' shown to be sufficient for CRISPR activity in previous experiments.

5.4.1.1. Proof of concept conclusions

As expected, the pET-GFP-TTAAT positive control vector cultures showed a low fluorescence:OD_{600} ratio, whilst the pET-GFP negative control cultures showed a high ratio. The distribution of the fluorescence:OD_{600} of these cultures showed almost no overlap, with almost all recorded negative control cultures displaying a higher ratio than all recorded positive control cultures. These data suggest that CRISPR activity from the pACYC-CLAU-CRISPR vector encoded mini Cas operon and crRNA reduces GFP expression from the positive control vector. It is hypothesised this is due to the presence of the

1.27 protospacer and 5'-TTAAT-3' PAM sequence described in section 5.3, with Cas3-mediated cleavage reducing copy number and therefore GFP expression. Conversely, lack of any crRNA-encoded target sequence on the negative control vector ensures a higher copy number and therefore higher fluorescence.

Distribution of the fluorescence:OD_{600} ratios of clonal cultures of the two control vectors shows a clear drop in GFP expression for plasmid harbouring a target sequence flanked by a correct PAM. The distribution of the degenerate pET-GFP-target vector library agrees with this, showing the majority of the library being more consistent with results for the negative control. This is due to the majority of cultures within this library harbouring plasmids without the correct PAM sequence, and therefore able to escape CRISPR activity and retain higher GFP expression.

This result confirmed the viability of the robotics approach for PAM screening described above. By returning to the cultures displaying low fluorescence:OD_{600} ratios within the degenerate library, the PAM sequences capable of enabling CRISPR activity on a crRNA encoded target region could be identified.

5.5. *C. autoethanogenum* engineered Cas operon (eCas) strain for transcriptomic control

To develop a chromosome-based CRISPR system for CRISPRi, the expression of Cas genes was analysed in using transcriptomic data compiled by Dr Anne Henstra. As shown in Dr Anne Henstra (personal communication, unpublished research), expression of Cas genes is consistently relatively low throughout a range of growth conditions.

For CRISPRi applications, replacing the endogenous Cas and CRISPR locus promoters would offer an ideal solution to weak expression of endogenous promoters. Although this would enable tuneable control of gene expression, to date no such system has been developed for use in *C. autoethanogenum*, however, development in this area is being pursued. In the meantime, a solution was planned using constitutively high expression promoters, in order to achieve proof of principle. In addition to modifying endogenous promoters, development of an eCas screen would require deletion of *cas3*, coding for the Cas3 nuclease component of the *C. autoethanogenum* CRISPR system.

5.5.1. Deletion of *cas3*

Deletion of *cas3* was achieved through use of the vector pMTL43151-CLAU-cas3. This vector was generated using pMTL43151-CLAU-pyrE as a backbone, with modifications made as described in section 4.4. For sgRNA insertion, CLAU-cas3-sg1-HiFi was used, and for homology arm cassette generation CLAU-cas3-LHA-F-AsiSI, CLAU-cas3-LHA-R, CLAU-cas3-RHA-F, and CLAU-cas3-RHA-R-AscI were used.

Conjugation of this pMTL843151-CLAU-cas3 resulted in ~30 colonies, 8 of which were screened as shown in Figure 5.7A. Screening resulted in the identification of 5 knockouts (colonies 1, 3, 5, 6, & 8). Colony 2 displayed a mixed population of wild type and knockout, with colony 7 displaying wild type genotype, and colony 4 failing to produce a PCR product. Following confirmation of gene deletion via colony PCR, mutants were verified via Sanger sequencing and underwent plasmid loss.

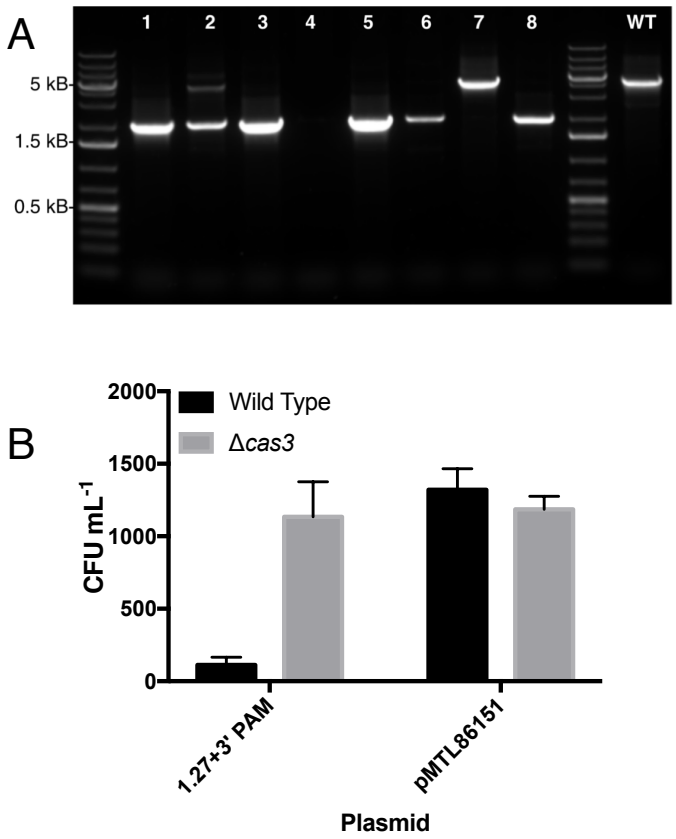


Figure 5.7: *C. autoethanogenum* Δ*cas3* colony PCR and CRISPR activity comparison
A) PCR of *C. autoethanogenum* pMTL43151-CLAU-Cas3 transconjugants generating using CLAU-Cas3-scr-F and CLAU-Cas3-scr-R. Lanes 1-8 represent colonies screened from selection plates, with "WT" lanes representing a wild type control. The wild type *cas3* locus is ~5 kB, with a band size of ~2 kB representing the size of the knockout locus. See Figure 4.1 for schematic explanation of colony PCR.
B) Both strains of *C. autoethanogenum* were used as conjugative recipients for Sexpress *E. coli* harbouring pMTL86151-1.27-3'PAM (1.27+3'PAM) and pMTL86151. Resulting CFU $mL^{=1}$ for each conjugation were calculated and shown above with black bars representing wild type and grey representing Δ*cas3*.

Following plasmid loss, loss of CRISPR-based nucleic activity was confirmed as shown in Figure 5.7B. Here, no observable reduction in CFU mL^{-1} is seen when the Δ*cas3* strain is conjugated with pMTL86151-1.27-3'PAM (containing spacer 1.27 immediately upstream of the confirmed 5'-TTAA-3' PAM) compared to the pMTL86151 control. In contrast, as shown in 3.3, wild type *C. autoethanogenum* (as previously illustrated in Figure 5.2) shows a significant reduction in CFU mL^{-1} when conjugated with a spacer/PAM containing plasmid.

5.5.1. eCas plasmid proof of principle

To develop the endogenous *C. autoethanogenum* CRISPR system for knockdown, the pMTL43151-CLAU-eCas vector series was designed. The rationale behind this approach is illustrated in Figure 5.8.

As shown in Figure 5.8, the eCas vector is designed to delete a 3.5 kB region immediately upstream of the Cas operon. This consists of CRISPR locus 1 (see section 5.2.2) and its native promoter, the Cas operon promoter, and a region of unknown function between the two. The homology arms for pMTL43151-CLAU-eCas are designed to replace this region with a synthetic CRISPR locus driven by *C. tetani* P_{fdx}, as well as the *C. acetobutylicum* P_{fdx} promoter for upregulation of Cas operon expression. Promoters selected based on high expression (see Appendix B).

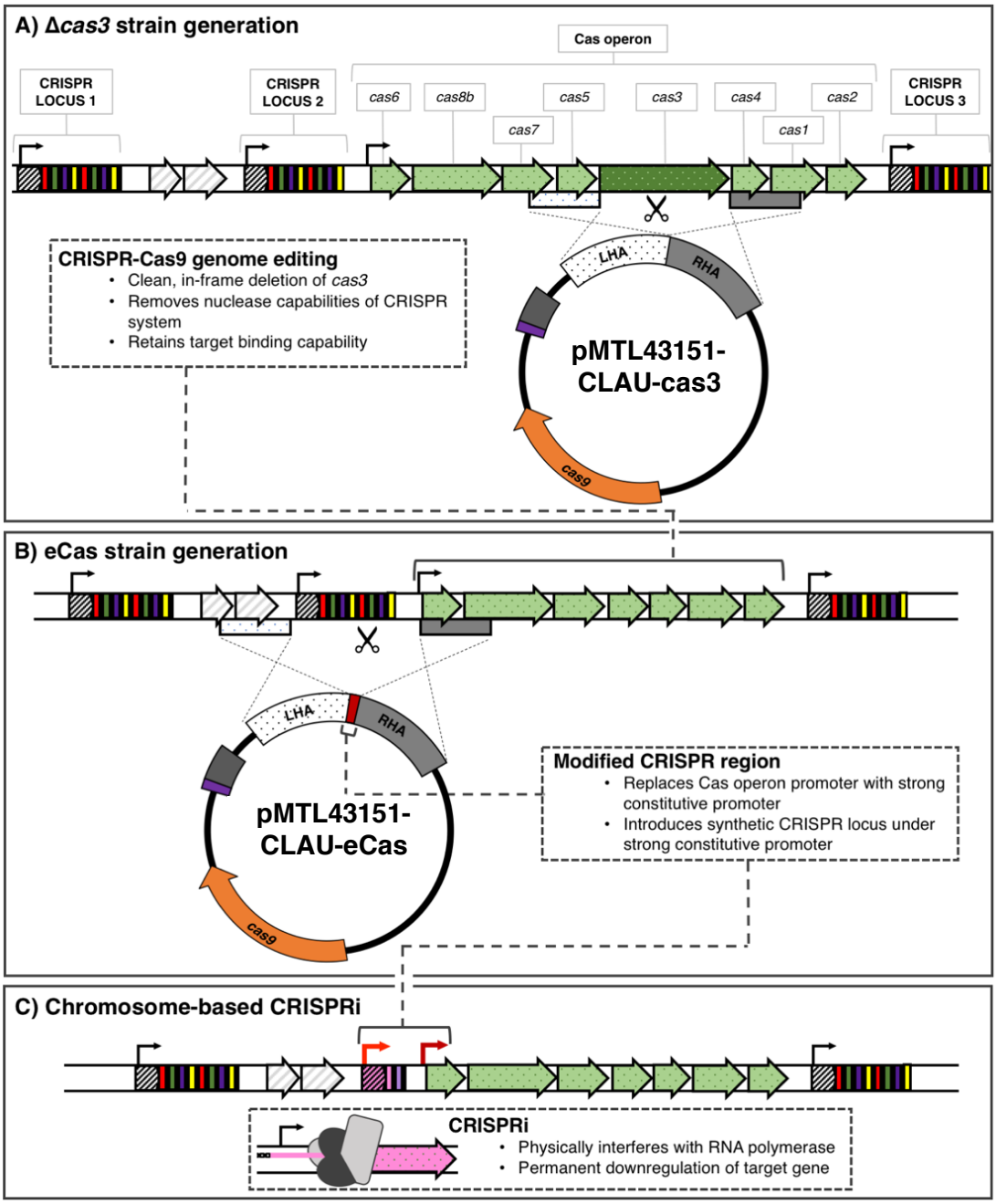


Figure 5.8: *C. autoethanogenum* eCas strain development
A) deletion of *cas3* via pMTL43151-CLAU-cas3 removes the nuclease capabilities of the *C. autoethanogenum* endogenous CRISPR system whilst leaving its target binding capability intact. B) Deletion of CRISPR locus 2 (located immediately upstream of the Cas operon) and Cas operon promoter is achieved with pMTL43151-CLAU-eCas, which replaces the removed section with a synthetic CRISPR locus under the control of a constitutively strong promoter, along with a second strong constitutive promoter for the Cas operon. C) The Cascade complex of the endogenous CRISPR system is guided to the target encoded by the modified CRISPR system, where is physically interferes with RNA polymerase activity, reducing transcription capacity at that locus.

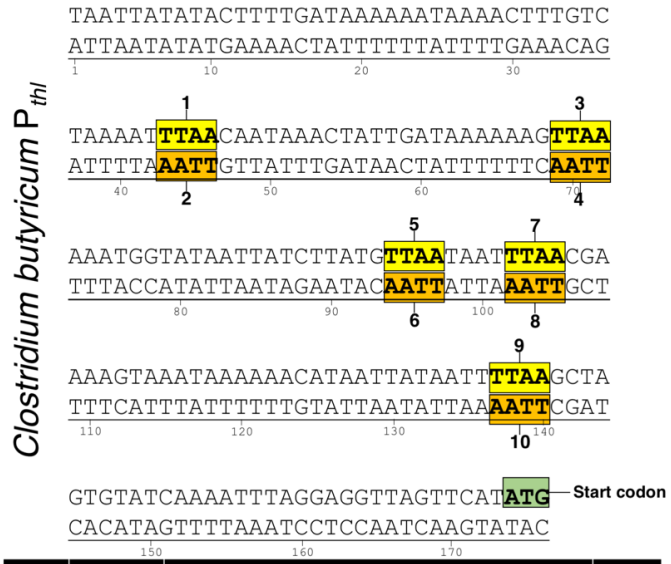
Within the synthetic CRISPR locus are two spacer regions, spacer 1.27 and an I-*Sce*I restriction site containing spacer. The latter was included to enable guide RNA modification using the NEB HiFi ssDNA discussed in section 4.4.1. This would allow easy reprogramming of targets for knockdown, with the parent pMTL43151-CLAU-eCas vector being used as a backbone for each new eCas vector. With regards to the former, the 1.27 spacer was included for the possibility of using the vector in a wild type (*cas3* intact) strain in order to repeat the top-down PAM assay (section 0), if the bottom-up assay (section 5.4) failed to yield significant results.

The pMTL43151-CLAU-eCas vector was constructed first by constructing the homology cassette via HiFi assembly. This was achieved through amplification of the eCas LHA with primers CLAU-eCas-LHA-F-AsiSI and CLAU-eCas-LHA-R, and the RHA with primers CLAU-eCas-RHA-R and CLAU-eCas-RHA-R-AscI. The final component of the homology cassette, containing the synthetic crRNA and promoters described above was amplified from CLAU-eCas-synRNA DNA2.0 GeneStrings synthetic DNA using primer CLAU-eCas-synRNA-F and CLAU-eCas-synRNA-R. All 3 homology cassette components were assembled via HiFi assembly, with the subsequent product being used as a template for amplification with CLAU-eCas-LHA-F-AsiSI and CLAU-eCas-RHA-R-AscI. This final PCR product was inserted into a pMTL43151-CLAU-pyrE backbone, replacing the existing homology cassette via classic molecular cloning using *Asc*I and *Asi*SI. Following this, the oligonucleotide CLAU-eCas-sg1-HiFi was used to insert the new seed region and Sanger sequence verified as described in section 4.4.1.

Prior to target specific modification of pMTL43151-CLAU-eCas, proof of principle of the vector's capability to modify the *C. autoethanogenum* CRISPR region must first be attained. However, proof of principle of the eCas approach for strain development was not achieved in this study due to lack of PAM sequence verification. As detailed in the discussion of this chapter, testing of this system delayed results of the bottom-down PAM assay.

However, using the putative 5'-TTAA-3' PAM described in section 5.3, spacers were designed to test knockdown via a CAT assay. Using this PAM sequence, the sequence of the *Clostridium butyricum* thiolase promoter was assessed[18], identifying 10 potential targets. The validity of these targets for use in CRISPRi were analysed manually. Each potential spacer was cross-referenced against the *C. autoethanogenum* genome to identify any high sequence identity, as shown by NCBI E-value score. The spacer sequence itself was also analysed for poly-adenine or poly-thymine regions known to hamper targeting in Cas9-based systems. Results of this analysis can be seen in Figure 5.9, identifying spacers 3, 8 and 9 as the best candidates.

18 Promoter selected due its high expression levels in *C. autoethanogenum*, and its lack of use in any of the functional components of pMTL43151-CLAU-eCas. Promoter strength data available in Appendix B.



PAM	Strand	Spacer sequence	Score
1	Positive	CAATAAACTATTGAT**AAAAA**AGTT**AAAAA**TGGTATA	33.7
2	Negative	ATTTTAGACAAAGTTTTA**TTTTT**TATCAAAAGTATA	31.9
3	Positive	AAATGGTATAATTATCTTATGTTAATAATTTAACGA	31.9
4	Negative	C**TTTTT**TATCAATAGTTTATTGTTAAATTTTAGACA	31.9
5	Positive	TAATTTAACGAAAAGTAAAT**AAAAA**ACATAATTATA	31.9
6	Negative	CATAAGATAATTATACCA**TTTTT**AAC**TTTTT**TATCA	33.7
7	Positive	CGAAAAGTAAAT**AAAAA**ACATAATTATAATTTTAAG	33.7
8	Negative	ATTATTAACATAAGATAATTATACCA**TTTTT**AACTT	33.7
9	Positive	GCTAGTGTATCAAAATTTAGGAGGTTAGTTCATATG	35.6
10	Negative	AATTATAATTATG**TTTTT**TATTTACTTTTCGTTAAA	33.7

Figure 5.9: PAM and spacer sequences available in the *C. butyricum* thiolase promoter
All instances of the *Clostridium autoethanogenum* 5' PAM (5'-TTAA-3') in the *Clostridium butyricum* thiolase promoter (P_{thl}). PAMs were named in order of appearance (coding strand, 5' to 3'), as shown on the promoter nucleotide sequence by yellow (positive strand) and orange (negative strand) boxes. The table below the promoter sequence lists the spacer sequence associated with each of the PAMs identified, with 3, 8, and 9 highlighted due to them being taken forward for knockdown analysis. Selection of spacers was achieved via cross referencing their sequences to the *C. autoethanogenum* genome, resulting in NCBI E-value scores. Poly A/T regions are highlighted in bold.

5.6. Analysis of the *E. limosum* endogenous CRISPR system

As mentioned in Chapters 3 and 4, the *E. limosum* ATCC 8486 strain was considered as a potential alternative to *C. autoethanogenum* due to DNA transfer issues. Further to the work on exemplifying the pMTL40000 exogenous CRISPR system, the endogenous CRIPSR system of *E. limosum* was also assessed for its potential use in genetic tools.

Through *in silico* analysis of the genome sequence generated by PhD candidate Robert Mansfield, a Type IB CRISPR locus was identified, along with 3 CRISPR loci (Table 5.6). Analysis of these loci resulted in the identification of several spacers with high sequence identity to sequences present in another strain of *E. limosum* (KIST 612). As detailed in Table 5.7 of the 6 spacers identified with >95% sequence identity to the KIST 612 strain, 5 were found in phage-related protein coding genes. Analysis of the 3' PAMs shows a conserved AG motif at positions two and three of the 3' PAM.

In order to identify the potential *E. limosum* ATC 8486 PAMs shown in Table 5.7 a series of vectors were constructed containing the stated putative protospacers and variations of possible PAMs. As outlined in Table 5.8, for each spacer, vectors were constructed with each possible PAM individually, with both possible PAMs, and without possible PAMs. Vectors were constructed through amplification of the pMTL87151 application specific module with the forward primer introducing the specific spacer and PAM(s), and cloning this PCR product into a pMTL87151 backbone using *Xho*I and *Asc*I restriction sites. Oligonucleotides used are detailed in table 2.2, with unique forward primers for each plasmid, and primer "ELM_spacer_R" as the universal reverse primer. Following library construction and Sanger sequence verification, the library was transformed into electrocompetent *E. limosum* ATCC 8486 cells. Unfortunately, time constraints meant the library could not be assessed for CRISPR activity.

Table 5.6: *E. limosum* CRISPR loci
Details of each of the 3 CRISPR loci (CR1-3) of *E. limosum* ATCC 8486 (ELM), including direct-repeat (DR) consensus sequence and length, and number of spacers and genomic position of locus.

CRISPR locus	DR consensus sequence	DR length	Number of spacers	Genome position
ELM CR1	GTTCGCTGGTTATCTTTGTTTTTCTGCCCTC	31	1	799908 - 800004
ELM CR2	ATTTAAATACATCTCATGTTAATCTTCAAC	30	15	2504039 - 2505051
ELM CR3	ATTTAAATACATCTCATGTTAATCTTCAAC	30	31	2843973 - 2846018

Table 5.7: Putative protospacers of *E. limosum* ATCC 8486
Spacers from *E. limosum* ATCC 8486 CRISPR loci (CR) with high sequence identity to sequences in *E. limosum* KIST 612. Details of the spacer, potential protospacer, and possible PAMs are listed, with subject gene of interest detailed in bold, and mismatches (MM) between spacer and subject sequence. Highlighted in grey is the conserved AG motif present at position 2 and 3 of the 3' potential PAMs.

ID	Spacer and subject sequence	Subject organism	MM (nt)	Possible PAMs	
				5'	3'
ELM CR2.08	AAGTTGTCCATTTTCAAAGGCATTCAAAACATCCAG ||||||||||| ||||||||||||||||||| AAGTTGTCCATTCTCAAAGGCATTCAAAACATCCAG	*Eubacterium limosum* KIST612 CP002273.2 **Phage tail protein**	1	AAGCC	AGACA
ELM CR2.12	CACGTTCGCCTTGCGGCCCGGTTTCCCCTTTCAG ||||||| ||||||||||||||||||| CACGTTCACCTTGCGGCCCGGTTTCCCCTTTCAG	*Eubacterium limosum* KIST612 CP002273.2 **Phage tail fibre protein**	1	TGGTC	AGACA
ELM CR3.11	GGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCA |||||||||||||||||||||||||||||||||| GGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCA	*Eubacterium limosum* KIST612 CP002273.2	0	GCACC	TGAAC
ELM CR3.28	TGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGA |||||||||||||||||||||||||||||||||| TGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGA	*Eubacterium limosum* KIST612 CP002273.2 **Phage tail terminase large subunit**	0	TTTTT	TGACT
ELM CR3.29	GGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGA |||||||||||||||||||||||||||||||||| GGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGA	*Eubacterium limosum* KIST612 CP002273.2 **Phage tail terminase**	0	TTCGT	TGATC
ELM CR3.30	GCCGATGAAGCCGGAGCCATGGACAGTTATCCGGT |||||||||||||||||||||||||||||||||| GCCGATGAAGCCGGAGCCATGGACAGCTATCCGGT	*Eubacterium limosum* KIST612 CP002273.2 **Phage tail terminase**	1	TTCTT	TGAGG

Table 5.8: *E. limosum* ATCC 8486 PAM analysis plasmids
Details of plasmids used for elucidation of *E. limosum* ATCC 8486 type IB CRISPR system. All plasmids based on the pMTL87151 vector (p19 Gram-positive replicon), with Protospacer/PAM sequences (as detailed) inserted at the application specific module.

Plasmid	5' PAM	Protospacer Sequence	3' PAM
pMTL87151-2.08-5'PAM	AAGCC	AAGTTGTCCATTCTCAAAGGCATTCAAAACATCCAG	N/A
pMTL87151-2.08-3'PAM	N/A	AAGTTGTCCATTCTCAAAGGCATTCAAAACATCCAG	AGACA
pMTL87151-2.08-5'+3'PAM	AAGCC	AAGTTGTCCATTCTCAAAGGCATTCAAAACATCCAG	AGACA
pMTL87151-2.08-noPAM	N/A	AAGTTGTCCATTCTCAAAGGCATTCAAAACATCCAG	N/A
pMTL87151-2.12-5'PAM	TGGTC	CACGTTCACCTTGCGGCCCGGTTTCCCCTTTCAG	N/A
pMTL87151-2.12-3'PAM	N/A	CACGTTCACCTTGCGGCCCGGTTTCCCCTTTCAG	AGACA
pMTL87151-2.12-5'+3'PAM	TGGTC	CACGTTCACCTTGCGGCCCGGTTTCCCCTTTCAG	AGACA
pMTL87151-2.12-noPAM	N/A	CACGTTCACCTTGCGGCCCGGTTTCCCCTTTCAG	N/A
pMTL87151-3.11-5'PAM	GCACC	GGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCA	N/A
pMTL87151-3.11-3'PAM	N/A	GGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCA	TGAAC
pMTL87151-3.11-5'+3'PAM	GCACC	GGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCA	TGAAC
pMTL87151-3.11-noPAM	N/A	GGAAGAAATCTTTGGTGTACTTATCTTTACTCGCCA	N/A
pMTL87151-3.28-5'PAM	TTTTT	TGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGA	N/A
pMTL87151-3.28-3'PAM	N/A	TGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGA	TGACT
pMTL87151-3.28-5'+3'PAM	TTTTT	TGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGA	TGACT
pMTL87151-3.28-noPAM	N/A	TGTTGTTGCGGTTCTGTGCACCAAGGCCAGAAGGGA	N/A
pMTL87151-3.29-5'PAM	TTCGT	GGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGA	N/A
pMTL87151-3.29-3'PAM	N/A	GGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGA	TGATC
pMTL87151-3.29-5'+3'PAM	TTCGT	GGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGA	TGATC
pMTL87151-3.29-noPAM	N/A	GGACATTGAAAAAGTACGCGCGTGCAAGATTAAGGA	N/A
pMTL87151-3.30-5'PAM	TTCTT	GCCGATGAAGCCGGAGCCATGGACAGCTATCCGGT	N/A
pMTL87151-3.30-3'PAM	N/A	GCCGATGAAGCCGGAGCCATGGACAGCTATCCGGT	TGAGG
pMTL87151-3.30-5'+3'PAM	TTCTT	GCCGATGAAGCCGGAGCCATGGACAGCTATCCGGT	TGAGG
pMTL87151-3.30-noPAM	N/A	GCCGATGAAGCCGGAGCCATGGACAGCTATCCGGT	N/A
pMTL87151	N/A	N/A	N/A

5.7. Analysis of other acetogenic and clostridial CRISPR systems

Following analysis of the *C. autoethanogenum* and *E. limosum* CRISPR systems, further analysis was conducted on other relevant members of the genus. Organisms were selected based on their potential for industrial applications, or health implications. A full overview of the systems identified in this study is shown in Table 5.9, with instances of CRISPR systems described elsewhere also shown.

Analysis of the endogenous CRISPR systems of these organisms was conducted as outlined in Chapter 2, with the CRISPRfinder tool used to identify CRISPR loci (Grissa, Vergnaud & Pourcel, 2007). For identification of Cas genes, the readily conserved Cas1 and Cas2 genes of *C. autoethanogenum* were used as queries for NCBI BLASTp.

In order to assess whether any commonly studied members of the genus *Clostridium* contained Class II CRISPR systems, a similar search technique was performed using the amino acid sequences of the Cas9 and CpfI endonucleases of characterised Class II CRISPR systems were used as query sequences to search against various databases. Results of this search found no Class II CRISPR systems present in any industrially or clinically relevant Clostridia.

However, using CRISPRfinder and BLASTp searches for Cas1/Cas2 homologs, a series of previously unidentified or unclassified CRISPR systems were analysed. Table 5.9 offers details of the type of CRISPR system identified and CRISPR loci present in the genome, with further details on each system described below.

Table 5.9: Analysis of clostridial CRISPR systems
Details of CRISPR system classifications from analysis performed in this study, as well as those identified in other studies. Strain information is provided along with GenBank ID. *Analysis source references: 1 (Mo *et al.*, 2015), 2 (Negahdaripour *et al.*, 2017), 3 (Boudry *et al.*, 2015), 4 (Makarova *et al.*, 2015), 5 (Pyne *et al.*, 2016b). **Multiple Type I and Type III system identified across several species. ***Work conducted in collaboration with PhD candidate Andrew Dempster, using genome sequence data generated by Andrew Dempster. ****Work conducted using genome sequence data generated by PhD candidate Robert Mansfield.

Organism	Type of CRISPR system	No. CRISPR loci	GenBank Reference sequence(s)	Analysis source
***C. aceticum* DSM 1496**	IB	2	CP009687.1 CP009688.1	This study
***C. acetobutylicum* ATCC 824**	N/A	0	AE001437.1 AE001438.3	This study
***C. acetobutylicum* GXAS18-1**	IIIA	7	JRWL01000001, JRWL01000049	1*
***C. autoethanogenum* DSM 10061**	IB	3	CP006763.1	This study
***C. beijerinckii* ASM83310v2**	N/A	N/A	CP010068.2	This study
***C. botulinum* [various**]**	Class I**	-	-	2*
***C. butyricum* KNU-L09**	IB	1	CP013251.1 CP013489.1	This study
***C. carboxidivorans* P7**	N/A	N/A	CP011803.1, CP011804.1	This study
***C. cellulolyticum* H10**	IB	3	CP00348.1	This study
***C. difficile* 630**	IB	14	AM1800355.1 AM1800356.1	3*
***C. kluyveri* DSM 555**	IB	5	CP000673.1	4*
***C. ljungdahlii* DSM13528**	N/A	N/A	CP001666.1	This study
***C. pasteurianum* ATCC 6013**	IB	2	CP009268.1	5*
***C. perfringens* ATCC 13124**	N/A	2	CP000246.1	This study
***C. saccharoperbutylacetonicum* DSM 14923**	IB	4	CP004121.1, CP004122.1	This study
***C. tetani* 12124569**	IB	8	AE015927.1	5*
***C. thermocellum* ATCC 27405**	IB	5	CP000568.1	5*
***C. scindens* ATCC 35704**	IC	4	Unpublished***	This study***
***A. woodii* DSM 1030**	IA	2	CP002987.1	This study
***E. limosum* ATCC 8486**	Type IB	3	Unpublished****	This study

5.7.1. CRISPR systems in acetogens

As shown in Table 5.9, Type IB CRISPR systems were identified in the following acetogens; *C. aceticum*, *C. autoethanogenum*, *A. woodii*, and *E. limosum*. The Cas operons identified in these organisms contained Cas genes in the same order as described by Makarova and colleagues (2015), for Type IB CRISPR systems; *cas6*, *cas8b*, *cas7*, *cas5*, *cas3*, *cas4*, *cas1*, and *cas2.*

No systems were found in either *C. carboxidivorans* P7 (GenBank: CP011803.1, CP011804.1), or *C. ljungdahlii* (GenBank: CP001666.1). Details of the Type IB CRISPR loci of *C. aceticum* and *A. woodii* are displayed Table 5.10, with details of their associated Cas operons detailed in the text below. Sections 5.2.2 and 5.6 provide more in-depth analysis of the CRISPR systems of *C. autoethanogenum* and *E. limosum* respectively.

Table 5.10: CRISPR loci of *C. aceticum* and *A. woodii* identified in this study
Details of the CRISPR loci (CR) for *C. aceticum* (Genbank: CP009687.1) (CACET) and *A. woodii* CP002987.1) (ACO), including direct repeat (DR) consensus sequence and length, the number of spacers, and the chromosomal coordinates.

CRISPR locus	DR consensus sequence	DR length	Number of spacers	Genome position
CACET CR1	ATTTAGAAAATACCTATAAGGGATTGAAAC	30	13	1146356 - 1147243
CACET CR2	ATTTACATTCTACTGTAGTTCTATTAAAGG	30	79	2305563 - 2310836
ACO CR1	ATTTACATTCCAATATGGATCTACTCAAAT	30	47	3377721 - 3380835
ACO CR2	GACGTAACACTGCTGGGTATCGTTAT	28	6	3904684 - 3905123

Analysis was performed on the *C. aceticum* DSM 1496 genome and plasmid sequence (GenBank: CP009687.1, CP009688.1) (Poehlein *et al.*, 2015). This revealed two CRISPR loci (CACET_CR1 and CACET_CR2) and a Type IB Cas operon located at position 1136914 – 1146129.

Using the *A. woodii* DSM 1030 (GenBank CP002987.1) genome sequence (Straub *et al.*, 2014), two CRISPR loci were identified with CRISPRfinder. Annotated as “ACO_CR1” and “ACO_CR1”, these loci (respectively) have direct repeat lengths of 30 and 28, and contain 47 and 6 spacers. Further analysis identified the presence of a Type IB Cas operon located at position 3389278 – 3389278 of the chromosome.

5.7.2. CRISPR systems in other industrially relevant *Clostridium*

As shown Table 5.5, several Type IB CRISPR systems were identified in several industrially relevant *Clostridium* species. An exhaustive search identified no CRISPR systems in the *C. acetobutylicum* ACC 824 type strain chromosome (GenBank: AE001437.1), or megaplasmid (GenBank: AE001438.3), despite the identification of a Type IIIA system in another strain (Mo *et al.*, 2015). This genome announcement of *C. acetobutylicum* GXAS18-1 (GenBank: JRWL01000001, JRWL01000049), identified 7 CRISPR loci, but failed to classify the type of CRISPR system. Re-analysis in this study resulted in the classification of a Type IIIA system, due to the Cas operon being arranged in the following order, consistent with the Makarova and colleagues’ 2015 classification; *cas6*, *csm1, csm2*, *csm3*, *csm4*, *csm5*, *cas02710*, *cas6*

(Peng *et al.*, 2014). Analysis of *C. beijerinckii* ASM83310v2 found no CRISPR system.

For the industrially relevant KNU-L09 strain of *C. butyricum*, analysis identified one CRISPR locus consisting of 57 spacers, as detailed in Table 5.11. Analysis of Cas genes identified a Type IB operon located at position 781962-785715 of chromosome 2. Similarly, 3 CRISPR loci and were identified in *C. cellulolyticum* H10 (GenBank: CP001348.1), along with a Type IB operon at position 1791651-1800472. For the hyper-butanol producing *C. saccharoperbutylacetonicum* (GenBank: CP004121.1, CP004122.1), 4 CRISPR loci were identified, as well as a Type IB Cas operon located at position 5281037-5289371 of the chromosome.

Table 5.11: CRISPR loci of industrially relevant clostridia identified in this study
Details of the CRISPR loci (CR) for *C. butyricum* (CB) (GenBank: CP013251.1), *C. cellulolyticum* H10 (GenBank: CP001348.1) and *C. saccharoperbutylacetonicum* (GenBank: CP004121.1), including direct repeat (DR) consensus sequence and length, the number of spacers, and the chromosomal coordinates.

CRISPR locus	DR consensus sequence	DR length	No. spacers	Genome position
CB CR1	ATTTACATTCTACTGTAGTTCTATTAAAGG	30	57	781962 - 785715
CCE CR1	GTCGCTCCTCTCGTAGGAGCGTGGATTGAAAT	32	5	1800644 - 1801010
CCE CR2	GTCGCTCCTCTCGTAGGAGCGTGGATTGAAAT	32	7	1802338 - 1802838
CCE CR3	GTTATAGCTCCAATTCAGGCTCCGATATGCTATAAT	36	8	3652357 - 3652923
CSP CR1	TGATTAACTTCAACATGAGATGTATTTAAATA	32	3	629945 - 630175
CSP CR2	AATTTAAAAGGAGCAGCTATGCTG	24	3	333451 - 3334724
CSP CR3	ATTTACATTCCTCATAGTTAATATAAAAC	29	25	5276899 - 5278564
CSP CR4	ATTTACATTCCTCATAGTTAATATAAAAC	29	33	5278666 - 5280855
CSP CR5	AACCAACCAGTTTTCATTGCTCCTGA	26	1	6007076 - 6007161

5.7.3. CRISPR systems in clinically relevant *Clostridium*

As shown in Table 5.12, this study analysed the genomes of *C. perfringens* and *C. scindens*. For the former, 2 CRISPR loci were identified, but extensive *in silico* analysis failed to identify any Cas genes. For the latter, 4 CRISPR loci were identified (see Table 5.8), along with a Cas operon. This operon consisted of *cas3*, *cas5*, *cas8c*, *cas7*, *cas4*, *cas1* and *cas2* genes consistent with the *Bacillus holodurans* C-125 Type IC CRISPR system described by Marakova *et al*. (2015).

Table 5.12 CRISPR loci of clinically relevant *Clostridium* identified in this study
Details of the CRISPR loci (CR) of *C. perfringens* (CPE) (GenBank: CP000246.1) *C. scindens* (CSC**)** (unpublished) including direct repeat (DR) consensus sequence and length, the number of spacers, and the chromosomal coordinates.

CRISPR locus	DR consensus sequence	DR length	Number of spacers	Genome position
CPE CR1	TTTTAAACAATATTTGCTTTTAGGT	25	3	1068681 - 1068882
CPE CR2	GTTAAAAAGATATTATGATAAATTT	25	3	1140440 - 1140660
CSC CR1	GTCGTTCCCTGCAATGGGAACGTGGATTGAAAT	33	9	190327 - 190963
CSC CR2	AGAAGCAGGATGATCCAGATACAGCTGT	28	1	840916 - 841024
CSC CR3	ATTTCAATCCACTCTCCCTATACAGGGAGAGAC	33	39	2050587 - 2053240
CSC CR4	CTTTCAATCCACTCCACCACACAGGTGGAGAC	32	13	2544798 - 2545694

5.7.4. Protospacer-like sequences identified from clostridial CRISPR spacers sequences

Analysis of the spacers of the CRISPR systems identified in this study resulted in the identification of several protospacer-like sequences. These are displayed in Table 5.3, with spacer sequences and their respective subject sequence for each instance of a hit identified. Details of the subject organism are provided, along with the possible PAMs identified. As described for *C. autoethanogenum* and *E. limosum*, the spacers from all CRISPR loci identified went through a pipeline of analysis to identify any highly similar DNA sequences in related organisms. Levels of sequence identity were lower than seem in *C. autoethanogenum* and *E. limosum*, with no spacers showing 100% sequence identity to any sequence in related organisms.

Table 5.13: Potential PAMs identified in *Clostridium*
Potential protospacer and PAM sequences identified through *in silico* analysis of CRISPR spacers, specified by CRISPR locus (ID) detailed in tables 5.5-8. Alignment of the spacer and potential protospacer (subject sequence) displayed, along with subject organism, interesting gene features (in bold), and number of mis-matches (MM). The 5 nt regions flanking the possible protospacers are detailed as 5' and 3' possible PAMs.

ID	Spacer and subject sequence	Subject organism	MM (nt)	Possible PAMs	
				5'	3'
CAC CR2.70	GATACGCCATACTTTTGCAACACATCTATTGAACAA ||| ||||||||||||| || ||||| |||||||| GATCCGCCATACTTTTGCGACCCATCTTTTGAACAA	*Plancoccus maratimus* CP016538.2	4	CATAT	TGGGG
AW CR1.36	ATTTATCCCGATATTGAAGCGGTTTTCATTGATAC || ||||| |||||||||||||| | |||| ATGTATCCGGATATTGAAGCGGTTTTCGTAAATAC	*Clostridium propionicum* DSM 1682 CP014223.1	5	GAGAA	CGGAC
CB CR1.13	ATTAAAATTACAGATGGAGAAAATGCAAAAGAAAAT | ||||||||||||| |||| || |||||| |||| AATAAAATTACAGATGAAGAATATTCAAAAGGAAAT	*Clostridium beijerinckii* ATCC 35702 CP006777.1	5	TATCT	ACTTT
CCE CR3.5	CTATTATCTATATTATAAATTATCATATAT |||||||||| ||| |||||||||||| TTATTATCTATTTTA-AAATTATCATATAT	*Anaerococcus prevotti* DSM 20548 CP001708.1	3	CAAAA	TTCTT
CSP CR3.7	GTTTCTTAATTTGCTTTTTCCTCCCATTCTAGTAAT ||||||||||| ||||||||||||||| |||| TTTTCTTAATTTACTTTTTCCTCCCATTCTATTAAT	*Clostridium pasteurianum* BC1 CP003261.1	3	ATGGT	AGGCG
CSP CR3.7	GTTTCTTAATTTGCTTTTTCCTCCCATTCTAGTAAT |||| || | |||||| |||||||||| ||||| TTTTCGTAGTATGCTTTTCCCTCCCATTCTCGTAAT	*Clostridium botulinum* CDC 53174 CP013247.1	6	ATTGT	CGGTG
CSP CR4.27	TTAAAAATACAACAGATA-CATTATTTATTTTTAATA ||||| |||||||||| |||||||||||| ||| TTAAATATACAACAGATAGCATTATTTATTTTAAATT	*Clostridium beijerinckii* B3 CP016090.1	3	AAACA	TTCTG
CSP CR5.1	TGGTTGTAAGTAATACCAAGTTCCATTGTCATTT || |||||||||||||||||||||||||| TGATTGTAAGTAATACCAAGTTCCATTGTCATTC	*Clostridium beijerinckii* B3 CP016090.1	2	CCTGA	AACCA

5.8. Conclusion

In silico analysis conducted in this study has identified several previously uncharacterised CRISPR systems in acetogens and other clostridia. Based on the most recent classification system proposed by Makarova and colleagues (2015), these systems were assigned classifications based on *cas* operon architecture. More in-depth analysis of the spacers present in the CRISPR loci of these systems revealed several instances of high sequence identity to phage-related genes in closely related species. Based on these sequences, potential PAMs were identified, and for *C. autoethanogenum* and *E. limosum*, further PAM analysis was conducted.

Separate to this study, Pyne and colleagues (2015) also identified the Type IB system of *C. autoethanogenum*, along with more in-depth analysis of the Type IB system of *C. pasteurianum*. Other studies have identified further Type IB systems in *C. difficile* and *C. kluyveri* (Boudry *et al.*, 2015; Makarova *et al.*, 2015). *In silico* analysis conducted in this study revealed yet more Type IB systems in the following species; *C. aceticum* DSM 1496, *C. autoethanogenum* ATCC 824, *C. butyricum* KNU-L09, *C. cellulolyticum* H10, and *C. saccharoperbutyacetonicum* DSM 14923. Examples of other systems identified include the Type IA and Type IC systems identified in *A. woodii* DSM 1030 and a Type IC system in *C. scindens* ATCC 35704 respectively. The only current example of a non-Type IB system identified in other studies is that of the Type IIIA present in *C. acetobutylicum* GXAS18-1 (Peng *et al.*, 2014).

Of the species studied, no CRISPR systems were identified in the following; *C. acetobutylicum* ATCC 824, *C. beijerinckii* ASM83310v2, *C. carboxidivorns* P7, *C. ljungdahlii* DSM13528, *C. perfringens* ATCC 13124. The lack of CRISPR systems in these species is not surprising, considering previous studies have failed to identify in ~50% of all bacteria. Also unsurprising is the prevalence of the Type IB systems in *Clostridium*, with 20% of systems identified in bacteria belonging to this classification (Makarova *et al.*, 2015). Other studies have shown that horizontal gene transfer dominates the evolution of CRISPR systems (Makarova *et al.*, 2011; Jackson *et al.*, 2014). This explains the difference in CRISPR systems between closely related organisms, such as the presence of a Type IB system in *C. autoethanogenum* and lack of any system in *C. ljungdahlii*. With this in mind, the prevalence of Type IB systems in the genus *Clostridium* is probably a reflection of shared environmental niches and the generally high prevalence of these systems in bacteria, rather than an evolutionarily conserved system from a shared ancestor.

5.8.1. Analysis of the *C. autoethanogenum* Type IB CRISPR system

At the beginning of this study, no work had been conducted in analysing the endogenous CRISPR systems of any member of the Clostridial genus. However, a recent study by Pyne and colleagues (2016) analysing the *C. pasteurianum* system also speculated on the likely PAM of the *C. autoethanogenum* based on *in silico* analysis. Independent *in silico*

analysis conducted during this study identified the same potential PAM, but disputed the location of the PAM as being on the 3' side rather than 5'.

This PAM was identified from spacer 1.27, identified as being 100% sequence identical to a phage related protein present in the *C. ljungdahlii* genome. Conjugation of vectors containing this protospacer sequence along with putative flanking PAM regions was performed. Despite DNA transfer stochasticity issues discussed in Chapter 1, a statistically significant decrease in conjugation frequency was observed with the 5' PAM sequence found upstream of the putative protospacer in the *C. ljungdahlii* genome. Further confirmation of this was seen in the increased conjugation frequencies observed when the nuclease encoding *cas3* was deleted.

Other attempts to assess the PAM were made through a spacer acquisition assay, similar to that previously described for analysis of the Type IC system of *E. coli* (Rollie *et al.*, 2015). In this study, *C. autoethanogenum* was repeatedly exposed to invasive DNA elements (in the form of a cocktail of pMTL80000 vectors). However, unfortunately, no increase in size was observed in any of the CRISPR loci. Unlike other published attempts at spacer acquisition assays, this study used conjugation as the DNA transfer technique (rather than electroporation). It could be speculated that perhaps the conjugation technique used in this study results in significantly less DNA transfer than those seen in electroporation protocols, and therefore decreases the likelihood of spacer acquisition. Another possible explanation could be lack of functional Cas1/Cas2 genes involved in spacer acquisition.

Although the spacer acquisition assay failed to provide any additional insight for the likely PAM sequence, further conjugation-based analysis did. This analysis refined the 5 nt 5'-TTAAT-3' sequence to likely 5'-TTVV-3' sequence. This was achieved through specific modification of each residue and resulting analysis of CFU mL^{-1}. Here the final residue was shown to not be essential in interference of conjugation, with consistently low CFU mL^{-1} observed regardless of which nucleotide this residue was modified to.

5.8.1. Analysis of the Type IB *E. limosum* CRISPR system

Using *E. limosum* ATCC 8486 genome sequence data generated by PhD candidate Robert Mansfield, 3 CRISPR loci were identified. Analysis of these loci found several spacers with high sequence identity to phage-related proteins of the KIST 618 strain. Unfortunately, transformation of *E. limosum* ATCC 8486 with the plasmid library constructed was not achieved due to time constraints. However, analysis of the 3' sequence showed a conserved "AG" motif at nucleotide residues two and three all of the putative protospacer 3' PAMs. Unlike other organisms analysed in this study, *E. limosum* had a high number of serendipitous spacers, with six sequences showing >95% sequence identity to phage-related genes in a highly-related organism. Although this was the case with spacer 1.27 of *C. autoethanogenum*, it was the only instance in this organism. The occurrence of so many spacers in *E. limosum* increases the likelihood of potentially correct PAMs, and also as described above provided the opportunity to identify a conserved motif.

5.8.2. Development of a robotics-based assay for PAM identification

In order to avoid the stochasticity issues associated with conjugation, and possibly to identify alternative PAMs used by the *C. autoethanogenum* endogenous CRISPR system, a degenerate PAM assay was designed. A similar approach to determining the PAM was taken to that described by Leenay *et al*. (2016). For this approach, the authors directed a Cas interference complex to the promoter of a reporter gene, with a measured decrease in expression indicating the presence of a correct PAM. Although this approach provides an accurate way to assess CRISPR targeting, taking this approach with the relatively uncharacterised *C. autoethanogenum* CRISPR system would still leave the functional activity of Cas3 in doubt.

In a modified approach to enable confirmation of Cas3 activity, the previously described putative protospacer sequence from spacer 1.27 was introduced to a GFP reporter gene vector (pET-GFP-target), flanked by 6 nt degenerate sequences. On a separate vector, the Cascade and Cas3 encoding genes, along with 1.27 encoding guide RNA were introduced (pACYC-CLAU-CRISPR). Both the GFP and CRISPR components were under the control of inducible promoters, with repression of this expression allowing transformation of the pTarget library into pCRISPR containing competent cells.

In order to assess the PAMs of plasmids showing decreased GFP expression, extremely high throughput measures would be required to gain enough coverage. This rules out conventional Sanger sequencing, instead the high-

throughput methods used for TraDIS could perhaps be adapted (Barquist *et al.*, 2016; Langridge *et al.*, 2009). Due to time constraints, sequencing of the plasmids showing lower GFP expression has not yet been performed. However, as discussed, early proof of concept experiments have shown a distinct difference in GFP expression between positive/negative controls, and the plasmid library, suggesting the screening method will be sensitive enough to detect CRISPR activity occurring.

5.8.3. Analysis of other Clostridial CRISPR systems

The acetogens, *C. aceticum, A. woodii,* and *C. carboxidivorans* have drawn interest for the reasons outlined in Chapter 1. The most recent published genome announcements for each have failed to mention the presence CRISPR systems (Poehlein *et al.*, 2015), however, *in silico* analysis performed in this study identified Type IB and IA systems (respectively), and several CRISPR loci for *C. aceticum* and *A. woodii*. Spacer analysis of these loci revealed a *C. aceticum* spacer with high sequence identity a recombinase gene found in the *P. maratimus* genome. Similarly, *A. woodii* spacer analysis identified a spacer with high sequence identity to a genomic region of *C. propionicum*. No systems were identified for *C. carboxidivorans* P7 and *C. ljungdahlii*.

As for other industrially relevant *Clostridium*, *C. cellulolyticum* holds promise as an industrial chassis for direct conversion of lignocellosic biomass to ethanol. Interestingly this species saw one of the earliest applications of CRISPR-Cas9 technology in the *Clostridium* genus, with the use of Cas9^{D10A} nickase shown to work efficiency. Analysis of the most recent *C. cellulolyticum*

genome shows the presence of a Type IB CRISPR system, with an operon partially annotated in the genome announcement. Full analysis of these genes identified them as being *cas6*, *cas8b*, *cas7*, *cas5*, *cas3*, *cas4*, *cas1*, and *cas2*, consistent with that of a Type IB system.

Analysis of clinically relevant *Clostridium* identified no CRISPR system in *C. perfringens*, an opportunistic pathogen and inhabitant of the avian and mammalian gastrointestinal tract. As previously detailed, the commonly hospital acquired *C. difficile* 630 has been shown to contain a Type IB CRISPR system (Boudry *et al.*, 2015). Although phage therapies for this organism are being trialled (Nale *et al.*, 2016), the presence of such CRISPR systems could interfere. Highlighted as another possible treatment for *C. difficile* infection, *C. scindens* has been shown to offer inhibition of *C. difficile* spore outgrowth (Buffie *et al.*, 2014). To date, no genetic engineering has been published in *C. scindens*, so the identification of the endogenous Type IA system in this species could offer a potential platform for genome editing.

Unlike *C. autoethanogenum* and *E. limosum*, no further PAM analysis was conducted in any of these species, although several possible protospacer-like sequences were identified, as outlined in Table 5.8. This lack of further study was two-fold; primarily due to time constraints, but also because spacer analysis identified protospacer-like sequences with far less sequence identity than those identified in *C. autoethanogenum* and *E. limosum*. The design of the robotics-based PAM screening assay described in section 5.4 could provide a better solution for elucidation of PAM sequences.

CHAPTER SIX

ANAYSIS OF ACIDOGENESIS MUTANTS

6. Analysis of acetate pathway mutants

6.1. Introduction

As discussed in Chapter 1, the two defining characteristics of acetogens are their ability to utilise C1 gas as carbon and energy sources, and their production of acetate (amongst other solvents). As might be suspected from the ubiquitous presence of both features in acetogens, they are intrinsically linked. As described below, acetate synthesis is crucial for acetogenic growth, as it results in the generation of 1 ATP, cancelling out the 1 ATP cost of the methyl branch of the Wood-Ljungdahl pathway.

6.1.1. ATP generation in acetogens

As the universal currency of energy in biology, ATP is fundamental to energy conservation in all life. Energy conservation can occur through either substrate level phosphorylation, or phosphorylation via a chemiosmotic ion gradient. For the former, a chemical reaction is directly linked to the generation of ATP (through phosphorylation of ADP), as seen in the 2 ATP gained in separate steps during glycolysis. In contrast, the latter can be seen in the coupling of an exergonic reaction to ionic gradient. Examples of such can be seen in the use of various membrane bound ATPases employing transmembrane ion gradients to provide the energy for ADP phosphorylation. As described in Chapter 1, acetogens are able to utilise C1 gases as carbon and energy sources, via the Wood-Ljungdahl pathway. Taken in isolation this pathway provides -1 ATP, however, acetyl-CoA generated in the Wood-Ljungdahl pathway is fed into the acetate pathway which produces 1 ATP.

During chemoautotrophic growth, acetogens lack ATP generation from glycolysis, and as described above, Wood-Ljungdahl coupled acetate production provides no net gain in ATP. In order to meet the demands of such a fine metabolic balance, acetogens adopt different strategies to make acetogenic growth possible. Unlike the conserved Wood-Ljungdahl and acetogenesis pathways, different acetogens utilise different ATPases. Until recently, acetogens were regarded as either Na^+-independent and proton-dependent, or Na^+-dependent and cytochrome containing. However, recent research has shown the acetogens such as *C. ljungdahlii* and *C. autoethanogenum* may use a reaction resulting in the synthesis of ATP.

In *C. autoethanogenum* electron transfer from reduced ferredoxin to NAD is thought to be coupled to translocation of protons across the cytoplasmic membrane, via the membrane-bound Rnf complex. The resulting proton gradient is then utilised by a membrane-bound ATPase for ATP generation. This proposed model predicts the generation of less than 1 mole of ATP, which although puts these organisms at the theoretical thermodynamic limits of life, does explain their ability to grow autotrophically without the presence of the Na^+-dependent and Na^+-dependant mechanisms seen in other acetogens.

6.1.2. Phosphotransacetylase and acetate kinase

As illustrated in Figure 6.1, *pta* (CLAU_3274) encoding phosphotransacetylase (PTA), catalyses the conversion of acetyl-CoA to acetyl-P, the immediate precursor to acetate. Catalysis of this second step is provided by acetate kinase (ACK), coded for by *ack* (CLAU_3275).

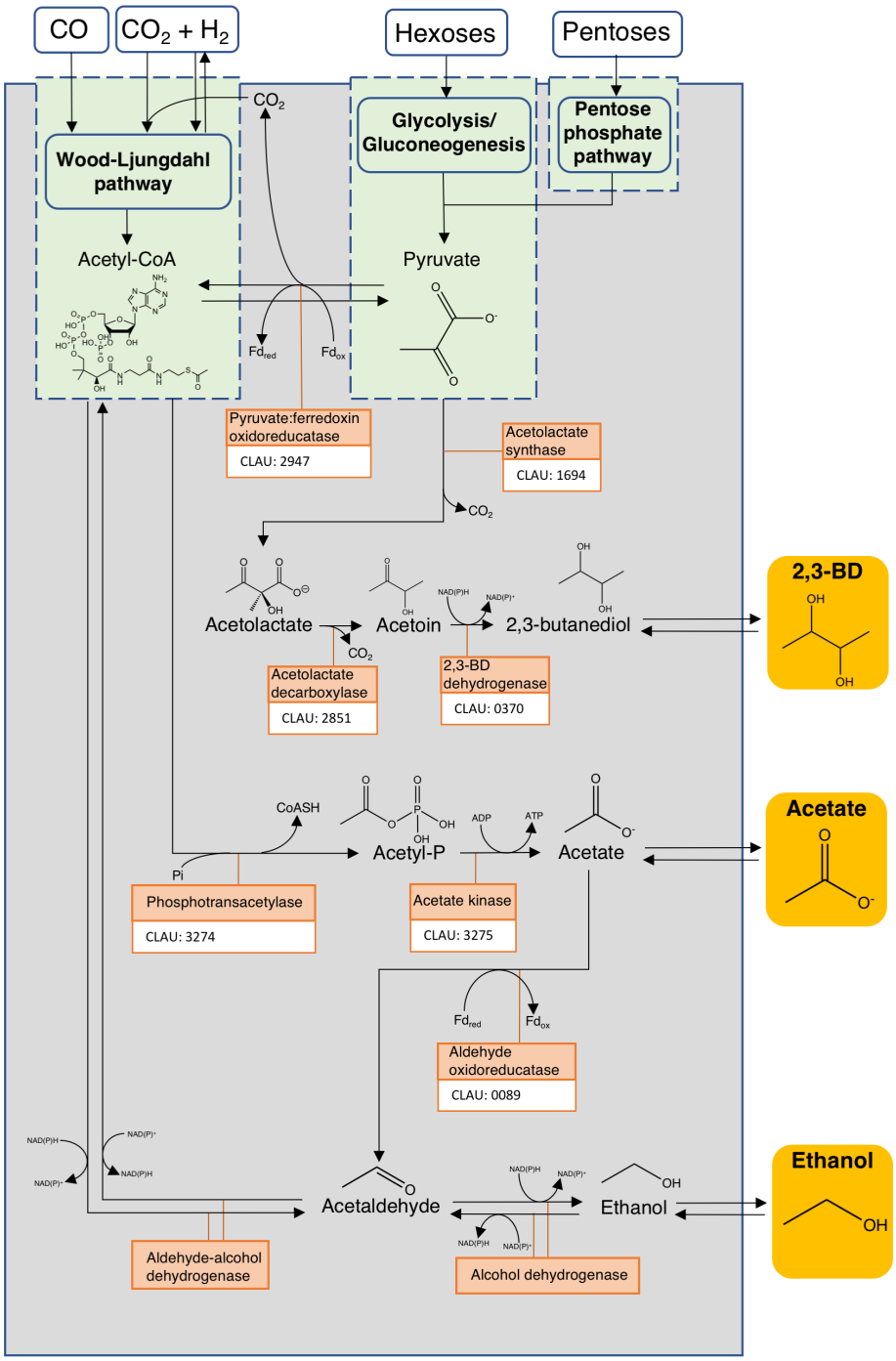


Figure 6.1: Acetate, ethanol and 2,3-butanediol generation in acetogens
(repeat of Figure 1.7)

With the theoretical yield of ATP from chemoautotrophic growth in *C. autoethanogenum* being under 1 ATP, knockouts of either *ack* or *pta* would in theory be deleterious if one were to assume they coded for the sole enzyme able to fulfil their role in acetogenesis. However, for *pta*, this appears to not be the case. As shown in *M. thermoaceticum*, conversion of acetyl-CoA to acetyl-P is performed by two propanediol utilisation protein (PduL) enzymes. As highlighted by PhD student Ronja Brietkopf, *C. autoethanogenum* has genes with high sequence identity to both PduL coding genes in *M. thermoaceticum*, with identical amino acid residues in the relevant domains.

With the likelihood of PduL enzymes being able at least partially able to fulfil PTA function, a *pta* knockout would in theory still generate ATP from conversion of PduL-mediated acetyl-P to acetate (via ACK). However, an *ack* knockout would have no known gene with a homologous function capable of ATP generation, and therefore theoretically a knockout would be more deleterious than *pta*.

6.2. Strain development

6.2.1. ACE and ClosTron mutagenesis

For both *ack* and *pta*, ACE mutagenesis was attempted in the Δ*pyrE* strain of *C. autoethanogenum* described in Chapter 3. The plasmid vector pMTL-ACE-CLAU-del-pta was generated for ACE knockout of *pta*, using pMTL-CLAU-del-pyrE as a backbone. For this, the primers CLAU-pta-ACE-LHA-F, CLAU-pta-ACE-LHA-R, CLAU-pta-ACE-RHA-F and CLAU-pta-ACE-LHA-R were used to generate the knockout cassette via SOEing PCR, which was subsequently cloned into the backbone (replacing the existing *pyrE* knockout cassette) using *Sac*II and *Asc*I restriction sites.

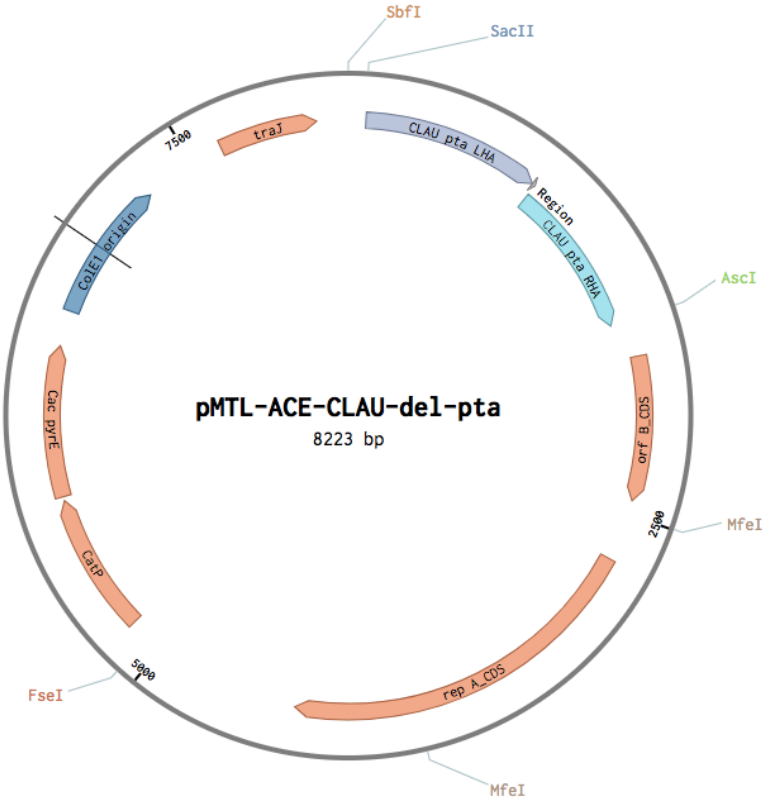


Figure 6.2: Plasmid map of pMTL-ACE-del-pta
The vector consists of a pCD6 Gram-positive replicon, CatP module encoding chloramphenicol/thiamphenicol resistance, ColE1 + *tra* Gram-negative module, and an ACE knockout cassette consisting of asymmetric homology arms (~0.5 kB left homology arm (LHA) and ~1.2 kB right homology arm (RHA)).

Once the final plasmid (as displayed in Figure 6.1) was sequence confirmed, it was transformed into *E. coli* CA434 which was used as a conjugative donor. For ACE knockout of *ack*, the pMTL-ACE-CLAU-del-ack plasmid was constructed as described above, with all design, cloning and conjugation work performed by PhD student Pawel Piatek.

In both instances, exhaustive attempts at deletion resulted in no detectable mutants. For *pta*, thiamphenicol resistant colonies were attained, however, no FOA resistant colonies were observed.

Multiple failed attempts at allelic exchange led to attempted ClosTron mutagenesis of both genes. For *pta*, the vector pMTL007-EC2-CLAU-pta was designed as described in Chapter 2, and synthesised by DNA2.0. For *ack*, the same process was followed by PhD student Pawel Piatek, using the pMTL007-EC2-CLAU-ack vector. More details on ClosTron vector construction can be found in research published by Heap *et al*. in 2007.

Initial selection for chloramphenicol resistance yielded several colonies for both *ack* and *pta*. These were used to inoculate liquid YTF culture supplemented with thiamphenicol, from which -80 °C stocks were taken, and from which ClosTron insertion was screened for on solid YTF supplemented with clarithromycin. In both instances, initial screening of liquid cultures yielded no clarithromycin resistant colonies. Revival of stocks and subsequent clarithromycin resistance screening was repeated on several occasions, each iteration of which failed to produce any knockouts.

6.2.2. pMTL40000-mediated Δ*ack* and Δ*pta* strain generation

Having developed proof of principle for *pyrE* deletion (see section 3.4), the pMTL43151-CLAU-pyrE vector was used as a backbone to generate pMTL43151-CLAU-ack and pMTL43151-CLAU-pta. For pMTL43151-CLAU-ack, the oligonucleotide primers CLAU-ack-LHA-F, CLAU-ack-LHA-R, CLAU-ack-RHA-F and CLAU-ack-RHA-R were used to generate the knockout cassette via SOEing PCR for *ack*, with the oligonucleotide primers CLAU-pta-LHA-F, CLAU-pta-LHA-R, CLAU-pta-RHA-F and CLAU-pta-RHA-R used for pMTL43151-CLAU-pta. Following homology arm SOEing, *Asc*I and *Asi*SI restriction sites were used to replace the backbone homology cassette with those of *ack* and *pta*, creating intermediate vector both still containing the *pyrE* sgRNA seed region.

Following Sanger sequence confirmation of homology arm insertion, the intermediate vectors were linearised via *Sal*I digestion. Using NEB HiFi ssDNA oligo assembly, as detailed in section 4.4.1, vectors were completed using the oligonucleotides ack-sg1-HiFi and pta-sg1-HiFi for pMTL43151-CLAU-ack and pMTL43151-pta, respectively. Finally, the vector sequence was confirmed via Sanger sequencing and transformed into *E. coli* Sexpress.

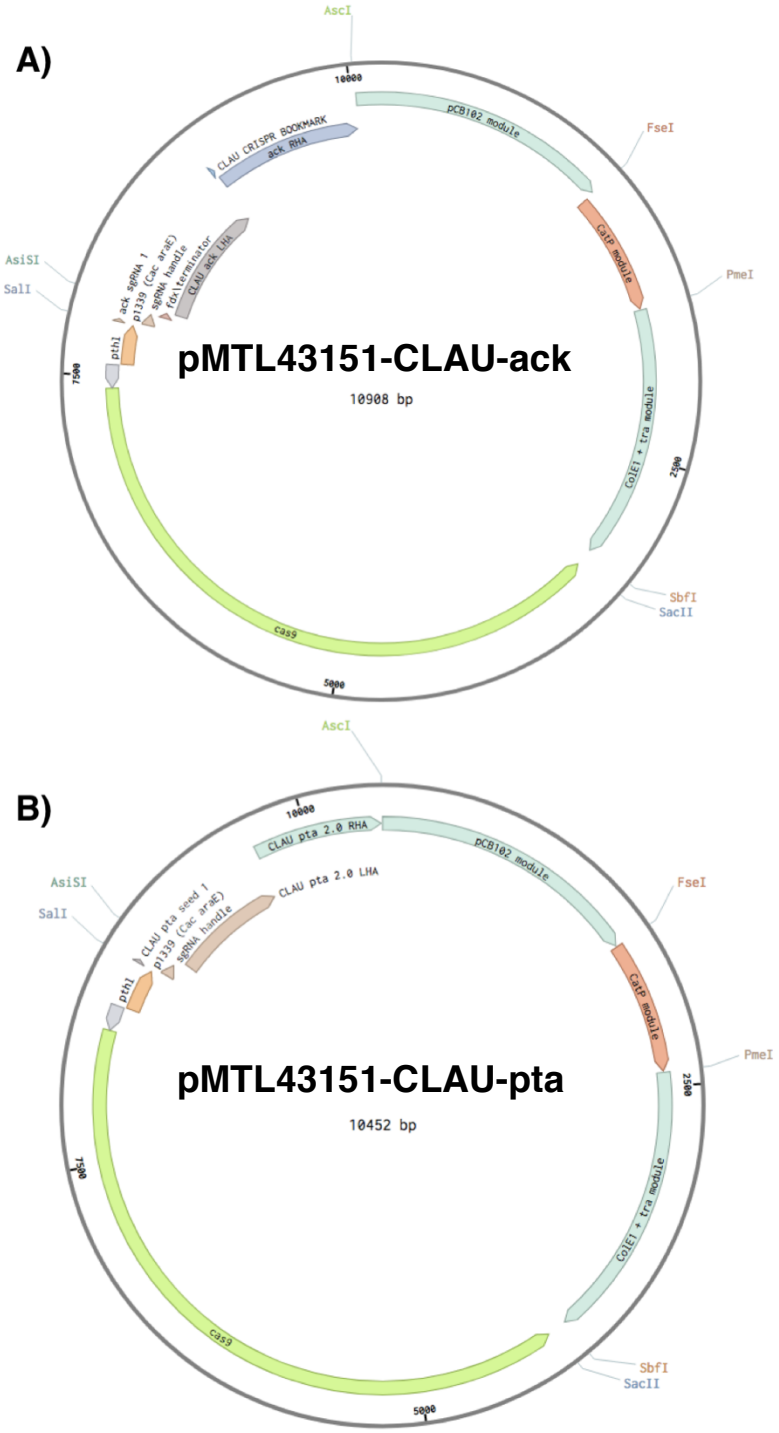


Figure 6.3: pMTL43151-CLAU-ack and pMTL43151-CLAU-pta
Both plasmids consist of a pMTL83151 backbone (pCB102 Gram-positive replicon module, CatP resistance module, and ColE1 + *tra* Gram-negative replicon module), with the pMTL40000 standard divergent Cas9 and sgRNA driven by *C. acetobutylicum* P_{thl} and P_{araE} respectively. The plasmids differ in terms of homology arms, each of which derived from the ~750 nt regions immediately upstream (LHA) and downstream (RHA) of the gene of interest. Details of target specific sgRNA seed regions can be seen in Chapter 2.

6.2.2.1. *ack* deletion

On the first attempt of conjugation with pMTL43151-CLAU-ack into the Col3 strain of *C. autoethanogenum*, no colonies were observed after 7 days of selection of solid YTF media supplemented with D-cycloserine and thiamphenicol. Repetition of this conjugation using triple the cell volumes yielded 15 colonies appearing after 6 days[19]. Colony PCR of each of these small colonies failed to yield any PCR products, but subsequent passage onto selective media yielded colonies from 12 original transconjugants that confirmed deletion of *ack* in each instance. This is illustrated in Figure 6.4. Following colony PCR, mutant strains were subjected to plasmid loss, confirmed via Sanger sequencing, and stored at -80 °C.

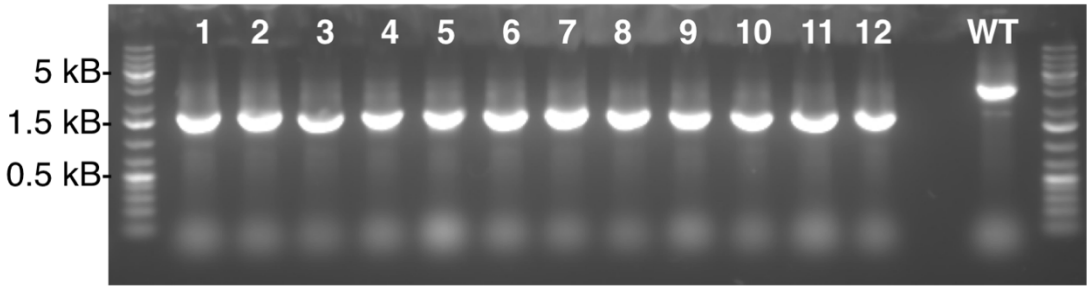


Figure 6.4: *C. autoethanogenum* Δ*ack* colony PCR
Lanes 1-12 represent individual colonies derived that appeared on selection plates 6 days after being passaged from colonies that appeared after 6 days on initial selection plates. All colonies showed ~1.5 kB bands, consistent with deletion of *ack*. "WT" lane shows PCR product from the conjugal recipient strain (wild type *C. autoethanogenum* Col3), displaying the ~4 kB locus size associated with intact *ack*. See Figure 4.1 for schematic explanation of colony PCR.

19 Twice the time taken for pMTL43151-CLAU-pta mutants to appear

6.2.2.2. *pta* deletion

Deletion of *pta* was attained on first attempt, with colonies appearing on selective YTF after 3 days. The analysis of PCR products obtained with pMTL43151-CLAU-pta transconjugants is shown in Figure 6.5. Following colony PCR, strains were subjected to plasmid loss, confirmed via Sanger sequencing and stored at -80 °C. As displayed in Figure 6.5, a clear size difference can be seen between the size of the PCR products obtained from mutants lacking the 1 kB *pta* gene, and the wild type strain.



Figure 6.5: *C. autoethanogenum* Δ*pta* colony PCR
Screening primer CLAU-pta-scr-F and CLAU-pta-scr-R used to amplify colonies after single passage on selective media. Lanes 1-21 represent individual colonies and "WT" lane shows the PCR product from the conjugative recipient. See Figure 4.1 for schematic explanation of colony PCR.

6.2.2.3. Bookmark mediated repair of *ack* and *pta* mutants

As described in section 4.10, pMTL43151-ack-repair and pMTL43151-pta-repair were generated and used to complement Δ*ack* and Δ*pta*. For repair of *pta*, conjugation of the repair plasmid resulted in >30 colonies appearing after 4 days. Colony PCR failed to yield any products for the initial transconjugants and 12 were passaged onto selective YTF plates. Colony PCR of these passaged clones resulted in wild type sized PCR products, confirming the knock-in of the *pta* gene, as shown in Figure 6.6.

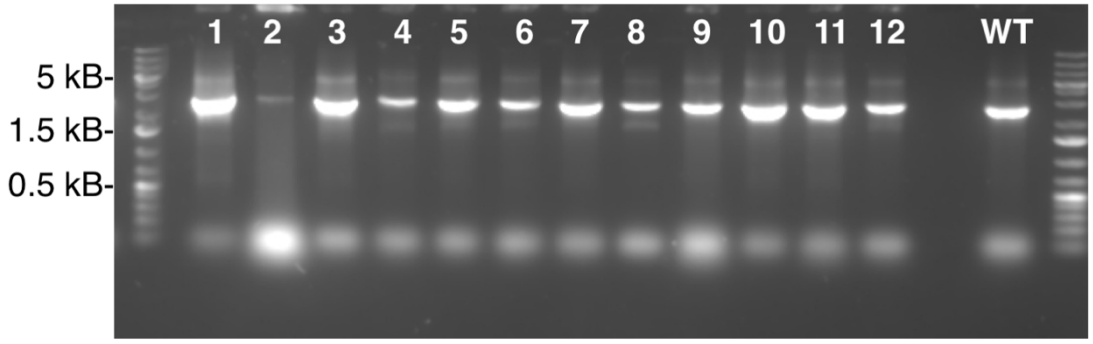


Figure 6.6: *C. autoethanogenum pta* repair colony PCR
Lanes 1-12 show the PCR products from colony PCR of 3 day old colonies resulting from passage from original conjugation selection plates. "WT" lane shows the PCR product from the conjugative recipient Δ*pta* strain. Each of the lanes shows a *pta* locus size consistent with the wild type size of ~2.6 kB. See Figure 4.12 for schematic explanation of colony PCR for Bookmark repair.

Despite multiple attempts, no repair of *ack* through this technique was possible. Although vector assembly provided no issues, multiple conjugation attempts with the pMTL43151-ack-repair vector failed. As detailed later in this chapter, this may be due to the deleterious effects of the *ack* deletion.

6.3. Strain comparison and solvent profile analysis

Initial growth and culturing of the Δ*ack* strain proved difficult. Consistent with the difficultly in generating the knockout, and subsequent failure to repair the knockout, revival from -80 °C stocks was problematic. After multiple attempts to revive the Δ*ack*, viable growth was observed on fructose and the strain taken forward for gas adaptation. Unlike the Δ*ack* strain, Δ*pta* stock revival was possible without issue, and it was also taken forward for gas adaptation.

6.3.1. Gas adaptation

Strains were grown in defined PET-C MES liquid media supplemented with fructose, prior to being using to inoculate PET-C MES media with CO as the only added carbon source, along with any carry over of fructose from the

inoculum. Following growth in this gas adaptation stage, gas growth cultures of PET-C MES with 1 bar CO overpressure were inoculated and analysed.

For the wild type Col3 and *pta* repair strains, gas adaptation was unremarkable, with OD_{600} of 0.4 being reached after 10 days. For Δ*pta*, an OD_{600} of 0.1-0.2 was reached during gas adaptation over a period of 12-16 days. Subsequent analysis of these cultures saw peak OD_{600} of >0.8 for wild type Col3 and *pta* repair strains, however, despite multiple repeated attempts the Δ*pta* failed to reach an OD_{600} value of >0.2. Despite multiple attempts at repeating gas adaptation, no detectable growth was observed for the Δ*ack* strain.

6.3.1. CO growth experiment

Following gas adaptation, gas growth experiments were conducted in triplicate as described in Chapter 2. As shown in Figure 6.7, the *pta* repair strain showed comparable growth and solvent profiles to wild type when grown autotrophically on CO. Despite multiple attempts, no gas growth was observed for Δ*pta.*

Due to focus on analysing the phenotype of the mutants during chemoautotrophic growth on C1 gas, no in-depth growth analysis was conducted on fructose. However, as described in 6.3.1, growth on fructose was possible for both knockouts.

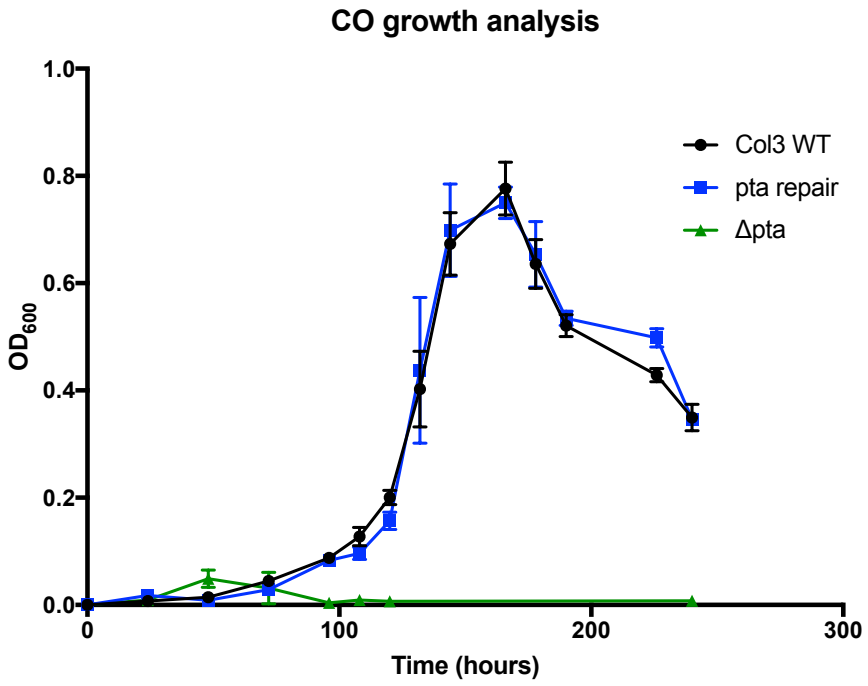


Figure 6.7: CO growth comparison of Δ*pta*, repaired *pta*, and Col3 strain of wild type *C. autoethanogenum*
Col3 wild type is represented as black line with black circles, *pta* repair as blue line with blue squares, and Δ*pta* as green line with green triangles.

6.4. Conclusion

This study looked to assess the impact of inhibition of the acetate production pathway of *C. autoethanogenum*, specifically with regards to its impacts on chemoautotrophic growth. Due to acetate being the most prominent chemical in the solvent production profile, redirection of carbon flux away from its production could potentially lead to higher yields of other solvents. Using a similar rationale, Kuit and colleagues (2012) showed that disruption of the *ack* gene in *C. acetobutylicum* resulted in a 63% increase in ethanol production. However, to date, no analysis of acetate pathway disruption in acetogens has been shown under chemoautotrophic conditions.

As outlined in section 6.1, the Wood-Ljungdahl pathway results in generation of acetyl-CoA, which used to generate acetyl-P via the activity of Pta, which itself is converted to acetate via Ack. Coupling of the two pathways results in a net ATP gain of zero, due to the cost of 1 ATP for the Wood-Ljungdahl pathway being cancelled out by the generation of 1 ATP via Ack activity. Due to lack of glycolysis-derived ATP during chemoautotrophic growth, *C. autoethanogenum* is thought to rely solely upon the Rnf complex to generate transmembrane proton gradient, enabling membrane-bound ATPase activity. As commented by Schuchmann & Müller (2014), although this generates sufficient ATP for autotrophic growth, the fine thermodynamic balance in acetogens means for a net yield of <1 ATP.

As the results in this chapter have shown, deletion of either gene involved in acetate production severely hampered strain performance and viability. For both knockout strains, no autotrophic growth was observed using CO as the

carbon source. This supports the claims by Schuchmann & Müller (2014) that Rnf complex in the closely related *C. ljungdahlii* yields <1 ATP, as the theoretical loss of 1 ATP from removal of Ack activity (in both instances), results in a net chemoautotrophic yield of negative ATP.

Despite chemoautotrophic growth being observed in neither knockout strain, differences were observed between the two mutants. Firstly, during the gas adaptation stage, Δ*pta* did show growth, whereas Δ*ack* did not. Further to this, Δ*ack* failed to yield any transformants when Bookmark complementation was attempted, in contrast to the relative ease with which Δ*pta* was transformed. These results suggest that deletion of *ack* is far more deleterious to cellular metabolism than *pta*, despite them both theoretically affecting the organism in the same manner, by interfering with acetate production. One possible explanation of this observation could be that the PduL activity described in section 6.1 fulfils part of the function of *pta*, thus enabling a small amount of acetate production and therefore generating ATP.

Unfortunately, time constraints meant that assessing the impacts of solvent profiles and growth during heterotrophic growth of *C. autoethanogenum* mutants on fructose could not be assessed, however as part of the work described in Chapter 3, one of the early successes in CRISPR-Cas9 tool development was deletion of *pta* in the closely related acetogen *C. ljungdahlii*. Here, a marked decrease of acetate was observed, as expected, however, ethanol yield was also severely hampered, with 0.2 gL^{-1} being produced in batch culture compared to 1.1 gL^{-1} in wild type. Although no gas growth was conducted for *C. ljungdahlii*, due to the high sequence similarity and presence

of the strains, it is likely that a similar lack of chemoautotrophic growth would be observed. Similarly, it is likely that the results attained for heterotrophic *C. ljungdahlii* growth on fructose would be observed for *C. autoethanogenum*. Together these results suggest that disrupting acetate production in Rnf-utilising acetogens is both inhibitory for chemoautotrophic growth, and deleterious to overall solvent production during heterotrophic growth.

Although limited characterisation of the Δ*pta* and Δ*ack* has left some unanswered questions, the value of these strains as the basis for potential chassis has been identified. Despite the theoretical potential for these mutants to enable shift of carbon flux towards formation of other products (as observed in *C. acetobutylicum* (Jang *et al.*, 2014)), no such result was observed. Instead the mutants appear to be deleterious to chemoautotrophic growth (as seen with *C. autoethanogenum*), and offer no increase in any other solvent production during heterotrophic growth (as seen with *C. ljungdahlii*).

As well as identifying the essential nature of the genes for chemoautotrophic growth, development of these mutants exemplified both the pMTL40000 vector series as a knockout tool, and exemplified its value in complementation through Bookmark mediated re-insertion of the *pta* gene (see Chapter 4).

As described in section 6.2.1, exhaustive attempts at creating knockouts of both genes were attempted through both allelic exchange, and ClosTron methodologies. Failure of these technologies led to the exploration of using CRISPR-Cas9 based technology, and ultimately to the creation of the pMTL40000 vector series. Although the value of CRISPR-based approaches over the allelic exchange methodology developed by Liew and colleagues

(2017) is clear, the improvement of knockout generation over ClosTron was unexpected. Allelic exchange offers no strong positive selection, with homologous recombination events screened for by increased colony size. For genes with no deleterious phenotype, this technique may suffice. However, if, as seen with *ack*, deletion of the gene of interest hampers growth, this screening is nullified. In contrast, ClosTron offers clear positive selection; insertion of the intron into the gene of interest enables expression of the retrotransposition-activated *ermB* marker gene, providing erythromycin resistance. Interestingly, at least in the context of these genes, using a CRISPR-Cas9 based approach enabled the generation of knockouts where ClosTron failed, despite having similar selection criteria. These results could suggest that the efficiency of homologous recombination to delete the gene of interest (and nuclease-mediated counter selection) with the pMTL40000 system is higher than the efficiency of intron insertion with ClosTron.

Bookmark complementation of *pta* restored the ability to grow on CO, with no notable differences to the wild type strain. This is to be expected, as in theory the strains should be sequence identical. The fact that no whole genome sequencing was performed for the Δ*pta* strain is mitigated by the use of Bookmark, as any potential mutations accumulated elsewhere in the genome were ruled out as contributing factors to the observed phenotype, due to reversion to wild type phenotype upon complementation. Unfortunately, Bookmark complementation was unsuccessful for *ack* restoration. The likely explanation for this failure is the sickly phenotype caused by the deletion of the gene decreasing conjugation efficiency.

CHAPTER SEVEN

CONCLUSION AND SUMMARY

7. Conclusion and summary

7.1. Addressing the Objectives of the Study

The initial objectives of the study were as follows:

1. Optimise DNA transfer methods and modify the *C. autoethanogenum* endogenous restriction-methylation systems to enable sufficient DNA transfer to enable CRISPR-based approaches to genome editing.

2. Using CRISPR-Cas9 technology, design and implement a modular plasmid system for efficient, in-frame knockouts of target genes in *C. autoethanogenum*. Using the modular nature of this system, apply it for genome editing applications in other members of the genus.

3. Analyse the endogenous CRISPR system of *C. autoethanogenum*, and identify its PAM. Once identified, use this PAM to enable modification of the CRISPR system for use in genome editing and knockdown applications.

4. Using tools developed in objectives 2 and 3, create mutants in the *ack* and *pta* genes involved in acetogenesis, as well as other relevant targets. Analyse and compare the mutant strains to wild type to assess the impact on solvent production.

As outlined in Chapter 3, DNA transfer issues presented a significant barrier to the development of *C. autoethanogenum* as an acetogenic chassis. Solving these issues subsequently became the primary objective of this study, and simultaneously work was conducted to assess the viability of *E. limosum* as an alternative chassis.

In depth *in silico* analysis of the RM system of *C. autoethanogenum* provided details on likely barriers to DNA transfer. The development of the Sexpress conjugative donor strain by Craig Woods enabled consistently high DNA transfer frequencies, as verified in this study. Here, by using the Dam^{+}/Dcm^{-} methylation profile afforded by Sexpress, Type IV endonuclease activity was avoided, resulting in higher frequencies of DNA transfer. Further work on improving DNA transfer was conducted during this study through deletion of the *caulI* restriction endonuclease identified by NEB REBASE.

The use of Sexpress as a donor strain raised DNA transfer frequencies to a high enough level to enable the use of CRISPR-Cas9 plasmids, allowing objective 2 to be addressed. Whereas prior to the use of Sexpress, a two-plasmid approach similar to that described for *C. acetobutylicum* was attempted (Wasels *et al.*, 2017), Sexpress opened up the possibility to incorporate all components onto one plasmid. As cited by the authors of the aforementioned study, the size of the SpCas9 and homology arms required for genome editing are limiting to DNA transfer. Although Wassels and colleagues (2017) state that these issues were overcome through splitting the CRISPR components onto two plasmids, this approach proved unsuccessful

for *C. autoethanogenum*. Instead, improvement of DNA transfer itself enabled the proof of principle of the pMTL40000 vector series in *C. autoethanogenum*. Following success in *C. autoethanogenum*, the modular design of the pMTL40000 vector series enabled easy adaptation for use in a range of other organisms. In a separate study published recently, SpCas9 genome editing was shown in *C. autoethanogenum*, however, the reported genome editing efficiencies were low and in some instances resulted in truncation (rather than in-frame deletion) of the gene of interest (Nagaraju *et al.*, 2016). In contrast, editing efficiencies of the pMTL40000 system were consistently high across a range of organisms, with no instances of inaccurate modification.

Although improved DNA transfer undoubtedly enabled better conjugation of the pMTL40000 vector series, other explanations may be found for the improved accuracy of genome editing. As highlighted in Chapter 4, sequencing of the pMTL40000 vector series highlighted the presence of an insertion mutation in a poly-thymine region at the extreme 5' end of the SpCas9 gene. Although this should in theory nullify any Cas9 activity, as discussed above, highly efficient genome editing was observed. As discussed by Qi and colleagues (2015), the use of the D10A nickase variant of SpCas9 proved more applicable for use in *C. autoethanogenum* and *C. beijerinckii*. Again, this provides another alternative to the Wassels and colleagues' (2017) two-plasmid approach, but the study also shows in both organisms nCas9^{D10A} provided higher efficiency than the wild type SpCas9. Through *in silico* analysis of the pMTL40000 vectors, a potential start codon was identified suggesting that a truncated version of SpCas9 (trCas9) lacking the same

domain mutated in nCas9^{D10A} could be expressed. If this were to be the case, the higher efficiencies observed compared to those reported by Nagaraju and colleagues (2016) could be due to this crucial difference in nuclease function.

As well as applications in genome editing, SpCas9 has been modified for use in knockdown of gene expression. Here as well as the RuvC nucleolytic domain disrupted in nCas9^{D10A}, the other nucleolytic domain (HNH) is disrupted. The resulting dead Cas9 (dCas9) is targeted to the 5' end or promoter of a gene of interest where it physically interferes with RNA polymerase activity (Gilbert *et al.*, 2013; Park *et al.*, 2014). The use of dCas9 has been shown in *C. beijerinckii* previously (Li *et al.*, 2016), however, in the context of acetogens it was deemed unsuitable. This unsuitability is due to the large nature of dCas9 harbouring plasmids and the typically slow growth rates of acetogens. To use *C. autoethanogenum* as an example, once a dCas9 plasmid vector was conjugated into the recipient strain, it would take several days (or weeks) to adapt the strain to gas growth, after which a batch growth experiment could take weeks, or a steady state fermentation could continue indefinitely. For such long conditions, plasmid selection and stability provide potential bottleneck. Alternatives to plasmid-borne dCas9 could be chromosomal integration, however its large size could provide issues. Instead, the proposed solution outlined in this study was the re-engineering of endogenous CRISPR systems.

As outlined in Chapter 5, *in silico* analysis identified the presence of several previously unidentified CRISPR systems. Although Pyne and colleagues (2015) independently identified spacer 1.27 as having 100% sequence

identity to a phage-related protein coding gene present in the *C. ljungdahlii* genome, prior to this study no work has been conducted in fully characterising the *C. autoethanogenum* CRISPR system. Similarly, several *E. limosum* ATCC 8486 spacers with very high sequence identity to phage-related genes in *E. limosum* KIST 618. The identification of these protospacer-like sequences was crucial to identifying the PAM.

In order to re-purpose CRISPR systems, their PAM must first be identified. No definitive study in to either of the *C. autoethanogenum* or *E. limosum* CRISPR systems had previously confirmed any PAM sequence for either. Through a top-down approach the sequence for each was identified, providing the foundation for using them as genetics tools. Development of the pMTL43151-CLAU-eCas vector has also provided the foundation for delivering engineered CRISPR components into a *cas3* deficient strain for knockdown.

Analysis of other clostridial species identified several previously unidentified CRISPR system, most of which belonging to the Type IB classification as described by Marakova and colleagues (2015). Unlike with *C. autoethanogenum* and *E. limosum* however, no spacers were identified with high sequence identity to any phage-related sequences in related organisms. Attempts at spacer acquisition as described in other organisms (Díez-Villaseñor *et al.*, 2013) proved unsuccessful in *C. autoethanogenum*. If this were to be the case in other organisms without the serendipitous presence of a high sequence identity spacer, PAM identification could prove difficult.

In a recent study, Leenay *et al.* (2016) analysed PAMs for a range of CRISPR systems through analysing knockdown capability of catalytically inactive

systems. Although the systems analysed in this study had known PAMs, use of degenerate sequences identified hitherto unknown additional functional PAMs (Leenay *et al.*, 2016). Through a similar method, a robotics platform based method for PAM analysis was designed in this study. Here rather than using catalytically inactive systems, the *C. autoethanogenum cas3* was included in order to verify functional activity. Although metadata analysis of this experiment was beyond the timeframe of this study, preliminary results showed that PAMs of plasmids showing reduced GFP expression were consistent with the PAM sequence described in Chapter 5. As shown in other studies (Hayes *et al.*, 2016), Type I systems can show PAM promiscuity and therefore full metadata analysis could provide information on other functional PAMs. Similarly, adaption of the pACYC-duet-CRISPR vector would provide the foundation for identifying the PAMs of CRISPR system identified in this study lacking any serendipitous spacers.

Although the broader scope of the GASCHEM to which this study was part of was to improve solvent production, the majority of this study, as described in Chapters 3-5, was dedicated to genetic tool development. Application of the pMTL40000 vector series provided *IIG* and *cas3* knockout strains. These knockouts aided improvement of DNA transfer and verification of the endogenous CRISPR systems, respectively. Further to this, *ack* and *pta* genes involved in the acetate production pathway were also deleted, providing application of the tools developed in this study towards the overall goals of GASCHEM.

Analysis of the *pta* and *ack* knockouts revealed that neither strain was able to grow on carbon monoxide. As discussed in Chapter 6, this is most likely due to the loss of ATP generated in the Ack-mediated conversion of acetyl-P to acetate. When growing on heterotrophically on fructose, net 2 ATP is generated through glycolysis. However, when growing autotrophically, the Wood-Ljungdahl pathway provides no net gain in ATP. Without an intact acetate pathway, the acetyl-CoA produced from the Wood-Ljungdahl pathway cannot be used for downstream ATP generation and therefore the cell's metabolism can no longer sustain autotrophic growth.

The results described for *ack* and *pta* characterisation show that the proposed redirection of carbon flux away from acetate does not lead to increased ethanol production, due to it being energetically unfavourable. As a result, alternative targets or strategies should be sought.

7.2. Future work

As a continuation of the work described in this study, further characterisation of the mutant strains developed would ideally be conducted. For the knockouts of *cas3* and *cauII*, further analysis would ensure that there was no major impact on the fermentation capabilities of these strains compared to wild type. With regards to tool development, the pMTL40000 vector series has already seen success in a broad range of organisms. However, the nature of the observed higher efficiency compared to other systems described in the literature is yet to be determined. Further efforts to compare trCas9 to the more conventionally used nucleases (SpCas9 and nCas9^{D10A}) could improve understanding of genome editing in clostridia. One such approach could be to

use riboswitch technology currently being developed at the SBRC Nottingham to attain tightly regulated, inducible regulation of nuclease expression and avoid potentially deleterious effects.

For the work described analysing the endogenous CRISPR systems of acetogens, using the known PAM sequences of *C. autoethanogenum* and the pMTL43151-CLAU-eCas vector, transcriptional modulation could be attempted. Again, through using an inducible promoter system, tuneable repression could be achieved for finer control of metabolic pathways. In addition, this approach could be applied to *E. limosum* using the PAM identified in this study.

Other CRISPR systems identified in this study lacking serendipitous spacers could be analysed through the robotics-mediated pipeline described in Chapter 5. Through minor modifications of the plasmids used, the functionality of the CRISPR systems and the identification of their PAMs could be used, opening up the possibility to modify them for molecular tools.

8. Appendix

A) Phylogenetic tree construction

Phylogenetic tree constructed using list of known acetogens published by Ref. (Schiel-Bengelsdorf & Dürre, 2012), using 16S RNA sequences accessed from SILVA rRNA database.

(available from: https://www.arb-silva.de/)

B) *C. autoethanogenum* promoter analysis

Promoter strength of a range of promoters in *C. autoethanogenum*. Research conducted by Bart Pander via CAT assay described in section 2.1.

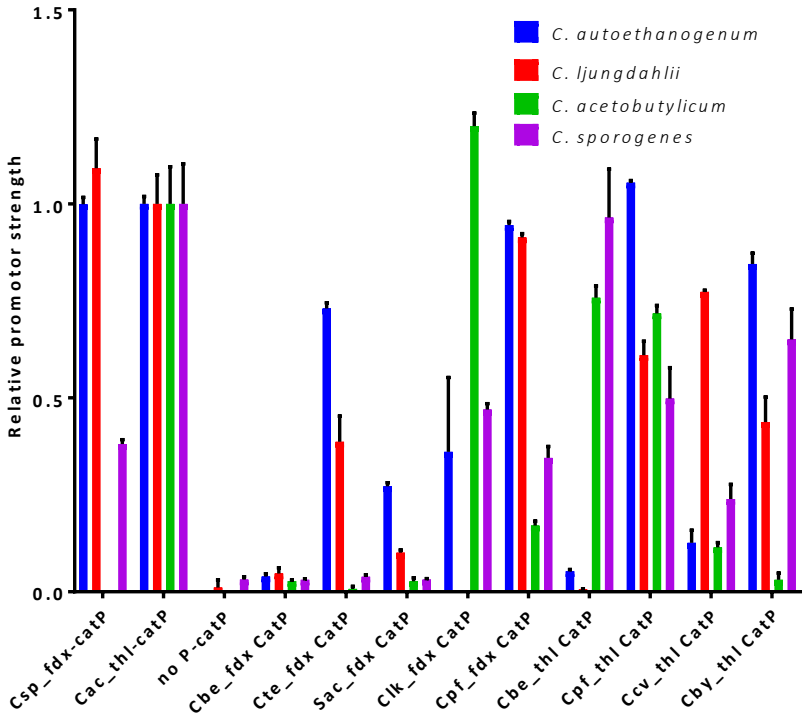


Figure 8.1: Promoter strength in *C. autoethanogenum*
Relative promoter strength of P_{fdx} and P_{thl} from; *C. sporogenes* (Csp), *C. acetobutylicum* (Cac), *C. beijerninckii* (Cbe), *C. tetani* (Cte), *C. saccharoperbutylacetonicum* (Sac), *C. kluykeri* (Clk), *C. perfringens* (Cpf), *C. carboxidivorns* (Ccv), and *C. butyricum* (Cby).

9. Bibliography

Abrini, J., Naveau, H. & Nyns, E. (1994) Clostridium autoethanogenum, sp. nov., an anaerobic

bacterium that produces ethanol from carbon monoxide. *Archives of microbiology*. 161, 345–351.

Altenbuchner, J. (2016) Editing of the Bacillus subtilis Genome by the CRISPR-Cas9 System. *Applied and environmental microbiology*. [Online] 82 (17), 5421–5427. Available from: doi:10.1128/AEM.01453-16 [Accessed: 13 July 2017].

Aro, E.M. (2016) From first generation biofuels to advanced solar biofuels. *Ambio*. [Online] 45 (1), 24–31. Available from: doi:10.1007/s13280-015-0730-0.

Auer, T.O., Duroure, K., De Cian, A., Concordet, J.-P., et al. (2014) Highly efficient CRISPR/Cas9-mediated knock-in in zebrafish by homology-independent DNA repair. *Genome research*. [Online] 24 (1), 142–153. Available from: doi:10.1101/gr.161638.113 [Accessed: 24 May 2014].

Barquist, L., Mayho, M., Cummins, C., Cain, A.K., et al. (2016) The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics (Oxford, England)*. [Online] 32 (7), 1109–1111. Available from: doi:10.1093/bioinformatics/btw022 [Accessed: 20 September 2017].

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*. [Online] 315 (5819), 1709–1712. Available from: doi:10.1126/science.1138140 [Accessed: 19 March 2014].

Barrangou, R. & van Pijkeren, J.P. (2016) Exploiting CRISPR-Cas immune systems for genome editing in bacteria. *Current Opinion in Biotechnology*. [Online]. 37 pp.61–68. Available from: doi:10.1016/j.copbio.2015.10.003.

Bentley, R. (2016) *Introduction to Peak Oil*. 1st edition. Charles AS Hall (ed.). Syracuse, NY, USA, Springer US.

Bhaya, D., Davison, M. & Barrangou, R. (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*. [Online] 45, 273–297. Available from: doi:10.1146/annurev-genet-110410-132430 [Accessed: 19 March 2014].

Boudry, P., Semenova, E., Monot, M., Datsenko, K.A., et al. (2015) Function of the CRISPR-Cas System of the Human Pathogen Clostridium difficile. *mBio*. [Online] 6 (5), e01112-15. Available from: doi:10.1128/mBio.01112-15 [Accessed: 3 September 2017].

Braun, M., Mayer, F. & Gottschalk, G. (1981) Clostridium aceticum (Wieringa), a microorganism producing acetic acid from molecular hydrogen and carbon dioxide. *Archives of Microbiology*. [Online] 128 (3), 288–293. Available from: doi:10.1007/BF00422532 [Accessed: 13 September 2017].

Breitkopf, R., Pander, B. & Piatek, P. (2014) *Transmission electron microscopy of Clostridium autoethanogenum*.

Breznak, J.A. (1994) Acetogenesis from Carbon Dioxide in Termite Guts. In: *Acetogenesis*. [Online]. Boston, MA, Springer US. pp. 303–330. Available from: doi:10.1007/978-1-4615-1777-1_11 [Accessed: 1 February 2017].

Brown, S.D., Nagaraju, S., Utturkar, S., De Tissera, S., et al. (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnology for biofuels*. [Online] 7 (1), 40. Available from: doi:10.1186/1754-6834-7-40 [Accessed: 4 May 2014].

Bruno-barcena, J.M., Chinn, M.S. & Grunden, M. (2013) *Genome Sequence of the Autotrophic Acetogen Clostridium autoethanogenum JA1-1 Strain DSM 10061 , a Producer of Ethanol*. [Online] 1 (4), 2010–2011. Available from: doi:10.1186/1471-2164-9-75.5.

Buffie, C.G., Bucci, V., Stein, R.R., McKenney, P.T., et al. (2014) Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature*. [Online] 517 (7533), 205–208. Available from: doi:10.1038/nature13828 [Accessed: 11 September 2017].

Cass, S.D.B., Haas, K.A., Stoll, B., Alkhnbashi, O.S., et al. (2015) The role of Cas8 in type I CRISPR interference. *Bioscience reports*. [Online] 35 (3). Available from: doi:10.1042/BSR20150043 [Accessed: 12 July 2017].

Celińska, E. & Grajek, W. (2009) Biotechnological production of 2,3-butanediol—Current state and prospects. *Biotechnology Advances*. [Online] 27 (6), 715–725. Available from:

doi:10.1016/j.biotechadv.2009.05.002 [Accessed: 12 July 2017].

Costantino, N. & Court, D.L. (2003) Enhanced levels of lambda Red-mediated recombinants in mismatch repair mutants. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 100 (26), 15748–15753. Available from: doi:10.1073/pnas.2434959100 [Accessed: 5 June 2014].

Cotter, J.L., Chinn, M.S. & Grunden, A.M. (2009) Ethanol and acetate production by Clostridium ljungdahlii and Clostridium autoethanogenum using resting cells. *Bioprocess and biosystems engineering*. [Online] 32 (3), 369–380. Available from: doi:10.1007/s00449-008-0256-y [Accessed: 16 May 2014].

Cummins, C.S. & Johnson, J.L. (1971) Taxonomy of the Clostridia: Wall Composition and DNA Homologies in Clostridium butyricum and Other Butyric Acid-producing Clostridia. *Journal of General Microbiology*. [Online] 67 (1), 33–46. Available from: doi:10.1099/00221287-67-1-33 [Accessed: 13 September 2017].

Datsenko, K. a, Pougach, K., Tikhonov, A., Wanner, B.L., et al. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature communications*. [Online] 3 (May), 945. Available from: doi:10.1038/ncomms1937 [Accessed: 30 April 2014].

Deng, M.D. & Coleman, J.R. (1999) Ethanol synthesis by genetic engineering in cyanobacteria. *Applied and environmental microbiology*. [Online] 65 (2), 523–528. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9925577 [Accessed: 3 September 2017].

Díez-Villaseñor, C., Guzmán, N.M., Almendros, C., García-Martínez, J., et al. (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli . *RNA Biology*. [Online] 10 (5), 792–802. Available from: doi:10.4161/rna.24023 [Accessed: 12 July 2017].

Dong, H., Tao, W., Gong, F., Li, Y., et al. (2014) A functional recT gene for recombineering of Clostridium. *Journal of biotechnology*. [Online] 173, 65–67. Available from: doi:10.1016/j.jbiotec.2013.12.011 [Accessed: 1 May 2014].

Dong, H., Zhang, Y., Dai, Z. & Li, Y. (2010) Engineering Clostridium strain to accept unmethylated DNA. *PLoS ONE*. [Online] 5 (2), 1–8. Available from: doi:10.1371/journal.pone.0009038.

Dower, W.J., Miller, J.F. & Ragsdale, C.W. (1988) High efficiency transformation of E. coli by high voltage electroporation. *Nucleic acids research*. 16 (13), 6127–6146.

Drake, H.L., Gößner, A.S. & Daniel, S.L. (2008) Old Acetogens, New Light. *Annals of the New York Academy of Sciences*. [Online] 1125 (1), 100–128. Available from: doi:10.1196/annals.1419.016 [Accessed: 1 February 2017].

Earp, M. (UK O. and G.A. (2015) UKCS Oil and Gas Production. *UK Oil and Gas Authority*. [Online]. (March). Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/249323/production_projections.pdf.

EIA, U. (US E.I.A. (2017a) *Short-Term Energy Outlook (STEO) Report*. [Online]. Available from: http://www.eia.gov/outlooks/steo/archives/jan17.pdf.

EIA, U. (US E.I.A. (2017b) *US Crude Oil First Purchase Price 1974-2017*. [Online]. Available from: http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=F000000__3&f=M [Accessed: 9 February 2017].

Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., et al. (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature methods*. [Online] 10 (11), 1116–1121. Available from: doi:10.1038/nmeth.2681 [Accessed: 21 March 2014].

Fagerlund, R.D., Staals, R.H.J. & Fineran, P.C. (2015) The Cpf1 CRISPR-Cas protein expands genome-editing tools. *Genome Biology*. [Online] 16 (1), 251. Available from: doi:10.1186/s13059-015-0824-9 [Accessed: 12 July 2017].

Feng, L. & Liao, W. (2016) Legislation, plans, and policies for prevention and control of air pollution in China: Achievements, challenges, and improvements. *Journal of Cleaner Production*. [Online] 112, 1549–1558. Available from: doi:10.1016/j.jclepro.2015.08.013.

Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Künne, T., et al. (2014) Degenerate target sites mediate rapid primed CRISPR adaptation. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 111 (16), E1629-38. Available

from: doi:10.1073/pnas.1400071111 [Accessed: 29 April 2014].
Fisher, A.K., Freedman, B.G., Bevan, D.R. & Senger, R.S. (2014) A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories. *Computational and Structural Biotechnology Journal*. [Online] 11 (18), 91–99. Available from: doi:10.1016/j.csbj.2014.08.010 [Accessed: 3 September 2017].
Fontaine, F.E., Peterson, W.H., McCoy, E., Johnson, M.J., et al. (1942) A New Type of Glucose Fermentation by Clostridium thermoaceticum. *Journal of bacteriology*. [Online] 43 (6), 701–715. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16560531 [Accessed: 3 September 2017].
Fraiture, C. de, Giordano, M. & Liao, Y. (2008) Biofuels and implications for agricultural water use: blue impacts of green energy. *Water Policy*. 10 (S1).
Fujita, T. & Fujii, H. (2013) Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochemical and biophysical research communications*. [Online] 439 (1), 132–136. Available from: doi:10.1016/j.bbrc.2013.08.013 [Accessed: 10 November 2013].
Gaj, T., Gersbach, C. a & Barbas, C.F. (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology*. [Online] 31 (7), 397–405. Available from: doi:10.1016/j.tibtech.2013.04.004 [Accessed: 6 November 2013].
Genthner, B.R., Davis, C.L. & Bryant, M.P. (1981) Features of rumen and sewage sludge strains of Eubacterium limosum, a methanol- and H2-CO2-utilizing species. *Applied and environmental microbiology*. [Online] 42 (1), 12–19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6791591 [Accessed: 13 September 2017].
Gilbert, L. a, Larson, M.H., Morsut, L., Liu, Z., et al. (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*. [Online] 154 (2), 442–451. Available from: doi:10.1016/j.cell.2013.06.044 [Accessed: 7 November 2013].
Goetzl, S., Jeoung, J.-H., Hennig, S.E. & Dobbek, H. (2011) Structural Basis for Electron and Methyl-Group Transfer in a Methyltransferase System Operating in the Reductive Acetyl-CoA Pathway. *Journal of Molecular Biology*. [Online] 411 (1), 96–109. Available from: doi:10.1016/j.jmb.2011.05.025 [Accessed: 24 September 2017].
Grissa, I., Vergnaud, G. & Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic acids research*. [Online] 35 (Web Server issue), W52-7. Available from: doi:10.1093/nar/gkm360 [Accessed: 19 March 2014].
Guerrero, F., Carbonell, V., Cossu, M., Correddu, D., et al. (2012) Ethylene Synthesis and Regulated Expression of Recombinant Protein in Synechocystis sp. PCC 6803 Brett Neilan (ed.). *PLoS ONE*. [Online] 7 (11), e50470. Available from: doi:10.1371/journal.pone.0050470 [Accessed: 3 September 2017].
Guo, H. & Kliesen, K.L. (2005) Oil price volatility and U.S. macroeconomic activity. *Federal Reserve Bank of St. Louis Review*. 87 (6), 669–683.
Guo, X., Zhang, T., Hu, Z., Zhang, Y., et al. (2014) Efficient RNA/Cas9-mediated genome editing in Xenopus tropicalis. *Development (Cambridge, England)*. [Online] 141 (3), 707–714. Available from: doi:10.1242/dev.099853 [Accessed: 19 March 2014].
Hayes, R.P., Xiao, Y., Ding, F., van Erp, P.B.G., et al. (2016) Structural basis for promiscuous PAM recognition in type I–E Cascade from E. coli. *Nature*. [Online] 530 (7591), 499–503. Available from: doi:10.1038/nature16995 [Accessed: 12 July 2017].
Heap, J.T., Kuehne, S. a, Ehsaan, M., Cartman, S.T., et al. (2010) The ClosTron: Mutagenesis in Clostridium refined and streamlined. *Journal of microbiological methods*. [Online] 80 (1), 49–55. Available from: doi:10.1016/j.mimet.2009.10.018 [Accessed: 29 April 2014].
Heap, J.T., Pennington, O.J., Cartman, S.T., Carter, G.P., et al. (2007) The ClosTron: a universal gene knock-out system for the genus Clostridium. *Journal of microbiological methods*. [Online] 70 (3), 452–464. Available from: doi:10.1016/j.mimet.2007.05.021 [Accessed: 27 May 2014].
Heap, J.T., Pennington, O.J., Cartman, S.T. & Minton, N.P. (2009) A modular system for Clostridium shuttle plasmids. *Journal of Microbiological Methods*. [Online] 78 (1), 79–85. Available from: doi:10.1016/j.mimet.2009.05.004.
Hedges, R.W. & Jacob, A.E. (1974) Transposition of ampicillin resistance from RP4 to other

replicons. *MGG Molecular & General Genetics*. [Online] 132 (1), 31–40. Available from: doi:10.1007/BF00268228 [Accessed: 3 September 2017].

Heidrich, N. & Vogel, J. (2013) Same Same but Different: New Structural Insight into CRISPR-Cas Complexes. *Molecular cell*. [Online] 52 (1), 4–7. Available from: doi:10.1016/j.molcel.2013.09.023 [Accessed: 10 November 2013].

Heler, R., Samai, P., Modell, J.W., Weiner, C., et al. (2015) Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature*. [Online] 519 (7542), 199–202. Available from: doi:10.1038/nature14245 [Accessed: 12 July 2017].

Henstra, A. (2016) *Unpublished analysis of steady state fermentation of Clostridium autoethanogenum*.

Hrle, A., Su, A.A.H., Ebert, J., Benda, C., et al. (2013) Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA biology*. [Online] 10 (11), 1670–1678. Available from: doi:10.4161/rna.26500 [Accessed: 12 July 2017].

Huang, H., Chai, C., Li, N., Rowe, P., et al. (2016) CRISPR/Cas9-Based Efficient Genome Editing in Clostridium ljungdahlii, an Autotrophic Gas-Fermenting Bacterium. *ACS Synthetic Biology*. [Online] 5 (12), 1355–1361. Available from: doi:10.1021/acssynbio.6b00044.

Huo, Y., Nam, K.H., Ding, F., Lee, H., et al. (2014) Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nature structural & molecular biology*. [Online] 21 (9), 771–777. Available from: doi:10.1038/nsmb.2875 [Accessed: 12 July 2017].

Ingram, L.C., Richmond, M.H. & Sykes, R.B. (1973) Molecular characterization of the R factors implicated in the carbenicillin resistance of a sequence of Pseudomonas aeruginosa strains isolated from burns. *Antimicrobial Agents and Chemotherapy*. [Online] 3 (2), 279–288. Available from: doi:10.1128/AAC.3.2.279.

Ishino, M., Sawada, Y., Yaegashi, T., Demura, M., et al. (1987) Nucleotide sequence of the adenovirus type 40 inverted terminal repeat: Close relation to that of adenovirus type 5. *Virology*. [Online] 156 (2), 414–416. Available from: doi:10.1016/0042-6822(87)90421-1 [Accessed: 12 July 2017].

Jackson, R.N., Lavin, M., Carter, J. & Wiedenheft, B. (2014) Fitting CRISPR-associated Cas3 into the Helicase Family Tree. *Current Opinion in Structural Biology*. [Online] 24 (1), 106–114. Available from: doi:10.1016/j.sbi.2014.01.001.

Jang, Y.-S., Im, J.A., Choi, S.Y., Lee, J.I., et al. (2014) Metabolic engineering of Clostridium acetobutylicum for butyric acid production with high butyric acid selectivity. *Metabolic engineering*. [Online] 23, 165–174. Available from: doi:10.1016/j.ymben.2014.03.004 [Accessed: 23 May 2015].

Jao, L.-E., Wente, S.R. & Chen, W. (2013) Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proceedings of the National Academy of Sciences of the United States of America*. [Online] 110, 13904–13909. Available from: doi:10.1073/pnas.1308335110.

Jiang, W., Bikard, D., Cox, D., Zhang, F., et al. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology*. [Online] 31 (3), 233–239. Available from: doi:10.1038/nbt.2508 [Accessed: 11 November 2013].

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., et al. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*. [Online] 337 (6096), 816–821. Available from: doi:10.1126/science.1225829 [Accessed: 10 June 2016].

Joung, J.K. & Sander, J.D. (2013) TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology*. [Online] 14 (1), 49–55. Available from: doi:10.1038/nrm3486 [Accessed: 28 April 2014].

Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., et al. (2013) crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. *RNA biology*. [Online] 10 (5), 841–851. Available from: doi:10.4161/rna.24203.

van Kessel, J.C. & Hatfull, G.F. (2007) Recombineering in Mycobacterium tuberculosis. *Nature methods*. [Online] 4 (2), 147–152. Available from: doi:10.1038/nmeth996 [Accessed: 5 June 2014].

Kim, H. & Kim, J.-S. (2014) A guide to genome engineering with programmable nucleases. *Nature reviews. Genetics*. [Online] 15 (5), 321–334. Available from:

doi:10.1038/nrg3686 [Accessed: 28 April 2014].
Kim, Y.-K., Park, S.E., Lee, H. & Yun, J.Y. (2014) Enhancement of bioethanol production in syngas fermentation with Clostridium ljungdahlii using nanoparticles. *Bioresource technology*. [Online] 159, 446–450. Available from: doi:10.1016/j.biortech.2014.03.046 [Accessed: 2 June 2014].
Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V. V, et al. (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*. [Online] 523 (7561), 481–485. Available from: doi:10.1038/nature14592.
Kopke, M., Held, C., Hujer, S., Liesegang, H., et al. (2010) Clostridium ljungdahlii represents a micrbiol production platform based on syngas. *Proceedings of the National Academy of Sciences*. [Online] 107 (34), 15305–15305. Available from: doi:10.1073/pnas.1011320107 [Accessed: 30 April 2014].
Köpke, M., Mihalcea, C., Liew, F., Tizard, J.H., et al. (2011) 2,3-Butanediol Production By Acetogenic Bacteria, an Alternative Route To Chemical Synthesis, Using Industrial Waste Gas. *Applied and environmental microbiology*. [Online] 77 (15), 5467–5475. Available from: doi:10.1128/AEM.00355-11 [Accessed: 10 May 2014].
Körner, H., Sofia, H.J., Zumft, W.G., H., H., et al. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiology Reviews*. [Online] 27 (5), 559–592. Available from: doi:10.1016/S0168-6445(03)00066-4 [Accessed: 25 July 2017].
Kovács, K., Willson, B.J., Schwarz, K., Heap, J.T., et al. (2013) Secretion and assembly of functional mini-cellulosomes from synthetic chromosomal operons in Clostridium acetobutylicum ATCC 824. *Biotechnology for biofuels*. [Online] 6 (1), 117. Available from: doi:10.1186/1754-6834-6-117 [Accessed: 28 May 2014].
Lander, E.S. (2016) The Heroes of CRISPR. *Cell*. [Online] 164 (1–2), 18–28. Available from: doi:10.1016/j.cell.2015.12.041.
Langridge, G.C., Phan, M.-D., Turner, D.J., Perkins, T.T., et al. (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome research*. [Online] 19 (12), 2308–2316. Available from: doi:10.1101/gr.097097.109 [Accessed: 20 September 2017].
Larson, M.H., Gilbert, L. a, Wang, X., Lim, W. a, et al. (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature protocols*. [Online] 8 (11), 2180–2196. Available from: doi:10.1038/nprot.2013.132 [Accessed: 7 November 2013].
Lederle, S. (2011) *Heterofermentative Acetonproduktion*. University of Ulm.
Leenay, R.T., Maksimchuk, K.R., Slotkowski, R.A., Agrawal, R.N., et al. (2016) *Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems*. [Online] Available from: doi:10.1016/j.molcel.2016.02.031 [Accessed: 13 September 2017].
Lepikhov, K., Tchernov, A., Zheleznaja, L., Matvienko, N., et al. (2001) Characterization of the type IV restriction modification system BspLU11III from Bacillus sp . LU11. *Nucleic Acids Research*. 29 (22), 4691–4698.
Levy, A., Goren, M.G., Yosef, I., Auster, O., et al. (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature*. [Online] 520 (7548), 505–510. Available from: doi:10.1038/nature14302 [Accessed: 12 July 2017].
Li, H. (2015) Structural Principles of CRISPR RNA Processing. *Structure*. [Online] 23 (1), 13–20. Available from: doi:10.1016/j.str.2014.10.006 [Accessed: 12 July 2017].
Li, Q., Chen, J., Minton, N.P., Zhang, Y., et al. (2016) CRISPR-based genome editing and expression control systems in *Clostridium acetobutylicum* and *Clostridium beijerinckii*. *Biotechnology Journal*. [Online] 11 (7), 961–972. Available from: doi:10.1002/biot.201600053 [Accessed: 11 July 2017].
Liew, F., Henstra, A.M., Köpke, M., Winzer, K., et al. (2017) Metabolic engineering of Clostridium autoethanogenum for selective alcohol production. *Metabolic Engineering*. [Online] 40, 104–114. Available from: doi:10.1016/j.ymben.2017.01.007 [Accessed: 12 July 2017].
Liew, F., Henstra, A.M., Winzer, K., Kopke, M., et al. (2016) Insights into CO 2 Fixation Pathway of Clostridium autoethanogenum by Targeted Mutagenesis. *mBio*. [Online] 7 (3), 1–10. Available from: doi:10.1128/mBio.00427-16.Editor.
Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., et al. (2011) Structural and functional

characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *The Journal of biological chemistry*. [Online] 286 (24), 21643–21656. Available from: doi:10.1074/jbc.M111.238485 [Accessed: 21 November 2013].

Luo, M.L., Mullis, A.S., Leenay, R.T. & Beisel, C.L. (2015) Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic Acids Research*. [Online] 43 (1), 674–681. Available from: doi:10.1093/nar/gku971 [Accessed: 19 September 2017].

Mackie, R.I. & Bryant, M.P. (1994) Acetogenesis and the Rumen: Syntrophic Relationships. In: *Acetogenesis*. [Online]. Boston, MA, Springer US. pp. 331–364. Available from: doi:10.1007/978-1-4615-1777-1_12 [Accessed: 1 February 2017].

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., et al. (2011) Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology*. [Online] 9 (6), 467–477. Available from: doi:10.1038/nrmicro2577 [Accessed: 8 November 2013].

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., et al. (2015) An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology*. [Online] 13 (11), 722–736. Available from: doi:10.1038/nrmicro3569.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., et al. (2013a) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology*. [Online] 31 (9), 833–838. Available from: doi:10.1038/nbt.2675 [Accessed: 20 March 2014].

Mali, P., Yang, L., Esvelt, K.M., Aach, J., et al. (2013b) RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*. [Online] 339 (6121), 823–826. Available from: doi:10.1126/science.1232033 [Accessed: 23 May 2014].

Marraffini, L.A. (2015) *CRISPR-Cas immunity in prokaryotes*. [Online] Available from: doi:10.1038/nature15386.

Marraffini, L. a & Sontheimer, E.J. (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (New York, N.Y.)*. [Online] 322 (5909), 1843–1845. Available from: doi:10.1126/science.1165771.

McCrone, A., Moslener, U., D'Estais, F., Usher, E., et al. (2016) *Global Trends in Renewable Energy*.

Mermelstein, L.D. & Papoutsakis, E.T. (1993) In vivo methylation in Escherichia coli by the Bacillus subtilis phage ??3T I methyltransferase to protect plasmids from restriction upon transformation of Clostridium acetobutylicum ATCC 824. *Applied and Environmental Microbiology*. [Online] 59 (4), 1077–1081. Available from: doi:10.1016/j.bbrc.2011.01.095.

Mermelstein, L.D., Welker, N.E., Peterson, D.J., Bennett, G.N., et al. (1994) Genetic and Metabolic Engineering of Clostridium acetobutylicum ATCC 824. *Annals of the New York Academy of Sciences*. [Online] 721 (1 Recombinant D), 54–68. Available from: doi:10.1111/j.1749-6632.1994.tb47376.x [Accessed: 29 May 2014].

Millard, J. (2017) *The Development of Forward and Reverse Genetic Tools in Eubacterium limosum (working title)*. University of Nottingham.

Miller, R.G., Sorrell, S.R., Lane, H., Kt, S., et al. (2013) The future of oil supply. *Phil. Trans. R. Soc. A*. [Online] 372. Available from: doi:10.1098/rsta.2013.0179.

Minton, N.P., Ehsaan, M., Humphreys, C.M., Little, G.T., et al. (2016) A roadmap for gene system development in Clostridium. *Anaerobe*. [Online] 41, 104–112. Available from: doi:10.1016/j.anaerobe.2016.05.011.

Minton, N.P. & Morris, J.G. (1983) Regeneration of protoplasts of Clostridium pasteurianum ATCC 6013. *Journal of Bacteriology*. 155 (1), 432–434.

Mladenov, E. & Iliakis, G. (2011) Induction and repair of DNA double strand breaks: the increasing spectrum of non-homologous end joining pathways. *Mutation research*. [Online] 711 (1–2), 61–72. Available from: doi:10.1016/j.mrfmmm.2011.02.005 [Accessed: 28 April 2014].

Mo, X., Pei, J., Guo, Y., Lin, L., et al. (2015) Genome Sequence of Clostridium acetobutylicum GXAS18-1, a Novel Biobutanol Production Strain. *Genome announcements*. [Online] 3 (2). Available from: doi:10.1128/genomeA.00033-15 [Accessed: 3 September 2017].

Mock, J., Zheng, Y., Mueller, A.P., Ly, S., et al. (2015) Energy Conservation Associated with Ethanol Formation from H2 and CO2 in Clostridium autoethanogenum Involving

Electron Bifurcation. *Journal of bacteriology*. [Online] 197 (18), 2965–2980. Available from: doi:10.1128/JB.00399-15 [Accessed: 12 July 2017].

Mojica, F.J.M., Juez, G. & Rodriguez-Valera, F. (1993) Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Molecular Microbiology*. [Online] 9 (3), 613–621. Available from: doi:10.1111/j.1365-2958.1993.tb01721.x [Accessed: 12 July 2017].

Mojica, F.J.M. & Montoliu, L. (2016) On the Origin of CRISPR-Cas Technology: From Prokaryotes to Mammals. *Trends in Microbiology*. [Online] 24 (10), 811–820. Available from: doi:10.1016/j.tim.2016.06.005 [Accessed: 12 July 2017].

Mosberg, J.A., Lajoie, M.J. & Church, G.M. (2010) Lambda red recombineering in Escherichia coli occurs through a fully single-stranded intermediate. *Genetics*. [Online] 186 (3), 791–799. Available from: doi:10.1534/genetics.110.120782 [Accessed: 3 June 2014].

Muller, M., Lee, C.M., Gasiunas, G., Davis, T.H., et al. (2016) Streptococcus thermophilus CRISPR-Cas9 Systems Enable Specific Editing of the Human Genome. *Molecular Therapy*. [Online] 24 (3), 636–644. Available from: doi:10.1038/mt.2015.218.

Muyrers, J.P., Zhang, Y., Buchholz, F. & Stewart, A.F. (2000) RecE/RecT and Redalpha/Redbeta initiate double-stranded break repair by specifically interacting with their respective partners. *Genes & development*. [Online] 14 (15), 1971–1982. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=316823&tool=pmcentrez&rendertype=abstract [Accessed: 5 June 2014].

Nagarajan, H., Sahin, M., Nogales, J., Latif, H., et al. (2013) Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of Clostridium ljungdahlii. *Microbial Cell Factories*. [Online] 12 (1), 118. Available from: doi:10.1186/1475-2859-12-118.

Nagaraju, S., Davies, N.K., Walker, D.J.F., Köpke, M., et al. (2016) Genome editing of Clostridium autoethanogenum using CRISPR/Cas9. *Biotechnology for Biofuels*. [Online] 9 (1), 219. Available from: doi:10.1186/s13068-016-0638-3.

Nakamura, C.E. & Whited, G.M. (2003) Metabolic engineering for the microbial production of 1,3-propanediol. *Current Opinion in Biotechnology*. [Online] 14 (5), 454–459. Available from: doi:10.1016/j.copbio.2003.08.005 [Accessed: 3 May 2014].

Nale, J.Y., Spencer, J., Hargreaves, K.R., Buckley, A.M., et al. (2016) Bacteriophage Combinations Significantly Reduce Clostridium difficile Growth In Vitro and Proliferation In Vivo. *Antimicrobial agents and chemotherapy*. [Online] 60 (2), 968–981. Available from: doi:10.1128/AAC.01774-15 [Accessed: 11 September 2017].

Negahdaripour, M., Nezafat, N., Hajighahramani, N., Rahmatabadi, S.S., et al. (2017) Investigating CRISPR-Cas systems in Clostridium botulinum via bioinformatics tools. *Infection, Genetics and Evolution*. [Online] 54, 355–373. Available from: doi:10.1016/j.meegid.2017.06.027 [Accessed: 3 September 2017].

Ng, C.Y., Jung, M., Lee, J. & Oh, M.-K. (2012) Production of 2,3-butanediol in Saccharomyces cerevisiae by in silico aided metabolic engineering. *Microbial Cell Factories*. [Online] 11 (1), 68. Available from: doi:10.1186/1475-2859-11-68 [Accessed: 12 July 2017].

Ng, Y.K., Ehsaan, M., Philip, S., Collery, M.M., et al. (2013) Expanding the repertoire of gene tools for precise manipulation of the Clostridium difficile genome: allelic exchange using pyrE alleles. *PloS one*. [Online] 8 (2), e56051. Available from: doi:10.1371/journal.pone.0056051 [Accessed: 6 May 2014].

Nikolskaya, A.N. & Galperin, M.Y. (2002) A novel type of conserved DNA-binding domain in the transcriptional regulators of the AlgR/AgrA/LytR family. *Nucleic Acids Research*. [Online] 30 (11), 2453–2459. Available from: doi:10.1093/nar/30.11.2453 [Accessed: 25 July 2017].

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., et al. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*. [Online] 156 (5), 935–949. Available from: doi:10.1016/j.cell.2014.02.001 [Accessed: 2 December 2016].

Oh, J.-H. & van Pijkeren, J.-P. (2014) CRISPR-Cas9-assisted recombineering in Lactobacillus reuteri. *Nucleic Acids Research*. [Online] 42 (17), e131–e131. Available from: doi:10.1093/nar/gku623 [Accessed: 13 July 2017].

Ollivier, B., Caumette, P., Garcia, J.L. & Mah, R. a (1994) Anaerobic bacteria from hypersaline environments. *Microbiological reviews*. 58 (1), 27–38.

Palmer, B.R. & Marinus, M.G. (1994) The dam and dcm strains of Escherichia coli — a review. *Gene*. [Online] 143 (1), 1–12. Available from: doi:10.1016/0378-1119(94)90597-5 [Accessed: 9 April 2018].

Park, A., Won, S.T., Pentecost, M., Bartkowski, W., et al. (2014) CRISPR/Cas9 Allows Efficient and Complete Knock-In of a Destabilization Domain-Tagged Essential Protein in a Human Cell Line, Allowing Rapid Knockdown of Protein Function. *PloS one*. [Online] 9 (4), e95101. Available from: doi:10.1371/journal.pone.0095101 [Accessed: 24 May 2014].

Peng, H., Wu, G. & Shao, W. (2008) The aldehyde/alcohol dehydrogenase (AdhE) in relation to the ethanol formation in Thermoanaerobacter ethanolicus JW200. *Anaerobe*. [Online] 14 (2), 125–127. Available from: doi:10.1016/j.anaerobe.2007.09.004 [Accessed: 12 July 2017].

Peng, L., Pei, J., Pang, H., Guo, Y., et al. (2014) Whole genome sequencing reveals a novel CRISPR system in industrial Clostridium acetobutylicum. *Journal of Industrial Microbiology & Biotechnology*. [Online] 41 (11), 1677–1685. Available from: doi:10.1007/s10295-014-1507-3 [Accessed: 3 September 2017].

van Pijkeren, J.-P. & Britton, R.A. (2012) High efficiency recombineering in lactic acid bacteria. *Nucleic acids research*. [Online] 40 (10), e76. Available from: doi:10.1093/nar/gks147 [Accessed: 23 May 2014].

Poehlein, A., Cebulla, M., Ilg, M.M., Bengelsdorf, F.R., et al. (2015) The Complete Genome Sequence of Clostridium aceticum: a Missing Link between Rnf- and Cytochrome-Containing Autotrophic Acetogens. *mBio*. [Online] 6 (5), e01168-15. Available from: doi:10.1128/mBio.01168-15 [Accessed: 2 September 2017].

Purdy, D., O'Keeffe, T. a T., Elmore, M., Herbert, M., et al. (2002) Conjugative transfer of clostridial shuttle vectors from Escherichia coli to Clostridium difficile through circumvention of the restriction barrier. *Molecular microbiology*. [Online] 46 (2), 439–452. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12406220.

Pyne, M.E., Bruder, M.R., Moo-Young, M., Chung, D.A., et al. (2016a) Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in Clostridium. *Scientific Reports*. [Online] 6, 25666. Available from: doi:10.1038/srep25666 [Accessed: 12 September 2016].

Pyne, M.E., Bruder, M.R., Moo-Young, M., Chung, D.A., et al. (2016b) Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in Clostridium. *Scientific reports*. [Online] 6 (December 2015), 25666. Available from: doi:10.1038/srep25666.

Pyne, M.E., Moo-Young, M., Chung, D. a & Chou, C.P. (2013) Development of an electrotransformation protocol for genetic manipulation of Clostridium pasteurianum. *Biotechnology for biofuels*. [Online] 6 (1), 50. Available from: doi:10.1186/1754-6834-6-50.

Ragsdale, S.W. & Pierce, E. (2008) Acetogenesis and the Wood-Ljungdahl pathway of CO2 fixation. *Biochimica et Biophysica Acta - Proteins and Proteomics*. [Online] 1784 (12), 1873–1898. Available from: doi:10.1016/j.bbapap.2008.08.012.

Ran, F.A., Hsu, P.D., Lin, C.-Y., Gootenberg, J.S., et al. (2013a) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*. [Online] 154 (6), 1380–1389. Available from: doi:10.1016/j.cell.2013.08.021 [Accessed: 8 November 2013].

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., et al. (2013b) Genome engineering using the CRISPR-Cas9 system. *Nature protocols*. [Online] 8 (11), 2281–2308. Available from: doi:10.1038/nprot.2013.143 [Accessed: 6 November 2013].

Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P.R., et al. (2015) Efficient programmable gene silencing by Cascade. *Nucleic Acids Research*. [Online] 43 (1), 237–246. Available from: doi:10.1093/nar/gku1257 [Accessed: 18 September 2017].

Richter, H., Zoephel, J., Schermuly, J., Maticzka, D., et al. (2012) Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic acids research*. [Online] 40 (19), 9887–9896. Available from: doi:10.1093/nar/gks737 [Accessed: 18 November 2013].

Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D. (2015) REBASE-a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Research*.

[Online] 43 (D1), D298–D299. Available from: doi:10.1093/nar/gku1046.
Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L., et al. (2015) Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *eLife*. [Online] 4 (AUGUST2015), 1–19. Available from: doi:10.7554/eLife.08716.
Rosencrantz, D., Rainey, F.A. & Janssen, P.H. (1999) Culturable populations of Sporomusa spp. and Desulfovibrio spp. in the anoxic bulk soil of flooded rice microcosms. *Applied and Environmental Microbiology*. 65 (8), 3526–3533.
Sakuma, T. & Woltjen, K. (2014) Nuclease-mediated genome editing: At the front-line of functional genomics technology. *Development, growth & differentiation*. [Online] 56 (1), 2–13. Available from: doi:10.1111/dgd.12111 [Accessed: 19 March 2014].
Schiel-Bengelsdorf, B. & Dürre, P. (2012) Pathway engineering and synthetic biology using acetogens. *FEBS Letters*. [Online] 586 (15), 2191–2198. Available from: doi:10.1016/j.febslet.2012.04.043.
Schuchmann, K. & Müller, V. (2014) Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nature Reviews Microbiology*. [Online] 12 (12), 809–821. Available from: doi:10.1038/nrmicro3365.
Selle, K., Klaenhammer, T.R. & Barrangou, R. (2015) CRISPR-based screening of genomic island excision events in bacteria. *Proceedings of the National Academy of Sciences*. [Online] 112 (26), 8076–8081. Available from: doi:10.1073/pnas.1508525112 [Accessed: 10 June 2016].
Sharan, S.K., Thomason, L.C., Kuznetsov, S.G. & Court, D.L. (2009) Recombineering: a homologous recombination-based method of genetic engineering. *Nature protocols*. [Online] 4 (2), 206–223. Available from: doi:10.1038/nprot.2008.227 [Accessed: 30 April 2014].
Shaw, W. V. (1975) Chloramphenicol acetyltransferase from chloramphenicol-resistant bacteria. *Methods in Enzymology*. [Online] 43, 737–755. Available from: doi:10.1016/0076-6879(75)43141-X.
Shen, Y., Brown, R. & Wen, Z. (2014) Syngas fermentation of Clostridium carboxidivoran P7 in a hollow fiber membrane biofilm reactor: Evaluating the mass transfer coefficient and ethanol production performance. *Biochemical Engineering Journal*. [Online] 85, 21–29. Available from: doi:10.1016/j.bej.2014.01.010 [Accessed: 2 June 2014].
Shmakov, S., Smargon, A., Scott, D., Cox, D., et al. (2017) Diversity and evolution of class 2 CRISPR–Cas systems. *Nature Reviews Microbiology*. [Online] 15 (3), 169–182. Available from: doi:10.1038/nrmicro.2016.184 [Accessed: 12 July 2017].
Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., et al. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *The EMBO Journal*. [Online] 30 (7), 1335–1342. Available from: doi:10.1038/emboj.2011.41 [Accessed: 12 July 2017].
Smale, S.T. (2010) Chloramphenicol Acetyltransferase Assay. *Cold Spring Harbor Protocols*. [Online] 2010 (5), pdb.prot5422-prot5422. Available from: doi:10.1101/pdb.prot5422 [Accessed: 27 September 2017].
Smargon, A.A., Cox, D.B.T., Pyzocha, N.K., Zheng, K., et al. (2017) Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Molecular Cell*. [Online] 65 (4), 618–630.e7. Available from: doi:10.1016/j.molcel.2016.12.023 [Accessed: 12 July 2017].
Smith, P.K., Krohn, R.I., Hermanson, G.T., Mallia, A.K., et al. (1985) Measurement of protein using bicinchoninic acid. *Analytical biochemistry*. [Online] 150 (1), 76–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3843705 [Accessed: 27 September 2017].
Steinert, J., Schiml, S., Fauser, F. & Puchta, H. (2015) Highly efficient heritable plant genome engineering using Cas9 orthologues from Streptococcus thermophilus and Staphylococcus aureus. *Plant Journal*. [Online] 84 (6), 1295–1305. Available from: doi:10.1111/tpj.13078.
Straub, M., Demler, M., Weuster-Botz, D. & Dürre, P. (2014) Selective enhancement of autotrophic acetate production with genetically modified Acetobacterium woodii. *Journal of Biotechnology*. [Online] 178, 67–72. Available from: doi:10.1016/j.jbiotec.2014.03.005 [Accessed: 3 September 2017].
Suzuki, Y., Kishigami, T., Inoue, K., Mizoguchi, Y., et al. (1983) Bacillus thermoglucosidasius sp. nov., a New Species of Obligately Thermophilic Bacilli. *Systematic and Applied*

Microbiology. [Online] 4 (4), 487–495. Available from: doi:10.1016/S0723-2020(83)80006-X [Accessed: 13 September 2017].

Tanner, R.S., Miller, L.M. & Yang, D. (1993) Clostridium ljungdahlii sp. nov., an Acetogenic Species in Clostridial rRNA Homology Group I. *International journal of systematic bacteriology*. 43 (2), 232–236.

Thapa, L.P., Lee, S.J., Yang, X., Lee, J.H., et al. (2015) Improved bioethanol production from metabolic engineering of Enterobacter aerogenes ATCC 29007. *Process Biochemistry*. [Online] 50 (12), 2051–2060. Available from: doi:10.1016/j.procbio.2015.09.007 [Accessed: 12 July 2017].

Tuominen, H., Salminen, A., Oksanen, E., Jämsen, J., et al. (2010) Crystal Structures of the CBS and DRTGG Domains of the Regulatory Region of Clostridium perfringens Pyrophosphatase Complexed with the Inhibitor, AMP, and Activator, Diadenosine Tetraphosphate. *Journal of Molecular Biology*. [Online] 398 (3), 400–413. Available from: doi:10.1016/j.jmb.2010.03.019 [Accessed: 25 July 2017].

Ueki, T., Nevin, K.P., Woodard, T.L. & Lovley, D.R. (2014) Converting carbon dioxide to butyrate with an engineered strain of Clostridium ljungdahlii. *mBio*. [Online] 5 (5), e01636-14. Available from: doi:10.1128/mBio.01636-14 [Accessed: 12 July 2017].

Ugoala, E., Ndukwe, G., Mustapha, K. & Ayo, R. (2012) Constraints to large scale algae biomass production and utilization. *J. Algal Biomass Utln*. 3 (2), 14–32.

Ugwu, C.U., Aoyagi, H. & Uchiyama, H. (2008) Photobioreactors for mass cultivation of algae. *Bioresource Technology*. [Online] 99 (10), 4021–4028. Available from: doi:10.1016/j.biortech.2007.01.046.

Upadhyay, S.K., Kumar, J., Alok, A. & Tuli, R. (2013) RNA-guided genome editing for target gene mutations in wheat. *G3 (Bethesda, Md.)*. [Online] 3 (12), 2233–2238. Available from: doi:10.1534/g3.113.008847 [Accessed: 19 March 2014].

Ventura, J.-R.S., Hu, H. & Jahng, D. (2013) Enhanced butanol production in Clostridium acetobutylicum ATCC 824 by double overexpression of 6-phosphofructokinase and pyruvate kinase genes. *Applied Microbiology and Biotechnology*. [Online] 97 (16), 7505–7516. Available from: doi:10.1007/s00253-013-5075-7 [Accessed: 3 September 2017].

Wakefield, N., Rajan, R. & Sontheimer, E.J. (2015) Primary processing of CRISPR RNA by the endonuclease Cas6 in *Staphylococcus epidermidis*. *FEBS Letters*. [Online] 589 (20PartB), 3197–3204. Available from: doi:10.1016/j.febslet.2015.09.005 [Accessed: 12 July 2017].

Wang, Y., Zhang, Z.-T., Seo, S.-O., Choi, K., et al. (2015) Markerless chromosomal gene deletion in Clostridium beijerinckii using CRISPR/Cas9 system. *Journal of biotechnology*. [Online] 200, 1–5. Available from: doi:10.1016/j.jbiotec.2015.02.005 [Accessed: 15 May 2015].

Wasels, F., Jean-Marie, J., Collas, F., López-Contreras, A.M., et al. (2017) A two-plasmid inducible CRISPR/Cas9 genome editing tool for Clostridium acetobutylicum. *Journal of Microbiological Methods*. [Online] 140, 5–11. Available from: doi:10.1016/j.mimet.2017.06.010 [Accessed: 25 July 2017].

Weenk, G., Fitzmaurice, E. & Mossel, D.A.A. (1991) Selective enumeration of spores of Clostridium species in dried foods. *Journal of Applied Bacteriology*. [Online] 70 (2), 135–143. Available from: doi:10.1111/j.1365-2672.1991.tb04439.x [Accessed: 3 June 2014].

Williams, R.D., Young, D.I. & Young, M. (1993) Conjugative plasmid transfer from Escherichia coli to Clostridium acetobutylicum. *Journal of General Microbiology*. 136 (1), 819–826.

Wilson, G.G. (1991) Organization of restriction-modification systems. *Nucleic Acids Research*. [Online] 19 (10), 2539–2566. Available from: doi:10.1093/nar/19.10.2539.

Wood, H.G. (1991) Life with CO or CO2 and H2 as a source of carbon and energy. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 5 (2), 156–163.

Woods, C. (2017) *Unpublished research*. University of Nottingham.

Xu, K., Ren, C., Liu, Z., Zhang, T., et al. (2015a) Efficient genome engineering in eukaryotes using Cas9 from Streptococcus thermophilus. *Cellular and Molecular Life Sciences*. [Online] 72 (2), 383–399. Available from: doi:10.1007/s00018-014-1679-z.

Xu, T., Li, Y., Shi, Z., Hemme, C.L., et al. (2015b) Efficient genome editing in clostridium cellulolyticum via CRISPR-Cas9 nickase. *Applied and Environmental Microbiology*.

[Online] 81 (13), 4423–4431. Available from: doi:10.1128/AEM.00873-15.

Xu, Y., Chu, H., Gao, C., Tao, F., et al. (2014) Systematic metabolic engineering of Escherichia coli for high-yield production of fuel bio-chemical 2,3-butanediol. *Metabolic engineering*. [Online] 23C, 22–33. Available from: doi:10.1016/j.ymben.2014.02.004 [Accessed: 28 April 2014].

Yao, S. & Mikkelsen, M.J. (2010) Identification and overexpression of a bifunctional aldehyde/alcohol dehydrogenase responsible for ethanol production in Thermoanaerobacter mathranii. *Journal of molecular microbiology and biotechnology*. [Online] 19 (3), 123–133. Available from: doi:10.1159/000321498 [Accessed: 12 July 2017].

Yosef, I., Goren, M.G. & Qimron, U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic acids research*. [Online] 40 (12), 5569–5576. Available from: doi:10.1093/nar/gks216 [Accessed: 12 July 2017].

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., et al. (2016) Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. *Nature Biotechnology*. [Online] 35 (1), 31–34. Available from: doi:10.1038/nbt.3737 [Accessed: 12 July 2017].

Zhang, J., Kasciukovic, T., White, M.F., Barrangou, R., et al. (2012) The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster Martin G. Marinus (ed.). *PLoS ONE*. [Online] 7 (10), e47232. Available from: doi:10.1371/journal.pone.0047232 [Accessed: 12 July 2017].

Zhang, Y., Xu, S., Chai, C., Yang, S., et al. (2016) Development of an inducible transposon system for efficient random mutagenesis in Clostridium acetobutylicum. *FEMS microbiology letters*. [Online] 363 (8). Available from: doi:10.1093/femsle/fnw065 [Accessed: 12 September 2017].