The University Of Sheffield.

Characterisation of *Clostridium saccharoperbutylacetonicum* for biotechnological applications

Laurence de Lussy-Kubisa

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Faculty of Science

School of Biosciences

February 2022

Declaration

I, Laurence de Lussy-Kubisa, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not previously been presented for an award at this, or any other, university.

LDLK

February 2022

Abstract

Clostridium saccharoperbutylacetonicum is an anaerobe that is growing in importance for biotechnological applications. In recent years, significant strides have been made in improving the genetic tractability and in understanding *C. saccharoperbutylacetonicum* fermentative biology. Despite this, multiple facets of the species' underlying biology are not well-characterised. In particular, there is little published work examining *C. saccharoperbutylacetonicum* using either forward or reverse genetics approaches. Particularly, the understanding of sporulation, germination, and quorum sensing in the species is severely lacking.

This thesis aimed to establish a pipeline for transposon insertion site sequencing, a forward genetics approach, in *C. saccharoperbutylacetonicum* to determine its essential genome. Whilst we were not ultimately successful, we did identify 677 open reading frames that are possibly essential to growth. We believe this list, along with a partial list of non-essential open reading frames, provides invaluable data with which to aid future *C. saccharoperbutylacetonicum* work.

We also examined sporulation and germination. These processes are a key to the life cycle of the species in its natural habitat. Sporulation is also intimately tied to the solventogenic stationary phase of vegetative cells that is key to biotechnological applications. Here we characterised for the first time the morphology and ultrastructure of *C. saccharoperbutylacetonicum* sporulating cultures. In addition, we screened and identified effective germinants and germination inhibitors to allow for control over those processes.

Finally, quorum sensing, the process by which bacteria sense population density, is likely extremely important in high density cell contexts. However, it has been little studied in *C. saccharoperbutylacetonicum*, despite its industrial relevance and routine fermentation. We present the first work partially characterising the accessory gene regulator quorum sensing system in *C. saccharoperbutylacetonicum*. We showed the importance of the system to both solventogenesis and sporulation and highlight pathways to a better understanding of this crucial but complex process.

Acknowledgements

Well, it's been an *interesting* journey. One that was bookended with brazen acts of aggression from Vladimir Putin and contained a hearty slice of pandemic. But it wouldn't have been possible without a myriad of people; those who made the project possible and those who made everyday life a joy.

Firstly, I must thank Rob Fagan and Liz Jenkinson for creating this project and entrusting me with it. Thanks as well must go for their help and advice during these four years, especially with the preparation of this manuscript. I'm also grateful for the scientific and project advice of Anna Baker who always approached my issues with customary energy and enthusiasm. I owe a not-insignificant science and personal debt to Mandy Nicolle and Sasha Atmadjaja for all their help, particularly during my placement. A big thanks too to Stéphane Mesnage whose principles and advice I've always benefitted from and respected – keep up the good fight. A hearty thank you to Emily Cooper for all her administrative help and patience. Finally, thanks as well to Roy Chaudhuri for vital data analysis.

To my F25/F13 lab friends and, dare I say it, colleagues, I owe a massive debt. Rob Smith has been a cornerstone of my time in Sheffield and I'm constantly inspired by his dedication, willingness to help a scientist in need, good humour and owe a lot to his late night lab session support. (Dr) Joe Kirk has been a fantastic support, supplying me with high quality science and life advice, many, many laughs and, most importantly, facts about ants and/or wasps. A huge thanks must also go to Jessie Davis and Hannah Fisher who have been so important to my getting through these last months and on whom I could always rely on to listen to my rants, give top notch advice, both scientific and otherwise, and, most importantly, for a laugh. A big thank you also to Sophie Irving whose wicked sense of humour and willingness to have just one more with me have resulted in some of my best nights in Sheffield and who has always tacitly understood the 'no mean banter' rule which has meant that a Manchester United fan and Liverpool fan can watch football together. Adam Brooks and Sophie McKie thanks so much for always having your apartment door open for me and for all the board game nights, especially at the end when I needed it most – keep on trucking! Thanks as well to Nadia Fernandes for being a key member of the TraDIS Support Club. Finally, thanks to Shauna O Beirne for sharing this wild ride and to Bartek Salamaga for his ever-present generosity.

I also would like to show some love to my pre-Sheffield friends. A huge thanks Lee Sewell (PhD) for always being willing to respond to even the stupidest messages, no matter what time of day – very much 'for me'. To George Sydenham for Peaks trips, heart-to-hearts and, most especially, lending me his spare trousers in Munich. To Jeremy Wikeley for making me aware of the power of

ressentiment and the always stimulating and amusing conversation (I *think* I know what poetry is now). To Harry Clayton for his calls, especially during the tough lockdown months and always having my back. To Charles Motraghi for uncovering the System together. To Helen Poulson for her support, humour and her generosity, especially when it came to needing a bed in London. To Li Sa Choo and Sam Ford for being Best Friends (and allowing me to tag along) and for sharing the highs and lows of our respective 'careers'.

Both literally and figuratively, none of this would have been possible without my family. I owe everything to Seran, Maya, Serge and Sylvie who made me who I am. Thank you so much for always supporting me, restoring my confidence and also for telling me, in the most loving way possible, when I'm wrong. Even through these difficult times and years apart, I have always felt your love, humour and support. I love you all very, very much.

A quick thanks to the metaphorical and literal top dogs: Lulu, Charcoal, Lucio and Felicia for their love and for keeping me on my toes (or should that be fingers Felicia?).

To Luz, I cannot thank you enough. You have been by my side on every step of this journey, listened to my frustrations over and over, shared the highs and been there to provide hugs, beers, snacks and humour for the lows. Your determination, resilience, and fierce intelligence inspire me every day. You made the first lockdown a breeze, creating a happy memory despite it all. Finally, thank you as well for introducing me to the myriad joys of Mexico, particularly the food! You are, quite simply, a delight.

Oh, it would be remiss of me not to mention Prof Alan Berry, without whom I would never have thought to drop all the dud experiments and just do the ones that work.

It would be, at best, a dubious honour to have this thesis dedicated to you, so I won't.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

4-MU	4-methylumbelliferone
ABE	Acetone butanol ethanol
agr	Accessory gene regulator
AIP	Autoinducing peptide
ATc	Anhydrotetracycline
ATP	Adenosine triphosphate
BHI	Brain heart infusion media
BLAST	Basic local alignment search tool
bp	Base pairs
Cas	CRISPR-associated protein
CFU	Colony forming units
CGM	*Clostridial* growth media
CoA	Coenzyme A
Cq	Quantification cycle
CRISPR	Clustered regularly interspaced short palindromic repeats
DMSO	Dimethyl sulfoxide
dNTP	Deoxynucleotide
DPA	Dipicolinic acid
EDTA	Ethylenediaminetetraacetic acid
Ery	Erythromycin
FDR	False discovery rate
gDNA	Genomic DNA
HITS	High-throughput insertion tracking by deep sequencing
HPLC	High-performance liquid chromatography
INSeq	Insertion sequencing
ITR	Inverted terminal repeat
LB	Lysogeny broth
Mbp	Million base pairs
MES	2-(N-morpholino)ethanesulfonic acid
NAD + H	Nicotinamide adenine dinucleotide + hydrogen
NEB	New England Biolabs
NHEJ	Non-homologous end joining
NTC	No template control
O/N	Overnight
OD600nm	Optical density at 600 nm
ORF	Open reading frame
PAM	Protospacer-adjacent motif
PBS	phosphate buffered saline
PCR	Polymerase chain reaction
PLG	phase lock gel tube
qPCR	Quantitative polymerase chain reaction
RBS	Ribosome binding site
RCM	Reinforced *Clostridial* media
RRNPP	Rap, Rgg, NprR, PlcR, and PrgX

rt	Reverse transcription
RT	Room temperature
SOC	Superoptimal broth with catabolite repression
STM	Signature-tagged mutagenesis
TBS	Tris-buffered saline
TEM	Transmission electron microscopy
Tm	Thiamphenicol
Tn-seq	Transposon sequencing
TraDIS	Transposon-directed insertion site sequencing
TraSH	Transposon site hybridisation
TRIS	Trisaminomethane
TY	Tryptose, yeast extract media
TYIR	Tryptone, yeast extract, iron sulphate, ammonium sulphate media
UPE	Upstream promoter element
v/v	Volume/volume
w/v	Weight/volume

Chapter I – Introduction

1.1 Clostridium saccharoperbutylacetonicum and solventogenic Clostridia

1.1.1 Biology

1.1.1.1 Overview

Clostridium saccharoperbutylacetonicum is a Gram positive, spore forming, acetone-butanol-ethanol (ABE) producing, obligate anaerobe with a long history of use in industrial biotechnology. Originally isolated from the soil (Hongo, 1960), *C. saccharoperbutylacetonicum* has since undergone multiple re-classifications and annotations, particularly during the 1990s and 2000s, as 16s rRNA and genome sequencing, and phenotypic characterisations became accessible and systematic (Keis et al., 1995, 2001; Poehlein et al., 2014). The species has been extensively studied due to its ability to produce ethanol acetone, and n-butanol, with the latter two being the most industrially targeted. As such, the majority of research conducted in the species has been focused on metabolic analyses, from gene alterations to fermentation, (e.g, Kosaka et al., 2007; Monaghan et al., 2021; Shaheen et al., 2000) or on fermentation engineering (e.g., Oshiro et al., 2010; Tanaka et al., 2012; Zheng et al., 2013). This work has generally shown *C. saccharoperbutylacetonicum* to fit the metabolic archetype of biphasic growth whereby high-energy yielding acidogenesis is used during logarithmic growth before a switch to lower energy yielded, acid-reabsorbing solventogenesis in stationary phase (Jones and Woods, 1986).

1.1.1.2 Metabolism

Under standard fermentation conditions – anaerobic, simple feedstock (e.g. glucose), controlled pH and temperature – *C. saccharoperbutylacetonicum* makes use of two core, overlapping metabolic pathways which are adequately described by the summary of Jones and Woods (Al-Shorgani et al., 2012a; Jones and Woods, 1986; Noguchi et al., 2013; Tashiro et al., 2007). During all growth phases, glucose is converted into pyruvate by glycolysis resulting in a net 2 ATP molecules produced per molecule of glucose (simplified in Figure 1.1). This process also produces reducing power in the form of 2 NADH molecules. If under stress, the cell can remove excess reducing power at this point through the production of lactate, however this does not occur under favourable conditions. Pyruvate is converted to two molecules acetyl-CoA in a manner coupled with the reduction of

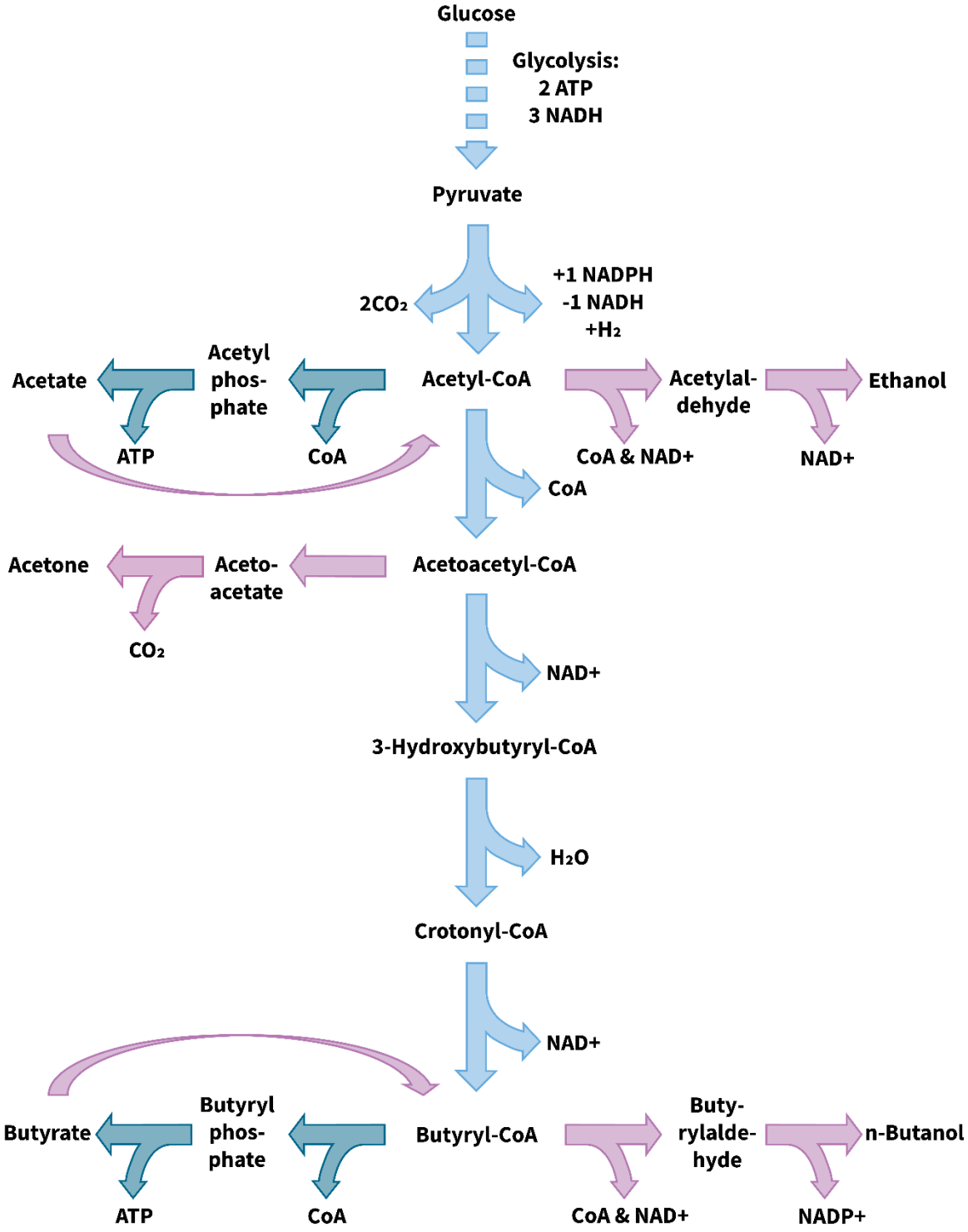

Figure 1.1 The core metabolism of solventogenic *Clostridia*. Glycolysis is shown as attenuated and proceeds as well-documented. Cyan shows pathways that are always active. Teal shows pathways that are highly active during acidogenesis. Mauve highlights pathways that are most highly active during solventogenesis. The reabsorption of acetate and butyrate is conducted by the same enzyme.

ferredoxin and NADP+ resulting in the release of H_2, a step which also releases two molecules of carbon dioxide.

During early growth phases, acetyl-CoA can be converted into any one of three different molecules – acetyl phosphate, acetylaldehyde, and acetoacetyl-CoA. These are the first molecules towards the production of acetate, ethanol, and butyrate respectively. The production of the acids predominates due to the higher ATP yield associated with their production. Either 2 molecules of acetate and 2 molecules of ATP can be produced from the 2 molecules of acetyl-CoA or1 molecule of butyrate, 1 molecule of ATP and 2 molecules of NAD+. Although the production of acetate increases the net yield of ATP significantly (4 net ATP vs 3 net ATP for butyrate), metabolic analyses show that both pathways are used with the overall production of acetate:butyrate being approximately 0.66:1 (Thauer et al., 1977). This is due to the need for the cell to maintain the internal redox balance by consuming the NADH generated during glycolysis. This mixed approach also explains the slight but consistent production of ethanol in early growth phases which also consumes 2 NADH per molecule of ethanol. Overall, the net ATP production during acidogenesis is approximately 3.25 moles per mole of glucose.

Naturally, the large-scale production of acids results in a drop in culture pH to levels that start to become dangerous for the cells. If left unchecked, acids can re-enter the cell and disrupt the proton gradient essential for cell survival. To deal with this, cells can switch to a solventogenic phase where butyrate and acetate are further processed to butanol and acetone. Butyrate and acetate are processed by the same enzyme – acetoacetyl-CoA:acetate/butyrate:CoA transferase – into butyryl-CoA and acetyl-Co-A respectively (Wang et al., 2017). From here, butyryl-CoA is converted in two steps into n-butanol, a process which also consumes reducing power. The fate of acetyl-CoA is more complex, being convertible in three steps to acetone, in two steps to ethanol and in 6 steps to n-butanol. Conversion to acetone requires no reducing power whilst ethanol production converts 2 NADH to 2 NAD+. Together these are likely form a sensitive dynamic able to respond quickly to the redox needs of the cell. Net ATP production drops to 2 moles ATP/mole of glucose during solventogenesis, i.e., only the output of glycolysis.

The solvents produced at this point are initially from reuse of the acids. However, once these have been consumed, they can continue to be produced directly without the need for acidogenesis. Solventogenesis allows the cells to persist for at least 96 h before mass culture death starts to occur. In *Clostridium acetobutylicum* solventogenesis is closely coupled with the initiation of sporulation and therefore seen as part of the species' long term persistence (Ravagnani et al., 2000; Santangelo

IY et al., 1998). Whilst, under the right conditions, this is also the case for *C. saccharoperbutylacetonicum* (Atmadjaja et al., 2019), solventogenesis always occurs whilst sporulation only occurs only under the right carbon-source conditions (personal communication, Sasha Atmadjaja). Whilst the general outline of the core components of acidogenesis and solventogenesis are relatively clear, much additional work is still required if more precise information is to be acquired as to the essential importance of each branch of the pathway and therefore how to anticipate and adapt to changes arising from genetic engineering, new feedstocks and fermentation techniques.

1.1.1.3 Phylogenetics

The focus of this thesis is not on the classification, genomic analysis or general phylogeny of *C. saccharoperbutylacetonicum* N1-4(HMT); the strain used in this study, however, it is useful to outline the relative relatedness of *C. saccharoperbutylacetonicum* to other solventogenic *Clostridia*. This summary is largely based on a thorough study undertaken to understand the genomics and phylogeny of solventogenic *Clostridia* (Poehlein et al., 2017) and two of their most pertinent figures are reproduced here (Figure 1.2).

In total, 44 genome sequences were compared from a variety of solventogenic *Clostridia* species and strains. From these, two broad clades emerged (Figure 1.2A), one containing *C. acetobutylicum*, *C. pasteurianum*, *C. felsineum*, *C. roseum* and *C. aurantibutyricum* and the other *C. puniceum*, *C. saccharobutylicum*, *C. beijerinckii*, and *C. saccharoperbutylacetonicum*. These differences were well represented by the distinctions in the solvent producing *sol* operon which are structured differently between the two clades. In *C. saccharoperbutylacetonicum* the operon structure is *bld-ctfA-ctfB-adc* which together encode three proteins key to the solventogenesis: the butyrylaldehyde dehydrogenase responsible for the conversion of butryl-CoA to butyrylaldehyde; the acetoacetyl-CoA:acetate/butyrate:CoA transferase key to the reabsorption of acids (encoded by *ctfA* and *ctfB*); and acetoacetate decarboxylase responsible for the final step of acetone production.

Bioinformatics confirmed the experimental evidence that *C. saccharoperbutylacetonicum* N1-4(HMT) is able to utilise a range of 5- and 6-carbon sugars for core metabolism including xylose, glucose and starch, sucrose, mannose and glycerol. Finally, this detailed analysis confirmed the experimental evidence suggesting that *C. saccharoperbutylacetonicum* contains all the genes necessary to produce the classical solventogenic products – acetate, butyrate, lactate, acetone,

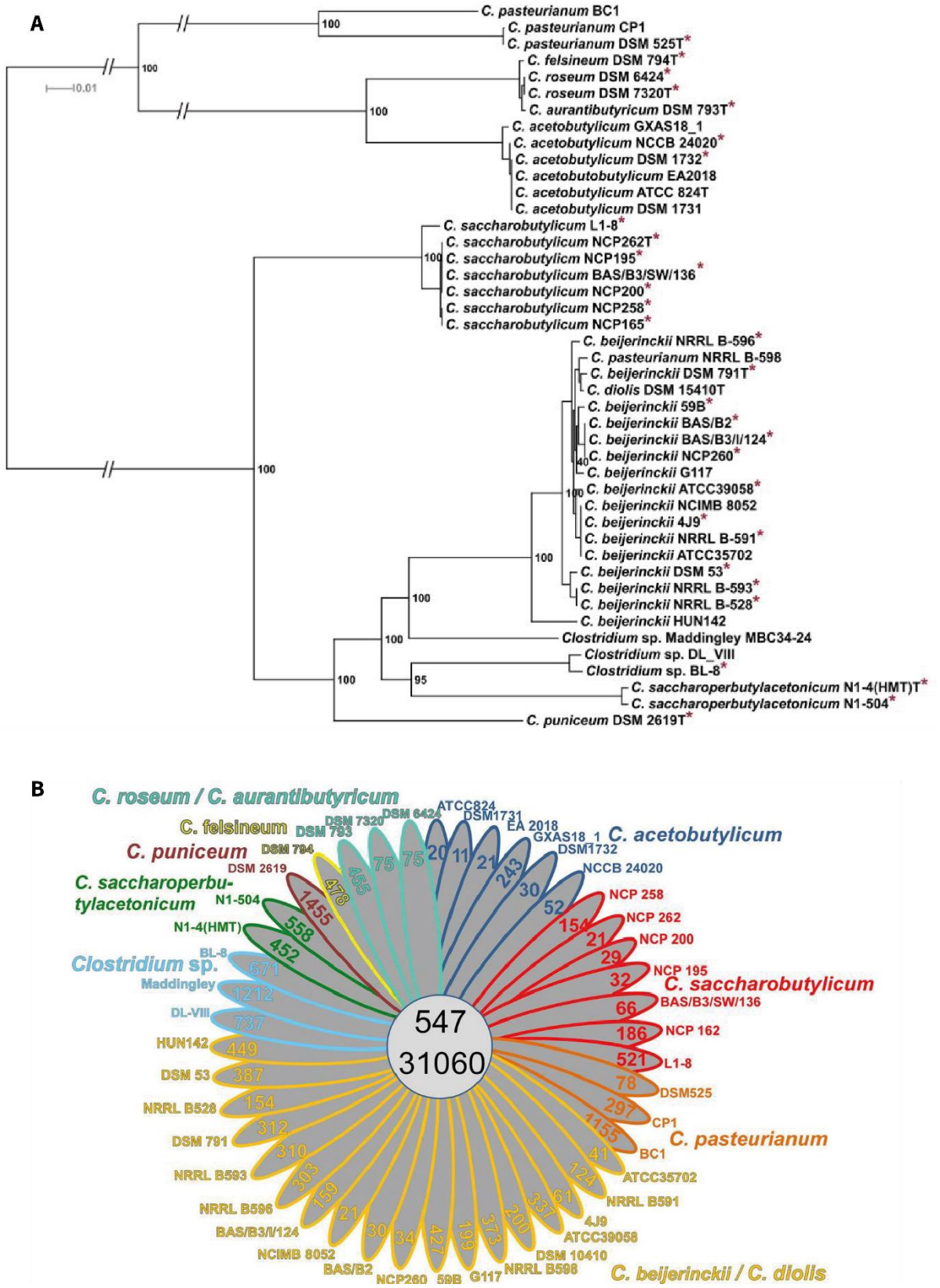

Figure 1.2 The phylogeny of *C. saccharoperbutylacetonicum* in comparison to other solventogenic *Clostridia*. A) Multilocus sequence analysis of 44 solventogenic *Clostridial*. The red asterisks denote genomes that were sequence by Poehlein, et al., The lines represent the

degree of divergence whilst the numbers indicate the likelihood (0-100) that that node would be generated by this data. *C. saccharoperbutylacetonicum* is located at the bottom right, grouped in a clade with *C. beijerinckii*. **B)** Illustration of the shared genome of the examined solventogenic *Clostridia*. The centre circle, listing two numbers, represents the core genome shared by all species and strains of 547 genes and the overall number of different genes between all species and strains. Each coloured petal indicates the number of genes unique to the listed species/strains. *C. saccharoperbutylacetonicum* is shown in green at approximately 11 o'clock. N1-4(HMT), the strain used in this study, contains 452 unique genes. This figure is a concatenation of Figure 2 (the pan genome) and Figure 3 (the MLSA tree) taken from Poehlein et al., 2017 and utilised under the terms of the Creative Commons Attribution 4.0 International License (**https://creativecommons.org/licenses/by/4.0/)**

butanol and ethanol – but does not contain the pathways necessary to produce common additional solvents such as isopropanol, 1,3-propanediol, 2,3-butanediol and isopropanol.

Despite varied and significant differences in the genomes – totalling 31,060 genes across all strains, a core genome of 547 genes shared by all could be determined (Figure 1.2B). *C. saccharoperbutylacetonicum* housed the largest genome of all strains analysed at 6.66 Mbp and the most genes with 5,937. The smallest genome was the 4.1 Mbp of *C. acetobutylicum* NCCB 24020 with approximately 4,000 genes. This neatly highlights the lack of any correlation between gene number (and genome size) and efficiency of growth, with *C. saccharoperbutylacetonicum* being the superior choice for industrial fermentation.

Further emphasising this point is the presence and absence of endogenous plasmids and megaplasmids amongst these strains. 13 of the 44 strains contained plasmids of some kind. Sizes vary massively from 192,000 bp of pSOL1, the *sol* operon containing megaplasmid of *C. acetobutylicum*, to the 2,936 bp of pNAK1 found in *C. saccharoperbutylacetonicum* N1-504 a sister strain of that used in this study. The 136,188 bp megaplasmid found in *C. saccharoperbutylacetonicum* N1-4(HMT) (the strain used in this study) is amongst the largest, yet surprisingly contains no obvious metabolic or defensive genes that might justify its maintenance. In a study where the megaplasmid was putatively cured from the strain, the cured strain showed slightly higher transformation efficiencies (Gu et al., 2019), suggesting that the megaplasmid may play a role in the defence against invading DNA, though how this occurs remains a mystery.

1.1.1.4 Fermentation

Fermentation of solventogenic *Clostridia* is typically conducted using one of three methods – batch, fed-batch and continuous. In batch fermentation, all feedstocks are already present at the time of inoculation and the culture is allowed to grow until the maximum amount of solvents – or other end products – are reached. In fed-batch, products are, again, left *in situ* until the end of the fermentation run, however feedstocks are added over the course of the fermentation run. The purpose of this is to avoid the substrate inhibition that can occur with some feedstocks, to reduce the effects of catabolite repression in mixed feedstocks and to aid in the post-fermentation processing by maintaining reasonable viscosity and water levels (Yamanè and Shimizu, 1984). The two forms of batch fermentation result in the separation of microbial growth and product purification. This means that product purification is typically easier but that extensive cleaning and sterilisation need to occur between batches in order to maintain pure cultures and consistent

batches. Finally, in continuous fermentation, feedstocks are continually added whilst products are continually removed. This has the advantage of allowing long fermentation processes and can help maintain cells in logarithmic growth (if desired), however this typically results in lower overall yields which must be balanced carefully with the longer fermentation run. In terms of solventogenic *Clostridia*, continuous fermentation needs to be designed to ensure the cells are maintained in stationary, solventogenic phase, however it does present the advantage of maintaining solvents, especially butanol, below toxic levels.

Since the inception of biobutanol production (Weizmann, 1919), extensive research has been conducted aimed at finding ideal batch and fed-batch conditions for *Clostridia*. Typically, fermentation studies examine several aspects of the fermentation process from equipment and environment, to feedstocks, culture regulation and removal of products (e.g., Liao et al., 2017, 2018; Monaghan et al., 2021b; Oshiro et al., 2010; Shaheen et al., 2000; Zheng et al., 2013). Industrial fermentation – in the production of biobutanol and bioethanol – usually uses feedstocks, such as maize or sugar cane derivatives, that can be described as competing with food production. This has hampered the case for these first generation biofuels (biofuels generated from edible biomass) and has led to strict criteria for the production of biofuels in the European Union (EASAC, 2012; European Parliament and Council, 2018). It is no surprise, therefore, that many studies focus on the improvement of fermentation using waste feedstocks (e.g., Al-Shorgani et al., 2011, 2012b, 2014; Ellis et al., 2012; Farmanbordar et al., 2018).

These waste feedstocks revolve around the potential of lignocellulose which is the core component of plant dry matter and forms a significant part of the agricultural, and other, waste. Composed of cellulose, hemicellulose and lignin, these represent huge stores of 5- and 6-carbon sugars. Pre-processing of lignocellulosics to remove lignin and partial breakdown cellulose is essential to allow Clostridia to break down these complex polymers into a usable carbon source (Ezeji et al., 2007; Mitchell et al., 1995). This pre-processing and the variety of inhibitory compounds present in the complex mix present the largest obstacles to their use large-scale industrial applications.

In recent years, fermentation research has expanded to include classical microbiological approaches such as gene deletion, insertion, modification, and overexpression. This has been helped by the development of genome modification techniques such as ClosTron (Heap et al., 2010), CLEAVE™ (Jenkinson and Krabben, 2015) and other CRISPR-based systems for *C. saccharoperbutylacetonicum* and other solventogenic *Clostridia* (Gu et al., 2019; Li et al., 2016; Wang et al., 2017) (reviewed in 1.2.4). These studies are diverse, from investigating the function of the

master regulator Spo0A (Atmadjaja et al., 2019), to the role of specific gene clusters (Kosaka et al., 2007; Nakayama et al., 2008), and from the role of autolysins in *C. saccharoperbutylacetonicum* (Jiménez-Bonilla et al., 2021), to the effect of the deletion of key metabolic genes (Monaghan, 2019; Monaghan et al., 2021a). The insights garnered from these studies are invaluable, however the reverse genetics approach is vulnerable to missing the role of genes that are not obviously annotated or located in stand-out operons or clusters.

The above constitutes the briefest outline of the current state of fermentation and feedstocks in the use of solventogenic *Clostridia* at an industrial scale. There remain multiple important obstacles to overcome as well as facets of fermentation that remain poorly understood. Particularly, the specific pressures of fermentation conditions, the nature of action of inhibitory compounds and the reliable triggers of solventogenesis remain to be clarified, particularly in *C. saccharoperbutylacetonicum*. Whilst transcriptomics has helped shed some light on these issues (Li et al., 2020), additional methods and approaches are necessary to further understand the underlying metabolomics and genomics underpin the key obstacles. One of the goals of this thesis is to apply transposon-directed insertion-site sequencing (reviewed in 1.6) to investigate the pressures of fermentation conditions as well as potentially examine the mode of impact of inhibitory compounds on *C. saccharoperbutylacetonicum*.

1.2 Genetic tools in solventogenic *Clostridia*

Despite their importance to human health, food safety and industrial applications, the genetic and molecular tools available for use in *Clostridia* has long lagged-behind that found in other important and model organisms. In solventogenic *Clostridia*, the situation was worse given the wide range of potentially useful species, each with their own biology to contend with. A particular obstacle to overcome was DNA transformation and native restriction systems (Gyulev et al., 2018). The latter, which degrade incoming foreign DNA that does not contain the correct methylation pattern, has first required the characterisation of the systems and the expression of exogenous methylases in *Escherichia coli* in order to make DNA permissive for transformation (Grosse-Honebrink et al., 2017; Lesiak et al., 2014; Mermelstein et al., 1993; Pyne et al., 2013). *C. acetobutylicum* has become one of the model organisms in the field, due, in part, to the early success with transformation (Mermelstein et al., 1993) and to being the first *Clostridial* species to have the complete genome sequenced (Nölling et al., 2001). In addition to fundamentals such as transformation, the challenges of

developing genome editing techniques has also hampered the field, though this is beginning to change with the development of CRISPR-Cas based approaches (see 1.2.4).

Non-coding genetic features such as reliable promoters, ribosome binding sites, replication origins and terminators are all at different stages of understanding and often created on an *ad hoc* basis and with functionality varying greatly between application in different species. All these issues can be adequately summarised by issues in four key areas: biological tractability; the huge differences in the biology of the different species utilised; *ad hoc* development; and a significant tranche of research being conducted by industry. For the latter issue, leaders in the field such as Green Biologics/Biocleave do generally share their positive results either formally (e.g., Atmadjaja et al., 2019) or informally through personal communication. However, a comprehensive understanding of experimental lines conducted is not always clear (though this is not an issue limited to industry). Notwithstanding this, companies such as GBL/Biocleave have been enormous drivers of research in the field, both directly and by providing a *raison d'etre* for much independent research. Indeed, advances other the last 10 years have been such that *Clostridial* toolkits have been expanded greatly and the field no longer lags as far behind others at it used to. Here, some of the key advancements are summarised and some gaps highlighted.

1.2.1 Shuttle vectors

Plasmid vectors are a universal tool of molecular biology being essential to numerous techniques fundamental to microbiology. The development of *E. coli* strains with extremely high and standardised transformation rates has prevented transformation, and therefore DNA proliferation, being a limiting factor in the manipulation of DNA and allowed the creation of a plethora of complex and low yield techniques that have revolutionised molecular biology. As for other species and genera, *Clostridial* research also makes use of *E. coli* as the chassis for the selection, proliferation, and long-term storage of plasmids. Shuttle vectors, plasmids capable of being transferred and replicated in both *E. coli* and the target species, are a cornerstone of this system. The development of the modular pMTL80000 series of shuttle vectors (Heap et al., 2009) has been very helpful to the field.

This study also employs the pMTL8000 system, though does not do so exclusively, so the system is briefly outlined here. The systems four modules: a Gram positive replicon, a selection marker, a Gram negative replicon and an application specific module. The last in this study was typically the most altered and the original plasmids used likely made use of the multiple cloning sites or spacer

regions which have subsequently been filled and altered based on previous needs. The genome editing technique used in this study, CLEAVE™, also employs the pMTL8000 series as the basis for the homologous recombination and targeting vectors (Atmadjaja et al., 2019; Jenkinson and Krabben, 2015). Vectors in this study made use of the *C. dfficile* replication origin pCD6 (Purdy et al., 2002), the pBP1 (Davis, 1999)and pCB102 (Minton and Morris, 1981). Either *catP* or *ermB*, conferring resistance to chloramphenicol/thiamphenicol and erythromycin respectively, were used as selection markers (Heap et al., 2007). Finally ColE1 or ColE1+tra were used as *E. coli* origins of replication (Chambers et al., 1988).

1.2.2 Genetic features

1.2.3 Promoters

The transcription of genes by RNA polymerase is initiated promoters, sequences recognised by sigma factors that bind to the transcriptional machinery including RNA polymerase and recruit it to DNA (Browning and Busby, 2004; Helmann and Chamberlin, 2003). The key sequences for σ70-based promoters are the sigma factor binding sites defined by the -10 region known as the Pribnow sequence or TATAAT box with an upstream -35 region with a spacer in between. These core components appear conserved in *Clostridia* (Sauer et al., 1995) with promoters calculated from prediction software such as BPROM (Solovyev, 2011) typically experimentally active. Indeed, the *Firmicutes* phylum generally contains promoters that closely match the consensus (Sinoquet et al., 2008). As our understanding of bacterial promoters has increased so has the complexity of bacterial promoters been revealed. Upstream promoter elements, sometimes hundreds of base pairs upstream of the -10 and -35 regions, are known to play important roles (Estrem et al., 1998). Repression and de-repression of promoters is important to their regulation which can be achieved through the binding of a repressor (Rojo, 2001) or even by alteration of the genome sequence itself (e.g., for *C. difficile*: Anjuwon-Foster and Tamayo, 2017; Emerson et al., 2009; Sekulovic et al., 2018).

Having a range of promoters with different expression levels and, potentially, repression mechanisms, is invaluable for a plethora of experimental applications. However, to create these panels of promoters, their relative expression levels must first be assessed. Typically, assessing expression levels can be conducted in three main ways: mRNA levels (through reverse transcription quantitative PCR, microarrays and RNA-seq, Lowe et al., 2017); fluorescent reporters (e.g. for anaerobes: Ransom et al., 2015, and in general: Shaner et al., 2005); and enzymatic reporters (e.g. in *C. acetobutylicum* Girbal et al., 2003; Tummala et al., 1999). mRNA quantification is by far the most

accurate method of assessing activity of the promoter, however these methods are typically laborious when focused on handful of promoters (as opposed to the whole genome/transcriptome) requiring the growth of cells, extraction of mRNA, conversion to cDNA and then assessment by RTqPCR, sequencing or microarray. These techniques, therefore, work best in assessing the transcriptome of a given species rather than the activity of a given, potentially exogenous, promoter.

Fluorescence and enzymatic methods both utilise the transcription and translation of a fluorescent/enzyme encoding gene as a proxy for expression. These proteins are very often exogenous, meaning they are not appropriate to all species. For example, common fluorescent reporters require oxygen to mature and so are inappropriate for measuring gene expression levels in anaerobes, leading to the creation of oxygen-independent fluorescent reporters (Drepper et al., 2007). Enzymatic reporters make use of the expression of enzymes that catalyse a reaction in an easily quantifiable manner, the rate of which is altered in a predictable way by the quantity of enzyme available. The most common of these have been the carbohydrate hydrolases, such as β-glucoronidase, which catalyses the breakdown 4-MU-β- glucuronide into 4-methylumbelliferone (4-MU) and β-glucuronide (Moberg, 1985). 4-MU emits light at 460 nm when excited by 365 nm wavelengths meaning it's presence can be detected by fluorimetry. This has been successfully applied in *C. acetobutylicum* (Tummala et al., 1999). Recently, luciferase, which catalyses the breakdown of luciferin resulting in the release of light, has been adapted for use as a secreted reporter in *C. difficile* (Oliveira Paiva et al., 2016).

Both enzymatic and fluorescent reporters suffer from being a measure of multiple factors: transcription levels, translational efficiency and the maturity/functionality of the protein/enzyme. They are not as sensitive as mRNA-based techniques, however, they are typically quicker and easier to use. In addition, their disadvantages are mostly an issue for comparison between work conducted in different laboratories and species at different times. For comparison between a suite of promoters conducted in the same manner at the same time, these techniques provide a quick and easy way to compare relative expression levels. Therefore, studies that examine a range of different promoters and their expression levels can add to the pool of promoters available which are capable of driving a range of expression levels depending on need.

1.2.4 Genome editing

1.2.4.1 ClosTron and homologous recombination

The ability to directly and specifically alter the genome of *Clostridia* has undergone incredible progress in the last 15 years. Very broadly, two main approaches have been taken to alter genomes in a directed manner in bacteria: targeted disruption by insertion (e.g. Karberg et al., 2001) and variations on homologous recombination (e.g. Datsenko and Wanner, 2000). It is natural, therefore, that analogous techniques should be developed in *Clostridia*. The first of these to prove adaptable and effective in multiple *Clostridia* has been ClosTron (Heap et al., 2007). Briefly, ClosTron exploits the ability of group II introns from *Lactococcus lactis* which can be targeted to insert themselves and a selectable marker into a gene. ClosTron was of enormous help to the field as it resulted in the stable inactivation of genes by insertion allow for their disruption and phenotyping. This technique was necessary due to the poor efficiency of previous attempts institute homologous recombination techniques. In addition, it worked in all tested *Clostridia*, requiring only transformation protocols to function.

Homologous recombination is a broad term encompassing the methods by which DNA can recombine with other DNA containing highly similar sequences. Variations on this process are found universally across all three domains and viruses. Typically, this process is used to create genome alterations by creating vectors with two regions that flank the region of interest. The region of interest will contain a modification: a deletion where little of the original region is present: a small change such as a single nucleotide polymorphism; or an insertion of additional DNA. The vector is then transferred into the target species where two recombination events can occur. Firstly, the entire vector is inserted into the region of interest thanks to the regions homologous to those in the genome. Then, one of two processes can occur: 1) the entire vector is re-excised by the inverse mechanism or 2) the vector and the original region of interest are excised instead leaving the mutant region.

This process is extremely complex and happens at very different rates in different bacteria. In *Clostridia* problems are due to low rates of the second recombination event, leaving single crossover insertion mutants with variable stability (e.g. O'Connor et al., 2006) meaning that, at best, insertional inactivation was possible but not the clean deletions, modifications and insertions possible for other bacteria. Hence the success of ClosTron in being able to reliably and easily make stable insertional inactivations. However, ClosTron comes with numerous downsides meaning the search for effective homologous recombination techniques continued. Firstly, it requires that a selection marker to be maintained on genome meaning that marker can no longer be used for

further ClosTron insertions or to select for future plasmid transfer. This limited construction of layered mutants to the number of markers available for the species. Secondly, there remains the possibility of polar effects of insertion whereby neighbouring genes are disrupted. Thirdly, precise modifications to examine specific domains or individual amino acids are not possible. Finally, complementation of the inactivated gene has to be conducted with a copy of the gene on a plasmid resulting in likely incorrect expression levels.

To overcome the limitations, several methods were created which incorporated techniques for selecting for the second recombination event (reviewed in Joseph et al., 2018). These take advantage of uracil metabolism genes (e.g., *upp*, *pyrE* and *pyrF*) whose enzymes will process 5-fluoroorotic acid to 5-fluorouracil which is toxic to the cell. Cells lacking the expression of these enzymes will be uracil auxotrophs (i.e., need media supplemented with uracil) but will be resistant to 5-fluoroorotic acid, being unable to catalyse the formation of 5-fluorouracil. Hence this can form the basis for a powerful selection system independent of antibiotic resistance markers (Croux et al., 2016; Heap et al., 2012; Tripathi et al., 2010). These methods require an pre-engineered strain to function, however, which prompted the development of pseudosuicide vectors (Cartman et al., 2012). In this system, the recombination vector contains the *codA* gene. When expressed, CodA catalyses the conversion of cytosine to uracil and therefore also convert non-toxic 5-fluorocytosine to toxic 5-fluorouracil. After allowing for homologous recombination, cells are plated on media containing 5-fluorocytosine with cells that grow being either wild type reversions or mutants. These can then be readily screened by PCR. These methods have become the *de facto* method of gene modification in *Clostridia* over the last ten years.

1.2.4.2 Homologous recombination incorporating CRISPR-Cas

The characterisation of clustered regularly interspaced short palindromic repeats (CRISPR) and how they are used by CRISPR-associated (Cas) proteins (Barrangou et al., 2007) has proved revolutionary across biology. Providing bacteria and archaea with adaptive immunity (Brouns et al., 2008; Marraffini and Sontheimer, 2008) CRISPR-Cas has been repurposed for an enormous range of techniques able modify genomes and transcriptomes across all domains of life (Adli, 2018). In the simplest terms, challenge of the cell by foreign DNA results in the creation of spacer sequences that are inserted into the CRISPR locus by Cas1 and 2. This locus consists of a leader sequence, followed by direct repeats that flank spacers. This locus is transcribed and processed into CRISPR RNA (crRNA) which then serves as a guide for another Cas protein to bind target DNA containing the

complementary sequence and destroy it. There are three main types of CRISPR-Cas system classified with each working slightly differently and with many variants between species (Makarova et al., 2011). However, they all achieve the same goal of the destruction of potentially harmful DNA in a targeted manner and the incorporation of foreign DNA sequences into the CRISPR array for future targeting.

The diversity of CRISPR-Cas systems and their potential in a variety of genetic and epigenetic applications is being extensively exploited (Adli, 2018; Lee and Lee, 2021; Vigouroux and Bikard, 2020). Many bacteria, including *Clostridia*, have either non-existent or particularly inefficient methods of non-homologous end joining (NHEJ) (Cui and Bikard, 2016) meaning that double-stranded breaks can only be repaired with homologous recombination. When all sequences are targeted simultaneously, as with CRISPR-Cas, homologous recombination is no longer possible meaning double stranded breaks are lethal. Naturally, CRISPR-Cas presented itself as a strong, markerless and metabolism-free selection system, particularly in this context. In addition, the system is highly targetable requiring only the presence of a 3 bp protospacer-adjacent motif (PAM) adjacent to the target sequence. This was first implemented based on the *Streptococcus pyogenes* CRISPR-Cas9 system (Jinek et al., 2012).

Most CRISPR-Cas genome editing systems developed in *Clostridia* have utilised the exogenous Cas9 from *S. pyogenes* (Bruder et al., 2016; Huang et al., 2016; Nagaraju et al., 2016; Wang et al., 2015b; Wasels et al., 2017). This a logical approach given the low number of components and well-characterised targeting system. However, this has come with obstacles, particularly the regulation of Cas9 levels, that can reduce the effectiveness of the system (Nagaraju et al., 2016). However, the harnessing of the endogenous CRISPR-Cas system present in *Clostridia* suggests at an even more practical route (Pyne et al., 2016). This approach has its own complications: unknown actions of the Cas proteins, unknown PAM sequences and the need to characterise spacers. However, if utilised correctly, it has the potential to be even more flexible and practical than Cas9 approaches. In *C. saccharoperbutylacetonicum* N1-4(HMT), an endogenous CRISPR-Cas genome editing technique was developed by Green Biologics (Jenkinson and Krabben, 2015), CLEAVE™, and has been shown to be able to make precise, targeted deletions, insertions and SNPs (Atmadjaja et al., 2019).

CLEAVE is a three-step process encompassing homologous recombination and selection by the targeting of wild-type sequences. In step one, the modification plasmid in transformed by electroporation in *C. saccharoperbutylacetonicum* containing the homology arms sandwiching the region of change. Transformants are screened and those containing the plasmid are selected for

passaging in the second step. Cells carrying the plasmid are then transformed in step three with a targeting plasmid containing all the key CRISPR components: leader sequence, direct repeats and the targeting spacer. This selects against wild type cells and results in a high percentage of desired mutants. This process was used in this study in Chapter V.

1.3 Sporulation in *Clostridia*

1.3.1 Background

Sporulation in bacteria is the process by which normal vegetative cells divide asymmetrical to produce metabolically inactive endospores. These endospores are typically dehydrated and surrounded by multiple protective layers. Together, these traits combine to create highly resistant structures that can persist in hostile environments, sometimes for 100s of years (Paul et al., 2019). This dormant, resistant state is essential to survival of the species and to the infective life cycle of pathogens such as *C. difficile* (Deakin et al., 2012). In solventogenic *Clostridia*, an intricate interplay exists between sporulation and industrially-valuable solventogenic phase (Diallo et al., 2021). In some species, including *C. saccharoperbutylacetonicum*, the master sporulation regulator Spo0A is involved in the initiation of both (Atmadjaja et al., 2019). These close ties are likely due to both processes being responses to increasingly hostile environments. Endospores are also a common method of long-term strain storage being both resistant to stressors and able to preserve the strain from degeneration (Jones and Woods, 1986). Therefore, the need to gain a more comprehensive understanding of sporulation in *C. saccharoperbutylacetonicum* is clear.

The *Firmicutes* phylum is home to many endospore-producing species. *Bacillus subtilis*, an aerobic member of the phylum, has long been considered the model organism for sporulation and has extremely well defined transcriptomic and physical sporulation pathways (Tan and Ramamurthi, 2014). However, studies in *Clostridia* have shown sporulation genetics, transcriptomics, proteomics and general endospore morphology to be highly diverse, no doubt with each organism adapting sporulation to meet the challenges of their niche (Al-Hinai et al., 2015). Of the *Clostridia*, *C. difficile*, *C. sporogenes,* and *C. acetobutylicum* account for the majority of studies with *C. acetobutylicum* being the model organism for solventogenic *Clostridia* (Al-Hinai et al., 2015). It is this model that is broadly explained in 1.3.2 and 1.3.3 with elements and observations from other *Clostridia* and *B. subtilis* incorporated.

1.3.2 Process

1.3.2.1 Triggers

Given that sporulation constitutes a major differentiation event for the cell, it is no surprise that initiation of sporulation is both tightly regulated and the product of complex interactions between trigger signals, transcription factors and the proteins that catalyse the process (Beskrovnaya et al., 2021; Diallo et al., 2021). The canonical sporulation trigger is starvation and nutrient deprivation of the cells (Schultz et al., 2009). However, this is very rarely the case for solventogenic *Clostridia* where sporulation occurs even with excess carbon and nitrogen sources (Dürre, 2014). Other factors such as secondary metabolites (Herman et al., 2017; Kuit et al., 2012), quorum sensing signals (Feng et al., 2020; Kotte et al., 2020; Ravagnani et al., 2000; Steiner et al., 2012) and type of carbon source (Basu et al., 2017) appear to have a far greater impact. This suggests that sporulation in solventogenic *Clostridia* is part of a bet-hedging strategy that anticipates rather than responds to hostile environments (Veening et al., 2008).

Little is known about the triggers for sporulation in *C. saccharoperbutylacetonicum*, however, Biocleave has shown that the nature of the carbon source is essential to the likelihood of sporulation (Jenkinson et al., 2019). In addition to this, different spore phenotypes and rates of sporulation have been observed for spores generated on solid media as opposed to liquid media (Personal Communication, Sasha Atmadjaja). Further work is required to fully elucidate why, when and how *C. saccharoperbutylacetonicum* sporulates and why spores may differ based on the media solidity.

1.3.2.2 Granulose accumulation

In most species of solventogenic *Clostridia*, the first visible step of sporulation is granulose accumulation and the formation of the *Clostridial* form by the mother cell (Stage I in Figure 1.3), which likely occurs concurrently with DNA replication in preparation for sporulation (Dürre, 2005). Granulose is a polyglucosan, similar to glycogen, which appears essential to sporulation and likely used for energy storage (Robson et al., 1974). This accumulation is triggered in *C. acetobutylicum* following the expression and activation of Spo0A following the triggering of phosphorylation cascades by orphan histidine kinases (Al-Hinai et al., 2014, 2015; Steiner et al., 2011). Phosphorylated Spo0A then activates and enhances the expression of sporulation-specific sigma factors. These, in turn, regulate the expression of genes necessary for all stages of sporulation (Bi et al., 2011; Jones et al., 2011).

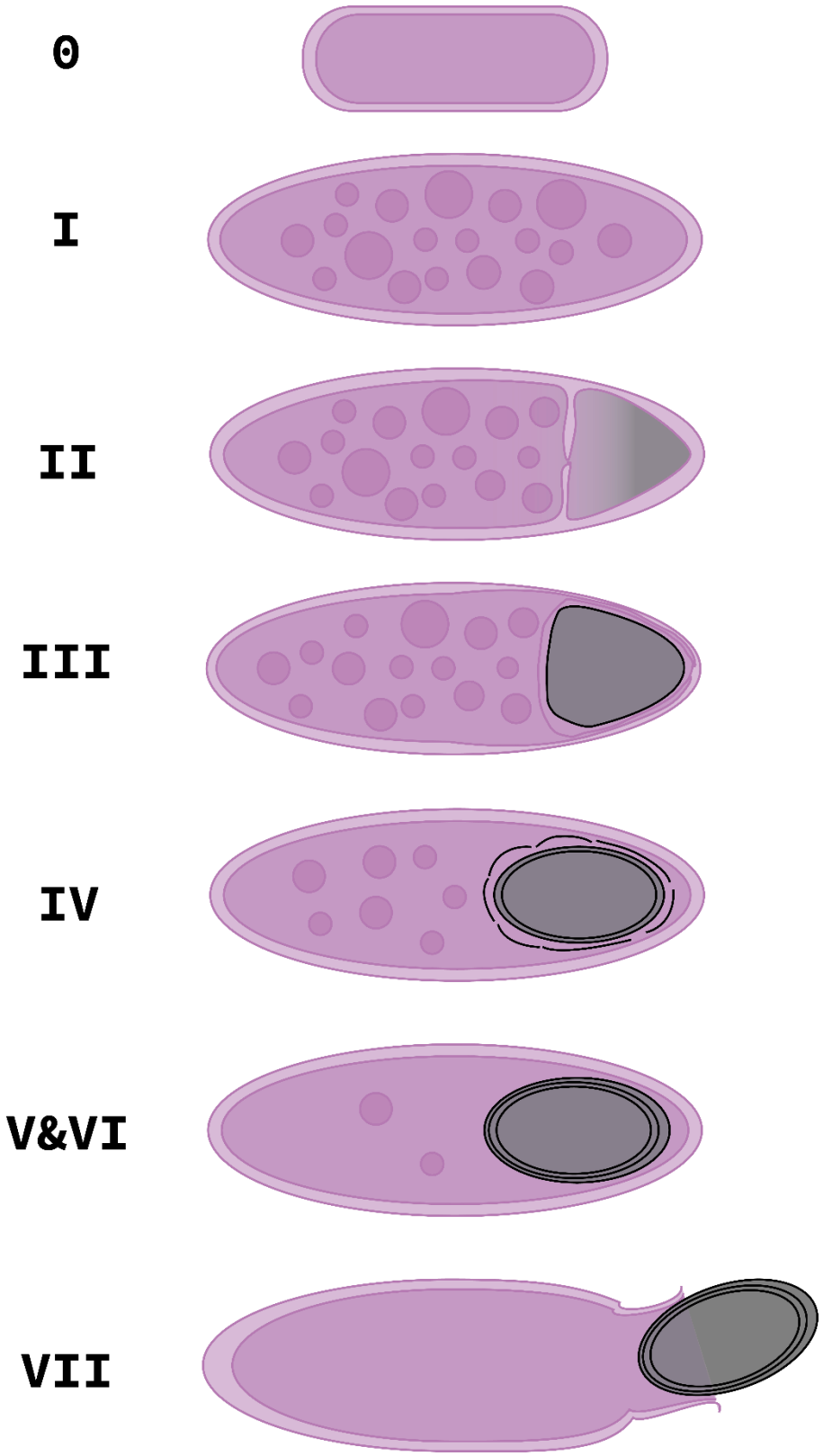

Figure 1.3 Simplified view of the stages of sporulation. Stage 0 is a vegetative cell. Vegetative cells will remain so unless they receive specific sporulation triggers. Stage I shows the development of the cigar-like *Clostridial* form and the accumulation of granulose (small circles).

Stage II depicts asymmetrical division with the developing forespore on the right. Stage III shows engulfment of the forespore (black) by the mother cell (mauve). Stage IV shows the development of the spore and the assembly of the spore coat (black dashed lines). Stage V&VI are spore maturation. During these steps, the energy stored as granulose is gradually used by the mother cell. Finally, Stage VII is the release of the mature spore from the mother cell. The latter dies in the process.

1.3.2.3 Asymmetrical division

Stage II of the visible processes of sporulation is the asymmetrical division of the mother cell (II in Figure 1.3). During this, a septum is formed close to one pole of the mother cell to create the initial forespore. A copy of the chromosome is transferred into the forespore at this point. Though this has not been studied in *C. acetobutylicum*, in *B. subtilis* chromosome transfer is mediated by DNA translocase SpoIIIE (Grainge, 2008).

1.3.2.4 Engulfment

Stage III is the engulfment of the forespore by the mother cell (III in Figure 1.3). The mother cell envelope extends to surround the forespore, creating a second membrane around the engulfed spore. This process is coordinated by the SigE sigma factor in *C. perfringens* (Harry et al., 2009) and *C. difficile* (Fimlaid et al., 2013), but this is not the case in *C. acetobutylicum* where SigE coordinates the sporulation process prior to asymmetrical division (Tracy et al., 2011). Generally, engulfment appears to be the step at which the mother cell and forespore are committed to the sporulation process (Al-Hinai et al., 2015).

1.3.2.5 Maturation

The development from double-membraned forespore to mature spore encompasses three key stages. In Stage IV, the cortex is formed between the two membranes and on top of the primordial cell wall (IV in Figure 1.3). This region, which typically appears white on transmission electron microscopy (TEM), is formed of a modified peptidoglycan (Al-Hinai et al., 2015), though the *C. acetobutylicum* cortex has not been characterised. During Stage V, the spore coat begins to be assembled, a process that has been characterised in the *Clostridia* pathogens (Shen et al., 2019) but practically unstudied in the solventogenic species (Diallo et al., 2021) (V in Figure 1.3). However, TEM suggested that assembly occurs in stages with parts of the coat(s) visible before a single, uniform structure is created (Al-Hinai et al., 2015). During these two stages, the spore core is also developing. The core contains the genome and the proteins necessary for the growth of the cell post-germination (e.g., ribosomes). During its maturation, dipicolinic acid (DPA) is produced in the mother cell and transported into the forespore core where it binds Ca^{2+} and displaces H_2O in the core (Piggot and Hilbert, 2004). The dehydration of the spore core and replacement by Ca^{2+}-DPA and small-acid soluble proteins is essential to the spore resistance to heat (Jamroskovic et al., 2016;

Piggot and Hilbert, 2004) Finally, Stage VI constitutes the final maturation step where the coat is fully assembled, the core tightly packed and the cortex expands to its fullest extent.

1.3.2.6 Spore release

Constituting Stage VII, spore release is seen as the final step in the development of the spore (VII in Figure 1.3). The mother cell lyses, releasing the spore into the environment. In the solventogenic *Clostridia*, the process by which this occurs is not well-understood but appears to require certain autolysins (Liu et al., 2015). It is worth noting that spore maturation appears to continue in *B. subtilis* post-release with isolated spores being less resistant to heat and UV than those isolated from later timepoints (Henriques and Moran, 2007; Sanchez-Salas et al., 2011). This suggests that spore release is not truly the final Stage and that further maturation still occurs.

1.3.3 Spore Ultrastructure and function

As this study intends to examine *C. saccharoperbutylacetonicum* spores under thin-section TEM, a brief overview of spore ultrastructure is necessary to understand how these structures are typically visualised (Figure 1.4). The *C. acetobutylicum* spore will be used as an exemplar, though spore ultrastructure can vary widely between species and even strains (Berezina et al., 2012; Diallo et al., 2021; Jamroskovic et al., 2016).

1.3.3.1 Core

The spore core contains the components necessary for the cell to resume a vegetative life cycle. In all *Clostridial* spores, the core appears under TEM as either dark and dense or as completely white due to lack of stain penetration (examples in: Tracy et al., 2008). This is due to the high density and hydrophobicity of the core. In *C. acetobutylicum*, the core shape is oval much the shape of the overall spore and constitutes a significant percentage of the overall volume of the spore (Long et al., 1983; Tracy et al., 2008).

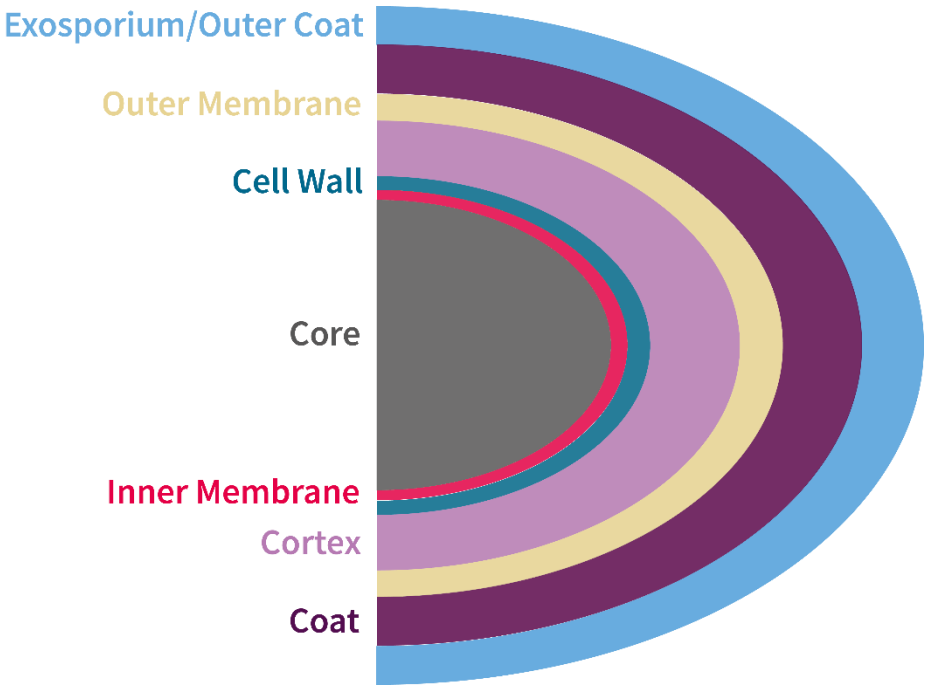

Figure 1.4 Simplified spore ultrastructure. The typical features present in *Clostridial* spores. The core, containing DNA and vital enzymes is metabolically inactive and surrounded by multiple protective layers. There is often an interspace between the exosporium/outer coat layer and spore coat of varying size. Not all features will be present in all spores and the image is not to scale

1.3.3.2 Inner membrane and cell wall

The inner membrane and primordial cell wall do not appear to be typically visible under thin-section TEM and are assumed to surround the core closely (Henriques and Moran, 2007). The cell wall forms the basis for the cell wall of the cell post-germination.

1.3.3.3 Cortex and outer membrane

The outer membrane, formed during engulfment, provides the scaffolding for the assembly of the cortex in all species (Henriques and Moran, 2007). The cortex is a broad peptidoglycan layer that is visualised as white on thin-section TEM (Long et al., 1983; Tracy et al., 2008). The constituent peptidoglycan is distinct from that found in the vegetative cell in *B. subtilis* (Atrih et al., 1999). The cortex is key to maintaining the core in a dehydrated state (Henriques and Moran, 2000).

1.3.3.4 Inner coat layers

The spore coat/coats – protein layers between the cortex and external coats – are an integral part of the spore structure, accounting for more than half the spore proteins. They appear as dark layers and laminations under thin-section TEM. Multiple layers are seen which are formed in a genetically distinct manner (Henriques and Moran, 2007). The number, structure and formation appears of these coat layers appears to vary greatly between species (Diallo et al., 2021). However, in all cases, the coat provides the crucial function of filtering and selecting for the chemicals that are able to reach the core. Germinants, for example, are clearly able to permeate, but the coats have been shown to provide vital protection against enzymes and is able to react with dangerous chemicals before the reach the core (Henriques and Moran, 2007).

1.3.3.5 Interspace

The region between the coat layers and the outer layer/exosporium is titled the interspace. This region varies greatly in size and composition even between spores and conditions (Driks, 2003; Westphal et al., 2003). This region has not been extensively studied in general and not at all in solventogenic *Clostridia*.

1.3.3.6 Outer coat layers and exosporium

Naturally, all spores contain an outermost layer. In *C. sporogenes* and *Bacillus spp* this takes the form of an exosporium, a paracrystalline layer sometimes containing hairy extensions (Henriques and Moran, 2007). This layer is important for further protection but also the adhesion of spores. Under thin-section TEM, the layer often appears to resemble a baggy structure, and, under TEM, it will cause refraction of the electrons due to a crystalline arrangement of the constituent proteins. In solventogenic *Clostridia*, a putative exosporium has been assigned (Diallo et al., 2021; Long et al., 1983; Tracy et al., 2008) for *C. acetobutylicum*, though it is unclear if this structure was formed of a paracrystalline layer. In *C. acetobutylicum*, no hairy structures were observed suggesting this was not a feature of solventogenic *Clostridial* spores. If the outer layer is not paracrystalline, it is perhaps more accurate to refer to the outer layer as the outer spore coat until further experimentation can clarify its composition.

1.4 Germination in *Clostridia*

Germination is the process by which spores regain metabolic activity and return to a vegetative state. The process is a crucial component of the life cycle in all spore-forming species providing the means for the organism to rapidly colonise the more favourable environment. In *Clostridia*, the process has largely been studied only in *C. difficile* and *C. botulinum* (and its non-pathogenic model *C. sporogenes*) as the germination process is essential to the pathogenic capabilities of these species (Setlow et al., 2017). Even within these species, multiple crucial steps are yet to be elucidated, particularly understanding how the germ cell wall is remodelled to accommodate the expanding cell, how lipid membrane mobility is recovered and how the core achieves full hydration.

In solventogenic *Clostridia*, the process has barely been studied at all. Germination has been used extensively as part of the fermentation process (Jones, 2001; de Vrije et al., 2013), especially given that spores were used for storage of strains between fermentation runs (Jones and Woods, 1986). This lack of attention is likely due to the ease of germinating solventogenic *Clostridial* spores in standard media. Multiple practical and theoretical questions are yet to be answered concerning the germination of *C. saccharoperbutylacetonicum*, particularly as concerns the compounds that can act as germinants and whether germination can be suppressed. Such questions will only grow in importance as *C. saccharoperbutylacetonicum* and other solventogenic *Clostridia* are increasingly used in biotechnological applications outside of solvent production by companies such as Biocleave and CHAIN.

Summarised here is a broad overview of germination in *Bacillus spp* and pathogenic *Clostridia* to outline the process and, particularly, the triggers.

1.4.1 Germinants

The environmental triggers for germination are known as germinants. These typically, but not exclusively, bind germination receptors found in the inner membrane of the spore. Binding initiates a cascade of signalling that results in the germination processes described in 1.4.2. Usually, germinants take the form of biologically active molecules, particularly, though not limited to amino acids. Minerals and salts also play a role in germination, often as co-germinants. Notwithstanding this, the most common germinant among *Clostridia* is L-alanine (Bhattacharjee et al., 2016), though L-cysteine, L-serine and a range of other amino acids are also common. The exact reason for prevalence of amino acids as germinants is not known. Evidently, amino acids are important biological molecules necessary for protein production, metabolism and cell walls. However, it seems unlikely that, as has been posited, L-alanine is prevalent due to its role in peptidoglycan composition. D-alanine, another key component of peptidoglycan, is an inhibitor of germination in many species (Brunt et al., 2014), suggesting that a narrow focus on immediate germinant utility is an insufficient explanation. Indeed, there are many nutrients and carbon sources for which a convincing argument could be made for their importance to a growing cell.

More likely, the nature of the germinant has been carefully honed by evolution to be closely correlated with the ideal conditions for growth – particularly a nutrient rich, anaerobic environment. The obvious example in support of this hypothesis is that *C. difficile* germinates in response to taurocholate and other bile salts (Sorg and Sonenshein, 2008). Whilst taurocholate can be used by the cell, the primary reason for its germinating activity lies in taurocholate indicating that the spore is in an environment that is nutrient rich, anaerobic and specific to the niche *C. difficile* has adapted to inhabit i.e., the gastrointestinal tract. If this concept is correct, it means that accurately predicting potential germinants is complex. Alternatively, the prevalence of L-alanine might be due to elevated importance of L-alanine in the early biological world (Kubyshkin and Budisa, 2019).

Interestingly, there are non-physiological compounds such as CaDPA, dodeclyamine (a surfactant) and even non-chemical triggers such as increased pressure on the cells that can induce germination (Setlow et al., 2017; Velásquez et al., 2014). However, these effects appear to be artifacts: CaDPA release from neighbouring spores is unlikely to be at high enough concentrations in biological

context; dodecylamine isn't abundant in nature and bypasses the usual routes to trigger sporulation (Vepachedu and Setlow, 2007); whilst high pressures (particularly osmotic) could plausibly play a biological role, it seems somewhat unlikely that this kind of pressure would represent a welcoming environment for a spore to germinate (Setlow et al., 2017).

In solventogenic *Clostridia*, few germinants have been recorded with only *C. roseum* and butyric acid producer *C. butyricum* having been studied (Bhattacharjee et al., 2016), though the original papers were from 1956 and 1973 respectively and this author was unable to recover the original manuscript to confirm. Assuming the secondary source is correct, L-alanine, L-arginine and L-phenylalanine were reported for *C. roseum* whilst L-cysteine, glucose and sodium bicarbonate was reported for *C. butyricum*.

1.4.2 Process

Even in the most-studied species – *Bacillus spp*, *C. difficile*, *C. perfringens* and *C. sporogenes* – the biochemical processes underlying germination are not well understood (Setlow et al., 2017). Biologically and physically, the following steps must occur: trigger of germination, release of CaDPA from the core, breakdown of the cortex, rehydration of the core and subsequent inner membrane and germ cell wall expansion and remodelling. However, even the order of these events is not the same between these relatively few species. For example, *C. difficile* hydrolyses the cortex prior to releasing CaDPA (Francis et al., 2015) whilst *Bacillus spp* follow the reverse order (Setlow, 2014; Setlow et al., 2017).

Germination receptors, found in the inner membrane, initiate the process upon binding to their cognate germinant, though this could be direct or via a signalling cascade. Even the presence and involvement of germination receptors is not essential to the process. *C. difficile* lacks any of the germination receptors found in *Bacillus spp* and instead uses Csp proteases present in the spore coat layers to activate the necessary spore cortex lytic enzymes (Bhattacharjee et al., 2016; Paredes-Sabja et al., 2014). *C. acetobutylicum* does appear to contain analogues to the germination receptors, though their function and purpose remains unknown (Bhattacharjee et al., 2016).

How the activation of germination receptors leads to the activation of spore cortex lytic enzymes and the release of CaDPA is not well understood. It is thought that in *B. subtilis* CaDPA is released by SpoVA, the same protein that is responsible for its accumulation (Li et al., 2012; Vepachedu and Setlow, 2007). Following CaDPA release and cortex hydrolysis, hydration of core, remodelling of the

cell wall and a reorganisation of the inner membrane to increase membrane fluidity occurs. Interestingly, this process requires no energy in the form of ATP (Magge et al., 2009). Indeed, it has been speculated that the release of CaDPA might even generate energy (Setlow et al., 2017).

Within a population of spores, not all spores will immediately sporulate in the presence of the germinant (Wang et al., 2011, 2015a). This seems to be modulated by the level of germination receptors available and, to some extent, random chance, and must confer the advantage of bet-hedging in case of a subsequent change of environment (Setlow et al., 2017). In industry, a homogenous response to germination is likely desirable so uncovering these steps and finding methods through which a greater proportion of spores can be made to germinate has potential application.

1.5 Quorum Sensing

Whilst it can be useful to think of bacteria as single, isolated clonal units, the biology is, naturally, far more complex. Quorum sensing is the process by which a population of bacteria is able to sense and respond to their overall density. Originally discovered as the method through which bioluminescence is mediated in *Vibrio fischeri* (Engebrecht and Silverman, 1984; Nealson et al., 1970), quorum sensing mechanisms have subsequently been identified in all investigated bacteria, including solventogenic *Clostridia* (Feng et al., 2020; Kotte et al., 2020; Piatek et al., 2022; Steiner et al., 2012; Verbeke et al., 2017). Our understanding of the importance of quorum sensing to bacterial life cycles has only increased in recent years, with the disruption of quorum sensing now seen as a potential avenue for treating bacterial infections (reviewed in e.g., Gray et al., 2013; Piewngam et al., 2020; Singh et al., 2016). As more research is conducted in the field, the complexity and diversity of systems both within a single species and across the kingdom becomes increasingly clear (Weiland-Bräuer, 2021).

Broadly, quorum sensing systems are constituted of signals which are secreted into the surrounding environment by each bacterium in the population. These are then sensed by bacterial cells which respond accordingly, depending on the concentration of the signal and often with a signal concentration threshold below which no response occurs. Signal detection results in the adjustment of transcription, often with far-reaching consequences. In Gram positive bacteria, all quorum sensing is mediated by signalling peptides, as opposed to the acyl-homoserine lactones seen in Gram negatives (Bassler and Losick, 2006). Summarised here is an overview of the Gram positive quorum sensing systems as they have been discovered to function in solventogenic

Clostridia, with some references to model organisms such as *Staphylococcus aureus* and *B. subtilis* as necessary.

1.5.1 RRNPP

1.5.1.1 Overview

Named after the protein families that are the key regulators of the system – Rap, Rgg, NprR, PlcR, and PrgX – RRNPP quorum sensing is common, though not ubiquitous, amongst *Firmicutes* (Neiditch et al., 2017). The system requires the production of a signalling peptide which is then cleaved and secreted from the cell by a transmembrane protease. The peptide, now present in the environment, must enter cells through a transmembrane oligopeptide permease (Perego et al., 1991). Once in the cell in this processed form, the oligopeptide is bound by one of the RRNPP receptor proteins. The binding site, defined by helix-turn-helix repeats that create a concave pocket, is strikingly conserved across all RRNPPs and allowed the identification of these peptides in multiple species (Declerck et al., 2007; Zeytuni and Zarivach, 2012).

Despite the structural homology of the oligopeptide binding domain, DNA sequence homology is extremely low between RRNPPs which likely reflects the diverse uses of these proteins across different species (Neiditch et al., 2017). The binding of the peptide typically results in the recognition of DNA target sequences by the RRNPP protein with diverse results. Examples of downstream effects are: the activation of virulence genes (Declerck et al., 2007; Grenha et al., 2013); the repression and activation of virulence and cell surface genes (Cook and Federle, 2014; Fontaine et al., 2015); the inhibition or induction of conjugation based on a two signal system (Cook and Federle, 2014); the activation of necrotrophic genes or the direct binding of sporulation proteins depending on the signal (Pomerantsev et al., 2009). In some cases, the RRNPP proteins initiate a signalling cascade with other proteins that then conduct transcriptional regulation (Perego et al., 1991; Perez-Pascual et al., 2016). As most species contain multiple RRNPP proteins, several of these systems can be present simultaneously controlling different processes (Neiditch et al., 2017).

1.5.1.2 In C. saccharoperbutylacetonicum

Two recent studies have investigated the presence and role of the RRNPP system in solventogenic Clostridia. One probed the system in *C. acetobutylicum* (Kotte et al., 2020) and the other in *C. saccharoperbutylacetonicum* (Feng et al., 2020). Feng et al identified five RRNPP receptor proteins in *C. saccharoperbutylacetonicum*. They were able to delete four of these using a CRISPR-Cas system

(Wang et al., 2017) whilst a fifth could only be knocked down, suggesting it may be essential. They found that all five RRNPP proteins were important in the transition from acidogenesis to solventogenesis. Growth rates of mutants were normal during the first 24 h followed by a sharp arrest of further growth beyond this point. Only low amounts of solvents were produced with acid levels being significantly higher than wild type for the duration of the growth period. Two of the five RRNPPs were directly involved in sporulation, showing reduced sporulation efficiency when deleted. All five deletions/knockdown had a significant impact on cell motility, increasing it substantially. *C. saccharoperbutylacetonicum* cells are motile during acidogenesis and lose that motility in solventogenesis. Motility assays conducted on solid media suggested that the wild type cells entered solventogenesis on the solid agar whilst the deletion mutants did not and continued to migrate. This appears to confirm that cells still undergo the two phases at roughly the same time, even in this more permissive environment.

Feng et al were unable to adequately complement their mutants using plasmid-based complementation, restoring only some phenotypes in some mutants. Whilst they did use the native promoters, they assumed that the copy number of the plasmid impacted overall expression levels. Overall, they were able to demonstrate that these RRNPP gene products are important in managing the transition to solventogenic phenotypes. Interestingly, their results also showed that sporulation, whilst interlinked with solventogenesis, is not lock-step with it in *C. saccharoperbutylacetonicum*. Clearly, quorum sensing has been overlooked as a key mechanism in fermentation and could well provide important avenues towards the precise control of growth phases in solventogenic *Clostridia*. The study also demonstrates the enormous complexity of interactions involved in quorum sensing. Despite the drastic phenotypes, there was no way to know through which mechanism(s) these proteins altered transcription.

1.5.2 Accessory gene regulator system

The accessory gene regulator (agr) system is a second quorum sensing system found across Gram positives (Wuster and Babu, 2008). It is often present in addition to RRNPP systems and has a distinct two-component signalling mechanism. Discovered in *S. aureus*, the canonical agr system consists of an agr locus containing four genes: *agrA*, *agrB*, *agrC*, and *agrD* (Ji et al., 1997; Kleerebezem et al., 1997). Together, these work to regulate gene expression (Queck et al., 2008) and especially the expression of RNAIII in *S. aureus* (Novick and Geisinger, 2008). As with the RRNPP system, a small oligopeptide (known as the auto-inducing peptide, AIP) – encoded by *agrD* – is processed and exported by transmembrane protein, AgrB. However, once exported, the AIP is no longer able to

enter the cell and, instead, can interact with a second transmembrane protein, AgrC. Upon binding, AgrC autophosphorylates and, in turn, phosphorylates the effector AgrA (Novick and Geisinger, 2008). AgrA is then capable of positively regulating the *agr* locus and altering gene expression directly and through RNAIII.

Agr locus analogues have been found in a wide variety of Gram positives including: *Enterococcus faecalis* (Qin et al., 2000); *Lactobacillus plantarum* (Bao Diep et al., 1994); *C. difficile* (Carter et al., 2005; Lee and Song, 2005); *C. perfringens* (Ma et al., 2015); *C. autoethanogenum* (Piatek et al., 2022); and *C. acetobutylicum* (Steiner et al., 2012). The diversity is not limited to that seen between species but also within species. *C. difficile* encodes multiple copies of agr proteins which are involved in motility, toxin production and sporulation (Ahmed and Ballard, 2022; Ahmed et al., 2020). However, their role as described in industrial solventogenic *C. acetobutylicum* will be explored here as the most relevant species to *C. saccharoperbutylacetonicum*. No studies have yet been conducted on the *C. saccharoperbutylacetonicum* agr system though it was noted to putatively encode one in its genome (Poehlein et al., 2013, 2017).

1.5.2.1 In *C. acetobutylicum*

The single study examining agr quorum sensing in *C. acetobutylicum* investigated the impact of insertional inactivation by ClosTron and vector complementation of three of the four genes (Steiner et al., 2012). They were unable to create a ClosTron mutant of *agrD*, likely due to its small size and proximity to *agrB*. They also attempted to restore phenotypes through exposure to wild type cells and to artificially synthesised AIPs. Finally, they conducted interesting comparative bioinformatics concerning the structure of the agr locus in different *Clostridia* and, particularly, the amino acid sequence of the unprocessed AIP. In *C. acetobutylicum*, the locus is organised with *agrB* and *agrD* overlapping and likely sharing a promoter and *agrC* and *agrA* again aligned with no intergenic space and likely sharing a promoter too. The overall structure is P-*agrB*-*agrD*-P-*agrC*-*agrA* where 'P' stands for promoter.

The insertional inactivation mutants were screened for growth kinetics, solvent and acid production and ability to sporulate. No significant differences were observed in growth rates with all mutants being able to reach the same final OD_{600nm} readings as the wild type. This is similar to the results seen for the *C. acetobutylicum* RRNPP system where insertional inactivation of genes encoding identified RRNPPs showed significant reduction in solvent formation in just one mutant with growth rate largely unaffected overall (Kotte et al., 2020). However, all three agr mutant strains were unable to

accumulate granulose and showed a significant reduction in, though not abolishment of, sporulation. Neatly, the *agrB* mutant, which should be unable to synthesis a signal but be capable of responding to one, did show signs of granulose accumulation when proximal to the wild type strain on solid media.

Complementation with the genes expressed on an exogenous vector under the control of the native promoter was largely successful, restoring granulose production and sporulation. However, the *agrC* mutant could only be complemented by a full *agrC-agrA* construct, indicating that ClosTron insertional inactivation had disrupted the expression of *agrA* and highlighting the pitfalls of ClosTron for the construction of gene disruptions.

The results were potentially extremely useful for the use of *C. acetobutylicum* in industrial processes. Sporulation and solvent production were considered inseparable in this species (Ravagnani et al., 2000; Santangelo IY et al., 1998) and it was shown here for the first time that it might be possible to genetically reduce sporulation without reducing solvent production. The agr system therefore presented itself as a potentially fruitful avenue for strain development in solventogenic *Clostridia*, including *C. saccharoperbutylacetonicum*. However, it should be noted that the lack of any significant obvious role for either agr or RRNPP quorum sensing in *C. acetobutylicum* is at odds with that seen in *C. saccharoperbutylacetonicum* and most other investigated species. Whilst a reduction in sporulation could be seen as a key role, it did not abolish with significant numbers of heat resistant CFUs recovered. This lack of phenotype is curious and suggests that these systems are involved in a more specific process that might not be detected under normal growth and sporulation conditions.

1.6 Transposon mutagenesis

1.6.1 Forward and reverse genetics

The gold standard for microbiological genetics has long been to accurately and precisely relate phenotype and genotype. This principle has been the method of assigning function to genetic material for nearly 80 years since the isolation of *E.coli* mutants that could not synthesise key amino acids by Joshua Lederberg and Edward Tatum (Lederberg, 1946; Lederberg and Tatum, 1946). Broadly, the process can be split between forward and reverse genetics. Reverse genetics is the process of manipulating the genotype, using methods such as those described in 1.2.4, and then working to characterise the phenotype arising from the new genotype. Forward genetics is the

inverse process whereby an isolated phenotype, such as Lederberg's amino acid auxotrophy, is accounted for by a change in genotype.

As can be seen by the example of Lederberg and Tatum, forward genetics was historically the first to be employed. This is because the only methods to manipulate genotype were based on random mutagenesis through chemical or physical mutagens (e.g., ultraviolet light). This method was enhanced by use of interrupted conjugation which was used to start mapping phenotypes to specific regions of the genome in *E. coli* (Wollman and Jacob, 1955). The invention of Sanger sequencing by Frederick Sanger further improved the method by including DNA sequences for the first time (Sanger et al., 1965). As the microbial toolbox advanced, reverse genetics became increasingly accessible and the most useful of the techniques for examining individual genes or loci.

Techniques continued to develop and, as it became clear that many genes have specific roles that may not yield obvious phenotypes under standard lab conditions, the need for high-throughput screening grew. Multiple methods arose, such as sophisticated replica plating techniques so as to rapidly test mutants in multiple conditions (Hasunuma, 2009). Such techniques were combined with large-scale mutagenesis, either random or targeted, to characterise the phenotypes of enormous numbers of genes and loci simultaneously.

The most famous library of targeted mutations is the Keio collection developed in *E. coli* (Baba et al., 2006). The library contains the individual deletion of every non-essential ORF annotated in *E. coli* in 2006. The Keio collection has proved invaluable to the characterisation genes in *E. coli* and has found numerous applications beyond (e.g., Typas et al., 2008). However, the creation of such collections in other species is severely limited by the difficulty in creating targeted gene deletions as outlined in 1.2.4. In addition, the approach is likely to miss smaller or unusual genetic elements and is dependent on the existence of accurate genome maps and annotations for a particular species. Forward genetic approaches employing random mutagenesis therefore offer a powerful method of rapidly characterising genomes of understudied or difficult to work with organisms.

1.6.2 Transposons in molecular biology

Transposons are mobile genetic elements that are able to move themselves between genomic loci. They were first discovered in maize by Barbara McClintock in her work describing how linear DNA fragments could insert themselves in different positions of the chromosome and inactivate the gene at the insertion site (McClintock, 1950; Ravindran, 2012). Since then, transposons have been

discovered in all kingdoms of life. The involvement of transposons in antibiotic resistance and their obvious application in manipulating DNA sequences resulted in a large body of research and applications (Berg et al., 1983; Clegg and Durbin, 2003; Kazazian et al., 1988; Kleckner et al., 1977).

Two classes of transposon have been described, based on the mechanism of transposition. Class I transposons use retrotransposition whereby the transposon is transcribed, then reverse transcribed and inserted into a different locus (Boeke et al., 1985; Bourque et al., 2018). In doing so, class I transposons duplicate themselves in a copy-and-paste method that explains their proliferation to account for significant percentages of large genomes, including ~40% of the human genome (Feschotte and Pritham, 2007). Well-characterised examples of these transposons include Ty elements, Alu elements, intra-cisternal A particles and copia-like elements found in yeast, primates, *Drosophila* and rodents respectively (Hamer et al., 2001).

The second group of transposons, class II, do not rely on transcription and instead employ a cut-and-paste mechanism. Here, DNA containing specific flanking inverted repeat sequences is excised by a transposase and relocated to a new location (Greenblatt and Brink, 1963; Rubin et al., 1982). As such, these transposons do not duplicate themselves during the transposition process and therefore their proliferation has been slower, though they are still able to dominate in species such *Caenorhabditis elegans* (Feschotte and Pritham, 2007). Class II transposons are also widely distributed across the kingdoms with well characterised examples being the Tn superfamily in bacteria, the original Activator/Dissociator transposons discovered by Barbara McClintock, and the Tc1/mariner superfamily and P-elements found in *Drosophila* (Bourque et al., 2018; Hamer et al., 2001). It is class II transposons that have the most relevance to this study where they will be applied as described in 1.6.3.2 and in Chapter III.

Transposons were first used to probe whole genomes in *Saccharomyces cerevisiae* with the aim of characterising recently discovered open reading frames (Smith et al., 1995). The Ty *I* transposon was used to create a random mutant pool. Primers were designed to amplify across the transposon junction, selecting for the insertion site. This pool was then exposed to different conditions and the insertion site amplified. Additional primers carrying fluorescent probes were designed with specificity to the uncharacterised ORFs of interest and annealed to the PCR products. Separation by agarose gel electrophoresis created a unique 'footprint' from which it was possible to identify when particular mutants disappeared from the pool under particular conditions. This highly accurate approach was adapted to multiple species including *E. coli* (Hare et al., 2001) and *Streptococcus pneumoniae* (Akerley et al., 1998) and included improvements such as controlling the expression of

the transposase and the use of an outward-facing promoter to reduce polar effects (Hare et al., 2001).

The genetic footprinting method is not high throughput owing to the need to design specific probes for each ORF of interest. The invention of DNA microarrays (reviewed in: Lenoir and Giannella, 2006; and Southern, 2001) paved the way for higher-throughput methods and led to the development of signature-tagged mutagenesis (STM) (Hensel et al., 1995) and transposon site hybridisation (TraSH) (Sassetti et al., 2001). STM uses an array of 96 barcoded transposons, each used to create a separate, random, library. Pools of 96 mutants, consisting of one mutant from each starting library, were assembled and then grown subjected to an experimental condition resulting in a selective bottleneck. The presence of mutants in the starting pool and after selection were then detected by hybridisation to an array of the 96 barcodes included in the original transposons. This process allowed for the screening of random mutants at a far higher rate than ever before.

TraSH improved on STM by making use of the explosion in genome sequencing around the turn of millennium driven by the Human Genome Project. Instead of barcoding the transposon, the microarray contained a binding site for every single predicted ORF in the *Mycobacterium bovis* genome (Sassetti et al., 2001, 2003). This was the first attempt to examine insertion sites amongst all predicted ORFs. However, both STM and TraSH has been associated with high false-positive rates and poor reproducibility (Tong et al., 2004).

1.6.3 High-throughput transposon insertion site sequencing

1.6.3.1 Development

The invention of high-throughput DNA sequencing made whole genome analyses rapid and increasingly cheaply available (Loman et al., 2012). Illumina sequencing (Bennett, 2004) has proven to be the cornerstone of advances in high-throughput transposon mutagenesis. The ability to sequence millions of DNA fragments in a short space of time led to the development of four similar methods for identify and mapping transposon insertions (Gawronski et al., 2009; Goodman et al., 2009; Langridge et al., 2009; van Opijnen et al., 2009). Together these techniques made it possible to screen potentially millions of transposon mutants simultaneously and assign some function to nearly every part of the genome.

The four techniques published in 2009 – transposon sequencing (Tn-seq; van Opijnen et al., 2009), insertion sequencing (INSeq; Goodman et al., 2009), high-throughput insertion tracking by deep

sequencing (HITS; Gawronski et al., 2009), and transposon-directed insertion site sequencing (TraDIS; Langridge et al., 2009) – all operate on the same basic principle. Transposition of a transposon containing a selectable marker is induced in as many cells as possible. Mutants are pooled and gDNA extracted. The gDNA is then fragmented, either enzymatically or using a physical process, and PCR used to select for transposon-genome junctions and add the DNA features necessary for Illumina sequencing. Illumina sequencing is then conducted, and results processed and mapped to the target genome. Broadly, where no insertions are present, that locus is considered essential to the functioning of the cell under that condition. Libraries can then be exposed to a selective condition and the same process conducted to determine conditional essentiality.

All four techniques show minor differences in the preparation of gDNA for Illumina sequencing. The greatest of these differences is in the method of gDNA fragmentation. INSeq and Tn-seq specifically employ the mariner transposon with a single nucleotide change in the inverted terminal repeats flanking the transposon. This change introduces an MmeI restriction enzyme recognition site and thereby a cut site 20 bp downstream. This ensures all the fragments containing the transposon also contain 16 bp of the adjacent genome sequence. Baring exceptional circumstances, 16 bp is typically sufficient to identify most loci in a bacterial genome. HITS and TraDIS utilise sonication for fragmentation which randomly shears the DNA, yielding a fragment size distribution dependent on the protocol used and the specific genome. All four then employ some method of DNA purification, adaptor ligation and PCR amplification across the transposon-genome junction. The main advantage of HITS and TraDIS is that fragmentation by sonication means that any class II transposon can be used, allowing for use of the most appropriate transposon/transposase system. INSeq and Tn-seq, by contrast, are limited to the use of the mariner system which can be a limitation, particularly for high GC genomes given that the mariner transposase inserts at AT sites. Now, however, the four techniques have largely been assimilated into the broader concept of high-throughput transposon insertion site sequencing.

1.6.3.2 Handling of sequencing data

Following Illumina sequencing, the raw data is checked bioinformatically for the relevant barcodes to differentiate samples when multiplexing multiple experiments. Subsequently, sequences are checked for the presence of the transposon sequence, allowing for some mismatches arising from random sequence mutations and sequencing errors. Short sequencing reads are removed as these

can be difficult to align accurately. Finally, sequences are aligned to a reference genome and calculations of insertion densities, both overall and in particular ORFs, are performed. This should output insertion sites, read counts and overall genome insertion summaries. These processes have been packaged in the Bio-Tradis program (Barquist et al., 2016). Mapped sequences can be viewed manually by examination through software such as Artemis (Rutherford et al., 2000).

ORFs with no insertions are deemed to be essential to the survival of the cell on the basis that insertions did occur but were lethal to the cell. Essentiality can be probed using statistical means with two main parameters. These are the all-important definition of essentiality and the normalisation of the number of insertions in a gene. Essentially, being a spectrum of fitness rather than an absolute, is always user-defined and can be adjusted depending on the stringency required for each experiment. Data can be normalised through one of two methods: by pre-defining the ORF length windows (e.g., Goodall et al., 2018); or through specifically defining the range size for every ORF (e.g., Chaudhuri et al., 2013). Either way, plotting a histogram of the number of insertions within the window/ORF yields a bimodal distribution with a steep left-hand (i.e., no insertions) peak and a broader right hand (i.e., ORFs with insertions) separated by a trough. Essentiality is assigned by the likelihood that a given ORF belongs to one of the modes when the number of insertions is above a certain threshold (typically 12 times more likely to be in one mode than the other). This process does leave some genes that are ambiguous and between the two modes, though the vast majority typically fall within one of the two.

1.6.3.3 Applications

High-throughput transposon mutagenesis has three obvious uses. Firstly, to define the essential genome of the species under standard laboratory conditions (e.g., Gawronski et al., 2009; Goodman et al., 2009; Langridge et al., 2009; Moule et al., 2014; van Opijnen et al., 2009). Essentiality is a somewhat fluid term but typically, the most permissive conditions – solid, rich media and ideal growth temperatures – are initially chosen to allow every mutant that can survive to grow. However, some natural variance amongst ORFs that are very important but do not induce direct lethality will exist between transposon libraries of a given species. Uncovering the essential genome is useful to for a variety of reasons but primarily to reveal the importance of previously un-investigated ORFs and as a roadmap for the accurate manipulation of the genome in the future. This can be particularly important in organisms with relatively little research history such as *C.*

saccharoperbutylacetonicum , but it's relevance to well-studied organisms has also been shown (Goodall et al., 2018).

Secondly, once a baseline essentiality is established, the same libraries of mutants can be passaged through specific conditions to screen for essentiality in those conditions. Conditions can be broad such as sporulating and germinating cells (Dembek et al., 2015), *in vivo* infection models (Chaudhuri et al., 2013; Subashchandrabose et al., 2016), and culturing with antibiotics (Jana et al., 2017). Any condition that is likely to add selective pressure and thereby cause the death of previously non-essential mutants can be used. For solventogenic *Clostridia*, it would be pertinent to screen for ORFs essential in fermentation conditions and those that may be involved in tolerance to solvents such as butanol or to inhibitors found in certain feedstocks.

Thirdly, transposon libraries can be constructed in strains with prior gene deletions and compared to wild type libraries. Such comparisons can reveal synthetically lethal pairings or the inverse, genes that are no longer essential in the deletion background (e.g., Fenton et al., 2016; van Opijnen et al., 2009). Alternatively, libraries can be generated in a condition which chemically inhibits a certain gene or pathway in order to probe that pathway (e.g., Santa Maria et al., 2014).

These three basic approaches were identified by the inventors of the four high-throughput transposon mutagenesis techniques. Since then, a variety of novel methods of comparing mutant fitness that are not exclusively based on competitive growth have been developed (reviewed in Cain et al., 2020). Genes essential for motility were assessed by harvesting faster moving mutants (Kakkanat et al., 2017). Cell sorting through a variety of methods has also been deployed to identify efflux pumps in *Acinetobacter baumannii* (Hassan et al., 2016) and capsule production in *Klebsiella pneumoniae* (Dorman et al., 2018). Even a limitation of high-throughput transposon mutagenesis – that mutants are always compared to the population – is being bypassed through the use of microfluidics to isolate individual mutants which indicated that up to 3% of *S. pneumoniae* mutants have a different fitness profile when grown individually (Thibault et al., 2019). These approaches highlight the flexibility of high-throughput transposon mutagenesis and the utility of such forward genetics approaches.

1.7 Project aims

As demonstrated, there is a significant dearth of biological knowledge concerning *C. saccharoperbutylacetonicum* compared to its growing industrial importance. High-throughput

transposon mutagenesis is a powerful technique for analysing a whole range of genotypes, especially in species where few reverse genetics studies have been conducted. Applying this technique to *C. saccharoperbutylacetonicum*, by adapting the method outlined for *C. difficile* by Dembek, et al., would reveal the essential genome under normal laboratory conditions. Doing so would provide avenues for investigating the pressures and stressors inherent in its industrial application. This was the primary aim of this project.

In addition, the lack of knowledge concerning sporulation, germination and quorum sensing were identified as key gaps that have the potential to improve industrial applications. These three areas were well-positioned to be investigated using the expertise in the Fagan lab at the University of Sheffield (e.g., cell envelopes and TEM) and those of Green Biologics/Biocleave (e.g., CLEAVE™, phenotyping and fermentation of solventogenic *Clostridia*). Therefore, these topics were selected secondary project aims. The sporulation phenotypes of *C. saccharoperbutylacetonicum*, in particular the putative difference between sporulation in solid and liquid media, would be investigated by phase contrast and fluorescence microscopy as well as thin-section TEM and basic phenotypic assays. The aims of the germination studies were to elucidate novel germinants and gemination inhibitors that could prove useful in industrial applications through germination assays using purified spores. Finally, quorum sensing as modulated by the agr system would be investigated through classical reverse genetics using the CLEAVE™ system followed by basic phenotypic characterisation.

This thesis, therefore, concerns the following main aims:

1) The elucidation of the essential genome of *C. saccharoperbutylacetonicum* under laboratory conditions utilising random transposon mutagenesis and insertion site sequencing (Chapter 3)

2) The exposure of transposon libraries generated in 1) to different conditions (e.g. small scale fermentation) to explore changes to the essential genome under those conditions and discover genes key to the process(es) (Chapter 3)

3) The investigation of the sporulation phenotypes of *C. saccharoperbutylacetonicum* in liquid and solid media through phase contrast, fluorescence and thin section transmission electron microscopy (Chapter 4)

4) The search for active germinants and germination inhibitors in *C. saccharoperbutylacetonicum* through the screening of candidate compounds (Chapter 4)

5) The characterisation of the role of accessory gene regulator quorum sensing in *C. saccharoperbutylacetonicum* by reverse genetics (Chapter 5)

6) The development of promoters for use in downstream research and industrial applications (Chapter 5)

In general, two overarching principles were applied to all aspects of this project:

1) To add to knowledge base of *C. saccharoperbutylacetonicum*

2) To provide practical results that have the potentially to be applied directly to industrial applications

Chapter II – Materials and methods

2.1 Growth of Bacteria

2.1.1 Strains and conditions

Liquid cultures of *Clostridium saccharoperbutylacetonicum*N1-4(HMT) (Hongo et al., 1968; Poehlein et al., 2014) were grown statically in Reinforced *Clostridial* Media (RCM; Oxoid), Tryptone Yeast extract Iron sulphate Ammonium sulphate (TYIR, Table 2.1), 50 mM 2-(N-morpholino)ethanesulfonic acid (MES) and 50 g/L glucose or in *Clostridial* Growth Media (CGM, Table 2.2) with 50 g/L glucose (Atmadjaja et al., 2019). Solid cultures were grown on RCM, TYIR (without MES) or CGM supplemented with 1.5% w/v agar. All cultures were grown at 32°C in an anaerobic cabinet (Don Whitley; 10% H_2, 10% CO_2, 80% N_2). Liquid media was reduced for a minimum of 3 h in the cabinet prior to use whilst solid media was reduced for a minimum of 30 min. When necessary, cultures were supplemented with the appropriate antibiotic: erythromycin (40 µg/mL), thiamphenicol (75 µg/mL), anhydrotetracycline (ATc; 4-500 ng/mL) and 50 µg/mL colistin sulphate.

Table 2.1 Composition of TYIR

Yeast extract	2.5 g/L
Tryptone	2.5 g/L
$(NH_4)_2SO_4$	0.5 g/L
$FeSO_4.7H_2O$	0.025 g/L
pH adjusted to 6.5	

Table 2.2 Composition of CGM

Yeast extract	5 g/L
K_2HPO_4	0.75 g/L
KH_2PO_4	0.75 g/L
$MgSO_4$	0.4 g/L
$FeSO_4$	0.01 g/L
$MnSO_4$	0.01 g/L
NaCl	1 g/L
$(NH_4)_2SO_4$	2 g/L
Asparagine	2 g/L
pH adjusted to 6.5	

Liquid cultures of *Clostridium difficile* were grown statically in TY broth (3% tryptose; 2% yeast extract (Bacto)) (Dupuy and Sonenshein, 1998). Solid cultures were grown on BHI-Agar (Sigma). All cultures were grown at 37°C in an anaerobic cabinet with the same environmental composition as above. All media was reduced for the same minimum duration as for *C. saccharoperbutylacetonicum*. Cultures were supplemented as necessary with the following antibiotics: thiamphenicol (15 µg/mL), ATc (4 - 500 ng/mL) and colistin sulphate (50 µg/mL).

Liquid cultures of *Escherichia coli* were grown in LB (Sigma) at 37°C with 225 rpm shaking. Solid cultures of *E. coli* were grown on LB-agar (Fisher) at 37°C. Cultures were supplemented as necessary with the following antibiotics: chloramphenicol (15 µg/mL), erythromycin (500 µg/mL) and kanamycin (50 µg/mL). *E. coli* strains NEB5α (New England Biolabs), DH10β and Top10 were used for cloning and routine maintenance of plasmid DNA. CA434 was used a conjugation donor for plasmid transfer into both *C. difficile* and *C. saccharoperbutylacetonicum*.

All species and strains were stored in 15% v/v glycerol at -80°C, unless stated otherwise and are listed in Appendix I.

2.1.2 Spore preparation and isolation

Two methods were used to generate spores of *C. saccharoperbutylacetonicum* as each method yields a different spore phenotype. Spores were generated in liquid cultures using a three-step process. In step one, 10 mL RCM was inoculated from a single colony and left to grow overnight (O/N). The optical density at 600 nm (OD_{600nm}) was measured and 1.5 mL used to inoculate 10 mL γ-cyclodextrin (50 g/L)-TYIR (Jenkinson et al., 2019). Growth was monitored until OD_{600nm} reached between 1.3-1.5. 2 mL of this subculture was then used to inoculate 20 mL fresh γ-cyclodextrin-TYIR. Cultures were left for up to 72 h and monitored by brightfield microscopy (x100 magnification) for the appearance of spores.

In the second method, 10 mL RCM was again inoculated from a single colony and left to grow O/N. 300 µL of these cultures were spread on γ-cyclodextrin-TYIR agar plates. Cultures were left to grow for up to 7 days and monitored for the appearance of spores as for liquid cultures.

Spores were isolated as described previously (Nerandzic and Donskey, 2013). Briefly, 2 mL of cell suspension at an OD_{600nm} of 1 were harvested by centrifugation at 4000 x *g*. The supernatant was discarded, and the pellet was suspended in 1 mL ice-cold water then centrifuged at 5000 x *g*. This was repeated five times. Cells were then resuspended in 500 µL 20% w/v HistoDenz (Sigma), layered

onto 1 mL 50% w/v HistoDenz and centrifuged at 15,000 x *g* for 15 min. The supernatant was discarded, the pellet suspended in 1 mL ice-cold water and centrifuged at 5000 x *g*. This was again repeated five times. Finally, the pellet was resuspended in 1 mL sterile water and stored at 4°C until use.

2.2 DNA manipulation

2.2.1 gDNA isolation

To extract genomic DNA (gDNA) for downstream processes, 1.5 mL of *Clostridium saccharoperbutylacetonicum* overnight (O/N) culture was harvested by centrifugation at 4000 x *g* for 10 min in a 1.5 mL microcentrifuge tube. The supernatant was removed by aspiration, the pellet resuspended in 200 µL phosphate buffered saline (PBS) and transferred to a fresh 1.5 mL microcentrifuge tube. All incubation steps were carried out in a water bath set to the stated temperature. Lysozyme was added (1 mg/mL) and the mixture was incubated for 1 h at 37°C. 10 µL of 20 mg/mL pronase was added and incubated for 1 h at 55°C. Subsequently, 80 µL of 10% N-lauroylsarcosine was added and incubated for 1 h at 37°C. Finally, 200 µL of 0.2 mg/mL RNase was added and incubated for 1 h at 37°C.

gDNA was separated from RNA and proteins through phenol-chloroform extraction. The sample and 500 µL of phenol:chloroform:isoamyl alcohol 25:24:1 were added to heavy phase lock gel (PLG; QuantaBio) tubes, mixed by inversion and centrifuged at 13,000 x *g* for 2 min. The upper layer was transferred to a fresh PLG tube, and the previous step repeated. The upper layer was again transferred to a fresh PLG tube and 500 µL chloroform:isoamyl alcohol 24:1 was added, mixed by inversion and centrifuged at 13,000 x *g*. This step was also repeated. The upper layer was transferred to a fresh 1.5 mL microcentrifuge tube, gDNA precipitated using 500 µL 100% v/v isopropanol at 4°C and left overnight at -20°C. gDNA was centrifuged at 4,000 x *g* at 4°C for 15 min. The isopropanol was removed, the pellet washed with 70% v/v ethanol and centrifuged at 4,000 x *g* at 4°C for 10 min. The ethanol was removed, the sample air-dried for 5 min at room temperature and then resuspended in 50 µL nuclease-free H_2O at 4°C overnight. Purity and quantity were assessed by absorbance at 260 nm and evaluation of the ratio of absorbance at 260 nm over 280 nm. Integrity was assessed though agarose gel electrophoresis and, when greater accuracy was required, concentration was assessed by Qubit fluorimetry (see 2.2.8).

2.2.2 Polymerase chain reaction

For PCR requiring high fidelity replication, Phusion polymerase high fidelity master mix (Thermo) was used according to the manufacturer's protocol. Typically, 1 ng of plasmid DNA was used as a template whilst for gDNA, 10-100 ng would be used. For colony PCR of *C. saccharoperbutylacetonicum*, either 1 µL of O/N culture or one colony suspended in PBS treated with proteinase K followed by heating at 80°C were used. The cycling conditions are listed in Table 2.3. 20 µL reactions were used as standard, but when a higher quantity of PCR product was required, larger reactions were prepared and split into 20 µL aliquots. A DMSO range of 1-9% v/v was used when required i.e., when primers with long A/T stretches or potential to form significant secondary structures were identified.

Table 2.3 Generic cycling conditions of Phusion PCR

Step	Temperature (°C)	Time (s)	Cycles
Initial denaturation	98	30	x1
Denaturation	98	30	x32
Annealing	2°C below primer Tm	30	
Extension	72	30 s/kb	
Final Extension	72	300	x1

For low fidelity PCR, primarily screening of cloned constructs, Taq polymerase was used. An in-house mixture was used (0.01% cresol red, 0.5 M sucrose, 400 µM of each dNTP, purified *Taq* polymerase 0.1 U/µl, 20 mM tris-HCl pH9, 100 mM KCl, 0.02% gelatin, 0.02% tween 20, 4 mM $MgCl_2$). The same quantities of template DNA were typically used.

Table 2.4 Generic cycling conditions of Taq PCR

Step	Temperature (°C)	Time (s)	Cycles
Initial denaturation	95	180	x1
Denaturation	95	30	x32
Annealing	2°C below primer Tm	30	
Extension	72	60 s/kb	
Final Extension	72	300	x1

For *E. coli* screening, a single colony was picked with a pipette tip, patch-plated and the remainder mixed with the PCR mix. 10 µL reactions were used as standard. The cycling conditions are listed in Table 2.4. PCR was conducted in a T100 Thermo Cycler (BioRad). Primers, listed in Appendix III were synthesised by Eurofins.

2.2.3 Isolation of plasmid DNA

Plasmid DNA was extracted from *E. coli* using the GeneJET Plasmid Miniprep Kit (ThermoFisher Scientific) following the manufacturer's instructions. Briefly, 10 mL of O/N culture was harvested by centrifugation at 4000 x *g* for 10 min. Pellets were resuspended in 250 µL of resuspension solution and transferred to a 1.5 mL microcentrifuge tube. 250 µL of lysis solution was added and mixed by inversion until clear. 350 µL of neutralization solution was added and again mixed by inversion. The solution was centrifuged at 21,000 x *g* for 5 min to remove cell debris and precipitated genomic DNA. The supernatant was retained and transferred to a GeneJET Spin Column. This was centrifuged at 21,000 x *g* for 1 min and the flow-through discarded. 500 µL of wash solution was added and the column centrifuged at 21,000 x *g* for 1 min. This step was repeated and followed by drying the column by centrifuging at 21,000 x *g* for 1 min. Columns were transferred to a fresh 1.5 mL microcentrifuge tube, 50 µL nuclease-free H_2O was added to the column and left for 2 min at room temperature. They were then centrifuged at 21,000 x *g* for 1 min to elute the DNA. Purity and quantity were assessed by absorbance at 260 nm using a DeNovix DS11-FX. Plasmids used in this study are listed in Appendix II.

2.2.4 Agarose gel electrophoresis

Agarose gel electrophoresis was routinely used to separate DNA by size for various purposes. Between 0.8-2% w/v agarose was melted in TAE (40 mM tris-acetate pH8, 1 mM EDTA), cooled to ~55°C, poured into a cassette, and allowed to set. The percentage gel chosen depended on the size of DNA being separated with higher percentages being better at resolving small (<500bp) DNA fragments and lower percentages for larger fragments (>500 bp). DNA was dyed using either SyberSafe (final concentration 1:10,000; Invitrogen) in the agarose or UView (final concentration 1:10; BioRad) added directly to the sample. UView was used when the DNA required gel extraction (see 2.2.14). Gels were placed in gel tanks containing TAE buffer. Standard DNA loading buffer (NEB) or UView were added to the samples and the samples loaded into the gel wells. Gels were

electrophoresed at 100 V for 30-50 min. DNA was visualised and images taken using a BioRad ChemiDoc MP imaging system.

2.2.5 Gel extraction

After resolving DNA on an agarose gel (2.2.13), DNA was visualised using a UV transilluminator and the desired band excised using a sharp scalpel. Extraction of DNA from the gel was conducted with the GeneJet Gel Extraction Kit (Thermo) according to the manufacturer's instructions, eluted in 20 µL nuclease-free H_2O and quantified using absorbance at 260 nm.

2.2.6 PCR purification

When purified PCR product was required without the need for gel electrophoresis, the GeneJet PCR Purification kit was used following the manufacturer's instructions. DNA was eluted in 20 µL nuclease-free H_2O and quantified using absorbance at 260 nm.

2.2.7 Restriction endonuclease digestion of DNA

Endonuclease digestion of DNA was conducted using restriction enzymes according to the manufacturer's instructions (New England Biolabs). The enzyme and its buffer were added to the DNA samples in the appropriate proportions and incubated for 1 h at 37°C. Digested DNA was visualised as described in 2.2.13 and purified according to 2.2.14.

2.2.8 Qubit fluorimetry

Qubit fluorimetry was used to accurately assess DNA concentration. The method utilises a fluorophore that fluoresces when bound to DNA is a predictable and proportionate manner. The concentration reading is therefore not affected by contaminants. The Qubit dsDNA High Sensitivity assay kit (ThermoFisher) was used according to the manufacturer's instructions. Briefly, the sample DNA concentration range was estimated, and dilution conducted if the concentration was likely to be >100 ng/µL. 1-20 µL of sample DNA was added to 180-199 µL of Qubit assay reagent, vortexed and allowed to stand at RT for 5 min. 0 ng/µL and 10 ng/µL standards were prepared as per the manufacturer's instructions. Concentration was assessed using the DeNovix DS11-FX fluorimeter's programmed Qubit HS assay settings.

2.2.9 Ligation of DNA fragments

T4 DNA ligase (NEB) was used to combine purified DNA fragments according to the manufacturer's instructions. Buffer and enzyme were added samples containing the destination vector and the insert DNA. Specific molar ratios were used based on the standard addition of 50 ng of vector DNA to the reaction. 3:1 and 1:1 insert:vector were the most commonly used ratios with 5:1 and 1:3 also used when initial attempts were unsuccessful. Reactions were prepared in 10 µL volumes and conducted at either RT for 1 h or 16°C overnight. Ligated product was then used in *E. coli* transformations.

2.2.10 Gibson assembly of DNA fragments

Gibson assembly was used to assemble multiple DNA fragments into a single vector. This was conducted using the NEBuilder HiFi DNA assembly Cloning Kit (NEB). Destination vector DNA was linearised using either Phusion PCR or blunt-end endonuclease restriction digestion. Insert fragments were amplified using primers containing 30 bp overhangs with perfect homology to other fragments and/or vector DNA sequence. All assemblies contained fewer than three fragments so 25 ng of vector and a 1:2 vector:insert molar ratio was used as standard. NEBuilder HiFi DNA assembly mix was used according to the manufacturer's instructions. Reactions were incubated at 50°C for 30 min (NEBuilder). The reaction product was used immediately in *E. coli* transformations.

2.2.11 Production of chemically competent *E. coli*

O/N cultures of the relevant *E. coli* strain (different strains were prepared the same way) were grown as described above. These were subcultured to 1:100 and grown to OD_{600nm} 0.4-0.6 (log phase). Cells were then harvested by centrifugation at 4000 x g for 10 min. Pellets were suspended in 5 mL ice-cold 100 mM $CaCl_2$ and incubate on ice for 15 min. Cells were again centrifuged at 4000 x g for 10 min. Cells were then suspended in 1 mL 100 mM $CaCl_2$ in 15 % v/v glycerol and aliquoted into 50 µL volumes. After incubation, cells were flash frozen in liquid nitrogen and stored at -80°C.

2.2.12 Heat shock transformation of *E. coli*

50 µL aliquots of chemically competent *E. coli* or competent NEB5α *E. coli* purchased from NEB were thawed on ice. Once thawed, 25 µL was aliquoted to microcentrifuge tubes containing either 2.5 µL ligation or Gibson assembly product or 0.25 µL purified plasmid and incubated on ice for 30 min. Cells were heat shocked at 42°C for 45 s followed by incubation for 2 min on ice. 600 µL was of superoptimal broth with catabolite repression (SOC, NEB) was added and cells were grown for 1 h at 37°C and 225 rpm. 100 µL of cells were then spread on LB agar with the appropriate antibiotic selection. If the ligation/Gibson assembly was deemed low yield, the remaining cells were centrifuged at 4,000 x g for 5 min, all but 100 µL of the supernatant removed and the remaining 100 µL used to suspend the cells. This concentrated suspension was then spread on a fresh LB agar plate.

2.2.13 Sequencing of DNA

Sanger sequencing of DNA constructs was conducted through the Genewiz sequencing service, and the results analysed using the Geneious software (v7.1.9). Illumina sequencing was conducted using three different services. AmpliconEZ from Genewiz was used for guaranteed 50,000 reads/sample to check the library composition of the TraDIS libraries during method development. The Illumina MiSeq v2.0 2x150bp Nano, was also used to check method development, and the MiSeq v2.0 2x150bp, used to generate the complete libraries, were conducted by the Sheffield Diagnostic Genetics Service at the Sheffield Children's NHS Foundation Trust. Analysis of all Illumina sequencing was conducted by Dr Roy Chaudhuri (The University of Sheffield).

2.2.14 Conjugative transfer of DNA into *Clostridia*

Except during the generation of knockout mutants in *C. saccharoperbutylacetonicum*, all plasmids were transferred into *C. saccharoperbutylacetonicum and C. difficile* through conjugative transfer. *E. coli* CA434, transformed with the desired plasmid, was used as the conjugative donor. O/N cultures of *Clostridia* and *E. coli* donor were grown as described above. 1 mL of *E. coli* was harvested by centrifugation at 4000 x g for 5 mins, the supernatant removed, and the cells brought into the anaerobic cabinet. For *C. difficile,* the recipient culture was incubated at 50°C for 10 min prior to combination with the *E. coli* donor (Kirk and Fagan, 2016). 200 µL of *Clostridia* was used to gently suspend the *E. coli* pellet by pipetting and spotted onto BHI agar (*C. difficile*) or RCM agar (*C. saccharoperbutylacetonicum*). These were left to grow for 8-24 h after which they were harvested

using 1 mL TY (*C. difficile*) or RCM (*C. saccharoperbutylacetonicum*) and a spreader. 100 µL was plated on BHI agar or RCM agar with the appropriate antibiotic and 50 µg/mL colistin sulphate (for *E. coli* counterselection). Conjugative transfer using this method is very efficient in *C. saccharoperbutylacetonicum* so 1:100 and 1:10 dilutions of the scrapings were typically plated.

2.2.15 Processing gDNA for TraDIS sequencing

2.2.15.1 gDNA shearing

1 µg of gDNA dissolved in 130 µL TRIS-HCl pH8.0 was sheared by sonication using the 300 bp protocol of an S220 Covaris sonicator. The settings for this protocol are given as: frequency sweeping power mode; continuous degassing; 10% duty factor; 140 W peak incident power; 200 cycles per burst; 80 s cycles; 7°C. Sonication was conducted at the Sheffield Diagnostic Genetics Service at the Sheffield Children's NHS Foundation Trust.

2.2.15.2 Sample volume reduction

Following shearing, the sample volume is too high for the downstream steps, so volume reduction was required. A rotor vacuum was used with sample tubes being spun at low RPM under vacuum to cause the evaporation and removal of the excess H_2O. This was done until the volume was 50 µL or lower and topped up to 55.5 µL with nuclease-free H_2O as necessary. The sample was then transferred to a 0.2 mL PCR tube.

2.2.15.3 DNA end preparation

Sonicated DNA requires end repair prior to be useable in downstream reactions. The NEBNext End Prep Ultra I kit according to the manufacturer's instructions. 55.5 µL of fragmented DNA was combined with 3 µL End Prep Enzyme mix and 6.5 µL End Repair Reaction Buffer. These were mixed by pipetting and placed in a thermocycler at 20°C for 30 min followed by 65°C for 30 min. The reaction was then held at 4°C until use.

2.2.15.4 Adaptor ligation

The Illumina NEBNext Adaptor was then added by ligation to the ends of the fragments using components from the NEBNext End Prep Ultra I kit. 15 µL Blunt/TA Ligase Master Mix, 2.5 µL NEBNext Adaptor for Illumina and 1 µL Ligation Enhancer was added to the entire 65 µL sample from

2.2.15.3, mixed by pipetting and incubated at 20°C for 15 min. 3 µL of USER™ enzyme was added and the sample incubated at 37°C for 15 min.

2.2.15.5 Sample purification and size selection

Sample purification was conducted using the Agencourt AMPure XP magnetic beads. These have the dual advantage of yielding high purity product and providing size selection. These libraries contain short fragments from a variety of sources that would prove unproductive for sequencing. Equally, it is useful to remove large fragments such as poorly fragmented gDNA. A protocol designed to select for 300 – 400 bp fragments was utilised. These fragment sizes are dictated by the bead:sample ratio. Larger fragments bind preferentially to the beads and therefore higher ratios of beads will lead to the binding of a wider range of fragment sizes. This first clean up step is designed to remove both the upper and lower limits. Subsequent purification steps are only designed to occlude smaller fragments.

AMPure XP bead stocks were vortexed thoroughly to resuspend them and allowed to reach room temperature. The ligation reaction from 2.2.15.4 was topped up to 100 µL using 13.5 µL nuclease-free H_2O and transferred to a 1.5 mL microcentrifuge tube (Axygen). 55 µL of beads were added, mixed by pipetting, and incubated for 5 min at room temperature. Tubes were placed on a magnetic stand until the solution cleared (2-5 min). The solution was transferred to a fresh 1.5 mL microcentrifuge tube leaving the larger fragments bound to the beads. 25 µL of fresh beads were added to the transferred solution, mixed by pipetting, and incubated at room temperature for 5 min. The tube was again placed on a magnetic stand until the solution cleared. This time the solution was removed and discarded, eliminating the fragments smaller than 300 µL. The tubes were kept on the magnetic rack and 200 µL freshly prepared 80% ethanol was added to the tube, incubated for 30 s and then removed. This step was repeated. Residual ethanol was removed, and the beads air dried for 5 min. Tubes were removed from the magnetic rack, 17 µL TRIS-HCl added, mixed by pipetting incubated at room temperature for 2 min. Finally, tubes were placed back on the magnetic rack until the solution cleared and 15 µL was transferred to a fresh 0.2 µL PCR tube.

2.2.15.6 PCR amplification of transposon junctions

This PCR step is designed to amplify across the transposon insertion site. The forward primer binds to the transposon whilst the reverse is the adaptor attached in 2.2.15.4. The KAPA HiFI polymerase with the cycling conditions listed in Table 2.5 was utilised. the entire 15 µL from 2.2.15.5 was used as template and combined with 25 µL KAPA HiFI polymerase, 2.5 µL of each primer (RF1520 and RF1522) and 5 µL nuclease-free H_2O. Only 10 cycles are conducted to reduce the potential for PCR bias.

Table 2.5 Transposon junction PCR cycling conditions

CYCLE STEP	TEMP (°C)	TIME (s)	CYCLES
Initial Denaturation	98	180	1
Denaturation	98	15	10
Annealing	65	30	
Extension	72	30	
Final Extension	72	60	1
Hold	4	∞	

2.2.15.7 Restriction enzyme digest to remove plasmid

Following PCR, restriction digestion using BstXI was used to remove contaminating pRPF215 delivery plasmid. The 50 µL PCR mix was transferred to a fresh 1.5 mL microcentrifuge tube and combined with 6 µL NEBuffer 3.1 and 4 µL BstXI. The sample was left at 37°C overnight in a circulating water bath followed by denaturation the following morning at 80°C for 20 min.

2.2.15.8 Sample purification

The sample was purified from the reaction components and small DNA fragments using the Agencourt AMPure XP magnetic beads. Beads were prepared as before. 54 µL of beads were added, mixed by pipetting and incubated at room temperature for 5 min. The tube was briefly centrifuged and placed on the magnetic stand until the solution cleared. The supernatant was removed and discarded. 200 µL of freshly prepared 80% ethanol was added for 30 s and removed. This step was repeated once before air drying the beads at room temperature for 5 min. The sample tube was removed from the magnetic rack and 17 µL TRIS-HCl was added and incubated at room temperature for 2 min to elute the DNA. The tube was placed back on the magnetic rack until the solution became clear after which 15 µL was transferred to a fresh 0.2 mL PCR tube.

2.2.15.9 PCR amplification for Illumina sequencing preparation

This PCR step is designed to add the extra sequences required for Illumina sequencing. Again, the KAPA HiFi PCR kit was used for this amplification. The entire 15 µL sample was combined with 25 µL KAPA HiFi polymerase, 2.5 µL of each primer (custom inline index primer and Illumina primer from kit NEBNext Multiplex Oligos for Illumina) and 5 µL nuclease-free H_2O. The cycling conditions are shown in Table 2.6

Table 2.6 Cycling conditions for library amplification

CYCLE STEP	TEMP (°C)	TIME (s)	CYCLES
Initial Denaturation	98	180	1
Denaturation	98	15	20
Annealing	65	30	
Extension	72	30	
Final Extension	72	60	1
Hold	4	∞	

2.2.15.10 Sample purification

The sample was purified as in 2.2.15.8, with minor modifications to account for the different volumes. 45 ul of AMPure XP beads were added and 33 µL 10 mM TRIS-HCl pH8.0 was used for elution. 32 µL of the final library was stored at -20°C.

2.2.15.11 qPCR library quantification

KAPA Illumina quantification kit (KK4824) was used for precise quantification of the final library. A mastermix containing the polymerase, dNTPs, SYBR green dye and primers designed to amplify from the P5 and P7 flow cell adapter was kept in the dark and thawed on ice. Serial dilutions of the library were prepared with the aim of creating 5 x 10^{-4} and 5 x 10^{-5} dilutions. This were carried out using 10 mM TRIS-HCl pH8.0 and repeated three times for each sample tested. 2 µL/reaction nuclease-free H_2O was combined with 6 µL/reaction of the above mastermix. 8 µL per reaction was aliquoted into a semi-skirted 96 well PCR plate (Starlab). 2 µL of template was then added to each reaction. Six standards are included in the KAPA kit and loaded twice with each quantification. A minimum of three no template controls (NTC) with replaced sample with nuclease-free H_2O were used per quantification. The plate was sealed with optically clear 8-strip caps (Starlab). The cycling conditions used are shown in Table 2.7 and a melt curve conducted at the end of the run. The Bio-Rad CFX Connect Real-Time PCR Detection System thermocycler was used.

Table 2.7 The cycling conditions of library quantitative PCR

STEP	TEMP (°C)	TIME (s)	CYCLES
Initial Denaturation	95	300	1
Denaturation	95	30	35
Annealing and Extension	60	30	
Final Extension	72	60	1
Hold	4	∞	

Data was analysed using the template Excel file supplied by KAPA. The cycle number was calculated using the Bio-Rad CFX Maestro software defaults. Briefly, standards were plotted on a graph of cycle number against concentration, the R^2 correlation calculated using the Excel function (>0.99 was considered acceptable) and the equation of the line derived using Excel. The efficiency was calculated using a comparison of standards to programmed figures and deemed acceptable between 90-110%. Obvious outliers were removed and NTCs were checked for amplification. The difference in cycle number between consecutive standards was used to further assess the accuracy and precision of the experiment with a difference of 3.1-3.6 cycles considered within acceptable range. The equation of the line was used to convert cycle number to concentration and the final concentration of each library was calculated by multiplying out the dilution factor and sample volume to attain a final concentration in the nM range.

2.2.15.12 MiSeq sequencing

The sequencing process was conducted by Sheffield Diagnostic Genetics Service at the Sheffield Children's NHS Foundation Trust according to their standard protocols. Prior to handing the library to them, library concentrations were adjusted to 8 nM in 200 µL TRIS-HCl.

2.2.16 Plasmid copy number analysis

To ascertain the copy number of pRPF215 and the *C. saccharoperbutylacetonicum* megaplasmid relative to the genome, quantitative PCR was conducted using the PowerUp SYBR Green Master mix (Thermo). Primers were designed to adhere to the specifications of a short amplicon (<200 bp) and a 60°C annealing temperature. Four primer pairs were designed for the genome and the

megaplasmid, whilst two primer pairs were created for the significantly smaller pRPF215 plasmid. Primers were tested using the PowerUp SYBR Green Master mix in a normal thermocycler and analysed for the production of a bright, single band by agarose gel electrophoresis.

Cells containing the plasmid were grown O/N in 75 µg/mL thiamphenicol RCM. O/Ns were subcultured to OD_{600nm} 0.05 and samples taken during log phase (OD_{600nm} 0.5), stationary phase (OD_{600nm} 2.5) and the following day (OD_{600nm} 3-4) and OD_{600nm} adjusted to OD_{600nm} 1.

gDNA and plasmid DNA were extracted simultaneously using the method described in 2.2.1. DNA purity was measured by absorbance at 260 nm and total DNA content by Qubit (2.2.8). DNA concentration was adjusted to 100 ng/µL and serial dilutions made down to 1 ng/µL. 1 µL of 10 ng/µL and 1 ng/µL dilutions were used as template for the qPCR reaction. Serial dilutions of pRPF215 of 10, 1, 0.1, 0.01 and 0.001 ng/µL were used as standards to convert cycle number into an evaluation of concentration. No template controls were included to control for contamination with nuclease-free H_2O used instead of DNA. The reaction was assembled according to the PowerUp SYBR green protocol with 5 µL of 2x Master mix being combined with 1 µL of each primer (500 nM final concentration), 1 µL of template and 3 µL of nuclease-free H_2O. The cycling conditions used are listed in Table 2.8. The primers used for this are listed in Appendix III

Table 2.8 The cycling conditions of PowerUp SYBR Green PCR

Cycle Step	Temperature (°C)	Time (s)	Cycles
Uracil-DNA glycosylase activation	50	120	x1
Dual-Lock DNA polymerase activation	95	120	x1
Denaturation	95	15	x40
Annealing	58	15	
Extension	72	60	

Following the 40 cycles of the qPCR a melt curve step was conducted to analyse the conformity of the amplicons between repeats. The melt curve is conducted by increasing the temperature from 70°C to 95°C in 0.2°C increments and recording spectrophotometry at each increment. All PCR and melt curve steps were conducted using a Bio-Rad CFX Connect Real-Time PCR Detection System.

Data was analysed using Microsoft Excel. Briefly, the standards were plotted as a standard curve and the R^2 value used to check the correlation. The equation of the curve was generated and used to calculate the concentration of the different repeats. Any repeats with cycling numbers differing greatly from the average were excluded from the analysis. The genome was assumed to have one copy. The results from the megaplasmid and pRPF215 were then divided by the value found for the genome and this final value presented as the copy number. Variations on these calculations were also made and are discussed in Chapter III.

2.2.17 Electroporation of C. saccharoperbutylacetonicum

The electroporation of *C. saccharoperbutylacetonicum* is a proprietary technique developed by Green Biologics/Biocleave/BCL2020 (Atmadjaja et al., 2019). O/N cultures were grown in RCM directly from glycerol stocks, and checked for healthy growth through OD_{600nm}, brightfield microscopy at x40 magnification, and pH measurements. 6 mL of O/N was added to 54 mL CGM-glucose and allowed to grow to OD 1.2. The culture was harvested by centrifugation at $4^{O}C$ and 4,000 x g for 10 min. The supernatant was discarded, and cells suspended in 20 mL salt buffer (300 mM sucrose, 0.6 mM Na_2HPO_4, 4.4 mM NaH_2PO_4, 10 mM $MgCl_2$). They were again centrifuged as before, the supernatant discarded, and the cells suspended in 1 mL of no-salt buffer ((300 mM sucrose, 0.6 mM Na_2HPO_4, 4.4 mM NaH_2PO_4). 200 µL of cells were added to 2 mm electroporation cuvettes containing 1 µg of plasmid DNA and incubated on ice for 5 mins. Cells were electroporated at 1.5 kV with a Bio-Rad MicroPulser electroporator, 1 mL of CGM-glucose added and then left to recover overnight. The next morning, cultures were mixed by pipetting and 200 µL spread on a CGM-glucose agar plate with the appropriate antibiotic.

2.2.18 CLEAVE™ mutagenesis

CLEAVE™ is the propriety genome editing technique invented by Green Biologics whose patents are held by Biocleave (BCL2020) (Atmadjaja et al., 2019; Jenkinson and Krabben, 2015). The technique is based on homologous recombination and selection that uses the endogenous CRISPR system found in *Clostridia* and other non-highly recombinogenic bacteria. Briefly, for deletion mutations, homology arms of between 500-1200 bp were designed flanking the desired knockout region. These were cloned into the Step 1 recombination plasmid (Atmadjaja et al., 2019). PAM sites were identified in the knockout regions and subsequently the downstream spacer sequence that is used to target the N1-4(HMT) genome (Poehlein et al., 2014). These were synthesised with flanking direct

repeats and cloned into the Step 3 plasmid containing the guiding apparatus for use with the endogenous CRISPR system.

The Step 1 plasmids were transformed into *C. saccharoperbutylacetonicum* by electroporation (see 2.2.16). Successful transformants are re-streaked and then grown O/N in RCM. Seven subcultures – conducted O/N and during the day – are made for each desired knockout. After the seventh subculture, a final O/N is made in preparation for the transformation by electroporation of the Step 3 targeting plasmid. Successful transformants were re-streaked and screened for knockout by PCR. One primer was designed to anneal to one flank of the knockout region within the homology region, whilst the other was designed to anneal outside of the homology region on the other flank.

2.3 Biological transposon mutagenesis libraries

The development of an effective transposon mutagenesis technique in *C. saccharoperbutylacetonicum* is discussed in more detail in Chapter III. Described here is the final method used to produce the libraries. The transposon delivery plasmid pRPF215 was transferred into *C. saccharoperbutylacetonicum* as described in 2.2.13. Transconjugants were selected for on agar containing 75 µg/mL thiamphenicol and 50 µg/mL colistin sulphate after 8 h. The following day, 10 colonies were selected for re-streaking on plates containing the same selective pressure and left to grow overnight. The next day, O/N cultures in RCM without selective pressure were set up for each transconjugant. The next morning, 100 µL of O/N culture were added to 20 x 20 cm plates containing 50 ng/mL ATc and 40 µg/mL erythromycin. Plates were spread using sterile glass beads. The following day, plates were harvested using 1.5 mL RCM with 500 µg/mL erythromycin and L-shaped spreaders. 300 µL from each plate was pooled into a pool for that transconjugant and an overall pool. Once every plate was harvested the pools were mixed and 1 mL taken from each pool for glycerol stocks. The remaining cell volume was frozen at -20°C.

2.4 Phenotypic analyses

2.4.1 Growth analysis

Two main methods of monitoring growth were employed. In both methods, the strain of interest was streaked from glycerol stocks or conjugations and single colonies used to inoculate O/N cultures in RCM. The OD_{600nm} of the culture was determined and the strain subcultured into the test condition to OD_{600nm} 0.05 unless otherwise stated. The two methods are differentiated by how OD

was subsequently monitored. In the first, growing culture was transferred to cuvettes, the OD_{600nm} spectrophotometer calibrated with the growth media and the OD_{600nm} recorded periodically, typically every 45-60 min. Beyond an OD_{600nm} of 1, the spectrophotometer becomes increasingly inaccurate so, when OD_{600nm} approached 1, samples were subsequently diluted in the growth media prior to measurement. The dilution factor depending on the probable OD with lower dilution factors preferred to reduce the error rate.

The second method utilised a Stratus automated plate reader (Cerillo). An optically clear 300 µL 96 well plate (Thermo) was used. The lid was treated with 0.05% Triton X in 20% v/v ethanol to prevent condensation. Media was supplemented with Antifoam (Sigma) to a final concentration of 0.0023% to prevent problematic foaming when shaken (this concentration is regularly used by Biocleave in fermentations). 196 µL of each test condition was aliquoted onto the plate. The optical density at 600 nm of the O/N was recorded and 4 µL of overnight added into each well. For continuous reads, the plate reader must blank based on the initial OD_{600nm} of the wells. The plate could not be removed after this point without disrupting the measurement process and so the blank OD_{600nm} value for all wells included the initial subculture OD_{600nm}. The plate reader than took OD_{600nm} measurements every 3 min. The raw numbers generated with this method are therefore not directly comparable to that generated by cuvette measurements. However, the growth rate should remain unchanged and is comparable between conditions using this method.

2.4.2 Quantifying colony forming units

Colony forming units (CFU) per mL is used as a definition of a single bacterium (or pair etc depending on species). It is a measure based on the counting of colonies derived from a single source e.g., a growing culture. To obtain the CFU value at different optical densities, a cuvette-based growth curve in RCM was set up as described in 2.4. 1.. OD_{600nm} was measured every 45 min. Starting at 45 min post-inoculation and every 2 h subsequently, samples were taken for CFU counts. For these counts, a serial 10-fold dilution in a 200 µL volume of reduced PBS was conducted for each sample. Depending on the OD, a range of 3 dilutions were chosen for spotting non-selective RCM agar plates with 4 x 10 µL of each dilution. The dilution range was selected with the aim of yielding 10-50 colonies/spot for the middle dilution. This was estimated based on preliminary tests. Plates were left to grow overnight before being removed and the number of colonies counted. The CFU/mL was then calculated using the following formula (x100 accounts for the scaling of a 10 µL spot to final mL volume):

$$CFU\ per\ mL = colonies\ counted\ \times\ dilution\ factor\ \times\ 100$$

2.4.3 Colony forming units per colony

In order to estimate the CFUs within a single, overnight colony, a CFU/CFU test was conducted. Wild type *C. saccharoperbutylacetonicum* was streaked from glycerol stocks on non-selective RCM-agar and left to grow overnight. 1 mL pipette tips (Starlab) were cut with scissors approximately 1 cm from the end. The tips were then sterilised by autoclaving at 121°C. Random colonies were selected and removed from the plates by creating a plug from the agar using the cut tips. The agar slice containing the colony was then transferred to a microcentrifuge tube containing 300 µL of RCM. The sample was vortexed for 1 min to separate the cells within the colony. This 1:300 dilution was taken as the baseline condition. A serial dilution as described for 2.4.2 was then conducted, 10^{-1}, 10^{-2} and 10^{-3} dilutions spotted with 4 x 10 µL volumes onto non-selective RCM agar, and the plates left to grow overnight. Colonies were then counted, and the CFU/CFU/mL was calculated as below (x300 accounts for the initial dilution of the colony):

$$CFU\ per\ mL = colonies\ counted\ \times\ dilution\ factor\ \times\ 300 \times\ 100$$

2.4.4 Transposition frequency

The pRPF215 transposon delivery system requires the induction of the expression of the mariner Himar1 transposase through the repression of the TetR repressor using ATc. To accurately determine the scale of experiments required to obtain a large range of transposon mutants, it was necessary to first calculate the rate of transposition upon induction. Cultures of *C. saccharoperbutylacetonicum* were serially diluted as previously described in 2.4.2. 4 x 10 µL of undiluted, 10^{-1} and 10^{-2} dilutions were spotted onto RCM containing ATc (variations tried) and 40 µg/mL erythromycin. 4 x 10 µL of 10^{-5}, 10^{-6} and 10^{-7} dilutions were also spotted onto non-selective RCM plates. Finally, 100 µL of undiluted culture was spread on ATc and 75 µg/mL thiamphenicol plates which should indicate the rate of clearance of the plasmid upon induction (see Chapter III for details). Plates were incubated overnight, and colonies counted the following morning. The number of CFU/mL for all conditions was calculated as in 2.4.2 and the transposition rate was taken as the average of the counts for the ATc/erythromycin plates divided by the counts on non-selective plates.

2.4.5 Optimisation of cryopreservation

This assay was designed to investigate the impact of freezing *C. saccharoperbutylacetonicum* in different cryopreservants on the survival rate. O/N cultures of *C. saccharoperbutylacetonicum* were serially diluted and 4 x 10 µL of dilutions 10^{-5}, 10^{-6} and 10^{-7} were spotted onto non-selective RCM agar. OD_{600nm} was recorded and 1 mL of cells was added to 2 mL cryo tubes (Sarstedt) prepared with candidate cryopreservants. These tubes were then stored at -80°C overnight, thawed, and surviving CFUs determined as before. Colonies of pre- and post-freezing cells were counted, and the CFU/mL calculated as in 2.4.2.

2.4.6 Sporulation efficiency

To test for sporulation efficiency as a proportion of total CFU/mL at OD_{600nm} 1, spores were prepared on either liquid or solid media as described in 2.1.2., At the desired timepoint, cells were removed either by pipetting or by harvesting from plates using 1 mL of reduced PBS and an L-shaped spreader. The OD_{600nm} was measured, and all samples were adjusted to OD_{600nm} 1 in PBS. Samples were split into tubes to be heated and controls to be left at RT. The test samples were heated to either 60°C, 70°C or 80°C for 15 min in a water bath. All samples were then serially diluted and 4 x 10 µL of the undiluted, 10^{-1} and 10^{-2} dilutions were spotted onto non-selective RCM. For the RT controls, 4 x 10 µL of 10^{-4}, 10^{-5} and 10^{-6} dilutions were spotted. Cells were left to grow overnight, and the colonies counted. The CFU/mL was calculated as described in 2.4.2.

2.4.7 Germination assays

Assay tubes were prepared containing the background media in addition to the candidate germinant. Spores purified as specified in 2.1.2 were heated to 60°C for 15 min in a water bath and the OD_{600nm} measured. The purified spores were evenly distributed between the assay tubes with a target final OD_{600nm} of between 0.05-0.15. Initial OD_{600nm} was recorded, and the tubes were left overnight (except where otherwise stated in Chapter IV). The following day, OD_{600nm} was measured regularly and brightfield microscopy at x40 magnification was occasionally conducted to monitor for contamination. The duration of the measurement period depended on delay in observing growth for each assay tube, but all assays were halted after 48 h.

2.4.8 Determination of promoter strength

Luciferase assays were employed as readout for promoter strength using the system described by Klass Smits (Oliveira Paiva et al., 2016). Briefly, promoters of interest were amplified by PCR from the *C. saccharoperbutylacetonicum* genome and megaplasmid using primers containing KpnI and SacI overhangs. These were cloned by restriction ligation into the plasmid containing the *BitLucOpt* gene and transferred into *C. saccharoperbutylacetonicum* by conjugative transfer. O/N cultures in RCM of the strains containing the plasmid and W/T controls were subcultured into RCM and grown to exponential phase. OD_{600nm} was measured and samples adjusted to OD_{600nm} 0.5. The assay kit (Promega) was thawed, and the assay reagent prepared as per the manufacturer's instructions. 50 µL of cells were added to 50 µL of assay reagent in a white 96 well plate (Starlab) and incubated at RT for 15 min. Samples were assayed on a Hidex Sense using the programmed luminescence settings and again 15 min later.

2.4.9 Bottle screens

To understand the growth kinetics of *C. saccharoperbutylacetonicum* strains, bottle screens were conducted. O/N of the test strains were grown in RCM supplemented with 500 µg/mL erythromycin when appropriate. These were subcultured using a 10% inoculum into 20 mL TYIR 50 mM MES 50 g/L glucose and grown for 24 h. A 10% inoculum was again subcultured into 30 mL screen bottles containing the same media. pH and OD_{600nm} was recorded, and 1 mL of culture transferred to a sterile 1.5 mL microcentrifuge tube. The cells were spun at 10,000 x g for 10 min, the supernatant transferred to a fresh 1.5 mL microcentrifuge tube and frozen at -20°C prior to analysis to determine the concentration of relevant solvents, acids, and glucose. OD_{600nm} and pH were recorded every hour during the first 6 h to monitor for healthy growth and then at 24 h, 30 h and 48 h. x40 brightfield microscopy was also occasionally conducted to monitor the health of the culture. Samples for sugar, solvent and acid analysis were taken at 0 h, 6 h, 24 h, 30 h and 48 h.

2.4.10 HPLC sugar, solvent and acid analyses

Sugar, solvent and acid concentrations were determined using HPLC by Dr Victoria Green and Sasha Atmadjaja at Biocleave using their proprietary methods. Chromeleon software was used to process the data.

2.4.11 High density quorum sensing

Cells were grown O/N and subcultured to OD_{600nm} 0.01 in 50 mL RCM. These were left to grow to OD_{600nm} 0.3. Cultures were then centrifuged at 4000 x *g* for 10 min, the supernatant discard and cells resuspended in either 50 mL RCM (control) or 10 mL (test). OD_{600nm} was monitored for 7 h, samples taken for potential solvent analysis and brightfield microscopy used to assess the health and growth phase of the culture.

2.5 Microscopy

2.5.1 Brightfield microscopy

Cultures were monitored for healthy appearance under brightfield light microscopy. 5 µL of cells were added to a microscope slide and a cover slip added. Cells were viewed under either using the x40 air lens or a x100 oil emersion lens. For the latter, a drop of immersion oil was added to the cover slip prior to use. Signs of a healthy culture were taken to be rod-shaped, motile, single cells during the logarithmic growth phase and rod-shaped, largely non-motile, cells clustered in pairs or chains. Cultures were also examined for the absence of signs of contamination from other species.

2.5.2 Cell fixation

Cells were fixed for epifluorescence, phase contrast and electron microscopy. 120 µL of fixation solution containing 100 µL of 16% paraformaldehyde, 20 µL of 1 M $NaPO_4$, 2 µL of 50% glutaraldehyde and 2 µL reverse osmosis H_2O was added to 1 mL of OD_{600nm} ~1 cells and incubated for 30 min in the anaerobic cabinet. Cells were removed from the anaerobic cabinet, incubated for 15 min on ice and centrifuged for 2 min at 5000 x g. The supernatant was removed and 500 µL of tris-buffered saline (TBS) used to suspend the cells. This step was repeated 3 times before a final suspension in 30 µL TBS. Fixed samples were stored at 4°C until use.

2.5.3 Epifluorescence and Phase contrast

For epifluorescence, cells were stained post-fixation with Nile Red (Thermo Fisher). 0.5 µL of Nile Red was added to the fixed sample and incubated in the dark on a rotary shaker at RT for 15 min. Samples were centrifuged at 5000 x g and resuspended in 30 µL TBS. 10 µL of sample was pipetted onto a glass microscope slide and dried. Excess sample was washed with reverse osmosis H_2O and dried as before. 5 µL of Diamond SlowFade (Thermo Fisher) was used to mount the cover slip and

clear nail polish used to seal the edges of the cover slip. Slides were stored in the dark at 4°C until imaging.

For phase contrast, fixed cells were dried onto glass microscope slides as described above. Cover slipes were mounted and sealed as before. Both epifluorescence and phase contrast samples were imaged using a Nikon Widefield Ti Eclipse inverted microscope at x100 magnification. ImageJ was used to process the images.

2.5.4 Thin-section transmission electron microscopy

Samples fixed in TBS were prepared for thin-section TEM by Christopher Hill in the Electron Microscopy facility in the School of Biosciences at The University of Sheffield.

2.6 Bioinformatics

Geneious v7.1.9 was used for routine DNA sequence analysis, sequence alignment and *in silico* cloning. Primer annealing temperature and hypothesised secondary structures were analysed using the Eurofins tool Oligo Analysis tool (https://eurofinsgenomics.eu/en/ecom/tools/oligo-analysis/). NEBuilder (nebuilder.neb.com) was used to design primers for Gibson assembly. SoftBerry BPROM (Solovyev, 2011) was used to predict promoters for characterisation using Luciferase assays (2.4.8) and during the analysis of target gene sequences. The analysis of TraDIS libraries was conducted by Dr Roy Chaudhuri at The University of Sheffield. The method is discussed in more detail in Chapter III but, briefly, sequences were sorted by sequence tag and trimmed. The output was used as input for the Bio-TraDIS package (https://github.com/sanger-pathogens/Bio-Tradis). The sequences were then visualised using Artemis and Dalliance software. Gene enrichment analysis was conducted by the author utilising the web program ShinyGO 0.76 (Ge et al., 2020) with the following settings: a false discovery rate (FDR) cut-off of 0.05; a minimum pathway size of 2 genes; redundancy removed; the gene list run against the STRING_db *C. saccharoperbutylacetonicum*N1-4(HMT) annotated genome; and the results sorted by best average FDR and fold enrichment. MUSCLE was used to align promoter sequences (Rice et al., 2000) and visualised using Jalview (Waterhouse et al., 2009).

2.7 Statistical analyses

Graphpad Prism 9 (GraphPad Software Inc) and Microsoft Excel were used to conduct routine statistical analyses using their respective in-built functions. Statistically analyses included two-tailed t-tests were conducted using Welch's correction, Brown-Forsythe and Welch ANOVAs, nonlinear regression analyses and Comparison of Fits analyses.

Chapter III – Developing plasmid-based transposon mutagenesis in *C. saccharoperbutylacetonicum*

3.1 Introduction

This chapter focuses on the development of a method for random mutagenesis and transposon-directed insertion site sequencing in *Clostridium saccharoperbutylacetonicum*. This would allow for the identification of an essential genome for the species which could then be probed under a variety of conditions commonly used in the field, for example fermentation conditions. To enable this process, elements were imported and adapted from already-established protocols. This can be broadly separated into the creation of biological libraries of transposon mutants and the DNA sequence libraries used for the sequencing process. Multiple steps of each process were optimised to account for the fundamental biology of *C. saccharoperbutylacetonicum*.

The transposon delivery system created by Dembek et al., 2015 was selected for use in this study. This is a plasmid-based system that was previously used in *Clostridiodes difficile* for this purpose. Phylogenetically, *C. saccharoperbutylacetonicum* and *C. difficile* are not especially closely related, however they do share fundamental characteristics and limitations that make the use of this plasmid-based system attractive. In particular, the resistance to electroporation means that many traditional methods of transposon mutagenesis are not available to either species. Whilst, in contrast to *C. difficile*, it is possible to electroporate *C. saccharoperbutylacetonicum*, the efficiency can be inconsistent, and scale up dependent on the equipment available, particularly in terms of anaerobic workspace. The primary issue with electroporation, however, is that *C. saccharoperbutylacetonicum* has been shown to resist the transfer of linear fragments of DNA (Personal Communication, Elizabeth Jenkinson). Most transposon mutagenesis systems make use of electroporation to transform linear transposons with transposases attached (or pre-transposed genomes in the case of *Streptococcus pneumoniae*), which evidently can't be used in this case (Cain et al., 2020) .

The pRPF215 delivery plasmid created by Dembek et al. therefore offers several advantages over electroporation in this context. To understand these, it is necessary to briefly cover how the system functions. The plasmid consists of several components key to its specific application – namely a custom transposon, inducible transposase, and a conditional replicon (Figure 3.1). The transposon is the mobile genetic element that will insert into the genome and consists of an erythromycin resistance gene flanked by identical inverted terminal repeats (ITRs). The gene allows for the selection of mutants whilst the ITRs are necessary to identify the sequence for transposition.

The Himar-1 mariner transposase is responsible for the transposition event, and its expression is tightly regulated by a promoter containing the tetracycline operator sequence. The Himar-1 mariner transposase catalyses the insertion of the transposon at 5'-AT-3' sites, ideal for an AT-rich genome such as *C. saccharoperbutylacetonicum*. Another tetracycline-inducible promoter controls the expression of the tetracycline repressor, which, in turn, regulates both inducible promoters. The latter is designed to prevent unwanted expression of the transposase promoter resulting from random de-repression – a known flaw of the tetracycline repressor. The final feature of the system is the use of a novel method of introducing plasmid instability. The *tetR* gene is located upstream of the plasmid origin of replication. With no terminator sequence at the end of the gene, upon expression of *tetR*, the transcriptional machinery is allowed to continue into the origin of replication. In *C. difficile*, this is sufficient to interrupt replication of the plasmid resulting in segregational instability whereby, after a small number of generations, there will no longer be sufficient copies of the plasmid to pass onto daughter cells. In this way, the plasmid is removed from the population, and it is possible to distinguish mutants from cells just carrying the plasmid. As a result, transposons that have remained on the plasmid (the vast majority) are not sequenced.

Previously, TraDIS in *C. difficile* was carried out in collaboration with the Wellcome Sanger Institute for processing and sequencing (Dembek et al., 2015). To allow greater control of the individual steps, which can be extremely useful in adding flexibility, and adaptability to problems or novel scenarios that may arise, an in-house method of processing extracted gDNA into libraries that could be sequenced by Illumina methods was established.

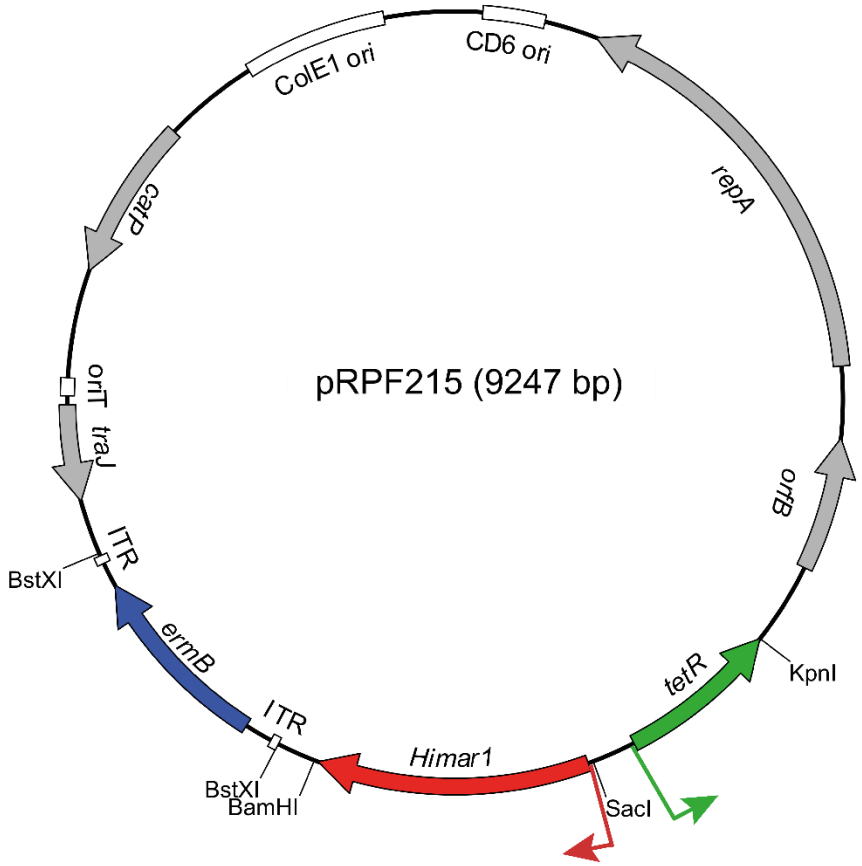

Figure 3.1 The plasmid of the pRPF215 transposon delivery system. Highlighted in colours are features integral to the transposon mutagenesis process. From left to right: In blue is the *ermB* erythromycin resistance gene flanked by inverted terminal repeats (ITR) specific for the *Himar1*/mariner transposase; in red is the *himar1* mariner transposase gene whose expression is driven by a promoter containing the *tetO* sequence (small L-shaped red arrow) which is recognised by the tetracycline repressor; in green is the *tetR* tetracycline repressor gene whose expression is driven by another promoter containing the *tetO* (small l-shaped green arrow). After the *tetR* there is no terminator sequence allowing transcription to run into the replicative machinery starting with *orfB*, continuining into *repA* and concluding with the pCD6 *C. difficile* origin of replication. ColE1 ori is the origin of replication utilised by *E. coli*. *catP* encodes chloramphenicol acetyltransferase which confers resistance to both chloramphenicol and thiamphenicol through the same mechanism: the attachment of an acetyl group to the antibiotic. oriT and *traJ* are required for conjugative transfer of the plasmid from *E. coli* into *Clostridial* strains.

3.1.1 Aims and objectives

The aims and objectives of this chapter were as follows:

1. Establish basic growth profiles of *C. saccharoperbutylacetonicum* under conditions relevant to creating transposon mutants.
2. Test the viability of pRPF215 in *C. saccharoperbutylacetonicum* as a method of creating transposon mutants.
3. Establish a protocol for the processing of gDNA from biological transposon libraries into sequences ready for massively parallel Illumina sequencing.
4. Sequence libraries using Illumina protocols and troubleshoot any issues.
5. Generate an essential gene list for *C. saccharoperbutylacetonicum* under laboratory conditions.
6. Test the library of mutants under different conditions that are of interest to the industrial microbiology field and generate new essential gene lists in those contexts.

3.2 Growth profiles of C. saccharoperbutylacetonicum

3.2.1 RCM

A 7 h growth curve was conducted to establish the basic growth dynamics of *C. saccharoperbutylacetonicum* in RCM during typical growth period (Figure 3.2A). A nonlinear regression analysis was conducted, based on the estimated exponential phase, which indicated a doubling time of 61.1 min.

3.2.2 In anhydrotetracycline

As ATc is used to induced transposition in the pRPF215 system, it was necessary to establish whether ATc has any impact on *C. saccharoperbutylacetonicum* growth. Cultures in RCM, and RCM supplemented with 50 ng/mL of anhydrotetracycline (ATc) were compared, starting at an inoculated $OD_{600\ nm}$ of 0.1, and monitored for 7 h (Figure 3.2B). Nonlinear regression analysis of the estimated exponential phases indicated a doubling time of 62 min and 67.7 min for RCM and RCM+ATc respectively, a difference that was considered significant by Comparison of Fits ($p=<0.0001$). A longer lag period before exponential growth was also observed for cultures grown with ATc, likely due to a higher drop in OD_{600nm} at the first timepoints. The addition of ATc therefore does significantly impact the growth of *C. saccharoperbutylacetonicum* compared to control conditions and could present a bottleneck in the transposon mutagenesis experiments.

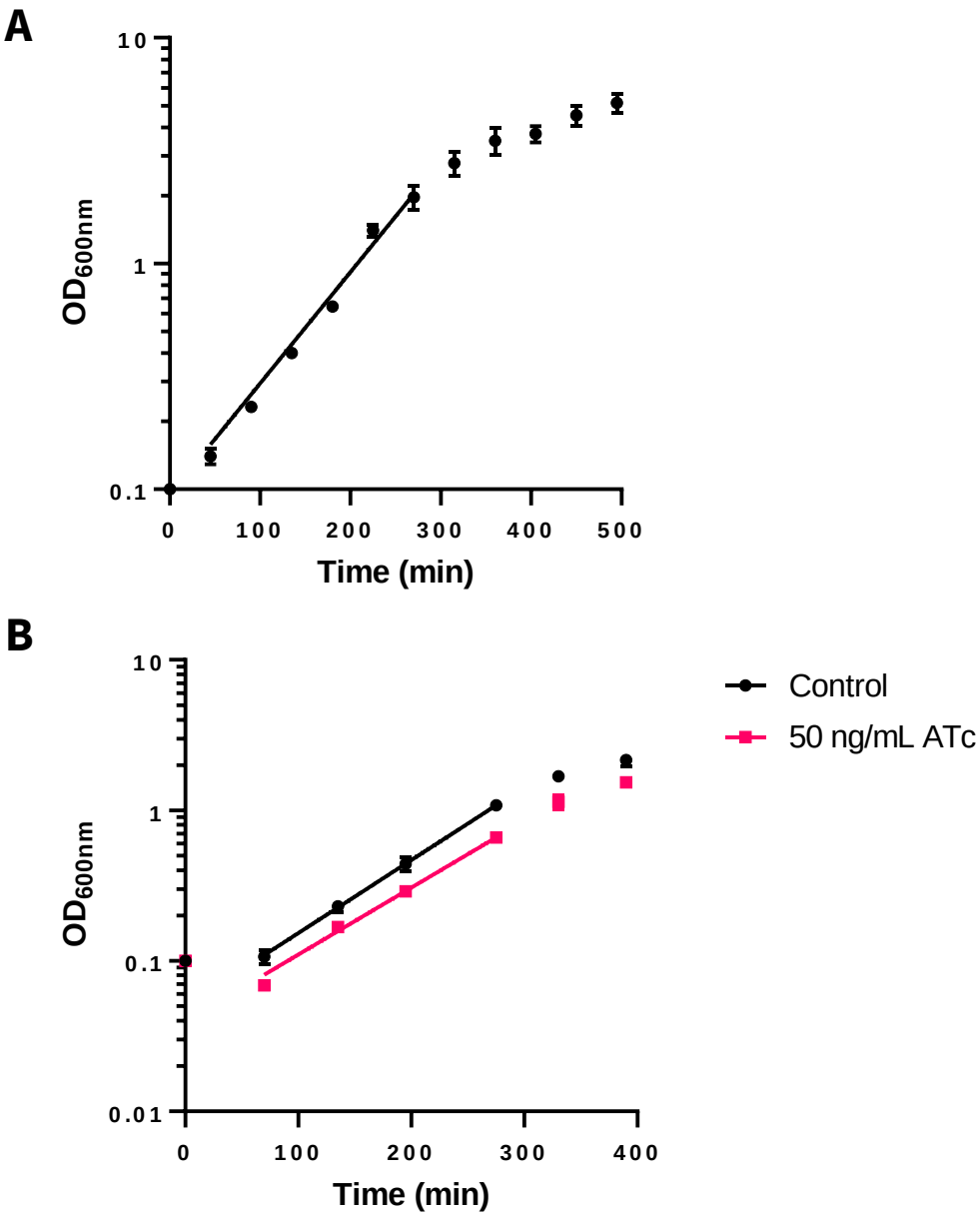

Figure 3.2 Growth of *C. saccharoperbutylacetonicum* in RCM and RCM supplemented with ATc. A) The growth of *C. saccharoperbutylacetonicum* in RCM over 8 h. The line represents a non-linear regression analysis to measure the growth rate during exponential phase. **B)** The growth of *C. saccharoperbutylacetonicum* in RCM (black) compared to RCM supplemented with 50 ng/mL ATc (magenta). The lines represent non-linear regression analyses to determine the growth rates of 62.05 min doubling time for the control compared with 67.69 min for the ATc condition, a difference which was considered significant (P=<0.0001) by comparison of fits. Growth in ATc exhibits a drop in initial OD_{600nm}. Unpaired t tests with Welch's correction at 135 min and 390 min showed significant difference between the condition at both timepoints (both P=<0.0001). Assays were performed on biological duplicates (A) or triplicates (B) and technical triplicates.

3.3 Colony forming units per mL

3.3.1 CFU/mL per $OD_{600\ nm}$

To correctly calculate the scale necessary during the creation of the biological transposon libraries, it was essential to understand the relationship between CFU/mL and $OD_{600\ nm}$ for this species. While characterising growth in 3.2.1, CFUs were also enumerated at each timepoint. The results indicate the expected rise of CFU/mL in line with the rise in $OD_{600\ nm}$ (Figure 3.3). Between OD_{600nm} 0.34 and 1.7, the CFU/mL rises exponentially from 5.9×10^5 to 3.5×10^7. Surprisingly, the CFU/mL then remains stable until OD_{600nm} 3.5. Despite the OD_{600nm} continuing to rise, the associated CFU/mL oscillates within 10^8 range. This likely reflects a combination of factors at play during stationary phase. It is possible that cells may be growing in size (comparatively larger cells were observed in stationary phase microscopy, as compared to exponentially growing cells, data not shown) but not dividing. It is also possible that cells are dying but not lysing so any new growth still adds to the culture turbidity but not to CFU/mL. Finally, stationary phase cells begin to form chains rather than remaining as single cells. It's possible that these chains contribute to a higher culture turbidity than the individual cells would otherwise. Following OD_{600nm} 3.5, there is a second increase in CFU/mL until OD_{600nm} 4 up to 4×10^8. This possibly reflects the biphasic growth profile of solventogenic *Clostridia*, with the production of solvents from acids allowing a second, shorter and slower growth period. The data presented is more variable than first thought once OD_{600nm} standard deviations were also plotted. Unfortunately, these standard deviations were not considered for analysis until this thesis was being written and so repeats were not possible.

3.3.2 CFU/colony

The later biological transposon libraries (see 3.7 and 3.10) were to be generated on agar plates, where each individual colony represented a potential transposon insertion mutant. Understanding the average CFU per colony could therefore enlighten as to copy number of each mutant. This information was necessary to determine if an individual mutant would be captured by the gDNA extraction process and what amount of a pooled biological library would be statistically representative of the entire library. Eight random colonies were taken from *C. saccharoperbutylacetonicum* streaked onto RCM agar and treated as described in 2.4.3. A single colony contains an average of 1.7×10^7 CFU/mL after approximately 17 h growth.

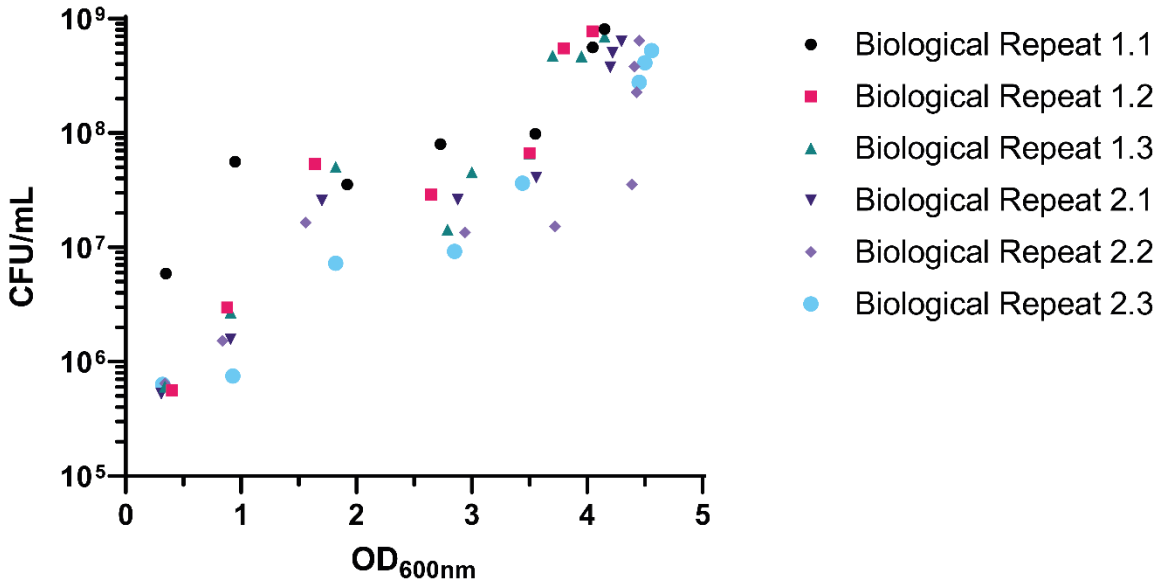

Figure 3.3 CFU against OD_{600nm} of *C. saccharoperbutylacetonicum* grown in RCM. The change in CFU/mL as OD_{600nm} changes with growth. Growth was monitored for 13 h to reach OD_{600nm} >4. The growth shows an increase in CFU/mL during as OD_{600nm} associated with logarithmic growth increases. CFU/mL then changes little during the OD_{600nm} associated with stationary phase before increasing again at OD_{600nm} >3.5. Each data point shows the CFU/mL and OD_{600nm} of a single repeat so as to reduce the confusion associated with error bars on both the X and Y-axis and to better demonstrate the general trends that are repeated across all repeats, despite different absolute values. The assay was performed with biological duplicates and technical triplicates.

3.4 Transposition frequency

The induction of the pRPF215 system with ATc results in the expression of the *Himar1* mariner transposase and therefore the initiation of transposition. However, it was necessary to understand the frequency of transposition in the population to correctly plan the scale of biological libraries. Both logarithmic and stationary cultures were considered for library construction and transposition frequency was determined accordingly. Cultures were grown on ATc/erythromycin, allowing enumeration of transposon mutants, and under non-selective conditions to quantify the background CFU/mL of the culture. Growth on ATc/thiamphenicol was presumed to be lethal as the induction by ATc should introduce instability in the plasmid, whilst thiamphenicol should select exclusively for the plasmid.

3.4.1 Logarithmic cultures

Subcultures of *C. saccharoperbutylacetonicum* were grown in non-selective RCM to $OD_{600\ nm}$ 0.5, serial dilutions performed, and plated on the three conditions (Figure 3.4A). The overall transposition efficiency was calculated to be 1.57×10^{-3} indicating a transposition event occurs in 1.57 of every 1000 cells on average. For the ATc/thiamphenicol condition, very small colonies were observed. However, due to their extremely small size, they could not be readily counted. Since the induction pRPF215 system only causes segregationally instability, it was hypothesised that these small colonies were the maximum extent of growth capable before there were no longer any copies of the plasmid to transmit from mother to daughter cells.

3.4.2 Stationary cultures

Cultures were grown O/N in non-selective RCM, the OD_{600nm} recorded, serial dilutions performed and plated onto the three conditions (Figure 3.4B). The transposition frequency of 1.22×10^{-5} was much lower than for logarithmic cultures, though it is compensated for by the higher overall CFU/mL in the cultures. Unlike for the logarithmic cultures, countable colonies were seen for the ATc/thiamphenicol plates and recorded as such. The appearance of these colonies could be due either to trouble removing the plasmid, or to early transposition. However, at the time of conducting

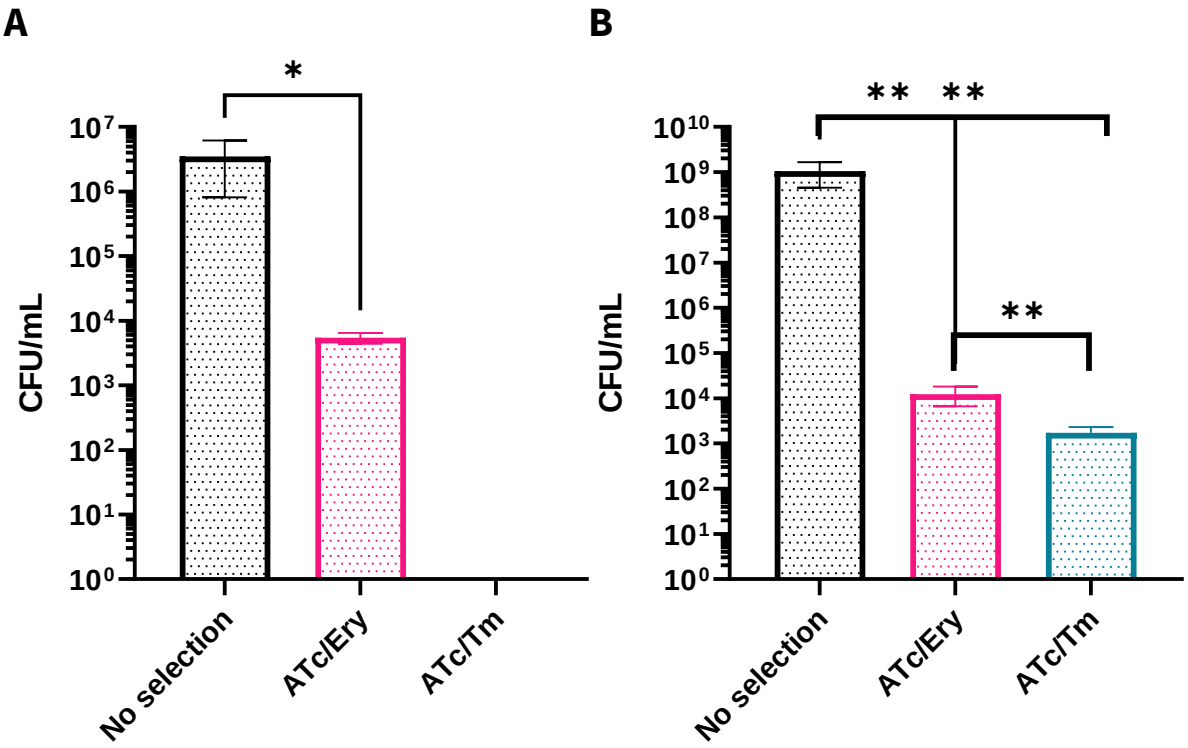

Figure 3.4 The transposition frequency of pRPF215 in *C. saccharoperbutylacetonicum*. Graphs showing the CFU/mL for no selection vs ATc/Ery vs ATc/Tm from **A)** logarithmic cultures plated at OD_{600nm} 0.5. An unpaired t-test with Welch's correction showed a significant difference between the no selection and ATc/Ery CFU/mL (P=0.0243). The ATc/Tm condition did show some colonies, but they were too small to count, and thus not included. The experiment was performed in biological duplicate with technical triplicates. **B)** CFU/mL from overnight cultures plated at an average OD_{600nm} 3.6. Unpaired t-tests with Welch's correction were conducted between each pair. There was a significant difference between both no selection vs ATc/Ery (P=0.008) and vs ATc/Tm (P=0.008). There was also a significant difference between ATc/Ery and ATc/Tm (P=0.0058). The experiment was performed in biological duplicate with technical triplicates

the experiment, I believed the issue to be with plasmid stability due to our long-term problems with plasmid persistence in the induced cultures.[1]

3.5 Effect of cryopreservation on cell viability

Throughout this study, *C. saccharoperbutylacetonicum* cultures were stored in 15% v/v glycerol at -80°C. Whilst this proved sufficient to be able to reliably recover pure strains, it was not known if storage in these conditions reduced cell viability in a way that might result in the loss of rarer mutants. Given that the transposon libraries were to be stored in this manner prior to later exposure to selective conditions, it was important to establish whether the cryopreservation of the species represented a selective condition that might lead to the creation of a bottleneck resulting in the stochastic loss of certain transposon mutants. To test this, O/N cultures, of known CFU/ml, were frozen in cryopreservant at -80°C O/N. The following day, cultures were thawed, and surviving CFUs counted again (Figure 3.5) and compared to the O/N culture CFU/mL (the 'Control'). Naturally, 15% v/v glycerol was tested, but also selected were two further concentrations of glycerol – 10% and 20 % v/v. These glycerol concentrations were chosen to probe whether the 'lab standard', 15% condition, is an optimal cryopreservant for *C. saccharoperbutylacetonicum*. Three concentrations of DMSO – 5%, 7% and 8.7% v/v were also tested, based on the generic storage concentrations used by ATCC (American Type Culture Collection, 2021).

The results suggested there was a 53.6% drop in CFU/mL for the routine condition of 15% v/v glycerol. 20% glycerol showed a similar drop, falling 56.4% from 2.69 x 10^9 CFU/mL to 1.17 x 10^9 CFU/mL, while 10% showed an even greater 67.7% drop in viability from 2.69 x 10^9 CFU/mL to 8.15 x 10^8 CFU/mL. The viability was greater for the 7% and 8.7% DMSO, showing a 28.7% decrease and a putative increase in viability respectively. The latter result indicates the natural variability in such an experiment. The consistency in the order of magnitude is potentially more meaningful than the number within that order of magnitude. 5% DMSO showed an 88.7% drop off in cell viability. Whilst DMSO performed better at both 7 and 8.7% than the best two glycerol conditions, the 8.7% results, suggesting an increase in viability, gives reason to doubt the accuracy of these results. It is clear from this experiment that cells will largely remain viable after freezing, however, it's difficult to decisively conclude which of the best conditions is preferable.

[1] This experiment was carried out 2 years after the logarithmic culture experiments and therefore after the plasmid work described later in 3.8.

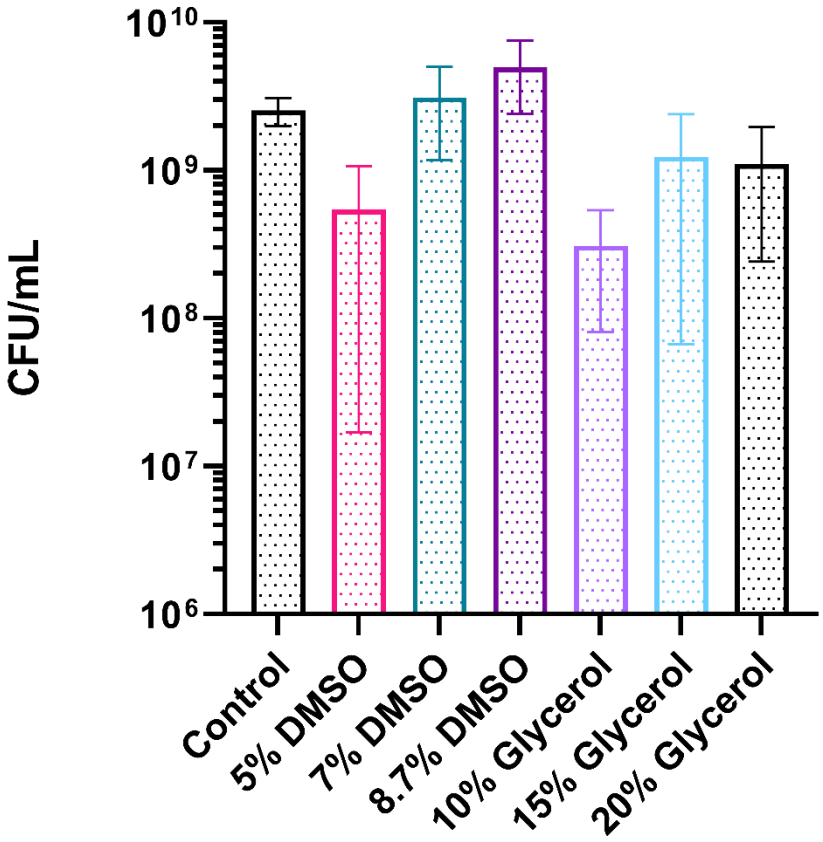

Figure 3.5 The survival of *C. saccharoperbutylacetonicum* at -80°C with different cryopreservants. The graph plots the mean CFU/mL of control (non-frozen) *C. saccharoperbutylacetonicum* along with the CFU/mL after freezing in different concentrations of DMSO and glycerol. The relative imprecise nature of the experiment can be seen in the high standard deviation of each condition (plotted as error bars), particularly for 5% DMSO and 15% glycerol. However, outside 5% DMSO and 10% glycerol, no one other condition presents as particularly better than the others. A Brown-Forsythe and Welch ANOVA showed a significant between all conditions (5% DMSO P=<0.0001; 8.7% DMSO P=0.0401; 10% glycerol P=<0.0001; 15% glycerol P=0.0082; 20% glycerol P=0.0002) and the control except for 7% DMSO (P=0.8420). All experiments were conducted in biological triplicate and technical quadruplicate.

Given that glycerol has long been the cryopreservation chemical of choice, I elected to continue its use for the biological transposon libraries, despite potentially slightly poorer performance. Glycerol does also have a practical secondary advantage over DMSO. By reducing the melting temperature of the cell suspension, glycerol stocks do not need to be completely thawed before they can be used to streak out cells. Conversely DMSO will raise the melting temperature and therefore the stock would always have to be completely thawed before use. Using glycerol likely makes the stocks more robust in the face of repeated freeze-thaw cycles.

3.6 Liquid broth transposon mutagenesis library

3.6.1 Biological library creation

3.6.1.1 Aims and background

Previous studies, both in *Clostridia* and in other species such as in *E. coli* (e.g. Goodall et al., 2018) and *S. aureus* (Santa Maria et al., 2014) used selection on agar plates for generation of transposon mutants. Typically, such a process involves transforming the cells with a DNA fragment(s) encoding transposon and transposase, or a mix of transposon DNA and transposase protein. These are usually optimised such that any positive transformants are highly likely to be mutants. Plates therefore offer the advantage of a highly controllable rate of mutant generation and the additional advantage of presenting a less competitive growth environment, more permissive of slower growth. As such, they represent less of a bottleneck that may result in the loss of slower growing transposon mutants. However, for anaerobic bacteria, the necessity to maintain an anaerobic environment represents a key limiting factor in the use of agar plates. In anaerobic cabinets, such as those manufactured by Don Whitely, space comes at a premium and it is not uncommon for labs working in the field to have access to only 1 or 2 cabinets (sometimes none). These provide only a limited room for plates, especially when working space is considered. Additionally, this limitation could mean that either all other anaerobic work must be postponed for the duration of the biological library preparation, or that fewer than the maximum number of cabinets can be used for the transposon libraries.

One advantage of using a transposon delivery system based on the induction of expression of a transposase, is the potential to control the moment of transposition. Therefore, introduction of the transposon delivery system and selection of mutants can be separated so it is not necessarily important for this induction to occur on agar plates. Liquid broth cultures reach much higher densities of cells for the volume they occupy in an anaerobic cabinet. Cells carrying the pRPF215 delivery plasmid could be grown to logarithmic phase as in 3.4.1 and induced with the addition of

ATc. Through this method, it should be possible to gain statistically significant libraries with much lower working space. In addition, the setup, induction, and harvesting process would be much less onerous than when using plates. The process from growth through to sequencing is outlined below (Figure 3.6).

3.6.1.2 Method

To generate the broth libraries, cultures of five transconjugants were prepared in 20 mL RCM. After O/N incubation these were then used to inoculate 5 x 480 mL RCM supplemented with erythromycin, allowed to grow to early logarithmic phase (OD_{600nm} 0.2-0.32), induced with ATc and then left until the next morning (~9 h growth). At this starting OD_{600nm} there were approximately 5 x 10^5 CFU/mL. With a transposition frequency of 1.57 x 10^{-3} and a volume of 500 mL, ~392,500 individual mutants per flask were expected. This is significantly higher than what could be achieved on plates. The O/N incubation time period was chosen to ensure that each mutant had grown sufficiently to provide enough material for sequencing, whilst also minimising the time for a competitive bottleneck to develop. Cultures were thoroughly mixed and 20 mL from each flask pooled. 1 mL was then stored as a glycerol stock, and the rest frozen as cell pellets (steps 1-3 in Figure 3.6).

3.6.2 Sequencing library optimisation

Following the pooling of cells, TraDIS requires the extraction of genomic DNA, the selection of transposon insertion sites within the genome, and the addition of the DNA sequences necessary for Illumina sequencing (Figure 3.6 and Figure 3.7). Whilst widely used, there is significant variation in specifics of each step and establishing a protocol for this process required a fine balance between different factors during each step. I collaborated with Professor Ian Henderson (The University of Queensland, Australia), then at the University of Birmingham, to adapt the method developed in their lab for use within our transposon mutagenesis system. As far as possible, I duplicated the details of their protocol, however, developing the method of use in both *C. saccharoperbutylacetonicum,* and *C. difficile* required certain changes and additions. This section describes the overall process and our initial attempts to adapt it for use in *C. saccharoperbutylacetonicum* which is summarised in Figure 3.6. For reference, the method listed in section 2.2.15 is the final method developed.

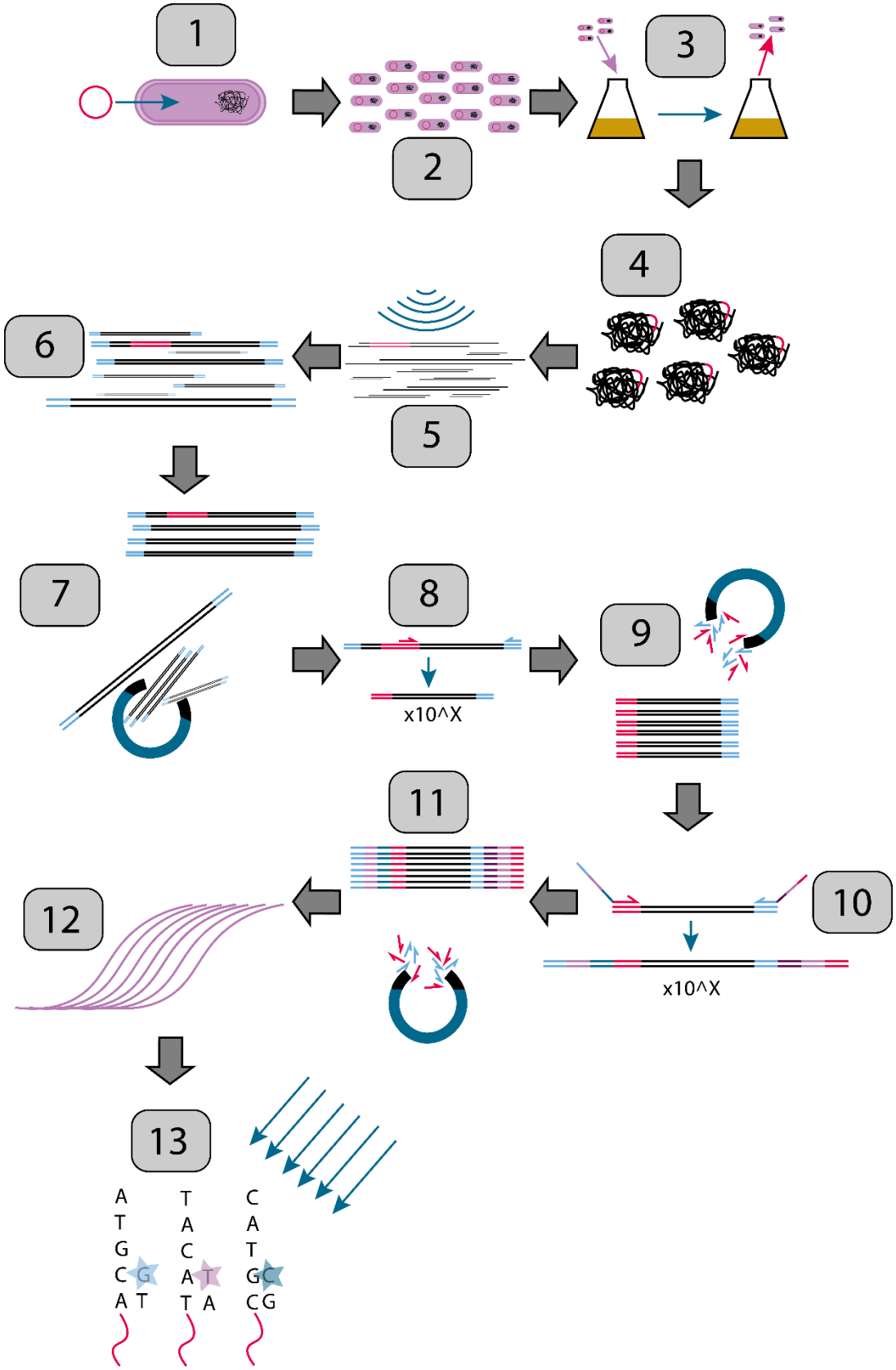

Figure 3.6 Schematic outline of the steps required to create a TraDIS library. The numbers indicate the step number. Images are purely representative and not to scale. 1) Transfer of delivery plasmid into cells. 2) Growth of plasmid-containing cells in pre-transposition conditions to expand

cell numbers. 3) Induction of transposition in broth cultures and removal of transposon mutants. 4) gDNA extraction. 5) Random shearing to an average size of 300-400 bp by sonication. 6) End repair and adaptor ligation. 7) Magnetic bead purification to remove fragments >500 bp and <200 bp. 8) PCR amplification between transposon insertion site and adaptor. 9) Magnetic bead purification to remove primers and small fragments. 10) PCR amplification to add sequences required for Illumina sequencing, barcoding and indexing. 11) magnetic bead purification to remove primers and small fragments. 12) Quantitative PCR to check the concentration and quality of the library. 13) Illumina sequencing by synthesis.

3.6.2.1 Genomic DNA extraction

The gDNA extraction protocol (2.2.1) was used as described throughout (step 4 in Figure 3.6). gDNA integrity was monitored using agarose gel electrophoresis and the purity of the final product was assessed by absorption at 280 nm, 260 nm, and 230 nm. TraDIS protocols require tight control of DNA concentration at multiple different steps, with 1 µg of gDNA recommended as the best starting point and Qubit fluorimetry was employed to ensure accuracy.

3.6.2.2 gDNA shearing

Breaking the gDNA is essential to creating the 20-300 bp fragments required for sequencing and mapping the location of the insertion sites. It is important, as far as possible, not to bias the breaking process to avoid artificially inflating the representation of certain mutants or of missing certain mutants entirely. Shearing by sonication is a useful method for achieving this aim. Sonication can never yield a pure sample of a specific length of DNA due to its inherently stochastic nature. However, multiple standard protocols exist for common commercial sonication equipment (e.g., Covaris) that attempt to specify the mean and mode fragment size. These protocols, however, can vary in their product depending on the source of DNA to be sonicated. Therefore, it was essential to test numerous protocols.

For this protocol fragments that fell largely in the 300-500 bp range were targeted. This was to shear the DNA sufficiently for downstream processing whilst also not losing significant amounts of DNA to small fragments, which were likely to be lost in sequencing (step 5 in Figure 3.6). Test shearing was conducted by the Sheffield Diagnostic Genetics Service at the Sheffield Children's NHS Foundation Trust with the 200, 300 and 400 bp Covaris protocols chosen initially to achieve our desired range. Sheared samples were analysed on a Tapestation capillary electrophoresis system (Figure 3.7) and based on these results, I opted for the 300 bp protocol which gave a good balance between the peak fragment size and the reduction in smaller fragments. Following sonication sample volume was reduced by using vacuum centrifugation, which follows the method used in the Henderson lab.

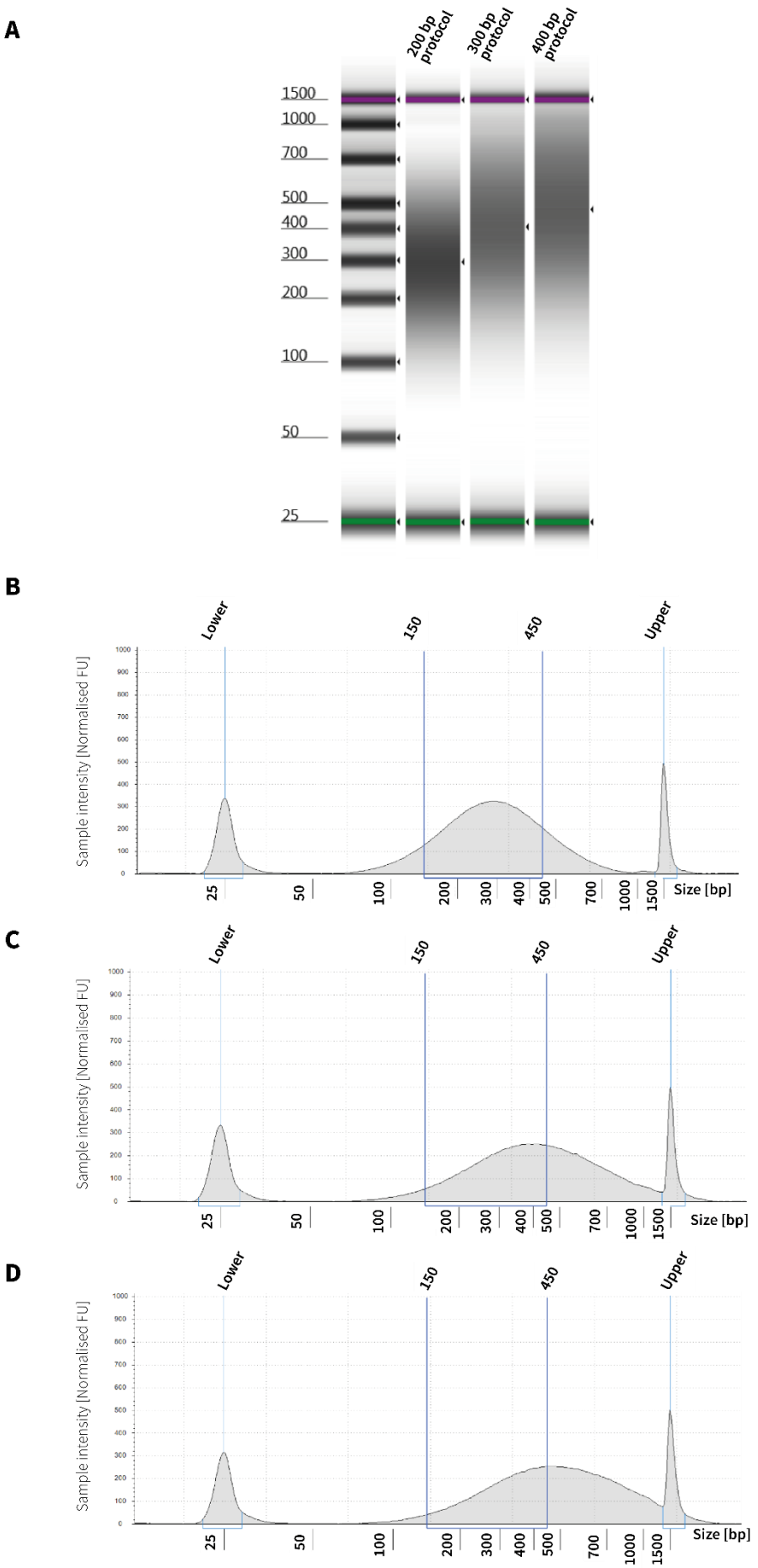

Figure 3.7 The results of shearing using three different Covaris protocols.

A) The overall Tapestation gel generated from the three shearing protocols. Each lane is presented in detail in B)-D) **B)** The size profile of the 200 bp protocol, Lower and Upper refer to the smallest and largest peak. Usable material is found between 150 bp and 450 bp, with a preference given to material falling between 300-450 bp. The peak of the 200 bp protocol is found in between these two bands. However, this protocol yields a greater amount of small fragments. **C)** The size profile of the 300 bp protocol. Again, Lower and Upper refer to the smallest and largest peaks. Usable material is defined as before. The peak of this protocol is closer to the 450 bp mark and there are few fragments below 200 bp compared to B) making it most suited for library processing. **D)** The size profile of the 400 bp profile. Again, Lower and Upper refer to the smallest and largest peaks. Usable material is defined as before. Whilst the 300-450 bp region does contain part of the peak for this protocol, a greater part is found above 450 bp cut off.

3.6.2.3 End repair and adaptor ligation of sheared fragments

Shearing of double stranded DNA by sonication causes breaks that result in overhangs and potential dephosphorylation. Therefore, end repair including phosphorylation and gap-filling is required (step 6 in Figure 3.6). For this, the NEBNext DNA Library Prep Kit (catalogue number: E7370) was used, which is widely employed for Illumina sequencing preparation.

In order to amplify DNA fragments containing the transposon, adaptors must be added to the end of the fragments (Figure 3.7). These contain a universal priming site used by Illumina and are blunt ligated to the ends of the fragments using the enzymes and buffers from the above kit. Following these steps, magnetic bead purification was carried out to select for 300 – 500 bp fragments. All these steps were unaltered from the Henderson lab protocol.

3.6.2.4 PCR to select for transposon junctions

Since whole gDNA is extracted from every mutant and only one transposon site was predicted per mutant, only a small proportion of DNA fragments generated above contain the transposon-genome junction for later analysis. To amplify these fragments a transposon-specific PCR was employed (step 8 in Figure 3.6 and explained in more detailed in Figure 3.8). A 48 bp primer (RF1520; Appendix III) annealing 18 bp inside the mariner inverted terminal repeat (ITR) was designed to pair with the 34 bp NEB adaptor (RF1522). The entirety of the purified product from 3.6.2.3 was used as template. A 10 cycle PCR using was performed using the KAPA HiFi Hot Start mastermix (catalogue number: KK2601). 10 cycles were used to adequately strike the balance between amplifying transposon junctions and not biasing the library based on PCR efficiencies. Fewer cycles risked not adequately amplifying all transposon junctions whilst more cycles would risk the mix being randomly dominated by a handful of mutants based on random differences in PCR efficiency. Following this PCR, a second magnetic bead purification step was conducted to remove primers and any other smaller fragments.

3.6.2.5 PCR to prepare library for Illumina sequencing

This second PCR was designed to add the sequences necessary to a) conduct Illumina sequencing and b) create a custom index allowing for efficient multiplexing and post-sequencing analysis (step 10 in Figure 3.6). The design of these index primers was based on that of the Henderson lab (Figure 3.9; Appendix III). The key difference is in the transposon-specific primer binding site which here was

designed to anneal 10 bp inside the ITR, and nested within the binding site for transposon junction amplification in 3.6.2.4. The primer was designed to anneal as close as possible to the end of the ITR. 10 bp before the end of the ITR was the minimum amount that could be left as these 10 bp were particularly A-T rich and therefore had the potential to reduce PCR efficiency. The reverse primer again bound to the Illumina adaptor added in 3.6.2.3.

The transposon-specific primer also contained the P5 flow cell adaptor, the forward sequencing primer site, and an inline index. The flow cell adaptor enabled binding of the library to the flow cell whilst the forward sequencing primer site was the binding site for the sequencing-by-synthesis Illumina process. These are both widely used across Illumina libraries. The inline index is a custom addition of the Henderson lab. This sequence constitutes an index tag that serves two purposes. Firstly, it can stagger the introduction of the transposon binding site and ITR in the sequencing process. This helps to reduce the chances of flooding the sequencer with a series of identical sequence reads early in the process which can lead to the device being unable to differentiate the clusters from different libraries. This was achieved by tagging each library to be sequenced with a different length index. None of these indices contain the same base at the same point in the sequence. Secondly, the indices allowed multiple libraries to be analysed simultaneously, with the option of analysing the data together or demultiplexing and looking at each data set separately. The reverse primer contained the P7 flow cell adaptor and the reverse sequencing primer site only.

Following this PCR, the library was purified for a final time using magnetic purification.

3.6.2.6 Quantitative PCR to measure library concentration

Despite multiple purification steps, it was inevitable that final library preparations would contain DNA fragments that do not have the sequences necessary for Illumina sequencing. It was not appropriate, therefore, to utilise a DNA quantitation method that measures whole DNA concentration such as spectrophotometry or fluorimetry. The only DNA quantitation method that can accurately measure the concentration of the library is quantitative PCR (step 12 in Figure 3.6). The KAPA Library Quantification kit (catalogue number: KK4824) which comes with prepared standards and a master mix containing the qPCR components and primers annealing to the Illumina P5 and P7 flow cell adaptors was used. The details of the qPCR protocol including the measurement of the standards and the calculations used to quantify the library can be found in Chapter II. The recommended concentration for sequencing is 8 nM and this preparation was adjusted accordingly.

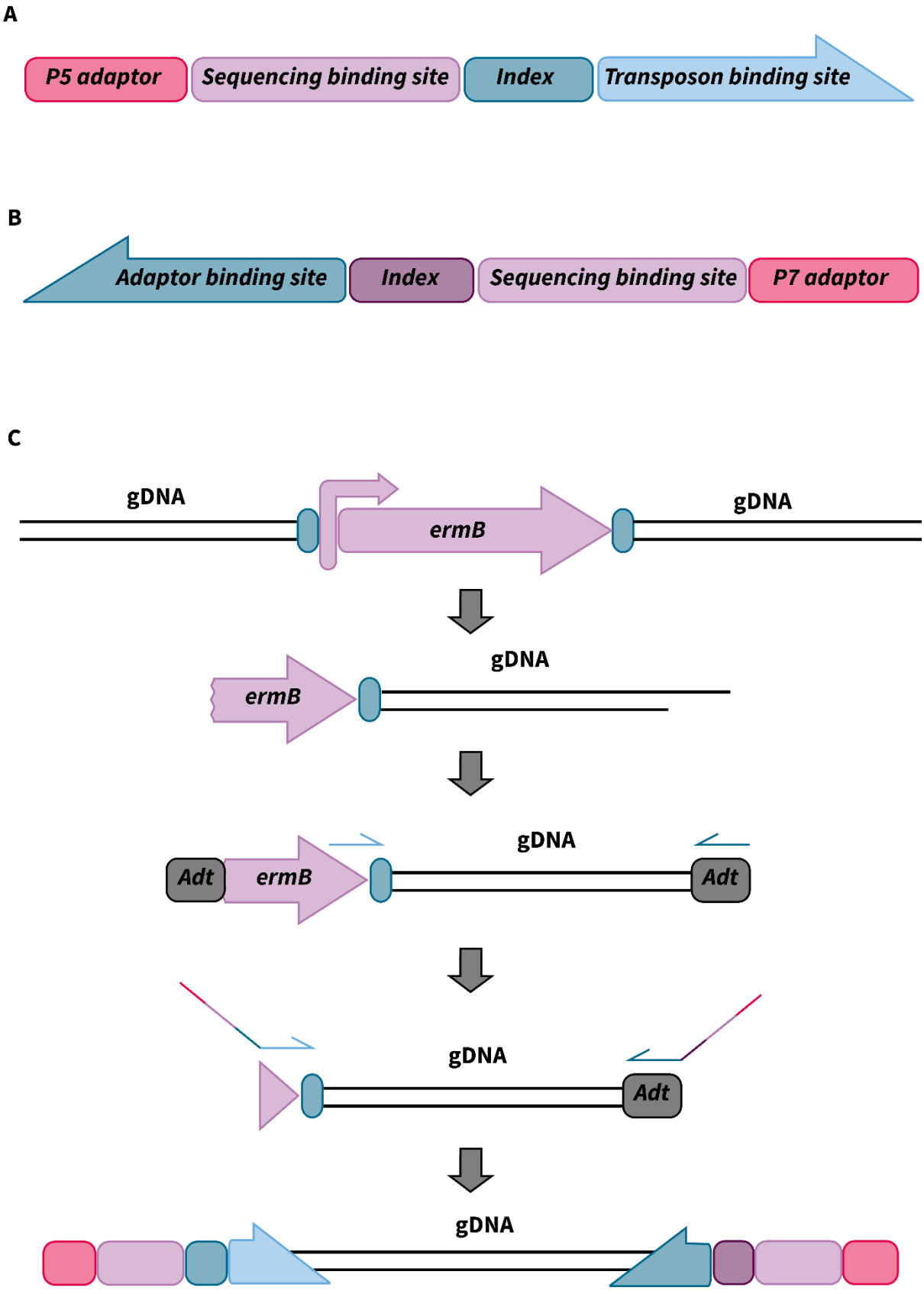

Figure 3.8 Primer design and schematic of library processing. A) The design of the custom primer annealing to the transposon used for the second PCR step (3.6.2.5). P5 adaptor is the flow

cell adaptor required for immobilising the sequence to the Illumina flow cell. The sequencing binding site for use during Illumina sequencing during amplification and sequencing by synthesis. The index is a custom inline index designed to allow further multiplexing of multiple libraries if necessary. A range of eight different indices of varying lengths (between 6-9 bp) were utilised. The transposon binding site is the priming region which anneals to 40 bp across the fdx terminator and 5' ITR of the transposon. **B)** The design of the standard primer utilised to amplify sequences from the adaptor added in 3.6.2.3. P7 adaptor is another flow cell adaptor required to immobilise the sequence to the Illumina flow cell. The sequencing binding site performs the same function as that in A). The index was used as another layer of multiplexing. Its sequence was variable depending on need but was one of those provided by the NEBNext Multiplex Oligos for Illumina kit. The adaptor binding site was designed to bind to the adaptor added to the end of fragments in 3.6.2.3. **C)** A schematic of the processing of gDNA to material that can be multiplexed and sequenced. The first image depicts a typical transposon insertion within the gDNA. The transposon consists of two ITRs in teal, *ermB* under a promoter in mauve (fdx terminator preventing transcriptional readthrough is not depicted). The gDNA is then sheared, yielding random fragments some of which will be similar to that depicted. The ends are repaired, and adaptors (in black boxes) added. PCR is then conducted with a primer specific to the transposon and to the adaptor (both depicted in cyan). A second PCR is conducted with the primers shown in A) and B). The final outcome of these processes is shown in the last image in which the colours match those in A) and B).

3.6.3 MiSeq Nano sequencing run

3.6.3.1 Method

A small-scale sequencing run, using the MiSeq v2 Nano 150 bp paired end kit, was deployed as a way of checking the composition, and likely statistical quality, of the library (step 13 in Figure 3.6). This kit would generate 2 million reads passing filter per lane. The lane was spiked with 25% PhiX DNA to increase read diversity and so, with two technical repeats, and two other samples for a different project, ~375,000 paired end sequencing reads/library were expected. Sequencing was performed by Dr Elsie Place at the Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust and the data analysed by Dr Roy Chaudhuri at the University of Sheffield. The analysis method is described in the material and methods and remained unchanged in this study.

3.6.3.2 Results

Tagged sequences were mapped to either the *C. saccharoperbutylacetonicum* N1-4(HMT) genome, the 100 Kbp megaplasmid, or the pRPF215 transposon delivery plasmid. A total of 469,998 reads passing quality control filters were obtained for repeat 1 and 413,408 for repeat 2. Both technical repeats showed that the vast majority of sequences mapped to pRPF215 – 410,511 for repeat 1, and 468,951 for repeat 2. The rest of the reads mapped to the either the chromosome (2,755 and 975 respectively) or the megaplasmid (142 and 72 respectively). Of reads mapping to pRPF215, 97% mapped to the transposon *in situ* and most of the rest to the opposite ITR, likely due to some mispriming. As mentioned, only 2,897 and 1,047 sequences from repeats 1 and 2 respectively mapped to the genome and megaplasmid combined. Naturally, this was far too few to be processed statistically, however, there did not seem to be any insertion sites that dominated beyond all others, suggesting that the underlying library may be viable.

3.6.4 Addition of restriction digestion post-shearing

To solve the plasmid contamination issue, two key ideas presented themselves. Firstly, improvements could have been made to the delivery plasmid to increase the effectiveness of the segregational instability upon induction. For example, creating a different induction system – i.e., using available xylose inducible promoters – may well provide better expression and therefore interfere more readily with plasmid replication. However, changing the delivery plasmid would entail re-testing all the parameters already established in terms of growth rates and transposition

frequency, as well as the associated non-trivial cloning. A simpler fix – digesting the plasmid out of our samples with a restriction enzyme – was investigated (step A1 in Figure 3.9).

For this method to work, the restriction enzyme would need to cut very close to the 3' ITR where the sequenced transposon junction is located. Additionally, it would need to exhibit a very low cutting frequency in the *C. saccharoperbutylacetonicum* genome to not bias the insertion site analyses. Four enzymes with recognition sequences just outside the 3' ITR were identified – ApoI, BstXI, EcoRI and SspI. None of these overlapped with the priming regions necessary for the first or second PCR. A search for the cut sites of each within the 6,530,257 bp *C. saccharoperbutylacetonicum* genome showed that the lowest number of cut sites was by BstXI (561 sites), then EcoRI (1896), SspI (14370) and finally ApoI (27044).

BstXI cut sites were predominantly located within genes; this is most likely due to the G-C rich nature of the recognition site. 39 genes were found to have two cut sites and three genes were found to contain three cut sites. Of those that contained two cut sites, the sites were generally >100 bp apart. Even in these cases, the chances that the cut site would lead to a gene being falsely assigned as essential was highly unlikely. Since gene length is important in assigning essentiality, these intragenic cuts are only likely to be relevant to categorising small genes. Overall, it was felt that the risks were relatively low and there was much to gain by introducing a restriction digest step with BstXI.

The digestion step was introduced after the end-repair and adaptor ligation steps but before the overall processing steps. By digesting then, pRPF215 would be eliminated prior to PCR amplification. This would reduce the PCR bias towards the dominating plasmid and increase the chances of amplifying every single insertion site. 5 µL of BstXI was added to 83.5 µL adaptor ligation mix with 10 µL of NEBuffer 3.1 and 1.5 µL nuclease-free H_2O and incubated overnight at 37°C in a water bath.

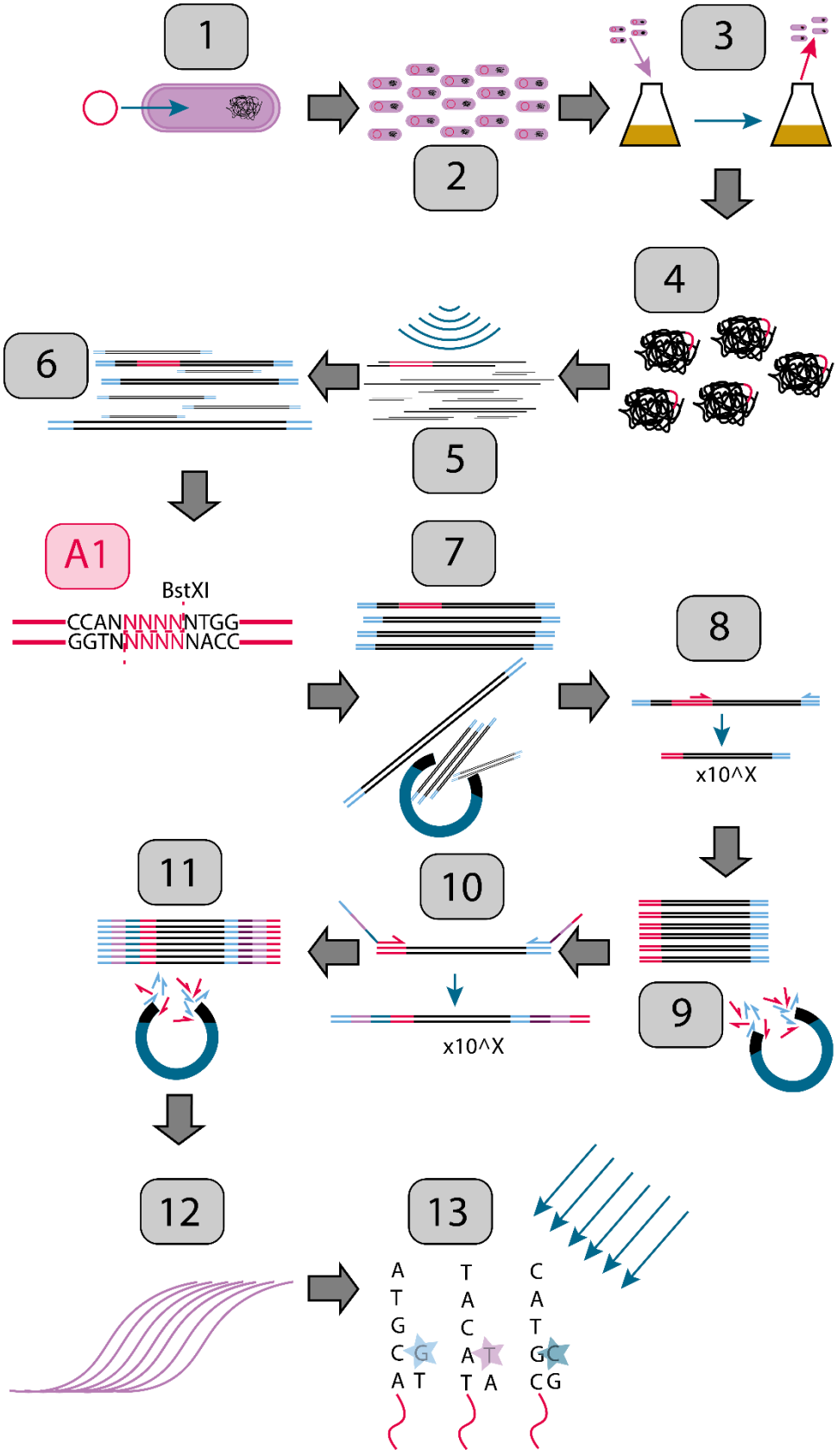

Figure 3.9 Schematic outline of the steps required to create a TraDIS library with digest step. The numbers indicate the step number. Images are purely representative and not to scale. Magenta rounded boxes indicate the altered step. 1) Transfer of delivery plasmid into cells. 2)

Growth of plasmid-containing cells in pre-transposition conditions to expand cell numbers. 3) Induction of transposition in broth cultures and removal of transposon mutants. 4) gDNA extraction. 5) Random shearing to an average size of 300-400 bp by sonication. 6) End repair and adaptor ligation. A1) Restriction digest of DNA by BstXI to remove plasmid contamination 7) Magnetic bead purification to remove fragments >500 bp and <200 bp. 8) PCR amplification between transposon insertion site and adaptor. 9) Magnetic bead purification to remove primers and small fragments. 10) PCR amplification to add sequences required for Illumina sequencing, barcoding and indexing. 11) magnetic bead purification to remove primers and small fragments. 12) Quantitative PCR to check the concentration and quality of the library. 13) Illumina sequencing by synthesis.

3.6.5 AmpliconEZ sequencing

3.6.5.1 Method

Following the re-processing of the libraries with the addition of the BstXI restriction digestion, an additional library quality control step was also introduced. Due to the expensive nature of MiSeq v2.0 Nano, Genewiz AmpliconEZ sequencing service was utilised to check library quality. This service provides Illumina-based sequencing with a guaranteed 50,000 sequencing reads per sample submitted. In addition, if a key element of our QC process was to check for plasmid contamination and dominating mutants, the Genewiz AmpliconEZ sequencing service did not require ~500,000 sequence reads to confirm this (as MisSeq v2.0 Nano does). To adapt the processed libraries for AmpliconEZ, 1 µL of the final libraries were amplified by nested PCR to contain overhangs requested by Genewiz. The PCR product was column purified and the samples prepared according to the Genewiz instructions.

3.6.5.2 Results

As before, AmpliconEZ data was analysed by Dr Roy Chaudhuri and mapped to the genome, megaplasmid and pRPF215. In total, sequencing yielded 96,015 reads, of which 94,346 mapped to the pRPF215 plasmid, 1638 to the genome and 31 to the megaplasmid, a slight drop in percentage reads mapping to pRPF215 (99.6% vs 98.3%). Clearly, however this version of the method did not eliminate the plasmid as intended. Given this information, it was decided not to proceed with further sequencing of the broth libraries and instead move to plate-based libraries as originally conducted using pRPF215 in *C. difficile*. Our assumption was that even a short growth period in broth introduced too great a bottleneck, allowing some mutants to dominate early. The solution to the plasmid contamination is listed in 3.7.2.

3.7 First plate transposon library

3.7.1 Library creation

For our first attempt at the creation of a biological library on plates, the process was conducted as close as possible to the method described previously (Dembek et al., 2015) whilst also taking into account the data generate from the above experiments. This differed in several key ways from the method described in 2.3 and so is explained in more detail here (step 3 in Figure 3. 10). The ultimate aim of the biological library creation was to generate a pool of mutants containing diverse, random

transposon insertions. Ideally, this would represent a number of insertions sufficient for each ORF to contain multiple insertions. The biological library was designed to maximise the number of individual mutants given our key constraint of anaerobic cabinet space. Two anaerobic cabinets – a DG250 and a MG500 – were used to construct the libraries. Taking into account the volume of the cabinets and the necessity to leave working room, 316 20 cm x 20 cm RCM agar ATc/erythromycin plates were prepared. Five transconjugants were prepared using the method described in 2.2.1. Each transconjugant was re-streaked twice to ensure purity. O/Ns in RCM supplemented with thiamphenicol were set up for each transconjugant. Each O/N was subcultured to OD_{600nm} 0.01 in RCM without selective pressure to ensure minimal thiamphenicol carryover onto the agar plates. These subcultures were staggered by 30 min to attempt to account for the plating time. At OD_{600nm} 0.6, 300 µL of cells was spread on each plate. At this OD_{600nm}, there is an estimated 1.2 x 10^6 CFU/mL. Accounting for plating volume, it was estimated that there would be approximately 600 colonies/plate resulting in 189,000 colonies in total. With each colony potentially representing a single mutant, this library could contain up to an insertion every 35 bp on average.

The following day, cells were harvested using 2 mL RCM/erythromycin (500 µg/µL) to cover the plate and an L-shaped spreader to scrape the colonies off the plates. Scrapings were mixed by aspiration and 400 µL from each was pooled, mixed thoroughly, and aliquots taken for glycerol stocks and cell pellets.

3.7.2 gDNA processing

All but one of the steps for processing the gDNA was kept the same as before. Following the AmpliconEZ sequencing run in 3.6.5, it was clear that either the BstXI site was methylated, or that cleavage was inhibited by known methylation of an overlapping EcoRI site (personal communication, Biocleave). To correct for this, the restriction digest step was moved to after the first PCR (step A1 in Figure 3.10). PCR products are always non-methylated and so the digest should work efficiently at this point. For this, 4 µL of BstXI was added to the 50 µL PCR product with 6 µL of NEBuffer 3.1 and again incubated overnight. The subsequent magnetic bead purification was adjusted to account for the differing volume.

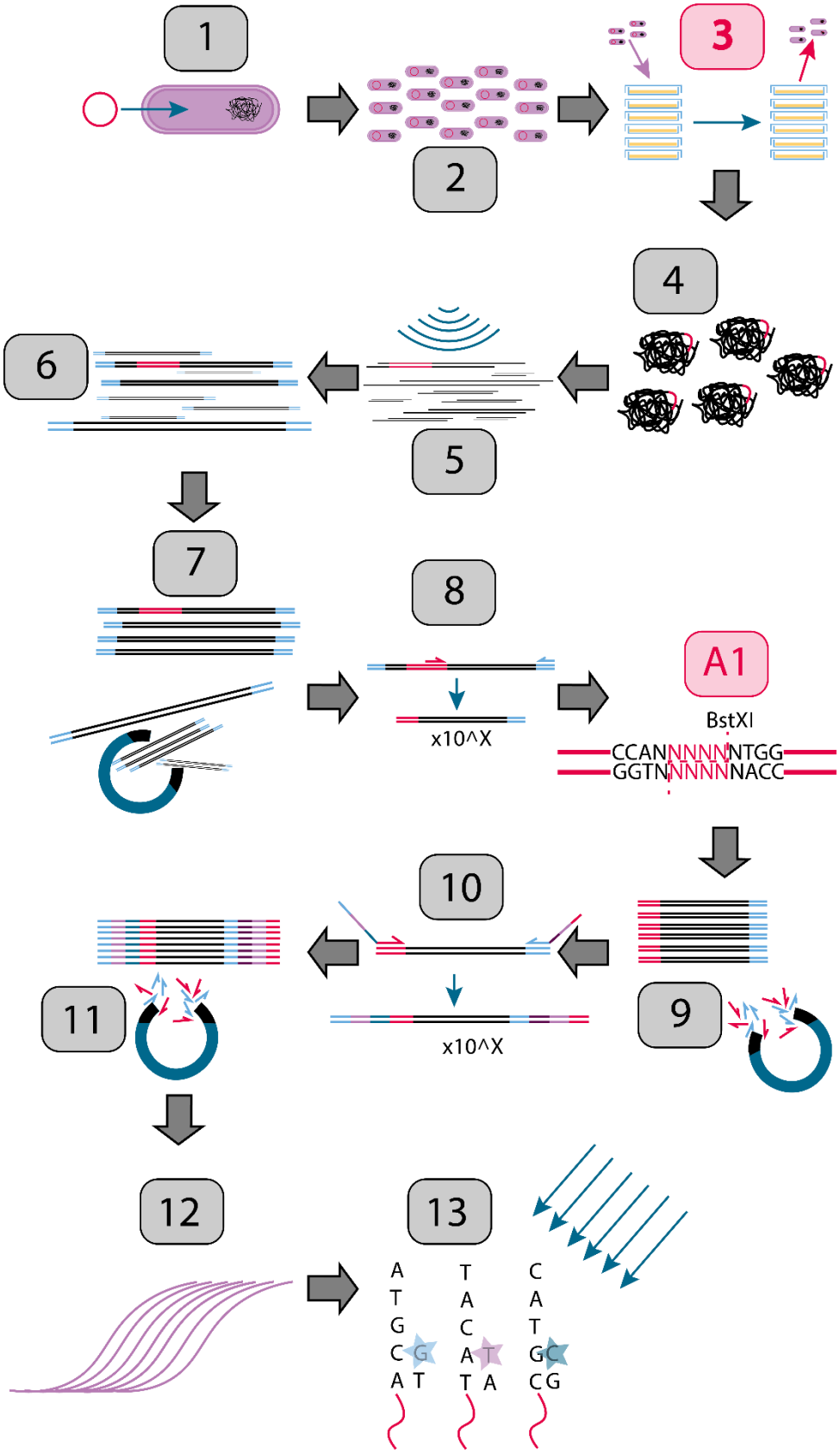

Figure 3.10 Schematic outline of the steps required to create a TraDIS library with plating and an altered digest step. The numbers indicate the step number. Images are purely representative and not to scale. Magenta rounded boxes indicate the altered step. 1) Transfer of

delivery plasmid into cells. 2) Growth of plasmid-containing cells in pre-transposition conditions to expand cell numbers. 3) Induction of transposition by plating on ATc/Ery and subsequent removal of transposon mutants by scraping. 4) gDNA extraction. 5) Random shearing to an average size of 300-400 bp by sonication. 6) End repair and adaptor ligation 7) Magnetic bead purification to remove fragments >500 bp and <200 bp. 8) PCR amplification between transposon insertion site and adaptor. A1) Restriction digest of DNA by BstXI to remove plasmid contamination after the first PCR. 9) Magnetic bead purification to remove primers and small fragments. 10) PCR amplification to add sequences required for Illumina sequencing, barcoding and indexing. 11) magnetic bead purification to remove primers and small fragments. 12) Quantitative PCR to check the concentration and quality of the library. 13) Illumina sequencing by synthesis.

3.7.4 HiSeq 2500 sequencing

3.7.4.1 Method

Samples were subsequently sequenced using on a HiSeq 2500 using the SBS V2 2x50bp reagent kit, which should provide approximately 600 million paired end reads. This would be excessive for a single TraDIS library, however, we were running multiple libraries from other sources as well as two repeats of this library, so the Hiseq 2500 represented the most cost-efficient approach. Sequencing was performed by Jennifer Dawe at the Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust and the data was analysed by Dr Roy Chaudhuri at the University of Sheffield.

3.7.4.2 Results

The sequencing run yielded a total of 1.79 x 10^7 reads containing the transposon tag between the two samples submitted. Of these 98% mapped to the genome and only 2% to pRPF215, indicating that our altered digestion step was successful. Further analysis using Bio-Tradis reduced the number of usable reads to a total of 1.15 x 10^7 for an average of 5.77 x 10^6. From these, 96.8% mapped to the genome and 3.2% to pRPF215. The mapping of insertion sites indicated a complex library with 168,826 unique insertion sites, enough for 1 in ~39 bp. There are 5,937 annotated open reading frames in *C. saccharoperbutylacetonicum*. However, just five ORFs accounted for 30% of reads. If the count is extended to include the 100 most sequenced ORFs, these account for 66.5% of all sequencing reads.

With the majority of reads being accounted for by a minority of ORFs, there are fewer reads available for genes with only a single insertion site. With a better distribution, Bio-Tradis is able to assign essentiality or non-essentiality to genes with one or two insertions based on gene size and the number of reads relative to other low-insertion genes. However, with many genes having insertions with very few sequencing reads, it becomes difficult to classify if the gene is essential or non-essential. For example, 1836 ORFs had five or fewer sequencing reads covering four or fewer insertions. A number of these will be clearly non-essential based on their small size, however, the majority will be either ambiguously classified or classed as non-essential with low confidence. In short, it is not possible, with just one read of one insertion to be sure that the insertion is genuine and not an artifact of the various molecular biology processes.

Further ambiguity is introduced when the intersection of essential genes of both repeats is introduced. Individually, repeat 1 yielded 389 ORFs with no insertions and repeat 2, 369. Based on the number of annotated ORFs in *C. saccharoperbutylacetonicum*, these figures together would be

on the lower end of the expected essential genome which is typically around 5-15% (Barquist et al., 2013; Dembek et al., 2015) However, when the intersection of these two gene sets is examined, only 170 ORFs are classified as essential in both. Whilst it can probably be assumed this represents a core set of essential genes, there are 410 ORFs whose essential status is left ambiguous. This is before other genes that are found to be ambiguous in both analyses are taken into account.

It is clear from this data that there must have been early transposition events that occurred prior to induction by plating. Therefore, I elected to develop measures that would reduce the chances of this occurring and increase our ability to remove populations with an early transposition event from the library.

3.8 Plasmid copy number

3.8.1 Aims

Whilst the plasmid contamination issues described in 3.6 were solved, it was pertinent to ask why there was such a high concentration of plasmid following mutant selection. The pRPF215 system should exhibit segregational instability upon induction, resulting in the eventual dilution of the plasmid from the population. This process was clearly unreliable leading to the majority of transposon-containing sequences in the sequencing libraries originating from the plasmid. Equally, however, the transposition frequency data combined with personal observations during the harvesting process suggested that the method is relatively effective at selecting for the growth of transposon insertion mutants. If this were not the case, it would be expected to find that colony counts are similar to those seen in uninduced controls for the former and lawns of cells on the plates of the latter.

There are several potential factors that could account for this phenomenon. It's possible that the issue lies with the conditional replicon, either the induction is lower than expected or the read-through of transcription into the *repA* is not as disruptive in *C. saccharoperbutylacetonicum* as it is in *C. difficile* or, of course, a combination of both. The observed transposition frequency is potentially indicative of low induction rates. However, the transposon frequency is determined by a combination of factors including efficiency of de-repression, transposase expression and activity, all of which are difficult to investigate in a meaningful context. Attempts to use a an enhanced Himar-1 transposase with higher rates of transposition (Lampe et al., 1999) failed to result in higher rates of transposition by colony count, suggesting that the amount of transposase is not a limiting

factor. However, plasmid copy number influences all these factors as well as the rate of plasmid dilution and the amount of plasmid in a single colony forming mutant.

Whilst segregational instability is induced with ATc, with more than one copy of the plasmid it takes many more generations before the plasmid is removed. Equally, this would imply that, in *C. saccharoperbutylacetonicum* , the copy number was high enough for the plasmid to constitute a significant presence in the cell. In such a scenario, it's likely that non-mutants placed on ATc/thiamphenicol plates would struggle to grow. Whilst the *catP* carrying plasmid is still present, the reduced copy number would presumably reduce the expression of antibiotic resistance and therefore the ability to grow on thiamphenicol. This would naturally result in very small or imperceptible colonies, as seen in 3.4. The same would also be seen for non-mutants during library production. Understanding the plasmid copy number therefore offered the most straightforward route to shedding light on the problem of plasmid contamination.

Equally, the large 136,188 bp megaplasmid presented as worthy of investigation. Previous genome annotation was unable to putatively identify any genes with obvious utility to the cell (Poehlein et al., 2014). In *C. acetobutylicum*, for example, the megaplasmid contains the *sol* operon necessary for the production of n-butanol and loss of the plasmid is a source of degeneration of the solvent-producing phenotype (Kashket and Cao, 1993). In *C. saccharoperbutylacetonicum*, all of the analogous genes are found on the chromosome. Elucidating the copy number of the megaplasmid was therefore important for three key reasons. Firstly, if the copy number was one, it would be possible to examine the megaplasmid for essential genes through TraDIS. Secondly, if the copy number was greater than one whilst being lower than that of pRPF215, the megaplasmid origin of replication might be suitable for use in the creation of a novel transposon delivery system that can quickly eliminate the plasmid from the population. Finally, the copy number could help to understand the kind of additional energetic burden the megaplasmid places on the cell and, in the long term, explain why such apparently extraneous DNA is maintained in the population.

3.8.2 Method

I utilised a method of calculating the copy number of both pRPF215 and the megaplasmid relative to the genome using qPCR, similar to that developed by Lee et al., 2006. Since, outside of exponential growth, the genome is assumed to be one, the value can be used to normalise the relative copy number for the two plasmids. This does mean that the final values are affected by the replication dynamics of both the genome and the plasmid. This will introduce extra error as it

cannot be assumed that the dynamics are the same for both. However, with our method, something of these dynamics can be inferred.

To account for variability in copy number depending on a DNA sequence's proximity to the origin of replication, multiple primer pairs were designed for each DNA molecule, with four primer pairs designed for the genome and megaplasmid and two for pRPF215. Circular bacterial genomes are typically closed at the site of the *dnaA* gene which is assumed to mark the origin of replication (Hunt et al., 2015). Primer pairs were designed at the site of *dnaA* (12 o'clock) and subsequently at approximately 3, 6 and 9 o'clock. A similar approach was used for the megaplasmid. Given the significantly smaller size of pRPF215, four primer pairs were unnecessary, so pairs were designed at just the *repA* site and a distal site opposite the ori. All primers were designed to anneal at 60^0C and yield amplicons ≤200 bp. Their specificity and effectiveness were checked by Taq PCR using the qPCR reaction conditions. 10, 20 and 30 cycle tests were conducted and run on an agarose gel to increase the chance of identifying if a primer pair was more or less effective than the rest. Serial dilutions of pRPF215 were used as concentration standards. Serial dilutions of the test samples were conducted and the 10 ng/µL and 1 ng/µL concentrations were used as template for the qPCR reaction. One 96 well plate was conducted for exponential phase gDNA, and the same conditions used to conduct another plate with the stationary phase gDNA. Cq values were determined, and a standard melt curve was conducted to ensure consistent and expected product size was amplified for each reaction.

3.8.3 Results

Following Cq value determination, standard curves were constructed using the Cq values from known concentration of pure pRPF215 and analysed using the logarithmic trendline function in excel. The resulting equation was used to convert Cq values from test samples into a concentration for each set of technical repeats (Figure 3.11). These were conducted separately for each primer pair and overall averages also calculated for the given DNA molecule. However, the three biological repeats were kept separate. The averages for the megaplasmid and pRPF215 were then divided by those of the genome for each biological repeat. The overall average for the two dilutions and the three biological repeats were then calculated along with the standard deviation for these values (Figure 3.11).

The overall results show a marked difference in between the copy number of pRPF215 in exponential (average: 30.6 copies) and stationary phase (average: 12.8). Notably both averages are

higher than those estimated for *C. difficile* of 6-10 (Purdy et al., 2002). During stationary phase, the standard deviation is markedly higher than in exponential phase. This is likely due to the high replication rate during exponential phase growth, meaning the difference between the biological repeats can be quite high depending on the exact dynamic of each population at the moment of extraction. Interestingly, the megaplasmid does not exhibit such great variance in copy number during exponential phase. This suggests two conclusions. Firstly, that the variance seen for pRPF215 in exponential phase is not due to poor experimental handling and is genuinely due to large natural variance in differing populations. Secondly, that the replication dynamics are much more stable for the megaplasmid. This, in turn, suggests much stricter regulation of replication and likely more stringent segregation of megaplasmid copies between cells.

If the data for each primer pair is separated and analysed individually, it is possible to infer the approximate location of the origins of replication of the genome and megaplasmid. Lower Cq values represent a higher sequence copy number at that location. Bacterial DNA replication typically involves multiple initiations of replication at the origin site (O'Donnell et al., 2013), particularly in fast growing cells. Therefore, if there are amplicon sites that are consistently exhibiting higher copy numbers, it can assumed that site is closer to the origin of replication. For the genome, the 12 o'clock amplicon located near the *dnaA* site was found to have the highest relative copy number at both exponential and stationary phase (Figure 3.11D). Interestingly, there is a significant difference in Cq between the 3 and 9 o'clock amplicons for the genome at both timepoints, suggesting either that the origin lies significantly closer to 3 than 9, or that DNA replication proceeds faster in that direction. The 6 o'clock amplicons consistently exhibited much higher Cq than all other sites suggesting that this site was furthest from the origin, as expected.

For the megaplasmid, the 6 o'clock amplicon consistently showed the highest copy number across both timepoints. Interestingly, the 12 o'clock site was equally consistent in being the second highest amplicon. There was no significant difference between the results of the 3 o'clock and 9 o'clock amplicons. These results imply that there are two origins of replication at opposite ends of the megaplasmid, in turn suggesting a tight regulation of replication. Though not well investigated, there is precedent for megaplasmids containing multiple origins of replication in Gram positives (Harris et al., 2018; Zhang et al., 2010). Finally, for pRPF215 and the pCD6 origin of replication, there was no dominant site in the exponential phase, but, during stationary phase, the amplicon nearest to the origin is present in higher copy. The former result is somewhat surprising but likely due to experimental inaccuracies caused by significantly higher rates of replication.

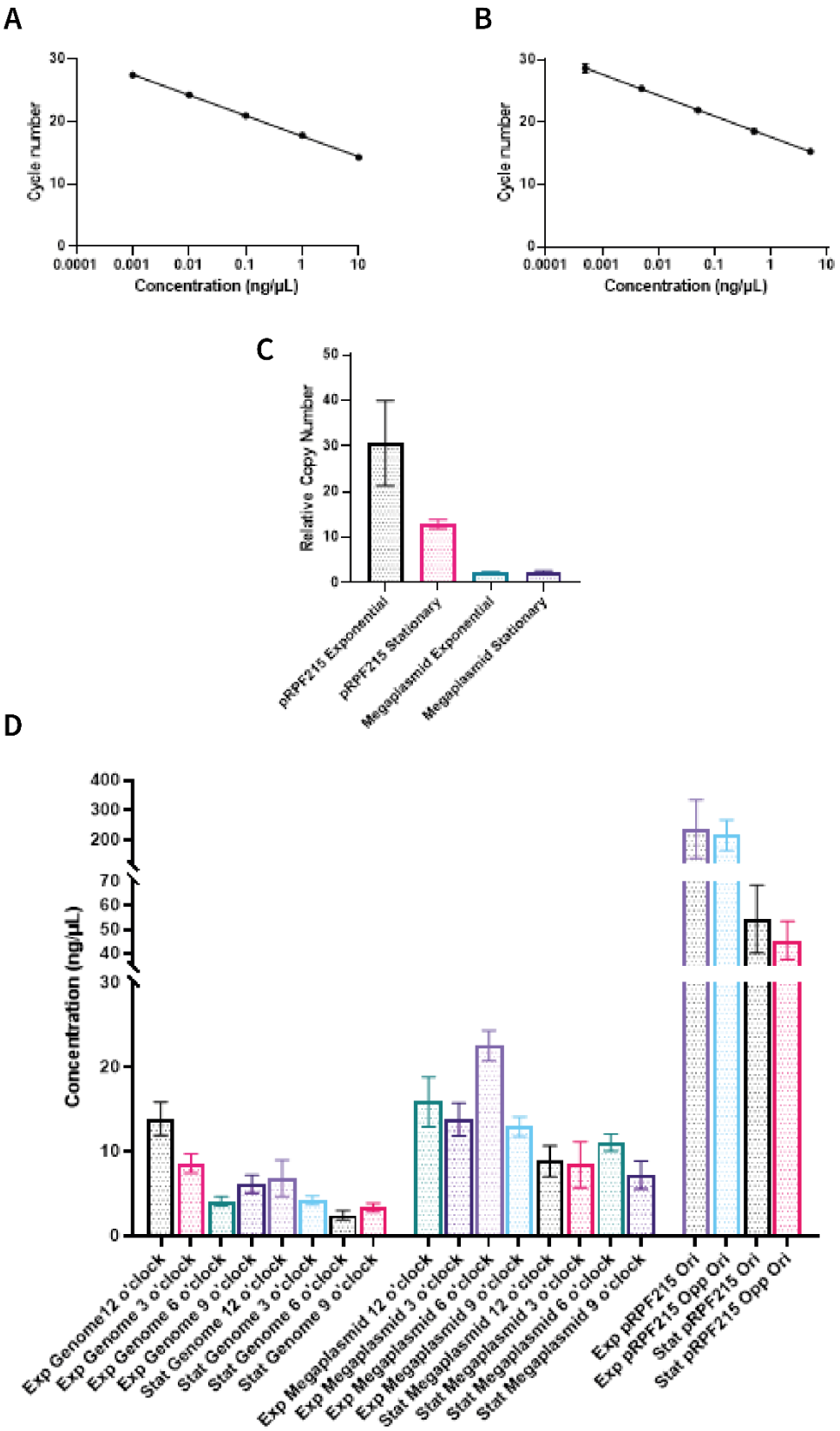

Figure 3.11 The results of qPCR to establish the copy number of pRPF215 and the megaplasmid. A) The standard curve conducted for the exponential growth phase extractions used to calculate concentration. A nonlinear fit of the points yielded an R^2 of 0.9996 and the

equation y=-1.426ln(X)+17.655. The curve was calculated from biological triplicates and technical duplicates. **B)** The standard curve conducted for the stationary growth phase used to calculate concentration. A nonlinear fit of the points yielded and R^2 of 0.9998 and the equation y=1.457ln(X)+17.581. The curve was calculated from biological duplicates and technical duplicates. **C)** The result of calculating the copy number after converting the cycle number for each pair to concentration and then dividing by the value found for the genome. An unpaired t-test of the pRPF215 conditions shows a significant difference between the copy numbers associated with exponential phase (avg = 30.6) and stationary phase (avg = 12.8), P=0.0053. In contrast, an unpaired t-test of the megaplasmid conditions showed no significant difference between the exponential phase (avg = 2.172) and stationary phase (avg = 2.270), P=0.5981. Both exponential and stationary phase experiments were conducted with biological triplicates and technical duplicates **D)** The concentration of each DNA structure when the primer pairs are calculated separately. Exp indicates the exponential experiment whilst Stat indicates the stationary experiment. O'clock indicates the position of the primer pair relative to the site of where the DNA sequence was closed (12 o'clock). Ori indicates the region of origin of replication in pRPF215.

3.9 Importing a novel plating method

During the harvesting process in 3.7.1, it became clear that the uneven spread of cells was a chronic issue. L-shaped spreaders typically don't produce completely even spreads due to the tendency to not absorb the last volumes and instead leave a significant streak. This is a greater problem when cell densities are higher as the culture becomes more viscous. In addition, with the need to spread >300 plates, it is even more difficult to avoid this issue without drastically increasing the time it takes to plate. Another common plating method is to place a small number of sterile glass beads (~3 mm diameter) on the agar plate, add the culture, and shake the plate laterally. The method would also have the added advantage of reducing the time it takes to plate due to the possibility of shaking multiple plates at once. Finally, glass beads are inherently reusable, reducing the plastic waste associated with the experiment.

To test the effectiveness of glass beads, a plating experiment was conducted that provided both quantitative and qualitative results was conducted. *C. saccharoperbutylacetonicum* containing the pRPF215 plasmid was grown overnight in triplicate. 100 µL, 200 µL or 300 µL were spread with either glass beads or L-shaped spreaders on 20 x 20 cm RCM agar supplemented with ATc/erythromycin and left to grow overnight. The following day, plates were photographed and 1/8 of each plate had colonies counted. Representative images of each clearly show the difference in colony distribution between glass bead (Figure 3.12A,C,E) and L-spreader spreading (Figure 3.12B,D,F). Large clumps were common across all L-shaped spreader plates but virtually absent from glass bead plates. Surprisingly, the use of glass beads also lead to higher colony counts in all conditions (Figure 3.13), though this was not statistically significant, likely due to a biological repeat that exhibited lower colony counts in all conditions.

The plated OD_{600nm} of all three were virtually identical (3.47±0.024). This could therefore be a sign of an early transposition in these two samples. This experiment was not conducted with fresh transconjugants but from glycerol stocks of the strain which may explain the higher tendency towards early transposition. Based on these results, I decided to use glass bead spreading in future library creation.

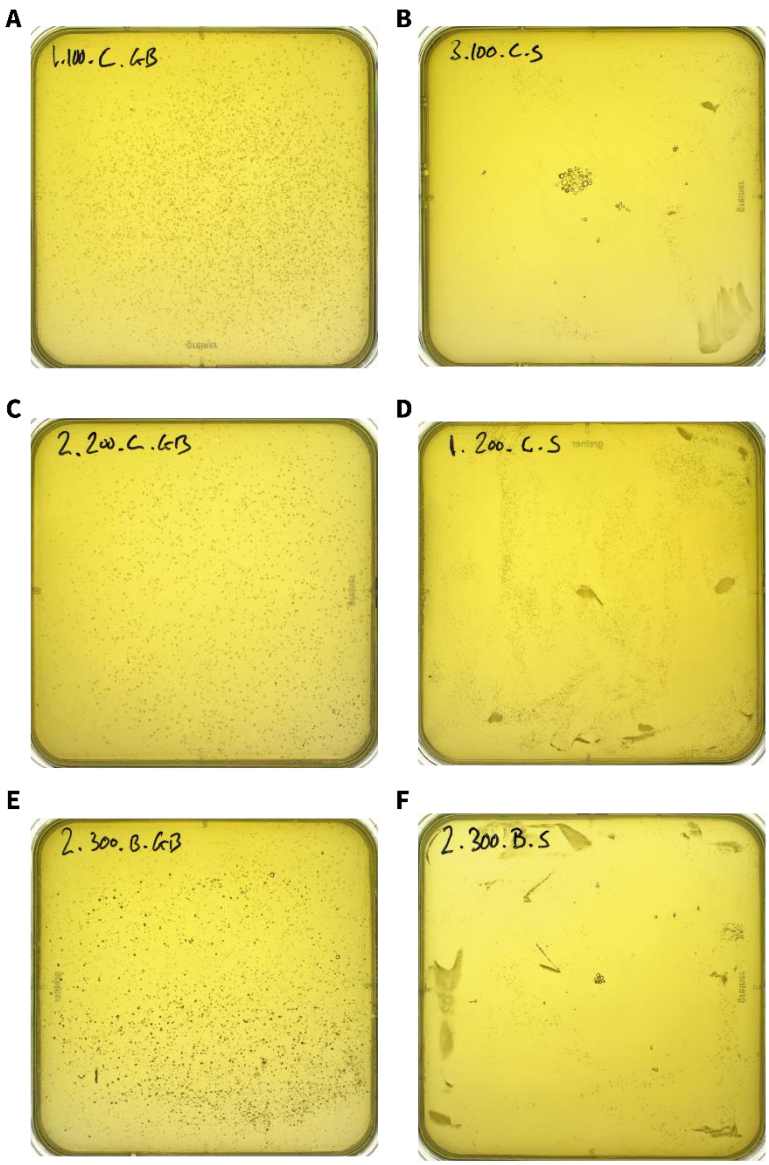

Figure 3.12 Exemplar images of the agar plates following the use of different culture volumes and spreading techniques. The first number refers to a given biological replicate, the second to the volume plated, the first letter to the technical replicate, GB stands for 'glass beads' and S for 'spreader'. **A)** The plating of 100 µL of overnight *C. saccharoperbutylacetonicum* culture using glass beads showing an even distribution of colonies and no obvious clumps or streaks. **B)** The plating of 100 µL of O/N culture using an L-shaped spreader showing relatively low cell densities and some clumping evident in the bottom right-hand corner. **C)** The plating of 200 µL of O/N culture using glass beads. A high density of colonies with an even distribution is easily visible. **D)** The plating of 200 µL of O/N culture using an L-shaped spreader. Whilst a reasonable cell density is seen, multiple clumps of cells are seen including the at the centre, top right, right and bottom of the plate. **E)** The plating of 300 µL of O/N culture with glass beads. An extremely high density is

seen with many larger colonies. At this density, colonies are starting to clump somewhat, particularly in the lower right corner. **F)** The plating of 300 µL of O/N culture using an L-shaped spreader. Cell densities are difficult to determine from this image as many colonies were smaller in size. Again, significant clumping is seen on three of the four edges and in the centre.

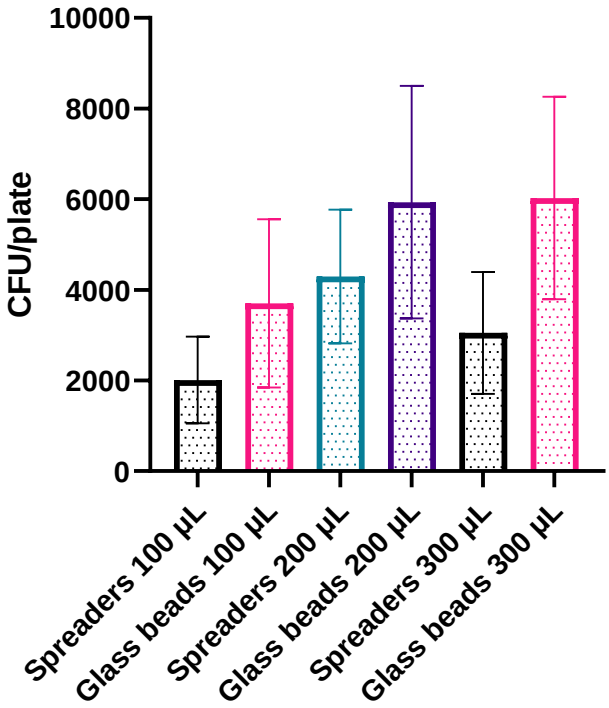

Figure 3.13 CFU/plate for each spreading condition and volume. The same volumes from different plating methods were paired together to allow for easier direct comparison. The error rate is relatively high and therefore not statistically different (P=0.1323) due to the one of the three biological repeats showing consistently lower numbers than the other two. However, the underlying numbers indicate the same pattern for all three repeats. A representative eighth of each plate was counted. The experiment consisted of three biological repeats for each condition.

3.10 Second plate transposon mutagenesis library

3.10.1 Biological library creation

The transposon library generation was further optimised to reduce the latency between conjugative transfer and the induction of the pPRF215 system and to maximise the number of mutants with four key changes (steps 1-3 in Figure 3.14). Firstly, the conjugative transfer protocol was reduced to approximately 36 h, by harvesting mixed conjugation plates after only 8 h and re-streaking successful transconjugants only once on agar containing 50 µg/mL colistin sulphate to ensure purity. Secondly, the first O/N culture of the fresh transconjugants was used directly for selecting mutants without a subculture step. Thirdly, 10 transconjugants were grown separately and their stacks of plates marked. Separate and mixed pools were taken to ensure the possibility of removing one of more transconjugant should they prove to be either contaminated or to contain an early transposition mutant. Finally, glass beads were used to ensure even spread and allow for higher overall colony counts.

3.10.2 gDNA processing

Pools of each transconjugant were used for gDNA extraction. Restriction digestion was conducted after the first PCR to eliminate the majority of plasmid-mapping reads (step A in Figure 3.14). This was conducted overnight in the PCR mixture and NEB3.1 buffer. The subsequent magnetic bead purification was adjusted to account for the change in volumes, but ratios of beads to sample and the process remained unchanged. A different inline index was used for each transconjugant in order to readily remove inadequate samples from the analysis. Two different Illumina indices were used, primarily to utilise the range of primers to hand and samples were sequenced on a MiSeq using the V2.0 2x150bp reagent kit. Sequencing was conducted by Timothy Wright at the Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust and data was analysed by Dr Roy Chaudhuri at the University of Sheffield.

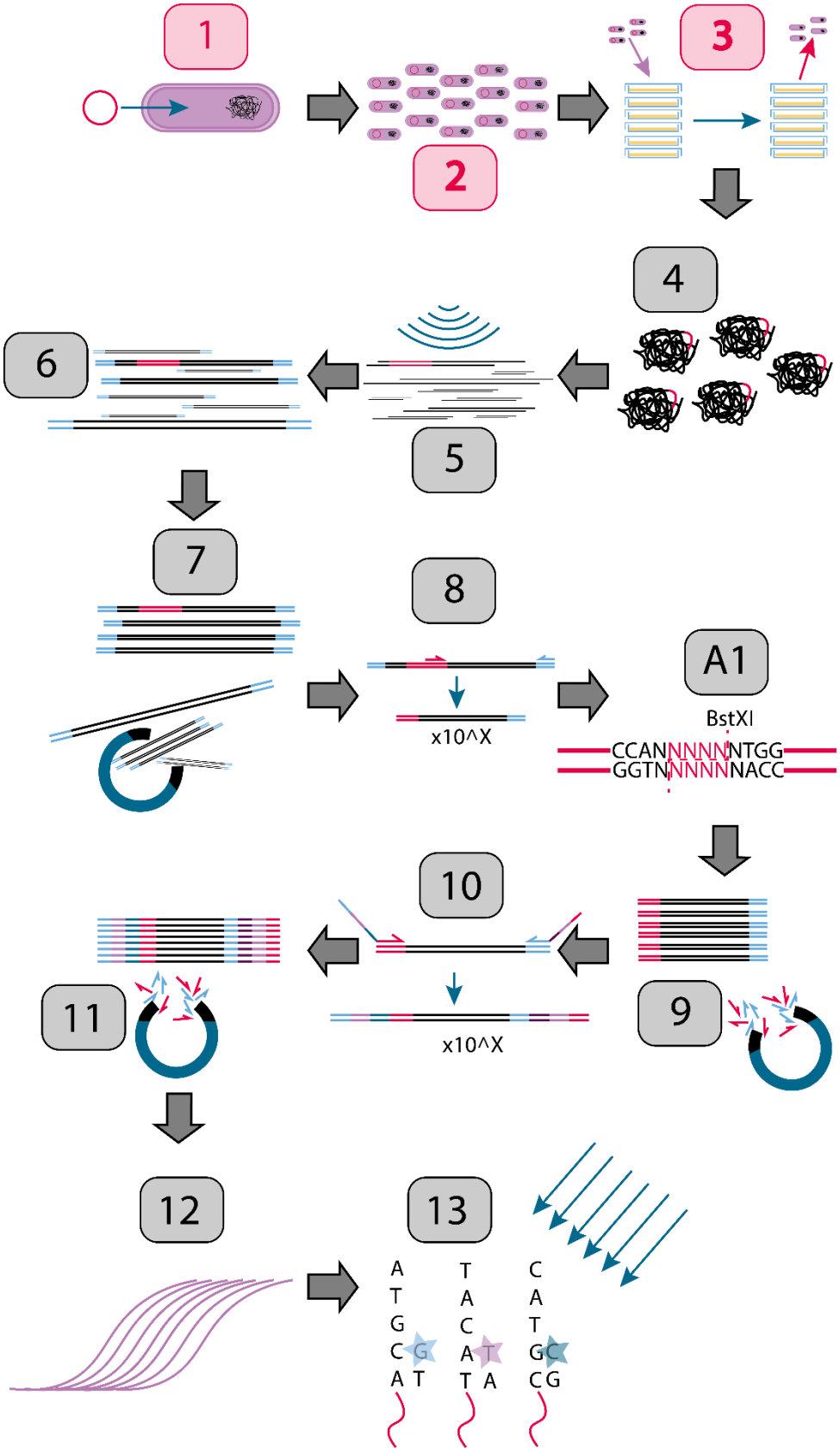

Figure 3.14 Schematic outline of the steps required to create a TraDIS library modified biological library creation steps highlighted. The numbers indicate the step number. Images are purely representative and not to scale. Magenta rounded boxes indicate the altered step. 1)

Transfer of delivery plasmid into cells. 2) Growth of plasmid-containing cells in pre-transposition conditions to expand cell numbers. 3) Induction of transposition by plating on ATc/Ery and subsequent removal of transposon mutants by scraping. 4) gDNA extraction. 5) Random shearing to an average size of 300-400 bp by sonication. 6) End repair and adaptor ligation 7) Magnetic bead purification to remove fragments >500 bp and <200 bp. 8) PCR amplification between transposon insertion site and adaptor. A1) Restriction digest of DNA by BstXI to remove plasmid contamination after the first PCR. 9) Magnetic bead purification to remove primers and small fragments. 10) PCR amplification to add sequences required for Illumina sequencing, barcoding and indexing. 11) magnetic bead purification to remove primers and small fragments. 12) Quantitative PCR to check the concentration and quality of the library. 13) Illumina sequencing by synthesis.

3.10.3 MiSeq sequencing run

Sequencing of the libraries yielded 10,457,808 reads mapping to the genome corresponding to 204,888 unique insertion sites. This latter result was highly promising as it corresponds to an insertion every ~32 bp, a sufficient average density to promise an insertion in all but the smallest ORFs. Statistical analysis was conducted using the Bio-TraDIS programme. Briefly, this takes the sequences with transposon tags, maps them to the genome and calculates the insertion index for each assigned ORF by dividing the number of unique insertions by the size of the ORF. Broadly, the higher the insertion index, the less likely the gene is essential whilst the lower the index the higher the odds of essentiality. In a typical analysis, the result will be a bimodal distribution, one mode with a small number of ORFs with very low insertion indices and another, containing the majority of ORFs, with a broad range of higher insertion indices. Bio-Tradis assumes that the former group constitutes the essential mode whilst the latter represents the non-essential mode. An ORF is considered essential if its log likelihood score was less than 12, i.e., that it is 12 times more likely to be essential than non-essential.

Issues arise with the Bio-Tradis method when there are problems with the underlying library. Particularly, when there is insufficient sequencing resolution on genes with few unique insertions and/or unique insertions with few sequencing reads. In this scenario, it becomes difficult to sort the data into the two modes and the bimodal distribution is broadly lost. The gamma fit of the distribution shows no drop in ORFs between the putative modes (Figure 3.15A), as is normally expected (e.g., Figure 1B in Langridge et al., 2009). Our data was again heavily influenced by dominant mutants. The most sequenced ORFs yielded 1,324,417 reads accounting for 12.7% of all sequencing reads. The 14 most sequenced ORFs all had a minimum of 100,000 reads and together accounted for 39% of the total. If the search is expanded out further, the next 157 ORFs account for another 33% of all reads.

All the separate repeats contained multiple varieties of dominant mutant suggesting this was not an issue of one early transposition event. Ultimately, 72% of all sequencing reads was spent sequencing only 54,276 insertion sites or 26.5% of unique insertions. The subsequent problem arising from this is the abundance of ORFs with fewer than ten sequencing reads, 1,308, and/or few insertions (1,465 have ≤5). This meant that Bio-Tradis over-assigned essentiality on the basis of these reads and insertions being insufficient to call the ORF non-essential and our library putatively contained 3,625 essential genes (Figure 3.15B). This is most likely a vast overestimation, accounting for 61% of all ORFs in the genome, much higher than any previous literature has shown (e.g., 11% in Dembek et al., 2015). It is highly unlikely, therefore, that all these genes are truly essential.

Ultimately, this library cannot be used for adequate investigation into gene essentiality under different conditions. The dominant mutants identified likely do not present a significant fitness cost and have a high chance dominating any condition subsequently tested. This, again, would result in being unable to identify whether the disappearance of a less well represented mutant was due to a fitness cost or random chance. Equally, the large number of assigned essential genes means it is not possible gain any understanding of the potential role of those genes in a different context, so, in effect, only 39% of the genome would actually be under investigation. Not enough information is likely to be derived from downstream experiments to justify their cost in both time and money.

Whilst these results mean that no downstream experiments could be conducted, there is still scope to provide a useful list of essential genes. If the statistical power of Bio-Tradis is dropped, and instead only ORFs with no insertions at all are examined, a list of 677 ORFs is generated and listed in Appendix IV. At 11.4%, this is far closer to the literature numbers previously quoted for an essential genome. This list, which contains genes expected to be essential such as ribosomal components, is likely to be useful as a rough guide to the essential genome, particularly when it comes to manipulating the genome or certain metabolic pathways, a common process for Biocleave/BCL2020 as well as for academics working with the species.

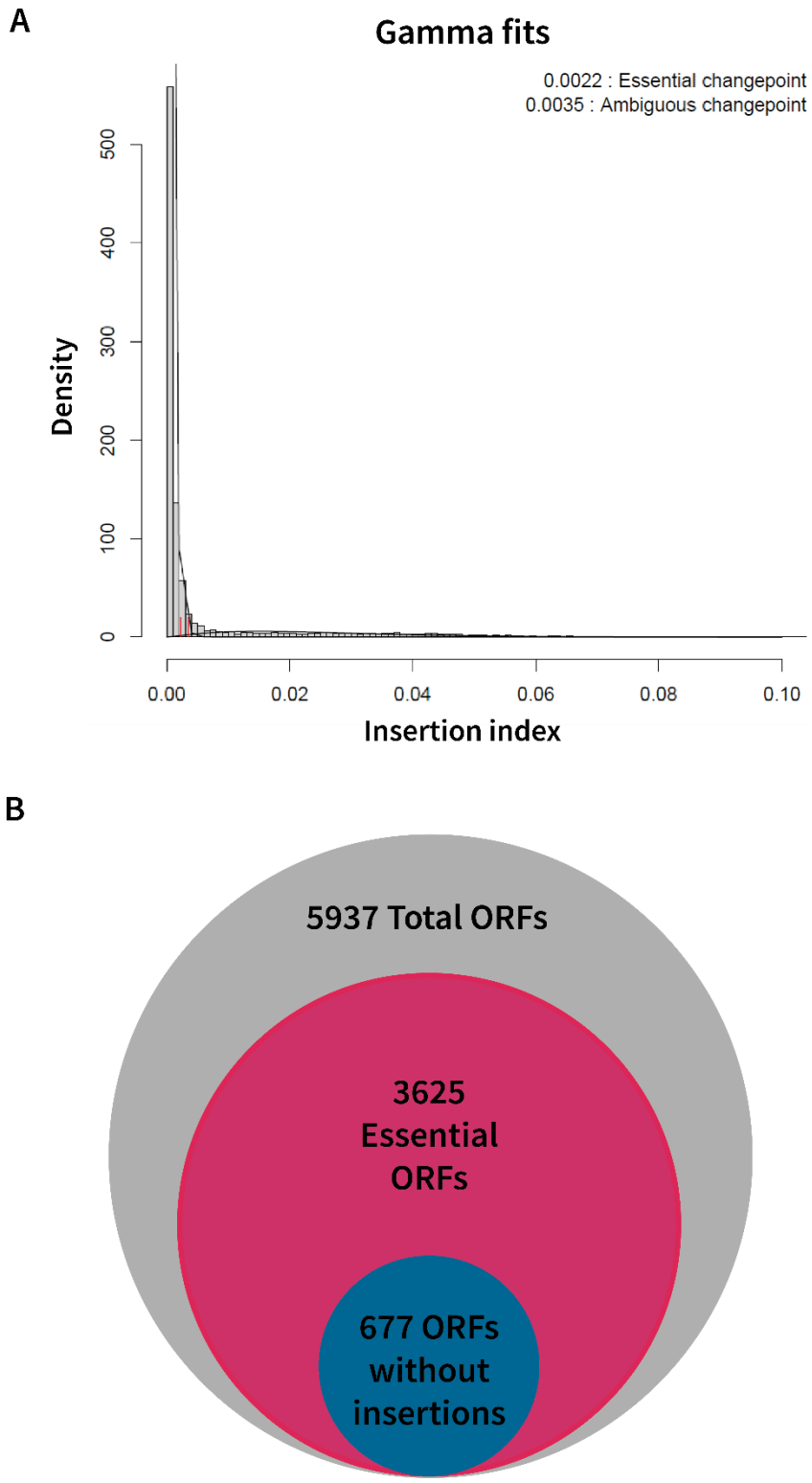

Figure 3.15 The results of MiSeq sequencing of the second plate transposon libraries. A) The gamma fit distribution of insertion indices. A unimodal distribution is seen with the majority number of ORFs having very low insertion indices. A small number of dominant mutants have an extremely high insertion index. **B)** A schematic summarising the output of 3625 essential ORFs as determined by Bio-Tradis (magenta circle) and 677 ORFs with no transposon insertions (teal circle) relative to the total number of 5937 ORFs.

3.10.4 ShinyGO gene enrichment analysis

To attempt to add greater validity the list of 677 ORFs without insertions, a gene enrichment analysis was conducted utilising ShinyGO (Ge et al., 2020). ShinyGO utilises publicly available genome annotations to compare a list of genes generated from an omics approach to an annotated genome and evaluate pathways that are likely overrepresented in that list. Where there is the information available, ShinyGo can generate plots to the most over-enriched pathways with the lowest false discovery rate (FDR) as well as network maps to readily visualise the interrelatedness of these maps. It can also link to KEGG and STRING databases to produce pathway diagrams and protein-protein interaction networks.

The ORF list generated in 3.10.3 resulted in 453 pathways being identified in total (Appendix VI). This was further refined to the 40 pathways as ranked by both fold enrichment, the percentage of genes belonging to the pathway divided by their percentage of the whole genome, and the FDR, the likelihood the enrichment is by chance. A lollipop bar plot was generated to show the over-enriched pathways along with the fold enrichment, FDR and number of genes involved in the pathway (Figure 3.16A). A network map showing the interrelatedness of the 40 over-enriched pathways was also generate (Figure 3.16B).

The information available on pathways in *C. saccharoperbutylacetonicum* is not especially detailed and resulted in somewhat generic pathway annotations (e.g., 'Helicase'). However, the over-enrichment was able to identify several pathways that would typically be assumed to be essential under typical laboratory conditions. Genes involved DNA replication are overrepresented 4.1-fold (FDR=0.0051), as are those in various branches of the membrane transport protein major facilitator superfamily by an average of 3.6-fold (average FDR=0.00034). Other clearly important pathways found to be over-enriched were: mur ligase genes necessary for cell wall synthesis (6.4-fold, FDR = 0.0028); sugar transport genes (2.8-fold, FDR=0.00071); glycosidases required for glycolysis and other core metabolic activities (2.4-fold, FDR=0.000034); DEAD/DEAH box helicases important in RNA metabolism (4.9-fold, FDR=0.0028); and aminotransferases necessary for the addition of amino acids to tRNAs (4.8-fold, FDR=0.0017)

The analysis is imperfect, resulting in several gene sets being categorised into putatively the same pathway listed under slightly different names. For example, the 6 mur ligase related genes are categorised as 4 separate pathways with only slightly differing names. This despite selecting the option to remove redundancy in pathway names from the analysis. This is like due to poor curation of pathways and low detail of annotation in *C. saccharoperbutylacetonicum*.

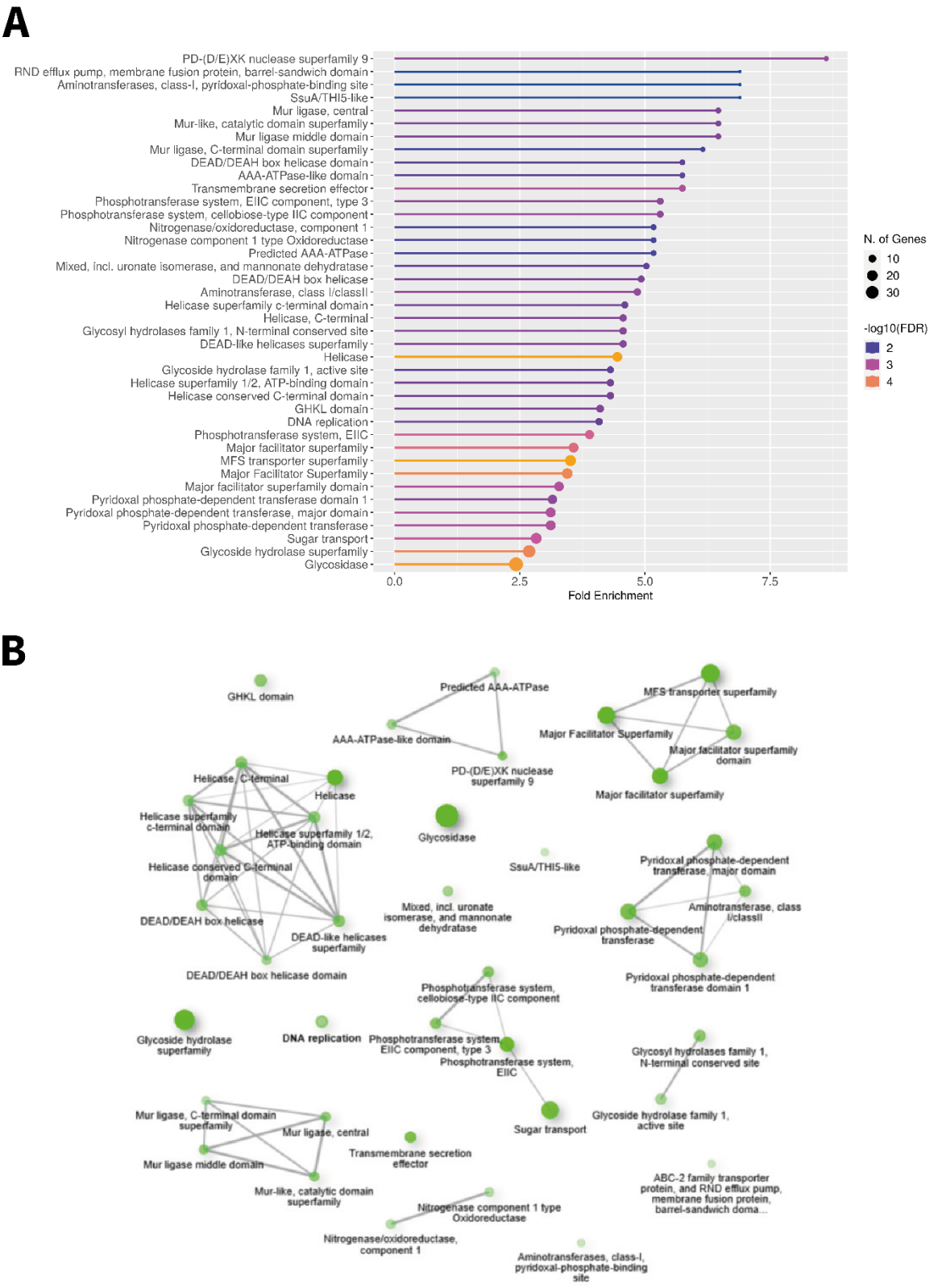

Figure 3.16 The results of enrichment analysis conducted on ORFs without insertions. A) Bar plot listing the 40 pathways with the combined highest fold enrichment and lowest false discovery rates. Length of the bar indicates fold enrichment. Colour represents the FDR order of magnitude with orange being the lowest and blue the highest. The circles indicate the number of genes in the pathway. **B)** Network map indicating the interrelatedness of the different pathways identified. Two pathways are connected if they share at least 20% of genes. Larger circles represent pathways with a greater number of genes. The darker circles represent more highly enriched pathways.

The ShinyGO analysis appears to identify pathways that would be expected to essential, enhancing the validity of the list of 667 ORFs without insertions as a resource for future *C. saccharoperbutylacetonicum* research. As the body of research on *C. saccharoperbutylacetonicum* continues to grow, it might be valuable to repeat analyses such as these to gain a more nuanced understanding of genes that are likely to be important to survival and laboratory conditions.

3.11 Discussion

Whilst these results are disappointing, they are, at least, conclusive. It is clear that the pRPF215 transposon delivery system cannot function sufficiently well in *C. saccharoperbutylacetonicum* in its current form. A combination of factors highlighted by and inferred from this study have contributed to this failure. In the timeframes necessary to transfer the plasmid by conjugation into the *C. saccharoperbutylacetonicum* and prepare the strain for scale application, multiple early transposition events are guaranteed. There are several possible explanations for this phenomenon. Transposase efficiency and the strength of repression by the tetracycline repressor are two key factors that haven't been thoroughly investigated here. The former is beyond the scope of this study whilst the latter has only been briefly probed through luciferase assays (see Chapter V). In the event, the sensitivity of the luciferase assays was not high enough to detect any leaky expression from the tetracycline repressed promoter, so this is less likely to be a key factor.

Underpinning all of these factors, and likely proving the key factor, is the copy number of the plasmid. The estimated copy number of the pCD6 origin in *C. difficile* is somewhere between 6-10 (Purdy et al., 2002). Even a small difference can increase the chances of early transposition significantly when the population scale is taken into account. Given that the copy number in *C. saccharoperbutylacetonicum* is potentially doubled to a minimum of 12 in stationary phase and is potentially as high as 30 during logarithmic growth, it is perhaps, ultimately, unsurprising that the odds heavily favour early transposition events. Even an extremely low level of de-repression could result in multiple transposon mutants. The data to support this is, in this author's opinion, conclusive. Ten different transconjugants all show evidence that transposition occurred multiple times prior to induction resulting in a handful of different mutant ORFs dominating the sequencing. When combined with the earlier similar finding in 3.7.4, it seems that early transpositions are all but inevitable.

The primary method of overcoming the issues would be to design a novel plasmid delivery system. This could incorporate up to two key changes, depending on the time allowed. Firstly, a change in

promoter could yield a system that is impervious to spontaneous de-repression. Whilst not discussed in detail here, this was attempted during this study. The P*xyl* promoter and *xylR* gene were cloned in the place of P*tet* and *tetR*. However, initial transposition efficiency tests yielded extremely low rates that were unusable at a workable scale. Further work to improve this is, no doubt, possible. Secondly, a different origin of replication could be cloned in the place of pCD6. The complexity in this lies in the difficulty in identifying appropriate origins of replication. The pMTL8000 series of vectors now includes nine possible Gram positive replicons, including pCD6 (Plasmidvectors.com). The understanding of the copy number of these plasmids is not well characterised and displays species variation. Of the nine, at least six are likely to be inappropriate with noted higher copy numbers: pCD6, pIM13 (Truffaut et al., 1989), pCB102 and pBP1 (Yu et al., 2012), pUB110 (Leonhardt, 1990), and pAMB1 (Pérez-Arellano et al., 2001). It would be worth investigating the copy number of p19, pIP404, and pCB101 in *C. saccharoperbutylacetonicum*. Alternatively, the current origin could be randomly mutated and selected for the desired copy number. Finally, the megaplasmid could be explored as a source of novel origins of low copy number given our findings in this chapter.

Several factors lay behind our decision not to further pursue the creation of a statistically significant TraDIS library. At the point at which we had processed the results in 3.10.3, there was relatively little time remaining for the project. Improving the delivery plasmid was a reasonable goal in this timeframe, however, in all likelihood, the best-case scenario would have yielded a statistically significant TraDIS library in the final weeks of the project. In the worst-case scenario, no library would be attained, and the project would appear short on practical results. Whilst the former would have been an extremely useful result, there are reasons why this alone was less useful than pursuing other lines of experimentation. The field of interest for this species is relatively small compared to most pathogens. Therefore, the generation of an essential gene list in laboratory conditions is less immediately useful than it might be in other species. Our 'rough' list of 677 ORFs without insertions, partially validated by pathway enrichment analysis, likely provides broadly the same usability to the field. Given that practical usability is the priority in a field such as this, it did not seem like a good use of resources to continue to pursue this line of investigation. To compliment this, the real interest in deriving such essential gene lists for industrial microbiology lies in the ability to examine the pressures placed on the cells during fermentations and other key steps in industrial applications (e.g., transformation, in the presence of inhibitors). Shorn of this possibility, pursuing TraDIS became much less interesting to both the author and the field in general.

Overall, this chapter was able to establish several useful practical results. Establishing the copy number of pCD6 in *C. saccharoperbutylacetonicum* will prove invaluable to future work. In addition, the practical establishment of the viability and utility of glass bead spreading shows a route to both better results and less plastic waste. The investigation into cryopreservation is also enlightening into the best way to preserve cells of long-term viability and provides reassurance that a significant number of cells do survive the process. Finally, the 677 ORFs without insertions found by this study should prove practical to the field moving forward.

Chapter IV – Sporulation and germination in *Clostridium saccharoperbutylacetonicum*

4.1 Introduction

4.1.1 Sporulation

Sporulation has long played an important role in our understanding and use of solventogenic *Clostridia,* with selection by heating being one the early methods used to isolate relatively pure cultures (Weizmann, 1919). In addition, the use of non-pathogenic *Clostridial* spores to deliver medical interventions has been mooted since the 1990s (Minton et al., 1995) and is currently being commercialised by CHAIN (Bradley et al., 2021). Despite this, relatively little work has been conducted to characterise sporulation across all solventogenic *Clostridia*, with particularly little available in the public domain. *Clostridium acetobutylicum* accounts for the majority of sporulation publications (e.g., Al-Hinai et al., 2014; Alsaker and Papoutsakis, 2005; Scotcher and Bennett, 2005; Steiner et al., 2011). Apart from some work conducted prior to modern naming classifications (e.g., Mackey and Morris, 1972), most investigations touching on sporulation are derived from examining transcriptomics across the growth cycle (e.g., Alsaker and Papoutsakis, 2005; Steiner et al., 2011) or from altering key sporulation genes (e.g., Jones et al., 2011).

In *C. saccharoperbutylacetonicum* the scope of published work is even narrower, consisting of just two studies that investigate sporulation in any capacity (Atmadjaja et al., 2019; Feng et al., 2020). However, the actual body of research conducted is likely to be much greater as companies such as Green Biologics/Biocleave have conducted internal studies. Despite this, there remain many aspects of the physical transformations, the structure and resistance to insults of the spores that are unknown. Information from Biocleave suggested that there may be two sporulation phenotypes depending on whether spores were generated in liquid broth or on solid media, which has not previously been reported for other species. This chapter aims to add to the knowledge of sporulation in *C. saccharoperbutylacetonicum* by examining the morphology and ultrastructure of sporulation and spores in the two conditions by phase contrast, fluorescence and transmission electron microscopy, and the characterising heat resistance.

4.1.2 Germination

As with sporulation, germination has received very little research attention if published studies are used as an indicator. In fact, to our knowledge, germination has been studied significantly less than sporulation. Likely, this is due to two key factors: 1) the relative ease of germinating spores on desired feedstocks and 2) the general desire to avoid sporulation, and therefore germination, when using solventogenic *Clostridia* to produce solvents. However, as the use of *Clostridia* outside of solvent production expands commercially – e.g., the business models of CHAIN and Biocleave – understanding germination, particularly how to induce it or prevent it, becomes more important. Since no study on germination has been published in *C. saccharoperbutylacetonicum*, the scope of potential work was great. Here, it was decided that the most practical course of action was to attempt to uncover a reliable germinant and a reliable germination inhibitor with which to manipulate *C. saccharoperbutylacetonicum* spores.

4.1.3 Aims and objectives

The aims and objectives of this chapter were therefore as follows:

1. Sporulate *C. saccharoperbutylacetonicum* in liquid broth and on solid media, monitor progress and collect samples
2. Establish the proportion of heat resistant CFUs over time in both conditions
3. Probe the morphology of sporulating cultures over time using phase contrast and fluorescence microscopy
4. Examine the ultrastructure of sporulating cells and spores using thin section transmission electron microscopy
5. Establish a method of purifying spores from sporulating cultures
6. Test candidate germinants on purified spores
7. Test candidate germination inhibitors on purified spores
8. Test the effects of candidate germinants and inhibitors on growing cultures

4.2 Sporulation

4.2.1 The successful use of Nile Red for fluorescence microscopy

To monitor sporulation progression, samples were examined under phase contrast and fluorescence microscopy. In the latter case, a membrane dye – Nile Red – was selected. The dye is able to insert into the membrane and fluoresce red under excitation by a green laser. Whilst there is no reason to suspect that the dye would be unable to function in *C. saccharoperbutylacetonicum*, it was important to verify that no interfering signal is generated by cells under the same conditions, as is seen for green autofluorescence in *C. difficile* (Oliveira Paiva et al., 2022; Ransom et al., 2015). Therefore, one sample was taken to prepare and visualise under the microscope prior to conducting a full-scale experiment. The dye to worked well, being able to indicate the cell membrane, invagination during cell division and engulfment during sporulation (Figure 4.1). There was low background fluorescence as evidenced by the no-stain controls. Therefore, the full-scale experiments were conducted with Nile Red.

4.2.2 Sporulation on solid media

4.2.2.1 Determination of sporulation rates and heat resistance

The rate of sporulation in broth and the heat resistance of produced spores has been previously investigated by Sasha Atmadjaja at Biocleave (personal communication). However, sporulation on solid media has not been previously studied. There were several ways these experiments could have been conducted, however, to equalise the difference created by different generation methodologies between liquid and solid media sporulation, I opted to control OD_{600nm} in both conditions. For the solid media sporulation process, cells were grown on RCM agar supplemented with 50 g/L γ-cyclodextrin and harvested at a given timepoint with PBS. The OD_{600nm} was then measured, and each repeat adjusted to OD_{600nm} 1 prior to heat treatment. Glucose appears to inhibit the sporulation of *C. saccharoperbutylacetonicum* and so cells grown on RCM agar supplemented with 50 g/L glucose were used as a non-sporulating negative control. Cells were treated at 60°C, 70°C or 80°C or left at room temperature on the bench for 15 min before growth and enumeration on non-selective RCM agar.

Over the course of the experiment, the proportion of heat resistant CFU/mL increased over time (Figure 4.2). The latter increases from just 0.00168% of total CFUs at 24 h to 15.17% at 120 h. After 24 h, the total CFU/mL in OD_{600nm} 1 of cells drops from 6.4×10^7 to 3.89×10^6. Interestingly, no such drop is

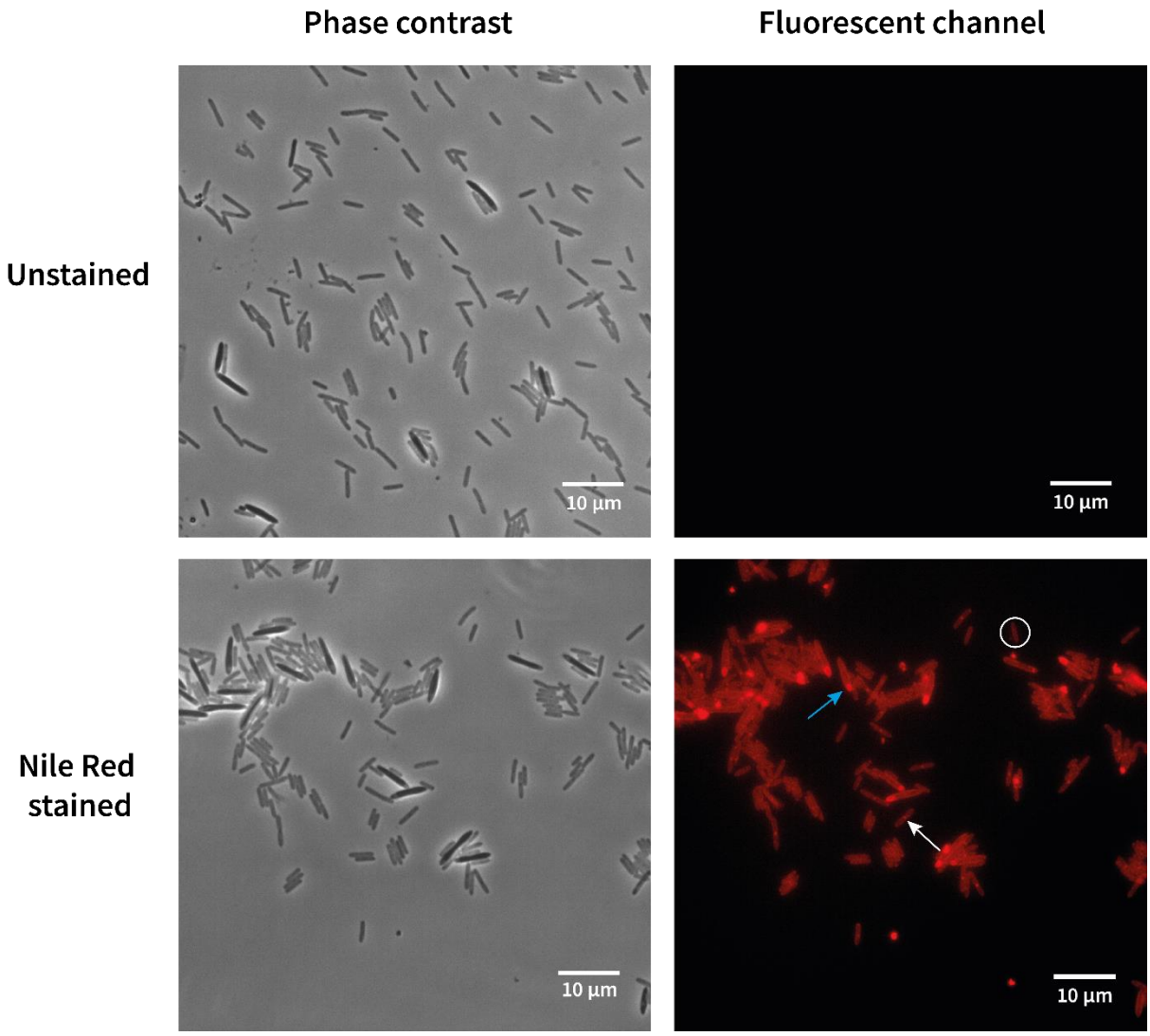

Figure 4.1 Sporulating cultures of *C. saccharoperbutylacetonicum* stained and unstained with Nile Red. Samples from the same sporulating culture either unstained or stained as visualised by phase contrast and fluorescence microscopy. No autofluorescence is seen for unstained samples. Clear staining is seen for samples prepared with Nile Red. Several features can be observed included a vegetative cell (white circle), dividing cells (white arrow) and sporulating cells with forespore (blue arrow).

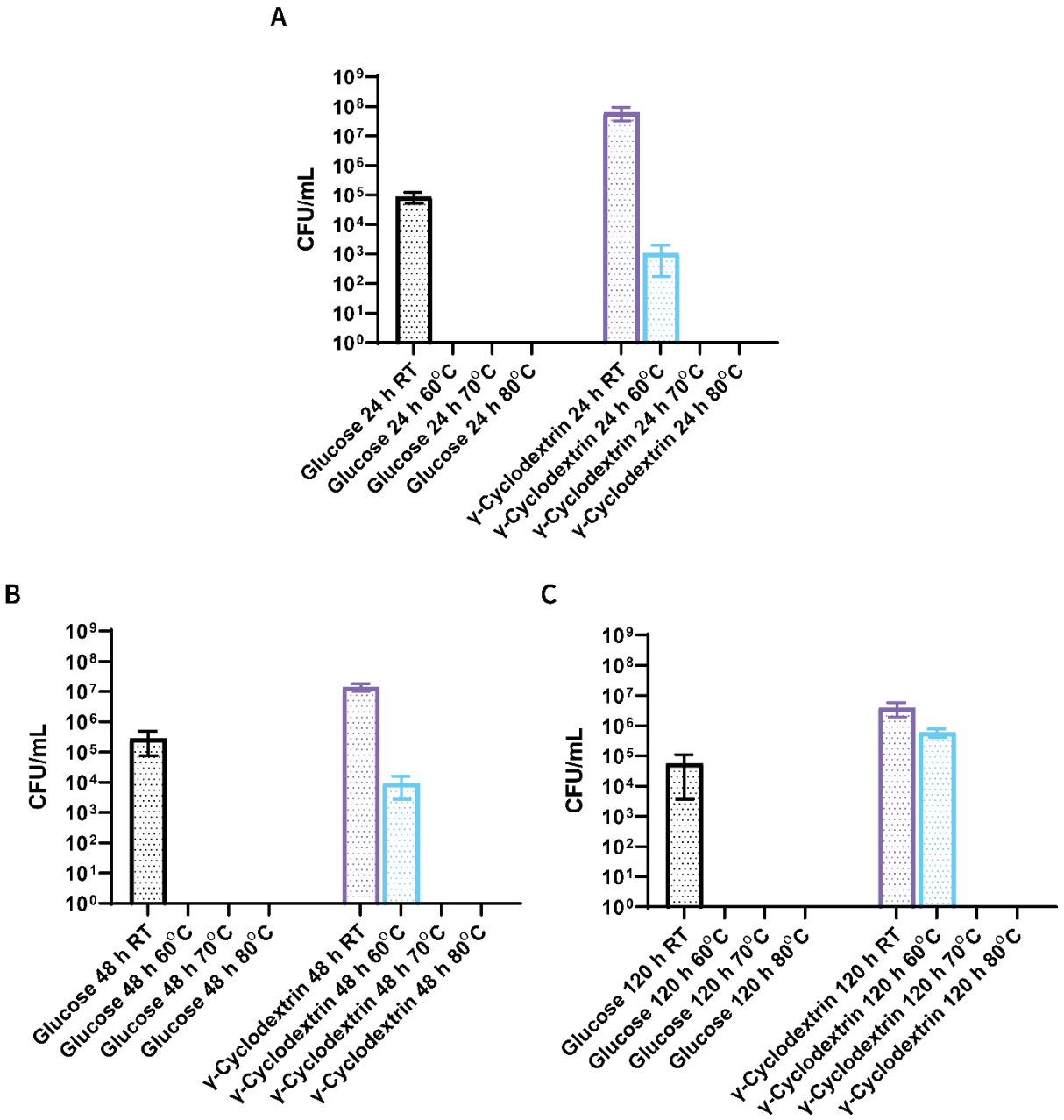

Figure 4.2 Heat resistance of *C. saccharoperbutylacetonicum* prepared on glucose or γ-cyclodextrin TYIR solid media over time. A) The average CFU/mL after 24 h in both conditions. **B)** The average CFU/mL after 48 h in both conditions. **C)** The average CFU/mL after 120 h in both conditions. The experiment was performed in biological duplicate and technical duplicate. Each technical repeat was spotted four times from three different dilutions.

seen for cells in the glucose condition which oscillate around the 1 x 10^5 mark for the duration of the experiment. The total CFU/mL in the glucose condition is consistently lower than that in the γ-cyclodextrin. Whilst this could be due to dilution error, it would be unlikely that this would be replicated consistently in the same manner across three different days and multiple repeats. The experiment seems to confirm the previously observed inability of cells grown on glucose to sporulate.

Spores generated in the presence of γ-cyclodextrin were heat resistant only up to 60°C under these conditions. This contradicted previous studies which suggested spores can survive up to 70°C. It must be noted that one repeat at 120 h did show growth after heating at 70°C, however, as this was the only example, the sample were considered contaminated, and the data excluded from the analysis. No growth was seen in any condition after heating at 80°C. Overall, *C. saccharoperbutylacetonicum* spores generated on agar plates were reliably heat resistant up to 60°C.

4.2.2.2 Morphology of sporulating cultures

To observe and monitor the progression of sporulation, phase contrast and fluorescence microscopy were deployed. For fluorescence, Nile Red, a membrane dye, was used to render cell membranes red. The key markers of sporulation progression were deemed to be enlargement of the vegetative cell, asymmetric membrane formation at one pole, complete engulfment of one pole and, finally, the presence of released spores. In phase contrast, the first of these can be easily observed as well as the development of a bright pole as the spore matures and optical properties change. Naturally, it was anticipated that the final released spore would also be bright under phase contrast as has been described for other Firmicutes. The experiment was conducted over seven days and samples prepared each day.

After the first day, a large population of the cells were already undergoing the initial stages of sporulation (Figure 4.3). They exhibited the classic *Clostridial* form of massively enlarged cell approximately double the length of a vegetative cell (~7 µm vs ~3.5 µm). Under fluorescence it was possible to see that, during the *Clostridial* form, membrane engulfment of the pre-spore was already underway. Cells that had progressed further along the sporulation pathway were exhibiting bulbous ends and some bright spots were apparent on phase contrast. The latter observation indicated a spore that is reaching maturity; however, no released spores were seen at this time point. Whilst the

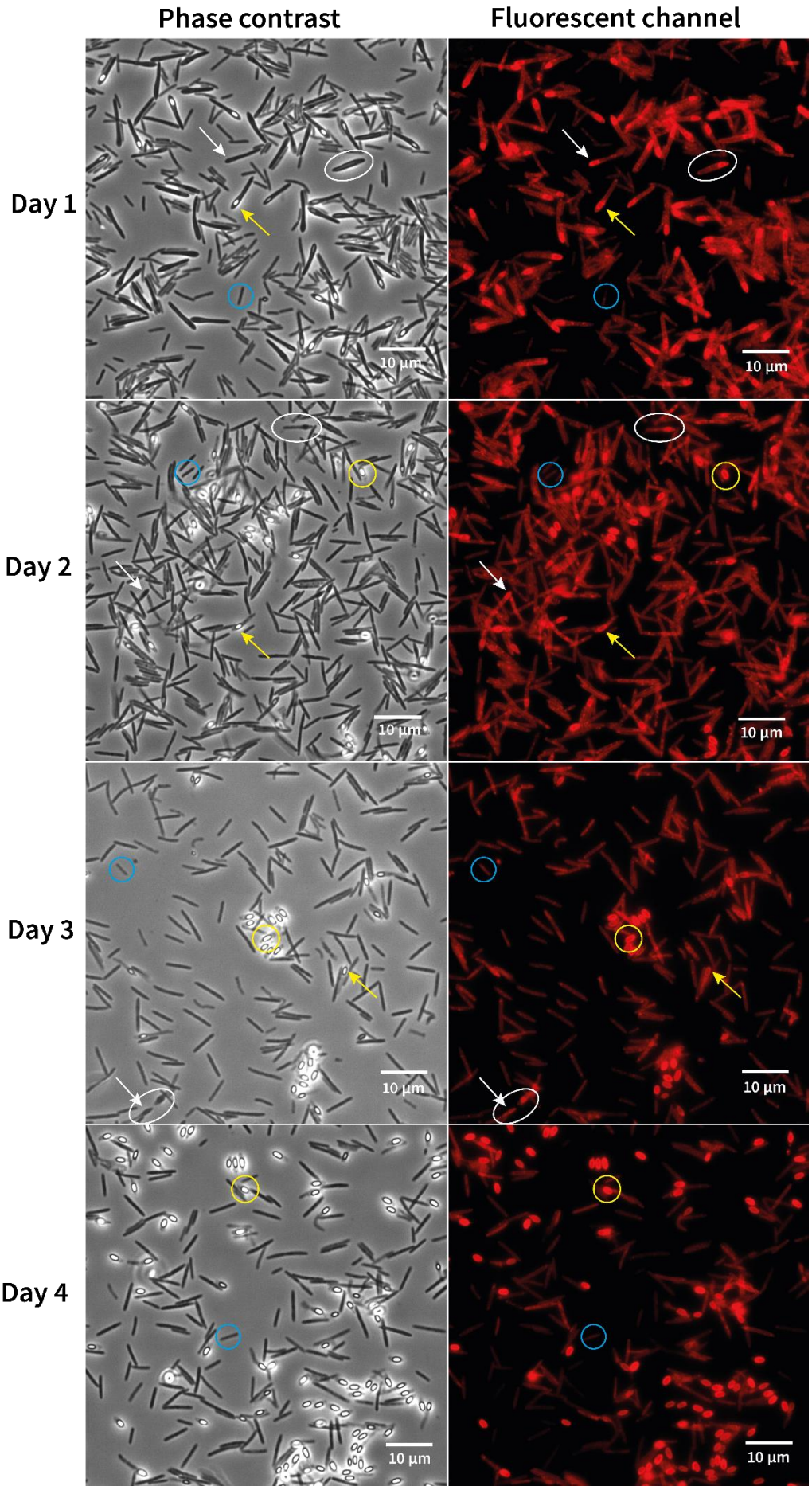

Figure 4.3 The morphology of day 1-4 cultures of *C. saccharoperbutylacetonicum*

sporulating on solid media. Phase contrast and fluorescence microscopy of sporulating cultures with examples of key features highlighted. Blue circles highlight vegetative cells, white ovals highlight large *Clostridial* form cells, white arrows indicate engulfment of forespores not yet phase bright, yellow arrows indicate forespores that have become phase bright, and yellow circles highlight released spores.

proportions were difficult to calculate from these images, the majority of cells at this timepoint retained vegetative morphology, reflecting their prevalence as estimated by the heat resistance assays in 4.2.2.1. Finally, it is worth noting that many of the *Clostridial* form cells appeared to contain membrane vesicles as indicated by small patches of fluorescence.

On day 2, it was possible to observe the first released spores. These appeared bright in both phase contrast and fluorescence channels. For the former, this is due to the dehydrated spore core whilst for the latter it is likely due to the two membrane layers that surround the spore as seen in other spores (Diallo et al., 2021). With phase contrast, it was possible to observe a wider, translucent outer layer surrounding the spore. These do not appear to be visible with fluorescence and so may indicate a non-membrane structure, possibly an exosporium. There already appeared to be fewer *Clostridial* form cells at this timepoint, although there were still sporulating cells with bright poles. The average non-spore forming cell appeared to be enlarged when compared to day 1, though smaller vegetative cells were also visible. This is consistent with observations during normal cell growth whereby stationary phase cultures appear to contain large cells as well as chains of cells.

Day 3 showed the emergence of two distinct populations: vegetative cells and spores. Few sporulating cells remained, with even fewer cells exhibiting the *Clostridial* form. Released spores again appeared to display the additional non-membranous outer layer to the spore. In addition, they appeared to be forming clumps, suggesting an adhesive quality to the spores that is common in other species (Paredes-Sabja and Sarker, 2012). Though caution should be taken however given the series of preparatory steps necessary prior to preparation of microscopy slides. Whilst it was only one image from one slide, it is worth noting that the proportion of spores to non-spore cells (12%; 26/~216) appears similar to that seen for the heat resistance assays after 5 days (15%) suggesting that the preparations were relatively representative.

Following the trend from day 3, day 4 showed only two populations – spores and vegetative cells. This strong differentiation suggests a tight control of the decision to not-sporulate that is not reversed once taken. Again, spores appeared to have a loose coat layer and displayed a tendency to clump, though somewhat less pronounced than previously. Days 5 was similar to day 4, though a lack of cell density across all images taken (including those not shown) suggest poor preparation (Figure 4.4). It is possible to see that some sporulating cells remained, in contrast to day 4. Day 6 again showed the same distribution; however, many of the vegetative cells now appeared to be elongated, with chains and irregular shapes also common. Together this suggests that more cells are dying at this stage, reflecting the CFU trend observed for day 5 of the heat resistance assays.

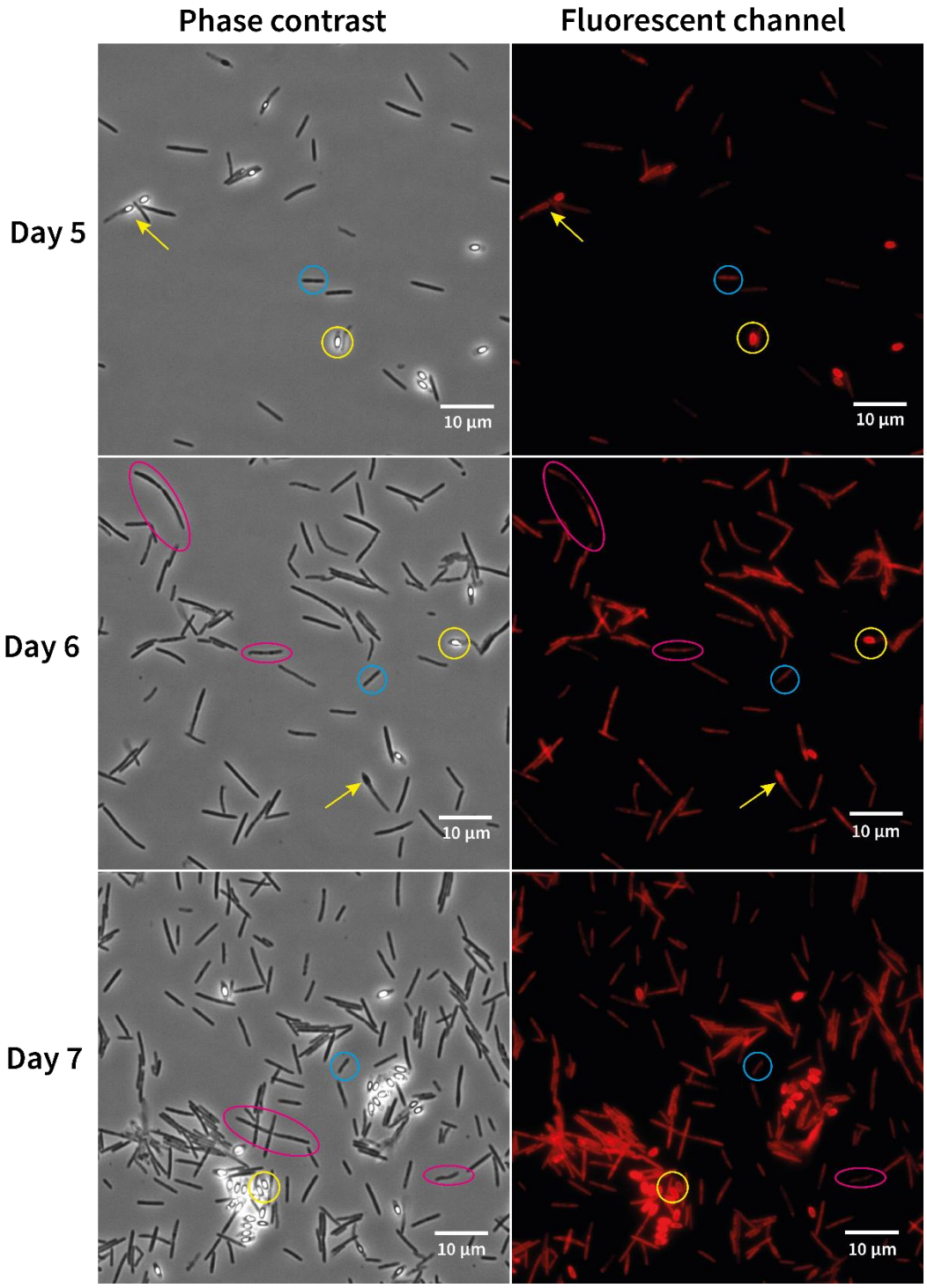

Figure 4.4 The morphology of day 5-7 cultures of *C. saccharoperbutylacetonicum* sporulating on solid media. Phase contrast and fluorescence microscopy of sporulating cultures with examples of key features highlighted. Blue circles highlight vegetative cells, white ovals highlight large *Clostridial* form cells, yellow arrows indicate forespores that have become phase bright, yellow circles highlight released spores and magenta ovals highlight distinctive irregular morphologies of putatively vegetative cells.

A single sporulating cell was seen for this timepoint. Once again, on day 7, clumps of spores and vegetative cells were seen but no sporulating cells. Again, the non-sporulating cells showed the forms seen on day 6.

4.2.2.3 The ultrastructure of sporulating cultures

Thin-section transmission electron microscopy was conducted on samples of sporulating cultures from day 1, day 3 and day 7. The available fields of view on day 1 showed the distribution of phenotypes similar to that seen for the above light microscopy (Figure 4.5A and D) with vegetative cells, *Clostridial* form cells, sporulating cells and spores all visible. Shown here are two developing spores at different stages of maturity (Figure 4.5B and C). In the first, the mother cell had already divided asymmetrically and engulfed the forespore. The core of the developing spore had not yet undergone the extensive dehydration and subsequent increase in density seen for more mature spores and is thus seen as larger and similar in density to the cytoplasm of the mother cell. The cell wall/ inner membrane of surrounding the core is clearly visible. The first signs of growth spore coat were also visible as dense, dark lines. The mother cell contained extensive depositions of granulose that likely could not be penetrated by the stain and hence showed as bright white globules.

The second image of a developing spore showed the latter stages of development within the mother cell (Figure 4.5C). The thin sectioning had cut the forespore across the top of the mother cell pole containing the spore. The forespore was still contained in the mother cell as evidenced by the granulose still visibly surrounding the spore. The core was denser which showed further progression towards maturity. A clear cortex, seen as a clear zone between the core and the spore coat, had developed. Neither the cell wall nor the inner membrane surrounding the spore core are visible, though this is not unusual for this method chemical fixation (personal communication, Dr Ainhoa Dafis-Sagarmendi). An additional feature, possibly consisting of protein, was visible between the cortex and the developing spore coat. Two potential coating layers were already visible – an inner spore coat and a middle spore coat (nomenclature takes into account the later visible outer spore coat). Neither spore coat appeared to be fully constructed as they presented in discontinuous segments at this cut and angle.

Finally, a released spores could be seen at this timepoint (Figure 4.5E). A transverse cut of the released showed no evidence of the mother cell (e.g., granulose, additional cell envelope). The core appeared dense, with lighter sections likely due to the stain being unable to penetrate the core. The cortex was plainly visible, though the additional feature seen previously and in future images

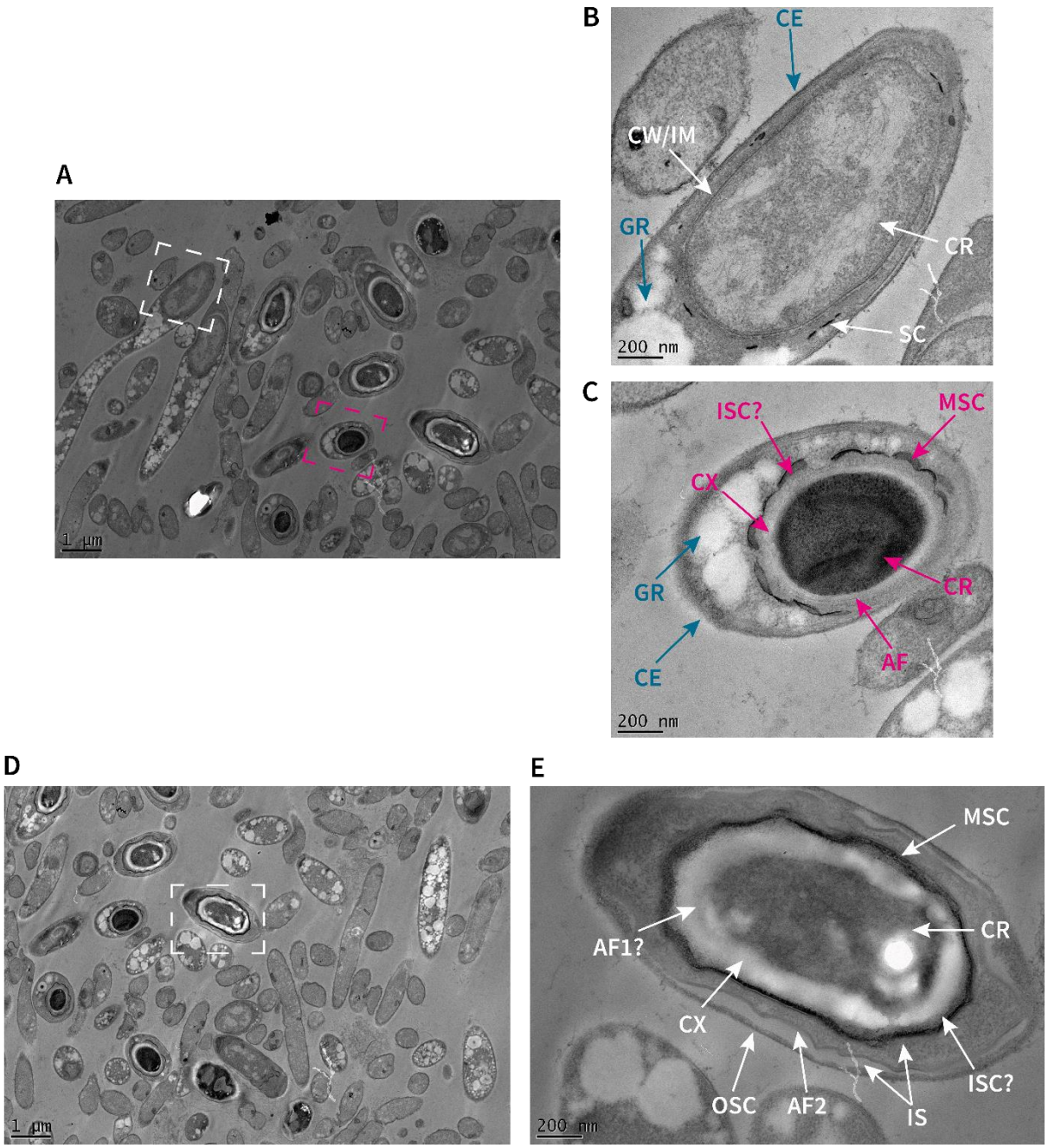

Figure 4.5 Ultrastructures on day 1 of sporulating cells in solid media. A) The field of view of (B) and (C). Multiple released spores are observable as well as developing spores and vegetative cells. The structural features in (B) and (C) are highlighted by the white and magenta boxes respectively. **B)** A transversal cut of developing spore. **C)** An endways cut of a developing spore at a later stage of maturity **D)** The field of view surrounding (E) again showed a mix of spores, sporulating cells and vegetative cells. (A) and (D) overlap. **E)** A transversal cut of a released spore. The fields of view are x1900 magnification whilst the highly magnified images are x11000. Labels in white and magenta highlight spore features whilst those in teal highlight features of the mother cell. CR: core; CX: cortex; ISC: inner spore coat; MSC: middle spore coat; AF: additional feature

(used to highlight a variety of prominent but unknown structures); IS: interspace; OSC; outer spore coat; CW/IM: cell wall/inner membrane; GR: granulose deposits; CE: mother cell envelope.

was not clearly observable. The densest spore coating layers appeared complete and contiguous round the whole spore core. It is not clear if there were two or only one spore coat at this location. An additional coating layer (additional feature 2) that appeared to contain denser contents is also visible before the outer spore coat. This additional layer resulted in two different interspaces being visible. The outer spore coat could have been a paracrystalline exosporium, however this cannot be confirmed from these images.

Overall, it is possible to recognise cells at all stages of sporulation at this timepoint and some of the key features of the *C. saccharoperbutylacetonicum* spore. This highlights the asynchronous nature of the sporulation amongst the population and how the decision to sporulate can occur significantly later for some cells. Notably, the spore coats are not close to the spore core as is the case in *C. difficile* (Pizarro-Guajardo et al., 2016). Finally, it appeared that it was possible for spores to reach a high level of maturity after 24 h sporulating on solid media, however, a comprehensive assignment of features was not possible based on the images recorded.

At day 3, the cells visible did not align closely to the morphologies seen in 4.2.2.2, however, released spores were visible (Figure 4.6A and C). Shown here are two released spores, one cut lengthways, retaining the core as an oval (Figure 4.6B), and the other endways, showing the core as a circle (Figure 4.6D). Both displayed ultrastructures similar to those seen previously. The dense core and clear cortex were both evident. The unknown additional feature could be seen in B but not D, likely due to the angle of cutting. Both showed clear outer and middle spore coats with an inner spore coat apparent on B. Again, the interspace between coat layers appeared to contain a variety of discontinuous features. Given the lack of significant differences between the released spores on day 1 and day 3, it is tempting to conclude these are fully mature spores. However, it is likely that full maturation of the spore coat takes several days as seen for other species (Henriques and Moran, 2007).

On day 7, the general morphologies on display were similar to those seen in 4.2.2.2, though not identical (Figure 4.7A and C). The released spores shown in B and D appeared to be sectioned transversally, but closer to the edge of the transverse than previously imaged. In D, the consequence was that a dense spore core is no longer visible with the transverse section likely cutting across the cortex rather than the core. Additionally, the core may not have stained due to lack of penetration from the stain. As for day 3, the spore layers were well-defined, suggesting maturity. However, there remained a number of ambiguities. The additional feature previously seen within the cortex was no longer visible. Instead, in D, an additional feature was seen close to the outer spore coat, potentially

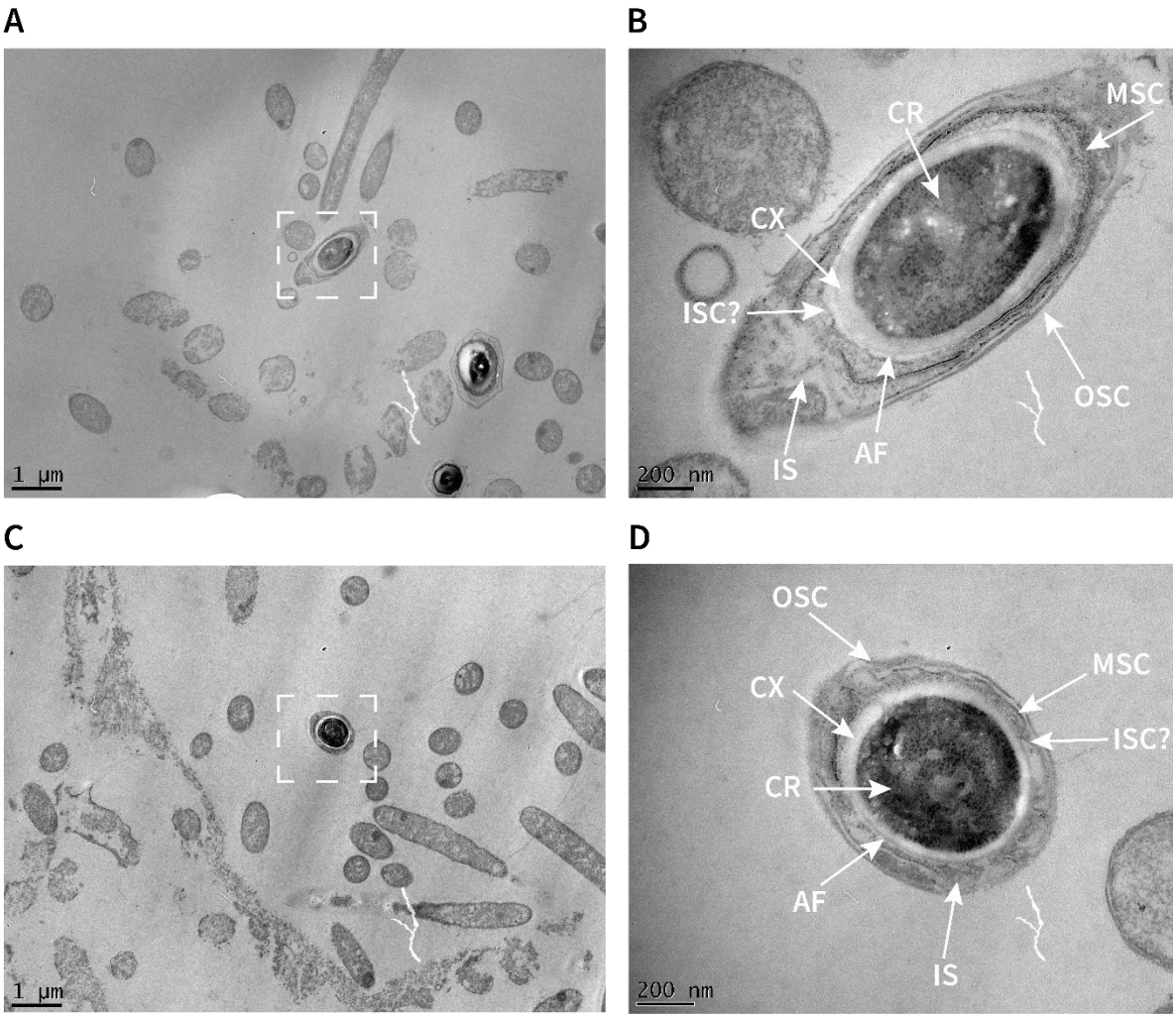

Figure 4.6 Ultrastructures on day 3 of sporulating cells in solid media. A) The field of view of (B) showing released spores, some vegetative cells and cell debris. **B)** A transversal section of a released spore. **C)** The field of view of (D) showing a released spore, vegetative cells and cell debris. **D)** An endways section of a released spore. The fields of view are x1900 magnification whilst the highly magnified images are x11000. Labels highlight the spore features. CR: core; CX: cortex; ISC: inner spore coat; MSC: middle spore coat; AF: additional feature (used to highlight a variety of prominent but unknown structures); IS: interspace; OSC; outer spore coat.

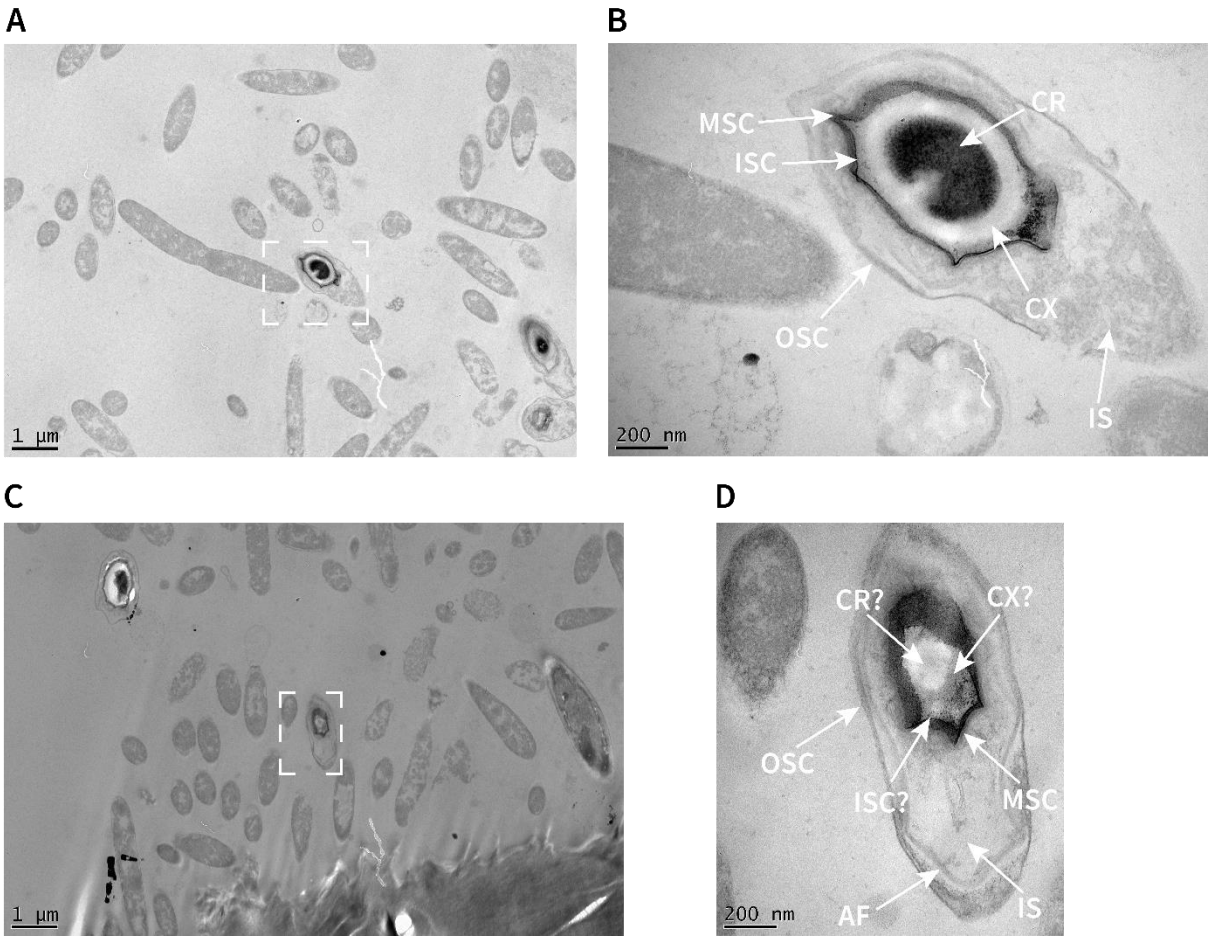

Figure 4.7 Ultrastructures on day 7 of sporulating cells in solid media. A) The field of view of (B) showing released spores, vegetative cells and cell debris. **B)** A transversal cut of a released spore the core and. **C)** The field of view of (D) showing released spores and vegetative cells. **D)** A transversal cut of a released spore that appears to exclude the core from the cut. The fields of view are x1900 magnification whilst the highly magnified images are x11000. Labels highlight the spore features. CR: core; CX: cortex; ISC: inner spore coat; MSC: middle spore coat; AF: additional feature (used to highlight a variety of prominent but unknown structures); IS: interspace; OSC; outer spore coat.

representing another matured spore coat. It may be that this coat was present on previous images but was more difficult to discern given its proximity to the outer coat, as it is in B. The inner and middle spore coats may actually have been a single uniform layer when judged on the day 7 images.

Together, these images have shown features seen in many *Clostridial* spores. However, the overall picture is unique when compared to other spore morphologies. The ultrastructures presented suggest that the released spores are able to reach a stable and mature state from at least day 3 and possibly even day 1. It is clear that the protection surrounding the core in *C. saccharoperbutylacetonicum* is formed of multiple layers that create large interspace(s) between the core and the outermost protective layers. It is also possible that the outer spore coat is a paracrystalline exosporium such as that seen in *C. sporogenes*.

4.2.3 Sporulation in liquid media

4.2.3.1 Determination of sporulation rates and heat resistance

To investigate phenotypic differences between spores generated on solid and liquid media, the same experiments described in 4.2.2.1 and 4.2.2.2 were conducted using a method for generating spores in liquid media developed by Sasha Atmadjaja at Biocleave. She found that it was necessary to introduce a subculturing step in γ-cyclodextrin in order for cells to reliably sporulate. Therefore, the experimental outline for liquid generated spores was slightly altered compared to generation on solid media and early sporulation timepoints were more difficult to sample. Due to working hour restrictions, early timepoints was missed. For similar reasons, 120 h samples were also not taken. As before, all samples were adjusted to OD_{600nm} 1 and heated for 15 min at the specified temperature.

Heat resistant CFUs were apparent after 24 h of sporulation in liquid media (Figure 4.8A). The proportion of heat resistance was lower than that seen on solid media (0.00054% vs 0.0017%). As expected, no heat resistant CFUs were observed for cells grown in the presence of glucose, despite comparable total CFUs (2.7×10^7 CFU/mL for glucose vs 3.6×10^7 CFU/mL for γ-cyclodextrin). There is a significant difference between the CFU/mL observed for growth on glucose in solid media and glucose in liquid media, though it is unclear why this might be the case for glucose and not for γ-cyclodextrin. Heat resistant spores did not survive incubation at 70°C or 80°C for the sporulating cultures.

At 48 h, both the amount and proportion of heat resistant CFUs in the γ-cyclodextrin condition increased up to 0.012% of the total culture, which is approximately 6-fold less than in the solid.

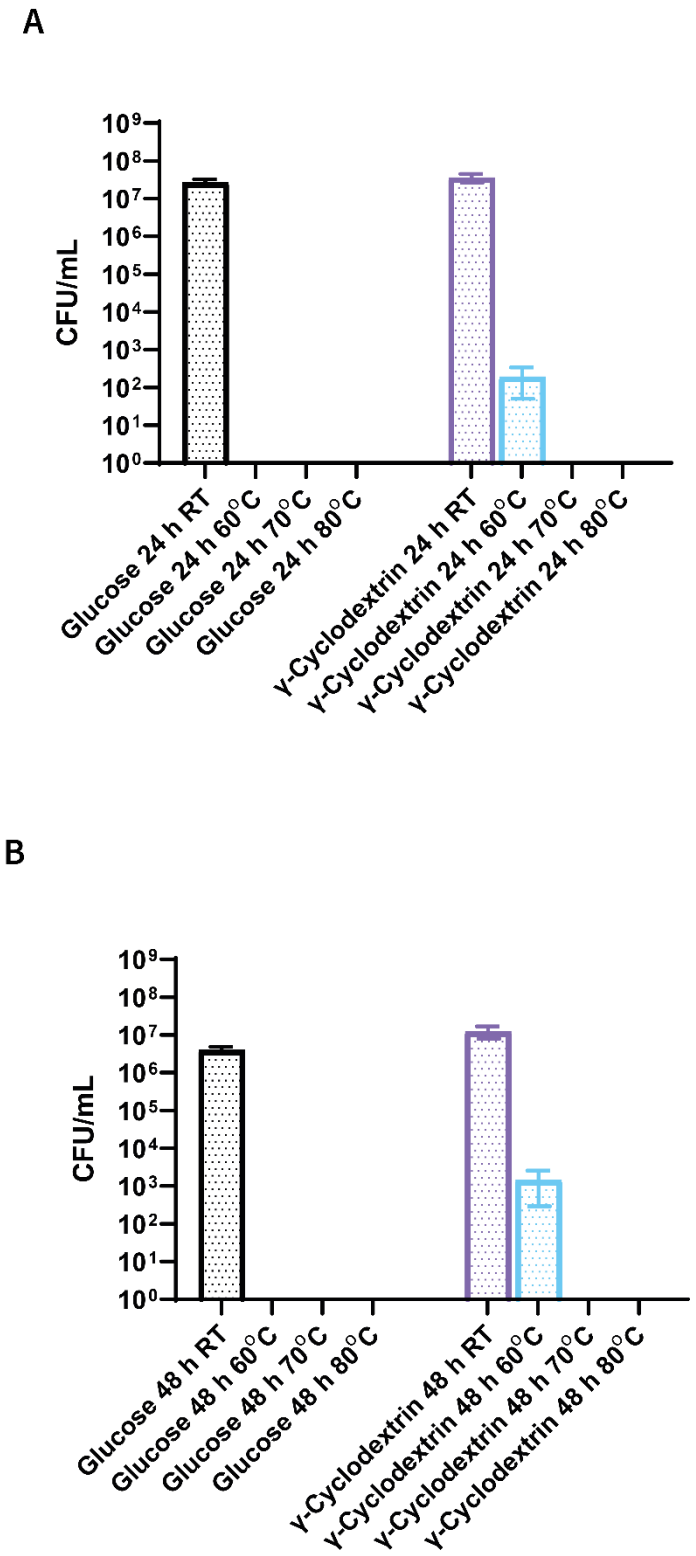

Figure 4.8 Heat resistance of *C. saccharoperbutylacetonicum* prepared on glucose or γ-cyclodextrin TYIR liquid media over time. A) The average CFU/mL after 24 h in both conditions. **B)** The average CFU/mL after 48 h in both conditions. The experiment was performed in biological duplicate and technical duplicate. Each technical repeat was spotted four times from three different dilutions.

media sporulation condition (Figure 4.8B). This represents a 21.5-fold proportional increase in heat resistant CFUs over the 24 h timepoint which is lower than that for solid media spores at the same stage (39.5-fold). The total CFU/mL count for dropped to 1.24 x 10^7, reflecting the increasing cell death associated with stationary phase cultures and also seen in the solid media experiments. This effect was also seen in the glucose condition at 48 h, with total CFU/mL dropping an order of magnitude to 4.03 x 10^6. Unlike in the solid sporulation experiment, no growth was seen after incubation at 70°C, nor any at 80°C, suggesting that the spores present may have lower heat resistance

4.2.3.2 Morphology of sporulating cultures

Phase contrast and fluorescence microscopy using Nile Red were again deployed to investigate the progression of sporulation in liquid media, with the same key markers of sporulation progression used to follow the process. The flexibility of liquid culture manipulation compared to plates allowed for earlier timepoints to be examined. Unpublished data suggested that spores created on liquid are not fully released from the mother cell and therefore may be phenotypically different (personal communication, Sasha Atmadjaja).

After 4 h in sporulation conditions, the initial stages of sporulation were already apparent (Figure 4.9). *Clostridial* form cells were present, with most at various stages of engulfment of the forespore as indicated by the fluorescence images. Phase contrast showed no bright poles indicating that spore development had not yet reached maturation at this time point. Both images techniques showed vegetative cells growing and dividing with the septum of the dividing cells easily identifiable.

At 6 h, the picture was similar to that seen after 4 h. *Clostridial* form cells were present in similar proportions and several showed putatively complete engulfment. No bright poles were observed, again suggesting that maturation of the spore requires more time. Dividing vegetative cells were still readily observable at this time, in line with the increasing OD_{600nm} over this time period.

8 h into the sporulation process, the *Clostridial* form cells were starting to increase in asymmetry. A prominent bulbous end at one pole developed in the cells that had advanced the most along the sporulation pathway. This timepoint also saw the emergence of bright spots of fluorescence within some *Clostridial* form cells, implying vesicle formation. Growing and dividing vegetative cells were still a significant presence and no phase bright structures were seen. However, after 10 h, new features could be observed in the sporulating population. The bulbous pole of some sporulating cells started to become brighter, indicating the forespore may be entering the final stages of

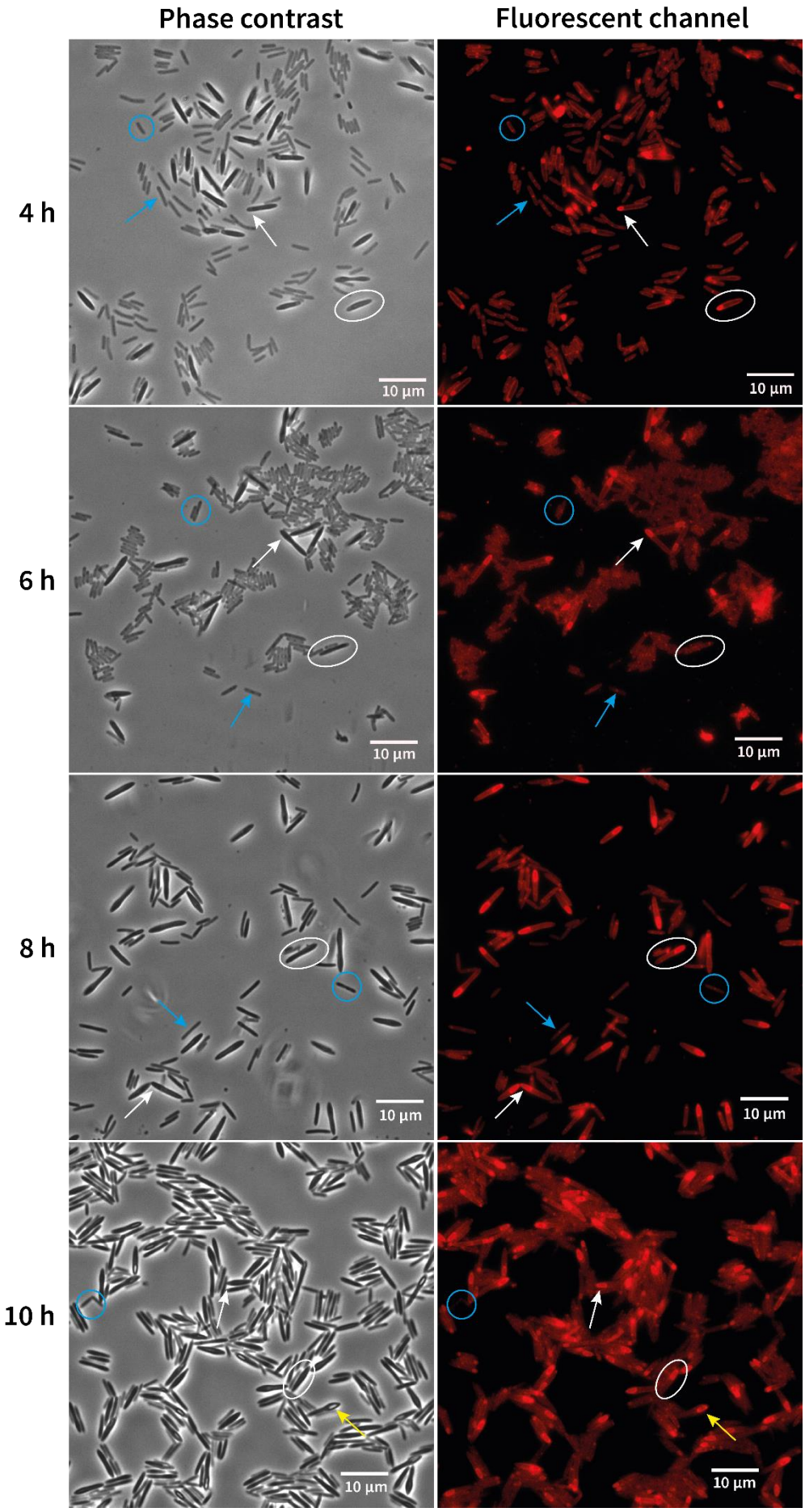

Figure 4.9 The morphology of 4-10 h cultures of *C. saccharoperbutylacetonicum*

sporulating in liquid media. Phase contrast and fluorescence microscopy of sporulating cultures with examples of key features highlighted. Blue circles highlight vegetative cells, blue arrows indicate the septum of dividing vegetative cells, white ovals highlight large *Clostridial* form cells, yellow arrows indicate forespores that have become phase bright.

maturation. By this timepoint, a large number of cells had developed the *Clostridial* form with even those retaining more symmetrical rod shape appearing to have enlarged compared to the vegetative cells seen earlier. It was again possible to see the separate vesicle-like spots at this timepoint.

After 12 h, there was an increase in the number of sporulating cells with bulbous, lighter poles seen in phase contrast (Figure 4.10). However, nothing resembling a mature spore was yet observable. A large number of cells appeared to either be in *Clostridial* form or otherwise enlarged, though many did not exhibit signs of engulfment despite appearing similar to those that did in phase contrast. Typical, small vegetative cells were still present, but in much lower quantities than at earlier timepoints. The ones that were present did not show up as brightly dyed by Nile Red, which may be indicative of cell death, and no obviously dividing cells were seen in this image. The vesicle-like structures can again be seen in some cells.

A difference between the liquid sporulation and solid sporulation process can be seen when the 24 h timepoint is compared. In liquid, the population appeared more uniform in shape with large *Clostridial* form cells dominating. However, the fluorescence highlighted more readily the heterogeneity of the culture in which a number of the enlarged cells contained neither a forespore nor appeared to be undergoing engulfment. Instead, cells that were clearly sporulating were starting to reach maturity with bright poles visible for several cells under phase contrast. Of the four cells that showed the latter phenotype, three were clearly still contained within the mother cell membrane, a phenomenon also observed for sporulation on solid media at this time. The fourth phase bright spore presented more ambiguously. It did not obviously appear to be contained in the mother cell, suggesting it had been successfully released. However, the potential spore did not resemble the released spores seen during the solid sporulation process. The core appeared somewhat elongated and the surrounding additional layer seemed to be dyed in the fluorescent image. This could indicate that the spore did, in fact, still contain the mother cell membrane around it but is orientated such that most of it is further from the focal point than the spore end. Observations of the images taken in the Z axis of this slide were, unfortunately, ambiguous on this point. Some smaller vegetative cells were still visible though, again, no dividing cells could be observed.

Despite the proportional increase in heat-resistant CFUs at 48 h seen in 4.2.3.1, there was no increase in the prevalence of spores – released or otherwise. Indeed, only a single phase bright spore could be seen in this image. This spore did appear to be released, though again, the morphology

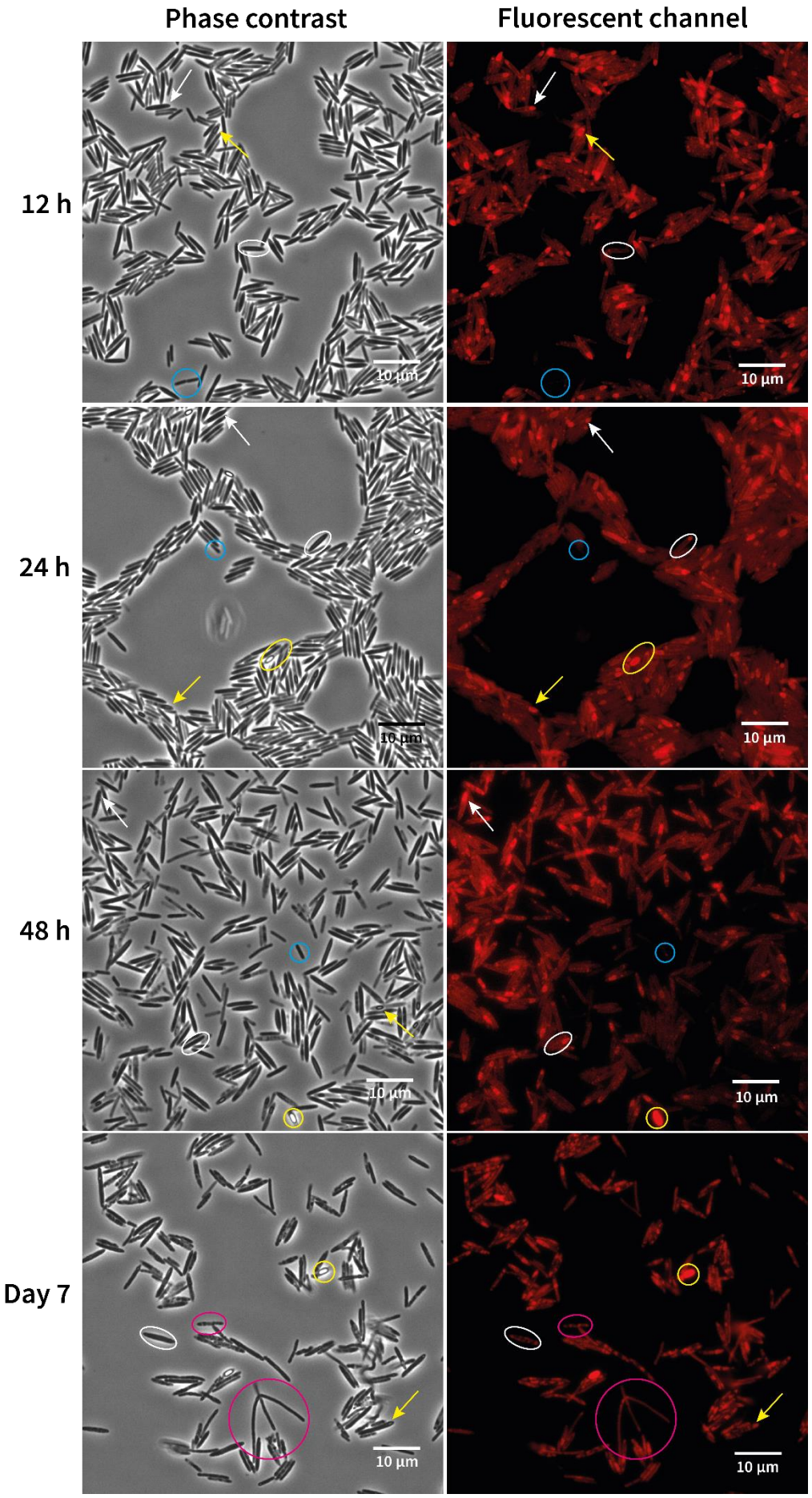

Figure 4.10 The morphology of 12 h – 7 day cultures of *C. saccharoperbutylacetonicum*

sporulating in liquid media. Phase contrast and fluorescence microscopy of sporulating cultures with examples of key features highlighted. Blue circles highlight vegetative cells, white ovals highlight large *Clostridial* form cells, yellow arrows indicate forespores, yellow circles highlight released spores and magenta ovals/circles highlight distinctive irregular morphologies of putatively vegetative cells.

seemed to differ from that seen for solid media, not exhibiting the pale outer layer seen in Figure 4.3 and 4.4, but rather this layer showed as darker on phase contrast and brighter on fluorescence. Again, this may simply be an issue of focus and plane, but, given the prevalence of spores in the solid media sporulation images, it is noteworthy that no spore in those images seemed to display a similar structure. There were sporulating cells with a bulbous, lightening pole though none appeared to be distinctly phase bright. Large cells were, again, ubiquitous, with a similar distribution of *Clostridial* forms and engulfments to that seen at 24 h.

Finally, after 7 days in liquid media, it is likely most non-sporulated cells had died, and distinct late stationary phase structures could be seen. These included overly large cells, chains of cells and distinct patchy cells visible both in phase contrast and fluorescence. These seem to imply a breakdown in cell structure, though these cells continued visibility in the fluorescence images suggests that the cell membrane is largely intact, in which case the damage to the overall cell must've been relatively small (though potentially no less fatal). Two spores could be observed that both appeared released from the cell, though again the structure is not exactly like those seen in the solid sporulation images. Light-looking forespore/spore structures could still be seen at the poles of some of the larger cells. Vegetative cells were no longer present.

4.2.3.3 The ultrastructure of sporulating cultures

To probe any differences in the ultrastructure of spores between those generated on solid and liquid media, thin section TEM was again conducted. The liquid media allowed for the capturing of earlier timepoints that would be difficult to sample for spores generated on plates. As for the light microscopy, a mixed culture of vegetative cells, *Clostridial* forms and actively sporulating cells could already been seen after 4 h in the sporulation media (Figure 4.11A and C). The higher magnification images showed sporulating cells at earlier stages than seen previously (Figure 4.11B and D). As before, the future core had not yet densified and had likely only recently been engulfed by the mother cell. Unlike the developing spore on day 1 on solid media, these spores did not yet show signs of building the spore coat layers. However, the emerging clarified region in both B and D suggested the development of the cortex was already underway. Interestingly, the immediate spore core envelope – typically consisting of inner membrane and cell wall – was clearly visible at this timepoint. D showed an additional feature between the mother cell envelope and the spore core envelope that was likely related to the construction of the spore coat. Not of direct relevance to this study but, nonetheless intriguing, was the presence, highlighted in white ovals, of unusual cell

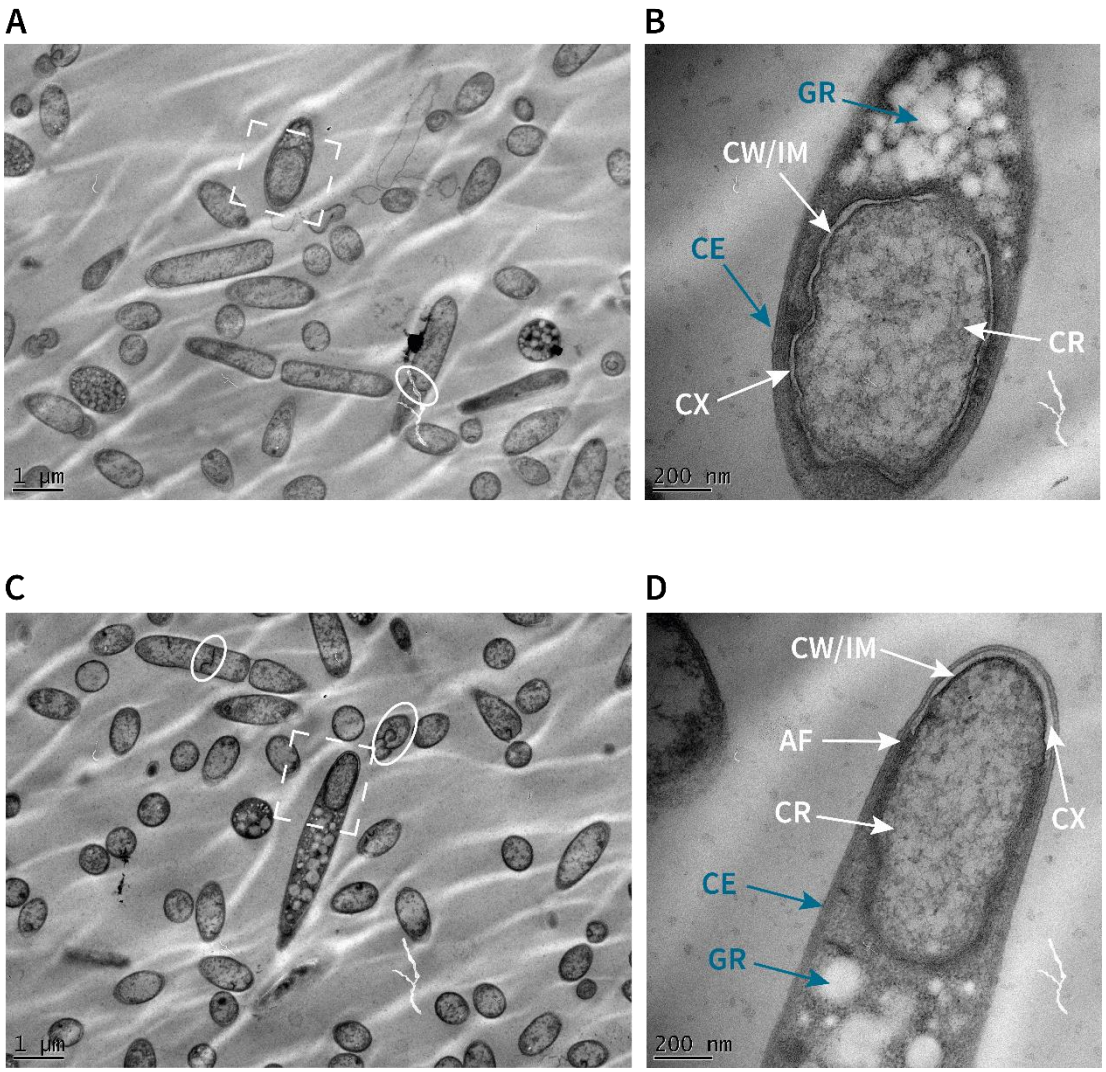

Figure 4.11 Ultrastructures at 4 h of sporulating cells in liquid media. A) The field of view of (B) showing largely vegetative cells with some *Clostridial* forms, developing spores and distinctive puzzle-like cell envelopes between dividing cells. **B)** A largely endways cut of a developing spore. **C)** The field of view of (D) showing a similar composition to that seen for (A). **D)** A largely transversal cut of a developing spore. The fields of view are x1900 magnification whilst the highly magnified images are x11000. White labels highlight the spore features, whilst teal highlights mother cell features. White ovals highlight examples of unusual cell envelopes in dividing cells. CR: core; CX: cortex; AF: additional feature (used to highlight a variety of prominent but unknown structures); CW/IM: cell wall/inner membrane; GR: granulose deposits; CE: mother cell envelope.

septums between presumably dividing cells. This puzzle-like structure was seen consistently, but not exclusively, between dividing cells suggesting it was more than an artifact whilst also not being an essential feature of growth.

After 12 h, the ultrastructures of spores showed increasing maturity and released spores were not observed, matching the morphologies seen in the light microscopy. (Figure 4.12A and D). Highlighted at higher magnification are two spores at difference stages of development (Figure 4.12B, C and E). This suggested that the decision to sporulate, whilst potentially occurring early, can take some cells longer. This is further emphasised in by the presence of an elongated vegetative cell that already appeared to be beginning to accumulate granulose (Figure 4.12F and G). In B, the forespore resembled those seen at 4 h with asymmetrical division and engulfment already having occurred and the development of the cortex visibly under way. In D, middle spore coat assembly was well underway, displaying the same apparently discontinuous process as seen for spores generated on solid media. The core was also increasing significantly in density with the cortex well developed. Additional features, likely later to form part of other spore coat layers could also be seen. Overall, the forespore in D appeared to be at a similar stage of development as seen for the forespore in Figure 4.5C. Highlighted by the oval was the only feature observed that potentially represents the engulfment of a forespore by a mother cell.

At 48 h, the population appeared to be similar to that seen in the light microscopy (Figure 4.13A, C and E). Vegetative cells were still present but a large number of sporulating cells at various steps of development dominate. Interestingly, released spores were easier to observe than would be assumed from the light microscopy (Figure 4.13D and F). The developing spore (Figure 4.13 B) showed a clear core, cortex and a spore coat that was closer to completion than previously observed. The mother presented with some interesting features, notably it was lacking in granulose, suggested its consumption during the spore assembly process. Additionally, it is possible to speculate that the image showed a spore in the early stages of exiting the mother cell. The disappearance of a discernible mother cell envelope to the right of the spore potentially indicates the beginning of this exit. Naturally, this could also have been due to the section cut and angle and the lack of visible spore coat in this region suggests this was possibility.

The released spores imaged appeared different to those seen from solid media (Figure 4.13D and F). Particularly, the discontinuous middle spore coat and lack of well-defined outer spore coat layers indicated that these spores were not yet fully mature. The difference is stark when compared to day 1 and day 3 released spores generated on solid media. Whilst these also appeared to undergo

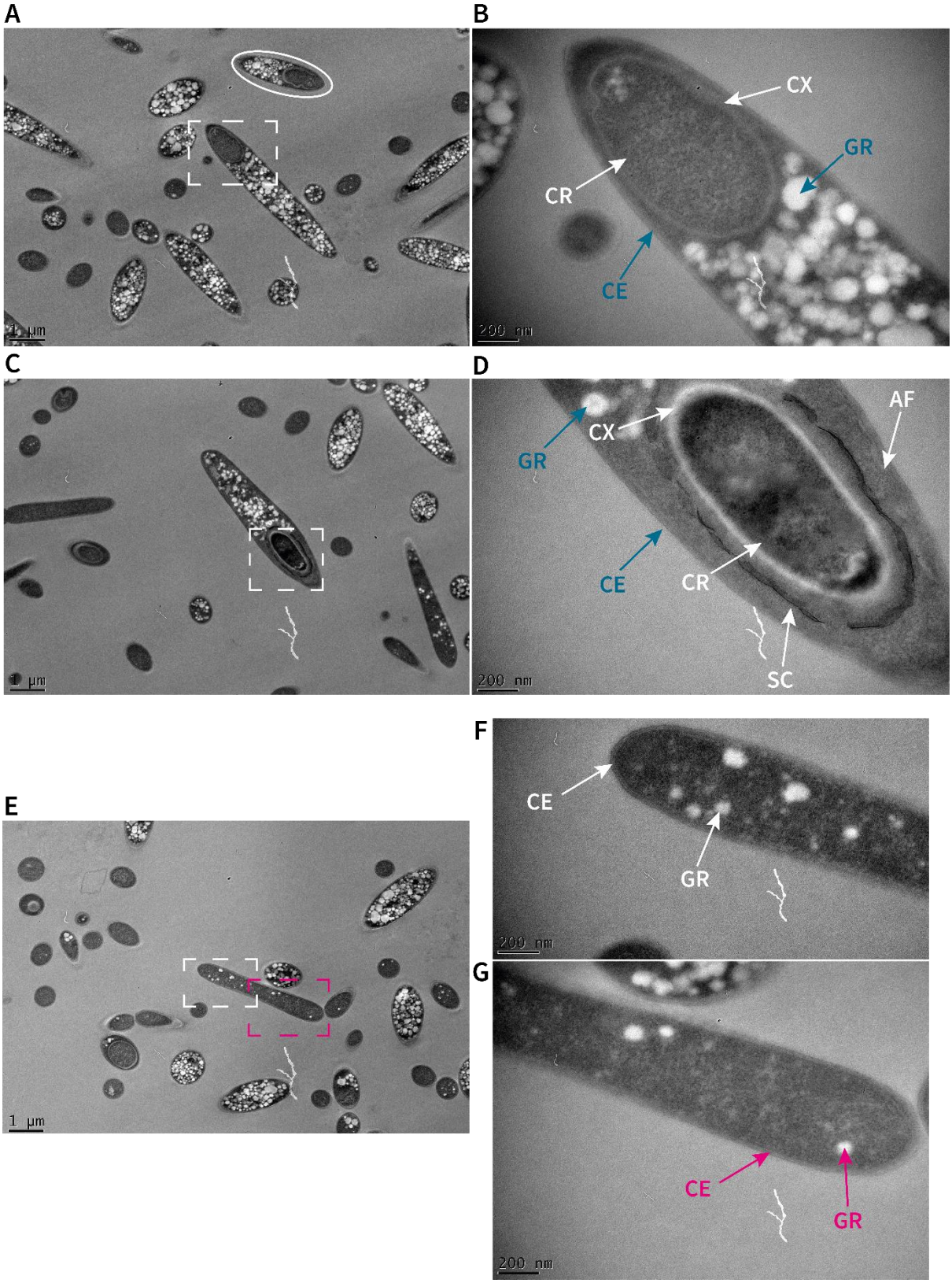

Figure 4.12 Ultrastructures at 12 h of sporulating cells in liquid media. A) The field of view of (B) including vegetative cells, sporulating cells, *Clostridial* forms and potentially the engulfment of a forespore. **B)** A transversal cut of a developing forespore. **C)** The field of view of (D) showing a similar composition as (A), without an engulfing mother cell. **D)** A transversal cut of a more mature

developing forespore. **E)** The field of view of (F) and (G) with similar composition to (A) and (C). **F)** The left pole of a transversal cut of a vegetative cell. **G)** The right pole of a transversal cut of the same vegetative cell as in (F). The fields of view are x1900 magnification whilst the highly magnified images are x11000. In (B) and (D), white labels highlight the spore features, whilst teal highlights mother cell features. In (F) and (G) white labels highlight vegetative cell features in (F) whilst magenta labels highlight vegetative cell features in (G). CR: core; CX: cortex; AF: additional feature (used to highlight a variety of prominent but unknown structures); SC: spore coat; GR: granulose deposits; CE: mother cell envelope.

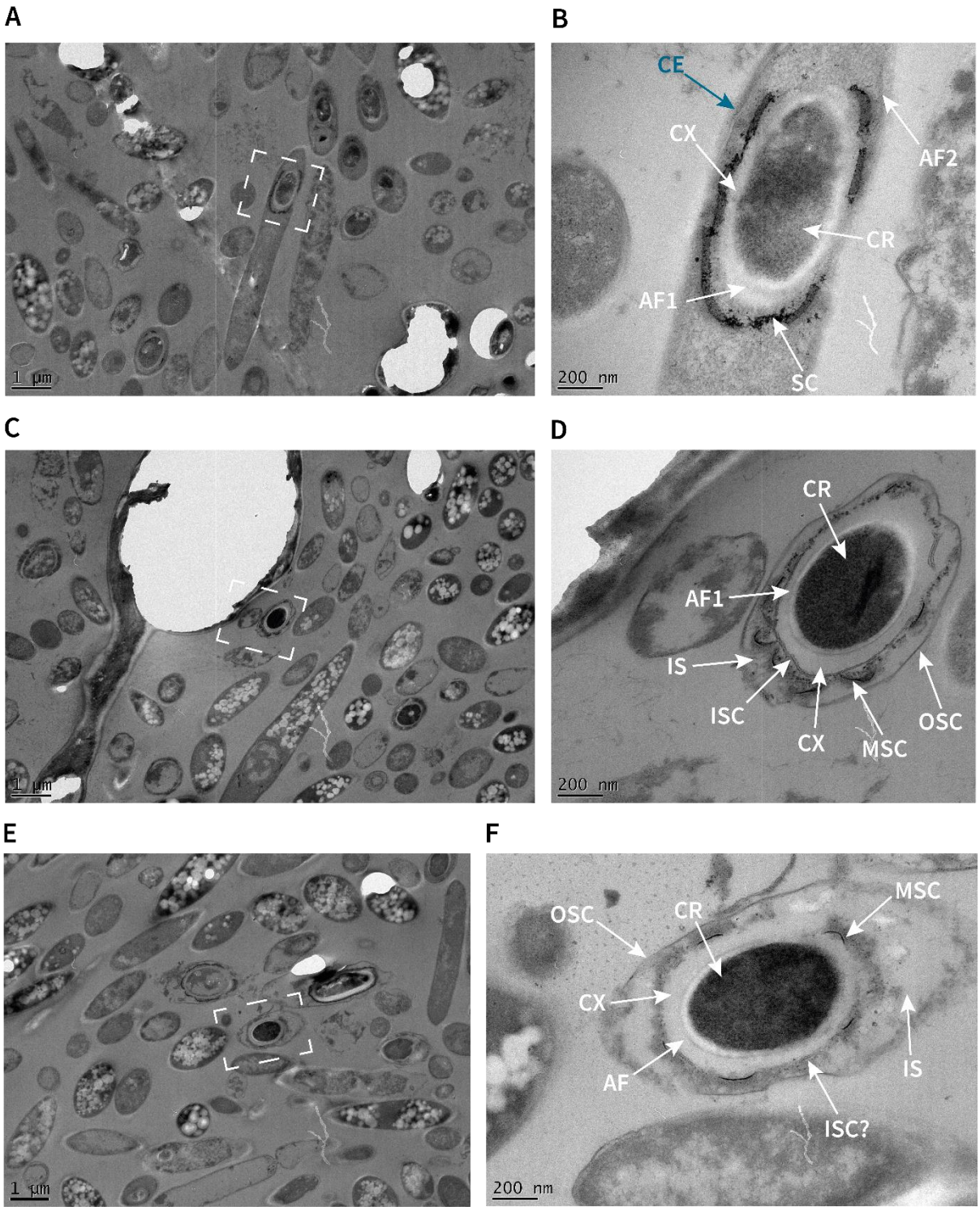

Figure 4.13 Ultrastructures at 48 h of sporulating cells in liquid media. A) The field of view of (B) showing released and unreleased spores, *Clostridial* forms, cell debris and some vegetative cells. **B)** A transversal cut of a developing forespore still contained within the mother cell **C)** The field of view of (D) with a similar composition to (A). **D)** Released spore with an ambiguous cut angle. **E)** The field of view of (F) with a similar composition to (A) and (C). **F)** A transversal cut of a released spore. Large white patches are due to tears in the resin created during sectioning. The

fields of view are x1900 magnification whilst the highly magnified images are x11000. Labels highlight the spore features. CR: core; CX: cortex; ISC: inner spore coat; MSC: middle spore coat; AF: additional feature (used to highlight a variety of prominent but unknown structures); IS: interspace; OSC; outer spore coat.

further spore coat maturations, the middle and outer spore coats were well defined and continuous around the whole spore. This was clearly not the case for either spore seen here. The additional feature 'within' the cortex is again seen here. Together the features seen at this timepoint indicate spores that either take longer or cannot reach full maturity. Either case could explain the lower numbers of heat resistant CFUs obtained from the liquid sporulating cultures. Later timepoints could help clarify which is the case, though the lack of spores visible in the light microscopy at these points suggests that the spores are unable to reach full maturity in liquid media.

4.3 Germination

4.3.1 Spore purification

To investigate the germination process of *C. saccharoperbutylacetonicum*, it was desirable to obtain high concentrations of pure spores. To purify spores from vegetative cells, dead cells and cell debris, I decided to utilise the spore purification protocol (Nerandzic and Donskey, 2013). This process utilises Histodenz to create a density gradient under centrifugation that is capable of isolating *C. difficile* spores from *C. difficile* cells on the basis that spores are denser than cells. The same protocol was used on *C. saccharoperbutylacetonicum* cultures. During the purification process, a lot of putative debris was consistently removed by both the cold washes and density gradient. However, examination of the purified samples under brightfield microscopy showed that the process did not yield a pure spore product, with many non-spore cells still present. This data was not captured at the time due to error on the author's oversight and, unfortunately, there was not sufficient time to repeat this investigation. Nevertheless, it seemed likely that the process was concentrating spores somewhat, allowing for a greater proportion of spores to vegetative cells which could putatively lead to more consistency in germination assays. It seems likely that adjustments to the protocol could yield a greater separation between spore and non-spore allowing for higher concentrations of spores to be purified. All the germination assays below used spores purified from solid spore preparations using this method, though not necessarily from the same batch.

4.3.2 Germination assays

4.3.2.1 Initial method of conducting germination assays indicated L-cysteine as a candidate germinant

4.3.2.1.1 Introduction

An important method for testing a potential germinant is to introduce it to spores in a buffer, such as PBS. If a germinant is present, it will bind to the spore's germination receptors, triggering a cascade that results in the release of dipicolinic acid, the breakdown of the spore cortex and the rehydration of the spore core. The sum of these effects is to rapidly decrease the refractory properties of the spore. This is readily observable by monitoring OD_{600nm} which should show a sharp decrease as the spores germinate. Without any additional nutrients, the resulting cell should be unable to grow and divide meaning there is minimal subsequent increase in OD_{600nm} due to culture growth. However, such a method relies on a high concentration of spores in the sample being tested so that a measurable response to a germinant can be measured. Given that these high concentrations could not be guaranteed, it was decided to attempt these preliminary germination assays both in PBS and in RCM supplemented with the candidate germinants. The latter condition allows for the germination and subsequent growth of a small amount of spores with a later rise, rather than a fall, in OD_{600nm} being indicative of germination. An earlier, or later, rise in OD_{600nm} could indicate the impact the candidate germinant was having on the spores. Germinants were chosen based on the available literature on related *Clostridia* with a particular focus on soil-based species (Hitzman et al., 1957; Mackey and Morris, 1972; Paredes-Sabja et al., 2008; Udompijitkul et al., 2014).

This approach meant that the results obtained from the assays were likely to be more ambiguous. Firstly, there is the composition of the medium itself. RCM is not a defined medium and could contain germinants or germination inhibitors that trigger or prevent germination. Secondly, using growth as a readout results in the possibility that a candidate germinant merely improves growth rate rather than germination rate. Finally, the longer time course and nutrient rich conditions increase the chances of contamination being an explanation for OD_{600nm} increase. This initial method also made a few poor assumptions that were later corrected in 4.3.2.2. Firstly, it was assumed that vegetative cells could not possibly have survived the spore purification process described in 4.3.1 due to a) the aerobic environment in which the washes were carried out and b) the storage in H_2O at 4°C over several days, sometimes weeks. Whilst not an unreasonable assumption, it was untested at the time of these initial experiments. Secondly, the pH of the germinant in either PBS or RCM was not measured and therefore not accounted for, which could have impacted on germination. Thirdly, no heat treatment was applied to the spores used in this section. This was primarily because heat resistance assays had not yet been conducted and a definitive temperature was not confirmed. This also means that it was not possible to investigate whether heat activation could play a role in *C. saccharoperbutylacetonicum* germination.

4.3.2.1.2 Results

As it was not possible to know how quickly visible germination might occur, the same experiment was conducted over multiple time ranges, first with a focus on 0 – 400 min followed by measurements between 1400 – 1800 min the next day (Figure 4.14A and Figure 4.15A) and then again focused on 5 – 35 min followed by 800 – 1400 min the next day (Figure 4.14B and Figure 4.15B). Neither the germination assays in RCM (Figure 4.14) nor in PBS (Figure 4.15) showed any obvious changes in OD_{600nm} either between 5 and 35 min or between 0 and 400 min. Using the lab brightfield microscope, no obvious activity was occurring during these timepoints with the cultures appearing to be of similar composition throughout these timepoints with spores still visible and no motile cells (data not shown).

In RCM, an increase in OD_{600nm} was seen from 760 min in all conditions except that containing dipicolinic acid (DPA) (Figure 4.14). The time taken by cultures containing different germinants to reach given densities varied, implying a difference in either growth rate or germination rate. L-cysteine, L-alanine and RCM with no additional germinant all showed growth earlier than RCM Glucose, with L-cysteine supplemented cultures appearing to do so slightly earlier. Equally, the lack of growth in RCM-DPA suggested the inverse, that DPA inhibited germination, growth or both. However, following this result from DPA, it was clear that I had not taken into account the role of pH in the process and measurements post-assay showed DPA supplemented media to be, unsurprisingly, very low in pH. Subsequently, the pH of all candidate germinants was adjusted prior to use.

In PBS, no significant differences were seen at any time in the experiment (Figure 4.15). This confirmed our hypothesis that the spore purifications did not yield spores in sufficiently high densities to see the drop in OD_{600nm} associated with the germination process. Following the results of these PBS assays, PBS was not used in subsequent germination assays.

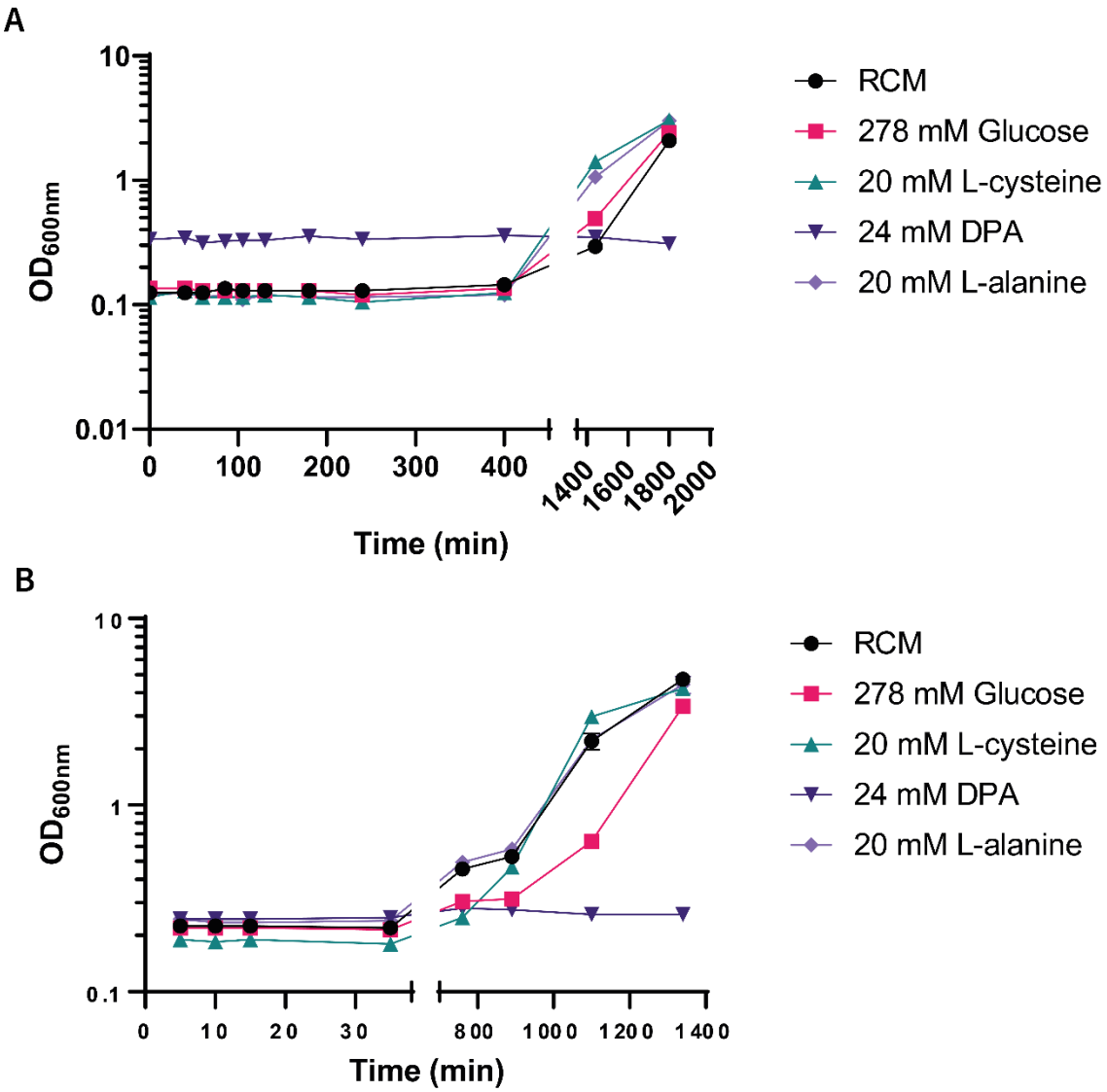

Figure 4.14 Initial germination assays in RCM. A) Assay conducted over the 0-400 min and 1400-1800 min time ranges. No change in OD_{600nm} is seen until after 1400 min. All conditions other than 24 mM DPA showed an eventual increase in OD_{600nm}. A Brown-Forsythe and Welch ANOVA was conducted comparing the conditions to RCM only at 400 min and 1800 min. No significant differences were seen at 400 min (Glucose P=0.6065; L-cysteine P=0.2469; DPA P=0.0607; L-alanine P=0.2469). Significant differences in OD_{600nm} at 1800 min compared to RCM only were seen for all (Glucose P=0.0445; DPA P=0.0272; L-alanine P=0.0183) except L-cysteine (P=0.1351) **B)** Assay conducted over the 0-35 min and 750-1400 min time ranges. Growth is seen for all conditions except 24 mM DPA. The 750-1150 min measurements show the difference in lag phase between the different conditions. A Brown-Forsythe and Welch ANOVA was conducted comparing the conditions to RCM only at 1100 min and 1340 min. No significant differences were seen at 1100 min (Glucose P=0.1196; L-cysteine P=0.1295; DPA P=0.0962; L-alanine P= 0.9974). Significant differences in OD_{600nm} at 1800 min compared to RCM only were seen for glucose (P=0.0146) and DPA (P=0.0211), but not L-cysteine (P=0.0950) and L-alanine (P=0.8422). Both assays conducted in technical duplicate. Where error bars cannot be seen, the error was too small to be plotted.

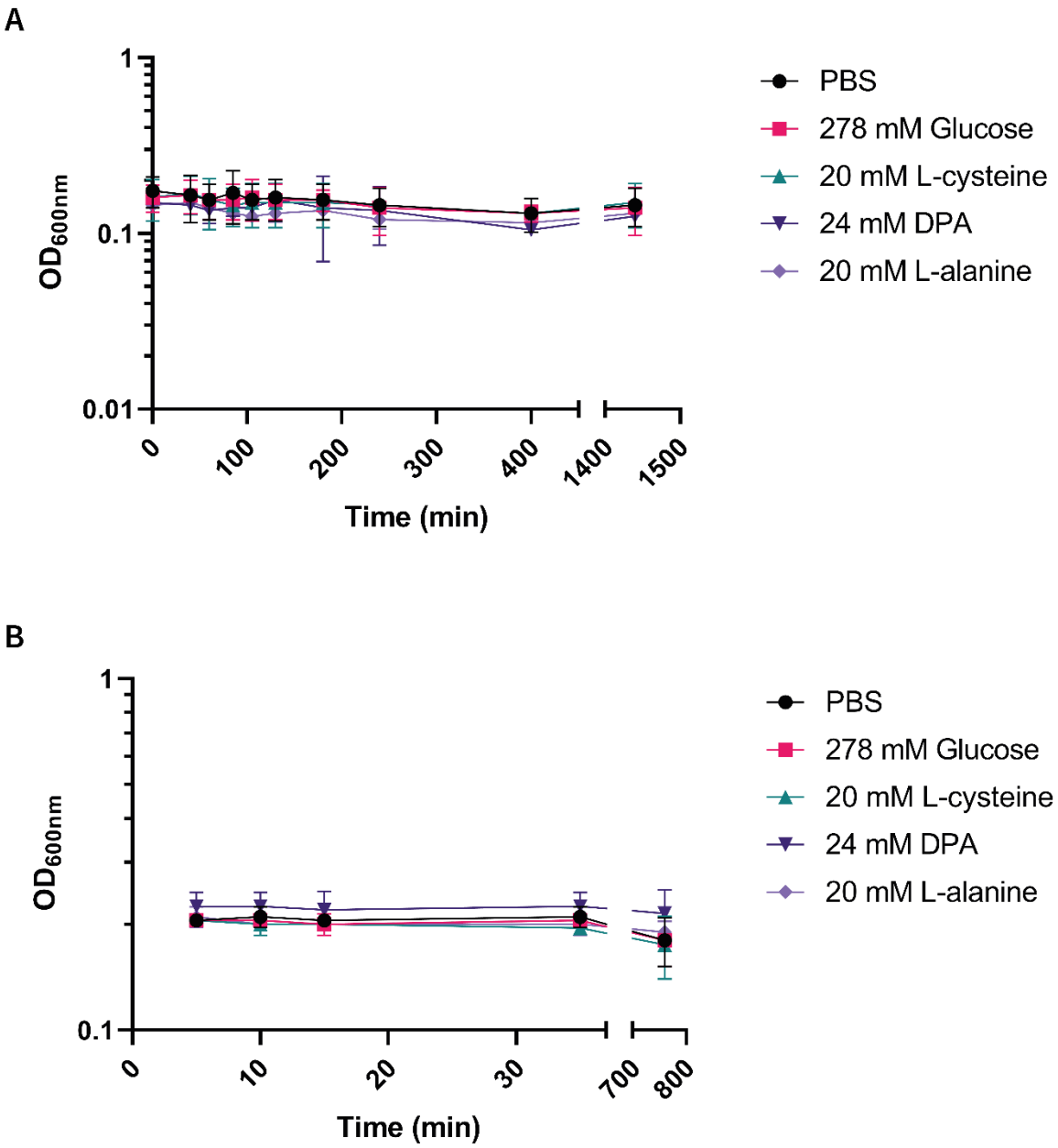

Figure 4.15 Initial germination assays in PBS. A) Assay conducted over the 0-400 min and 1400-1800 min time ranges. No significant OD_{600nm} changes can be identified for of the conditions when compared to the PBS control at the 1440 min by Brown-Forsythe and Welch ANOVA (Glucose P=0.9998; L-cysteine P=0.9998; DPA P=0.8763; L-alanine P=0.8763). **B)** Assay conducted over the 0-35 min and 750-1400 min time ranges. Again, no significant change in OD_{600nm} is seen for any of the conditions when compared to PBS control at 760 min by Brown-Forsythe and Welch ANOVA (Glucose P=>0.9999; L-cysteine P=0.9995; DPA P=0.7453; L-alanine P=0.9720). The experiments were conducted in technical duplicate.

4.3.2.2 Finalised method showed L-cysteine to be the most consistent germinant

Following the results from 4.3.2.1, the method was optimised in a number of ways. It was clear that no OD_{600nm} changes would be visible in the first 6 – 8 h, so the assays were always set up in the evening, OD_{600nm} recorded, and the cultures left until the following morning. This occasionally led to missing the early growth phase, but typically allowed us to capture the first indications of growth. As mentioned, pH was always adjusted to 6.5 prior to the assay. Finally, two other commonly used media – TYIR and CGM – were selected to examine whether there was a difference in germination between media types. L-serine was also added as another candidate germinant for these assays.

Whilst all these media contained at least one complex component (all have yeast extract), so rendering conclusive identification of a specific germinant molecule difficult, they have the advantage of representing the complex context in which a germinant may actually be used, unlike the artificial environment of PBS. Much of this work was conducted with an eye to direct practical application so these results were potentially more relevant than identifying the very best germinant in isolation. It is worth noting that DPA was not used in the assays conducted with TYIR and CGM. This was done due to pH adjusted RCM-DPA assays showing that DPA was not capable of inducing germination. Given that DPA is also poorly soluble, it did not present as a particularly practical candidate germinant.

In RCM, L-cysteine again emerged as the molecule capable of inducing the earliest visible growth (Figure 4.16A). However, all conditions, except DPA, showed visible growth within a similar time period, including RCM without any additions. RCM does however contain ~4 mM of cysteine in addition to peptone which contains a mix of amino acids, likely explaining this ability to induce germination on its own. Glucose was the weakest candidate germinant, displaying the longest lag period of the six conditions. However, statistically, no significant difference was seen between any of the conditions exhibiting growth and DPA at the final timepoint, likely due half the repeats in this experiment exhibiting a much greater lag than the other half.

In TYIR, growth was seen earlier than for RCM, suggesting a component of TYIR either causes faster growth or faster germination (Figure 4.16B). L-cysteine again displayed the highest ODs, implying the earliest initial rise in OD_{600nm}. This was followed by TYIR-glucose (TYIR by itself was not tested as this is never used as a growth condition) and L-alanine showing similar growth profiles. Finally, L-serine contributed to the latest initial growth. The error is high in this graph due to different biological repeats showing different growth latencies. This became a common problem in TYIR and CGM (though interestingly not in RCM), however the data were retained because the patterns of

earliest to latest initial growth were retained in all conditions for TYIR. ANOVA analysis at the final timepoint confirmed these inconsistencies showing no significant difference between any of the conditions.

In CGM, sporulation was incredibly inconsistent across all assays (Figure 4.16C), including the later inhibition assays (4.3.2.3). This was confirmed by ANOVA which showed no significant differences between any of the conditions at the final timepoint. The figure shows the combination of three biological repeats, each with a different outcome: one in which germination occurred followed by normal growth, one in which germination and initial growth occurred before all cultures ceased growing at OD_{600nm} ~0.6 and one in which no germination occurred in any sample. Where germination occurred, the pattern seen for TYIR appeared to hold with L-cysteine typically showing the highest initial growth. For one repeat, L-cysteine was the only condition to show any growth, for the other, initial growth was fairly similar across all conditions. Whilst presenting the data as in Figure 4.16C is not the most visually helpful, it did not seem appropriate to remove any one of the three results given the overall inconsistency. It is possible that there is a component of CGM, likely present in low concentrations, that is inhibitory to germination, but is not consistently distributed between all batches of CGM.

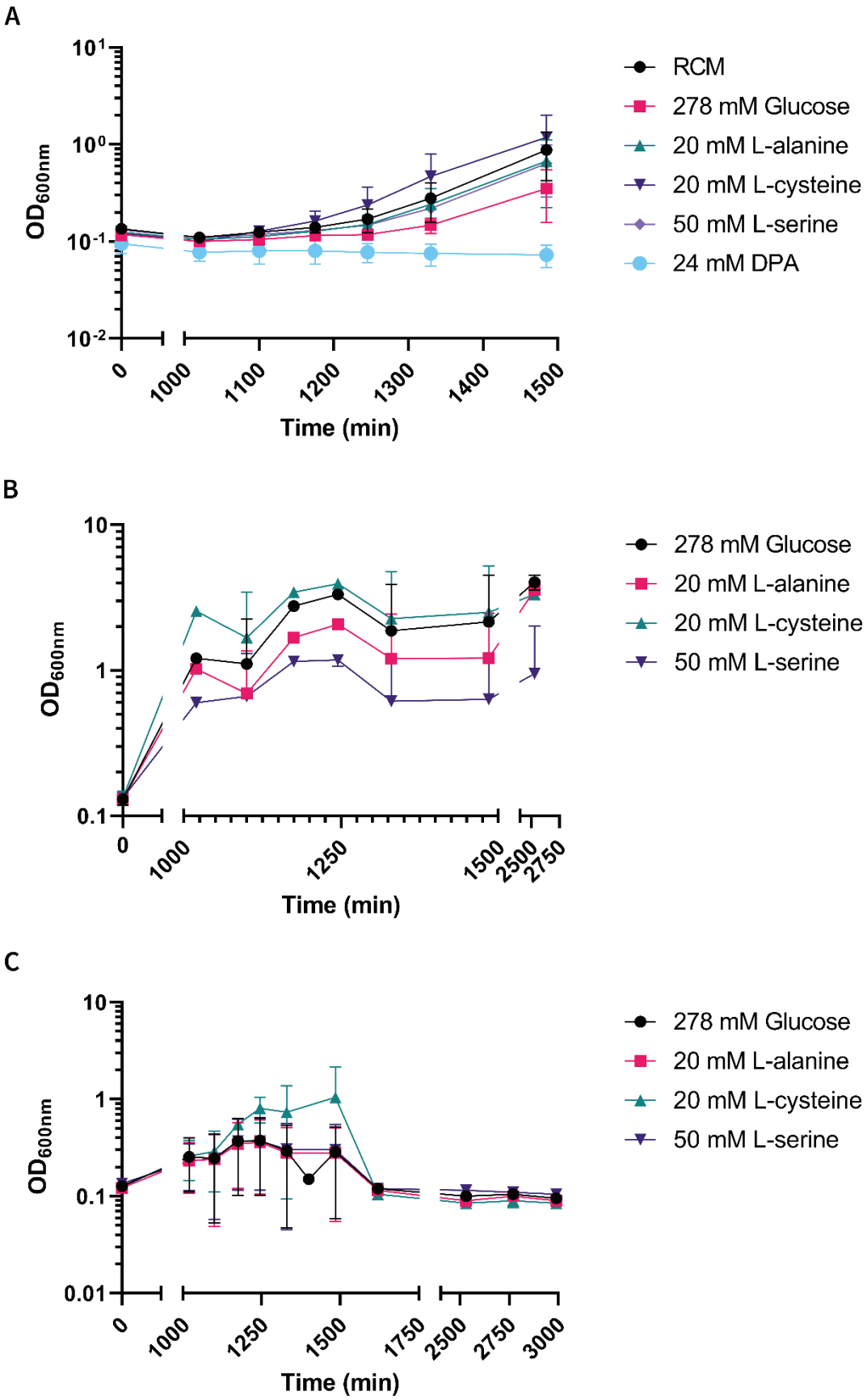

Figure 4.16 Finalised germination assays conducted in different base media. A) Assay

conducted using RCM as the base media for all conditions. Growth only becomes observable after 1100 min. After this point, all conditions apart from 24 mM DPA show growth with various degrees of lag phase. No significant OD_{600nm} differences can be identified for of the conditions when compared to DPA at the 1485 min by Brown-Forsythe and Welch ANOVA (RCM only P= 0.1216; Glucose P=0.2011; L-cysteine P= 0.2268; L-serine P=0.1529; L-alanine P= 0.2300). **B)** Assay conducted in TYIR. All conditions contain 278 mM glucose. Growth can be observed after 1100 min for all conditions. No significant OD_{600nm} differences can be identified for of the conditions when compared to Glucose at the 2530 min by Brown-Forsythe and Welch ANOVA (L-cysteine P=0.4786; L-serine P=0.2727; L-alanine P= 0.6523). **C)** Assay conducted in CGM. All conditions contain 278 mM glucose. Growth was inconsistent with only one of the three repeats showing growth except in for 20 mM L-cysteine which showed growth in two of the repeats. No significant OD_{600nm} differences can be identified for of the conditions when compared to Glucose at the 2990 min by Brown-Forsythe and Welch ANOVA (L-cysteine P=0.5411; L-serine P=0. 5411; L-alanine P= 0. 5411). TYIR and RCM experiments were conducted in biological duplicate and technical duplicate. The CGM experiment was conducted in biological triplicate and technical duplicate. Where error bars cannot be seen, the error was too small to be plotted.

4.3.2.3 Screening for germination inhibitors yielded two candidates

Controlling the timing of germination is useful to both the study of spores and their potential application. No inhibitors of germination had previously been described for *C. saccharoperbutylacetonicum* so I decided to investigate compounds that might be capable of inhibiting germination, even in the presence of a putative germinant. For the latter, L-cysteine was used given its most consistent performance and association with the earliest initial growth in most assays. Candidate inhibitors were chosen based on literature searches, availability at short notice and reasonable cost. Cost is important when considering such compounds and their potential use at a large scale. Two candidate inhibitors, satisfying these criteria, were chosen: potassium sorbate and D-alanine (Brunt et al., 2014). These were tested with and without L-cysteine and all assays included the media without any germinant and with L-cysteine as positive controls. Otherwise, they were conducted in the same manner as previously described.

For RCM, only one candidate completely blocked all growth (Figure 4.17A). Potassium sorbate was able to significantly block germination induced both by RCM alone and in the presence of L-cysteine ($P=<0.0001$ for both potassium sorbate conditions). The other conditions all showed growth with L-cysteine with D-alanine and L-cysteine showing the highest rates. D-alanine alone showed almost identical rates to that seen for RCM only.

The inhibitors behaved differently in TYIR where D-alanine displayed inhibitory capabilities (Figure 4.17B). Potassium sorbate did significantly delay growth in comparison to L-cysteine and TYIR-glucose positive controls. However, growth eventually occurred after 1300 min. D-alanine showed inhibitory effects both with and without L-cysteine throughout the majority of the experiment ($P=0.0012$ and $P=0.0014$ in D-alanine alone and D-alanine + L-cysteine respectively). The condition including L-cysteine did eventually show some growth after 2000 min, though the rise was slight and very late on. The results for TYIR raise the possibility of the different candidates acting in concert with other potentially inhibitory compounds present in TYIR.

CGM essentially showed no growth across all conditions until the 2000+ min timepoint (Figure 4.17C). This heavily implies that the repeats showing no germination in 4.3.2.2 were likely the true result with the other data being affected by some differences in media or potentially contamination with vegetative cells. This result also suggests that a component of CGM not present in the other two media is also inhibitory to germination. There are four compounds likely unique to CGM – K_2HPO_4, KH_2PO_4, $MgSO_4$ and $MnSO_4$ and any one of these could be providing an inhibitory effect and each could warrant testing as candidate inhibitors.

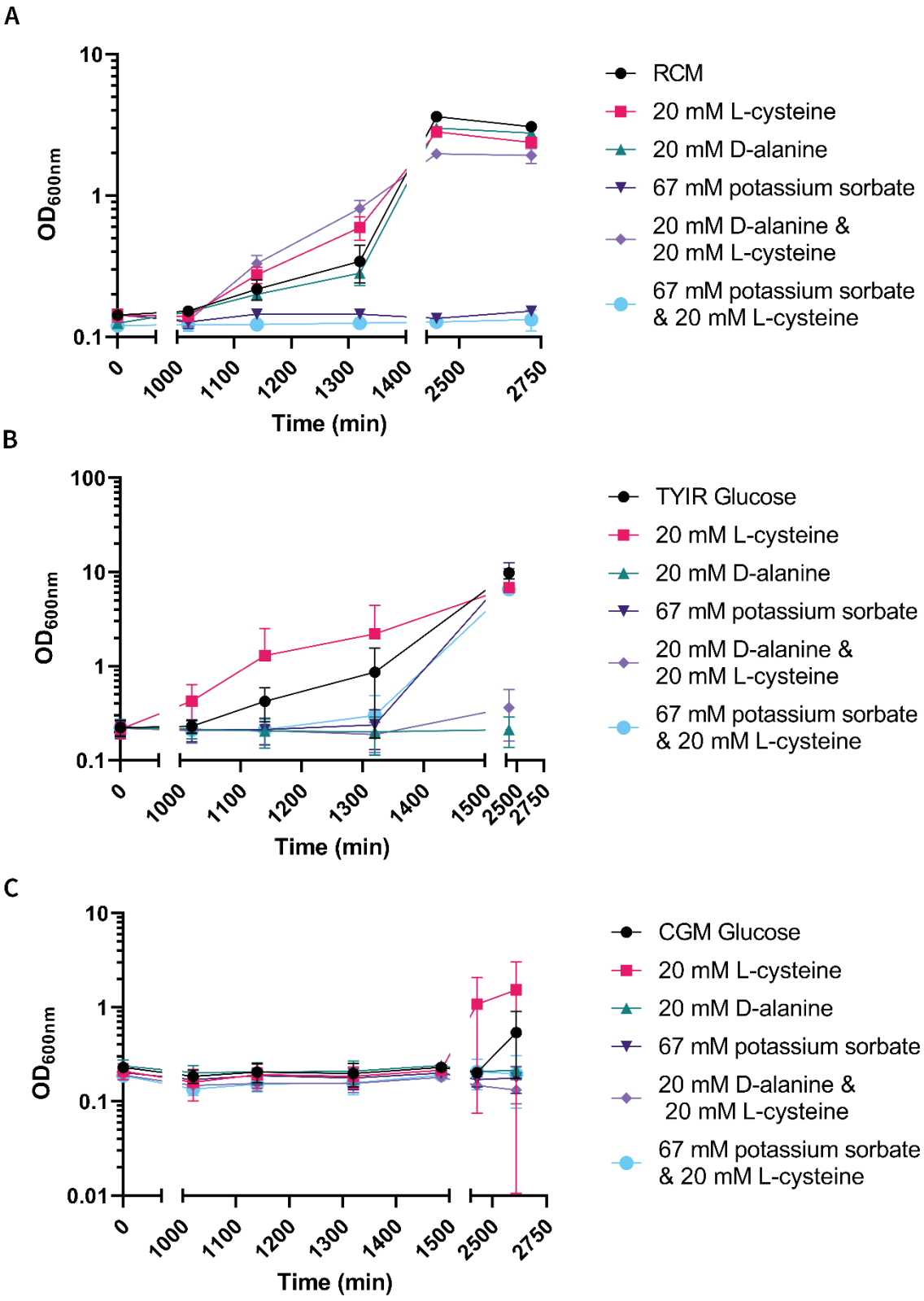

Figure 4.17 Germination inhibition assays conducted in three different base media. A) Assay conducted in RCM. No growth is seen for any condition until after 1140 min. Both conditions containing 67 mM potassium sorbate exhibit complete inhibition of growth whilst all other conditions demonstrate growth. A Brown-Forsythe and Welch ANOVA was conducted for the final

timepoint comparing RCM L-cysteine, the positive control, with all other conditions. It showed a significant difference between L-cysteine and potassium sorbate (P=<0.0001), L-cysteine and potassium sorbate + L-cysteine (P=<0.0001), L-cysteine and D-alanine (P=0.0267), and L-cysteine and RCM only (P=0.0034). No significant difference was observed between L-cysteine and D-alanine + L-cysteine (P=0.1080). **B)** Assay conducted in TYIR. All conditions contain 278 mM glucose. Growth is eventually observed in all conditions except those containing 20 mM D-alanine. Growth is significantly delayed for conditions containing 67 mM potassium sorbate. A Brown-Forsythe and Welch ANOVA was conducted for the final timepoint comparing TYIR L-cysteine, the positive control, with all other conditions. It showed a significant difference between L-cysteine and glucose (P= 0.0449), L-cysteine and D-alanine + L-cysteine (P= 0.0014), and L-cysteine and D-alanine (P= 0.0012). No significant difference was observed between L-cysteine and potassium sorbate (P= 0.5031), nor between L-cysteine and potassium sorbate + L-cysteine (P=0.9266) **C)** Assay conducted in CGM. All conditions contain 278 mM glucose. The only growth seen is in one of the two 20 mM L-cysteine repeats. A Brown-Forsythe and Welch ANOVA was conducted for the final timepoint comparing CGM L-cysteine, the putative positive control, with all other conditions. No significant difference was seen between L-cysteine and any of the other conditions (glucose only P=0.7050; D-alanine P=0.4949; potassium sorbate P=0.4759; D-alanine + L-cysteine P=0.4532; potassium sorbate + L-cysteine P=0.4862). All experiments were conducted in biological duplicate and technical duplicate. Where error bars cannot be seen, the error was too small to be plotted.

4.3.3 Growth assays confirmed the discovery of germinants and a germination inhibitor

The germination assays cannot adequately differentiate between earlier growth due to germination and that which is due to increased growth rate during logarithmic phase. Therefore, the need to grow vegetative cells in the germination conditions tested was clear. By doing so, it became possible to assign the differences seen in the germination assays to the effects of increased and/or earlier germination, rather than to an altered growth rate. Given the large range of conditions that required testing, our typical growth conditions of 5 mL broth and measurement by spectrophotometer in cuvettes was swapped for growth in 200 µL of media in 96 well plate and measurement by plate reader. The OD_{600nm} output from this was not comparable to previous growth curves, however, the growth rate during logarithmic phase should have remained relatively consistent. This method measured OD_{600nm} every 3 min resulting in hundreds of measurements over time. Therefore, to ease readability of the data, a continuous line has been plotted with the error also represented as continuous. The plate reader had a tendency to produce negative OD_{600nm} value when measuring during an extended lag phase. Where this has occurred, the lag phase has been excluded from the graphs presented to render the data more readable, however the lag phase still forms an important part of the analysis. Negative controls (media with no cells) were performed for all conditions and no growth was recorded.

Growth in RCM was relatively consistent across all conditions and the two biological repeats (Figure 4.18). Nonlinear regression analysis was conducted for each condition over a 200 min period from the start of measurable growth. Using this method, L-serine showed the shortest doubling time at 45 min, followed by RCM only at 46 min and then D-alanine (47 min), L-alanine (48 min), D-alanine and L-cysteine (48 min 30 s), potassium sorbate and L-cysteine (50 min), potassium sorbate (51 min), L-cysteine (51 min) and glucose (56 min). This result conclusively showed that neither the early growth phenotype of L-cysteine nor the inhibitory phenotype of potassium sorbate were due to their impact on growth rate. This was confirmed by a Comparison of Fits analysis showing the different growth rates to be significant ($P=<0.0001$). It is therefore likely that L-cysteine is a germinant for *C. saccharoperbutylacetonicum* whilst potassium sorbate is a germination inhibitor in this context.

For the germinant candidates in TYIR, growth rate appeared fairly consistent (Figure 4.19A). However, L-cysteine did demonstrate the fastest doubling time at 50 min. L-alanine was the next fastest at 52 min, followed by L-serine (55 min) and finally TYIR-glucose (61 min 30 s).

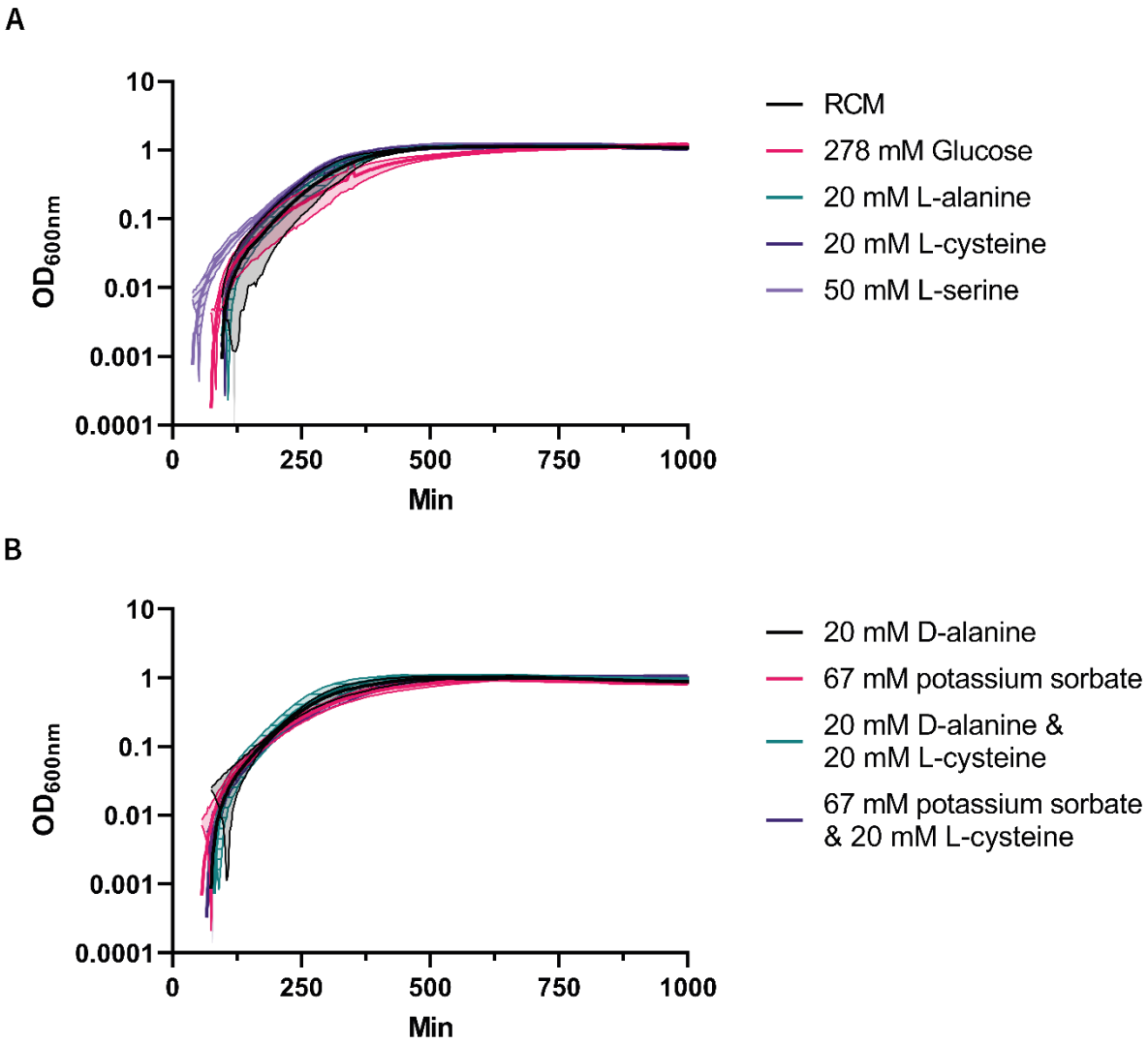

Figure 4.18 Growth of *C. saccharoperbutylacetonicum* in the conditions used for the germination assays using RCM. A) Growth in the conditions containing candidate germinants. **B)** Growth in the conditions containing candidate germination inhibitors. For both growth plots, the lag phase typically yielded negative OD_{600nm} values and so has been removed from all conditions for clarity. Individual data points are not plotted, instead the connecting lines are shown and the error displayed as a continuous window. The experiments were conducted in biological duplicate and technical duplicate

This left the status of L-cysteine as a germinant in this condition as somewhat ambiguous. The result of TYIR-glucose suggested there is an active germinant within the media as the slow doubling time in this media is not matched by the early growth seen in the germination assays. L-serine and L-alanine appear to be useful to growth but have no especial role in germination.

For the candidate germination inhibitors, the picture was more complex (Figure 4.19B). Potassium sorbate with L-cysteine showed the fastest doubling time (52 min), followed by D-alanine (57 min), potassium sorbate (63 min) and finally D-alanine with L-cysteine (88 min). These were deemed significant by Comparison of Fits analysis (P=<0.0001). However, both the D-alanine conditions showed a significant lag of 500 min before any signs of growth. This suggests some inhibitory growth effect is due to the presence of the compound which may also account for its inhibitory effect in the germination assays. Potassium sorbate again performed relatively well when compared to the TYIR-glucose control condition (Figure 4.19A). Although potassium sorbate did not perform as well as D-alanine in the germination assays, if the effect was entirely due to germination inhibition, it could still constitute a genuine germination inhibitor in this context.

As with the germination assays, CGM showed the greatest variability in results (Figure 4.20). In terms of doubling time, D-alanine allowed for the quickest logarithmic growth (49 min 30 s), followed by L-alanine (53 min 30 s), L-serine (56 min), L-cysteine (62 min), CGM glucose (66 min), potassium sorbate with L-cysteine (72 min), potassium sorbate (81 min) and finally, D-alanine with L-cysteine (123 min). Both D-alanine conditions again presented with long lag phases (740 min for D-alanine alone; 190 min for D-alanine with L-cysteine). In addition, significant error is seen for nearly all conditions owing to lag in one repeat but not the other. However, these differences in growth rates were deemed significant by Comparison of Fits analysis (P=<0.0001). It is difficult to conclude much from this data, but it is noteworthy that L-cysteine did not enhance the growth rate significantly. Potassium sorbate did slow the growth compared to the positive control, suggesting it could have an impact on growth rate under certain conditions. However, the margin of error for both these conclusions is high.

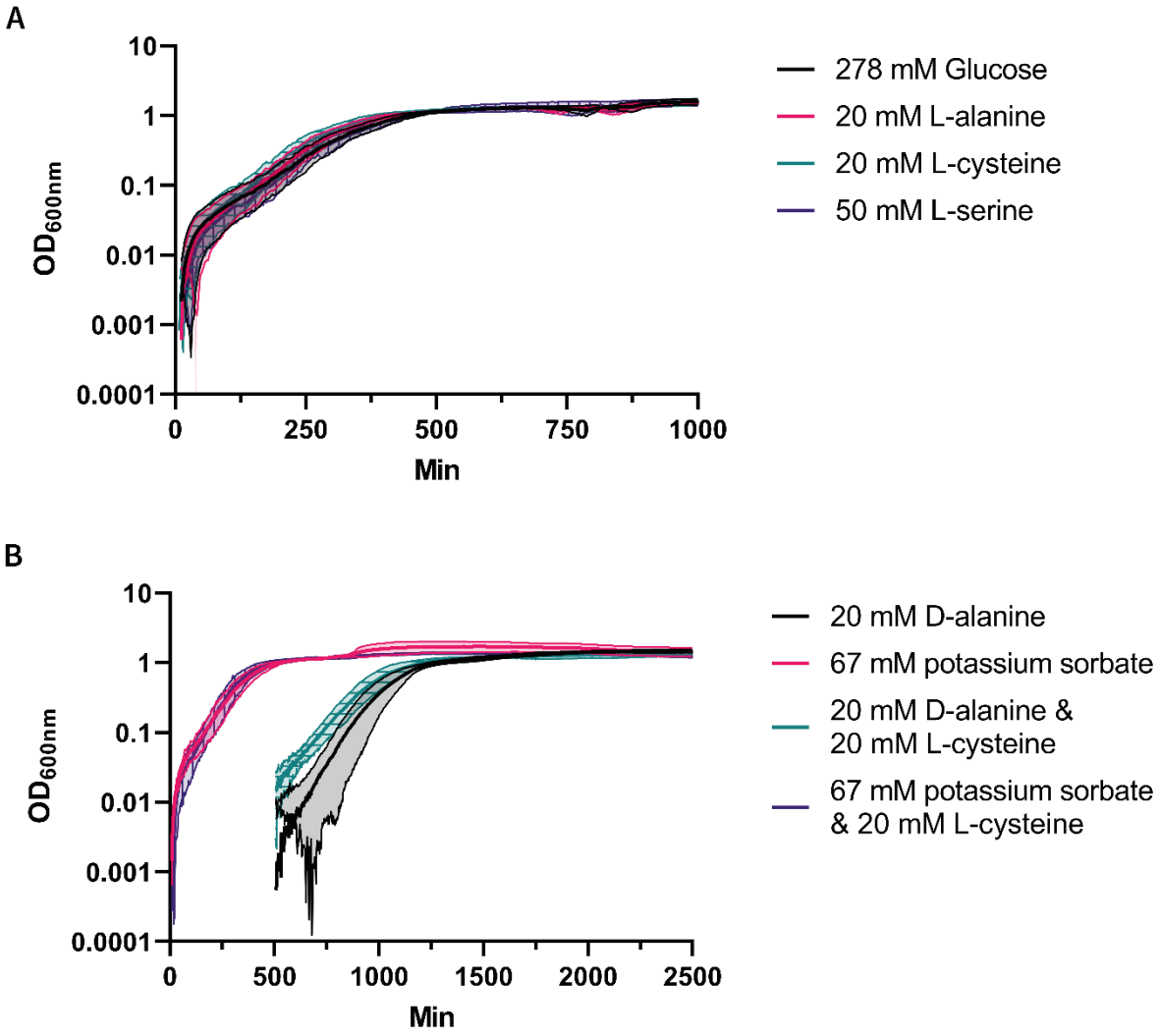

Figure 4.19 Growth of *C. saccharoperbutylacetonicum* in the conditions used for the germination assays using TYIR-glucose. A) Growth in the conditions containing candidate germinants. All conditions contain 278 mM glucose **B)** Growth in the conditions containing candidate germination inhibitors. Again, all conditions contain 278 mM glucose A significant lag phase was seen for the two conditions containing 20 mM D-alanine. The 20 mM D-alanine only condition showed high error in the early growth phase due to one repeat retaining a negative lag phase for longer than the other three. This was unusual for the plate reader experiments. For both growth plots, the lag phase typically yielded negative OD_{600nm} values and so has been removed from all conditions for clarity. Individual data points are not plotted, instead the connecting lines are shown, and the error displayed as a continuous window. The experiments were conducted in biological duplicate and technical duplicate. Sudden increases in OD_{600nm} error can be accounted for by the introduction of repeats that were previously read negatively by the plate reader

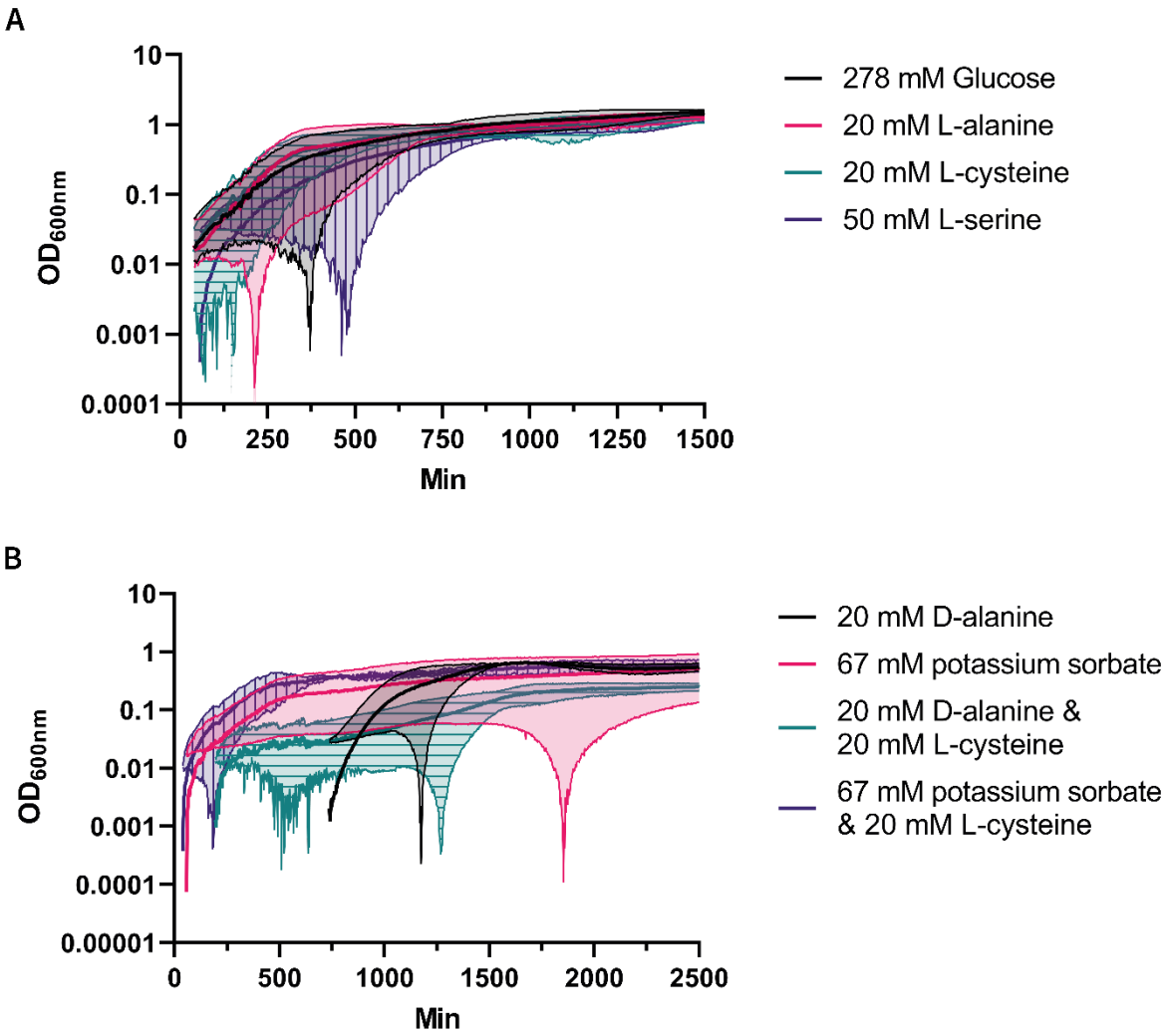

Figure 4.20 Growth of *C. saccharoperbutylacetonicum* in the conditions used for the germination assays using CGM-glucose. A) Growth in the conditions containing candidate germinants. All conditions contain 278 mM glucose **B)** Growth in the conditions containing candidate germination inhibitors. Again, all conditions contain 278 mM glucose. Error was high in all conditions in both A) and B) due to inconsistent growth between biological repeats. Sudden increases in OD_{600nm} error can be accounted for by the introduction of repeats that were previously read negatively by the plate reader. For both growth plots, the lag phase typically yielded negative OD_{600nm} values and so has been removed from all conditions for clarity. Individual data points are not plotted, instead the connecting lines are shown, and the error displayed as a continuous window. The experiments were conducted in biological duplicate and technical duplicate

4.4 Discussion

4.4.1 Sporulation

There were two key drivers for the investigations into sporulation in *C. saccharoperbutylacetonicum*: the general need to understand more about sporulation progression and the specific interest in probing the previously observed differences between the two methods of generating spores. For the former, this chapter is able to shed light on heat resistance, morphology and ultrastructure of *C. saccharoperbutylacetonicum* spores. For the latter, this chapter attempts to understand if and how there may be differences in spore phenotype and suggest why this might be the case.

Interestingly, in both conditions, the produced spores were only able to survive exposure to 60°C heat and not 70°C heat as previously reported (Sasha Atmadjaja, personal communication). There was a difference in heating time (15 min vs 10 min) that might account for this difference. Though, if this were the case, it would present an interesting finding given that most other investigated sporulating species are able to resist heat for longer (Jamroskovic et al., 2016). Alternately, the difference could be due to a difference in handling although this seems unlikely. Either way, the difference suggests the spores are not as resistant to either handling or heat as other species.

Differences were also observed between the two methods of generating spores. Over the equivalent time periods, spores prepared in liquid broth were less resistant to heat than those prepared on solid media. However, it is unclear from this work alone if this was due to a fundamental difference in heat resistance or whether spores mature more slowly in the liquid broth environment. When the general morphology displayed under phase contrast and fluorescent microscopy is taken into account, there does appear to be significant differences in between the two populations, even into late stages of growth. Cells grown in liquid broth appeared to form a more uniform population with most cells appearing as large, *Clostridial* form cells within a few hours and remaining that way until the end of the measured period. This contrasts with the morphology of spores generated on solid media which appeared to form two populations – vegetative cells and *Clostridial* form cells. Interestingly, whilst the majority of solid-generated *Clostridial* forms contained developing forespores (as visible with membrane stain), this is not the case for liquid broth-generated spores. Ultimately, it does appear that fewer spores are produced in liquid broth judging by their scarcity in later timepoints. However, the day 3 and day 4 timepoints are missing for liquid broth-generated spores which is, unfortunately, when most released spores were visible for solid-generated spores.

There were released spores visible in both the light and electron microscopy. There did appear to be differences in the quantity and both the overall morphology and ultrastructure of these spores.

This must be caveated with the acknowledgement that any estimations of quantity from microscopy may not be representative, and that thin section TEM may simply be displaying identical spores viewed from different cutting angles or may have caused damage to the spore during sectioning. However, under light microscopy, the liquid broth-generated released spores appeared to show membrane staining that matched their overall size as seen by phase contrast. This was not the case for the solid-generated spores which clearly showed membrane staining that did not encompass the entire structure. This potentially suggested that the membrane-bound elements (i.e., the spore core), were larger and less well-contained in liquid broth-generated spores.

When the ultrastructures were compared, differences were again apparent. These differences did not obviously extend to the size of the core, with different sizes observed primarily due to the cutting angle and not the real size. The clear differences were seen in the spore coat and other surrounding layers. For the spores observed in this chapter, the ultrastructures beyond the cortex consistently appeared more fragmented and thinner for liquid broth-generated spores than for solid-generated spores. In addition, the liquid broth-generated spore samples were the only of the two to contain an example of an unreleased mature-like spore at later timepoints. Where spores were unreleased in solid-generated samples, these were always still maturing. It therefore seems likely that spores have difficulty in maturing in liquid broth. This could potentially explain the difference between the numbers of sporulating visible at early timepoints and numbers of spores at later timepoints in the liquid broth condition. Whilst a difference was also seen for solid-generated spores, it appeared to be much less drastic.

C. saccharoperbutylacetonicum is a naturally found in the soil, a relatively solid environment that is only occasionally saturated with liquid. It is likely that one of the scenarios in which cells sporulate is during extend dry periods in the environment. In contrast, wetter periods likely represent opportunities to grow quickly, though full saturation may also not be ideal. It is plausible, therefore, that the species struggles to sporulate effectively in highly liquid environments having never had to adapt to doing so. Based on the data presented here, this results in spores struggling to reach maturity due to poor formation of important spore structures. This, in turn, could explain the lower overall heat resistance observed with the weaker structures being less able to withstand the heat. Though this could also be explained by lower spore numbers, itself a potential side effect of poor maturation ability. An additional factor could be the hydration of the spore core. The replacement of water with calcium-DPA in the spore core is a crucial factor in the heat-resistance of spores, preventing denaturation of key proteins contained within. None of the data presented here is able to give any indication of spore hydration, however, if the spore is not suited to sporulation in wet

conditions, it's plausible that the spore also struggles to fully eliminate water from its core resulting in greater susceptibility to heat. Finally, our data cannot rule out the possibility that spores germinate in later time periods, reducing their overall number, although this seems unlikely.

4.4.2 Germination

In slight contrast to the sporulation investigations described above, the goal of the germination research was to find suitable germinants and germination inhibitors for practical use. Candidates were selected based on the literature and their suitability for regular use (i.e., price and ease of handling). To this end, it was somewhat fortunate to have found plausible candidates in both categories given the limited range that were tested. Based on this work, it appears highly likely that L-cysteine plays a role in prompting the germination of *C. saccharoperbutylacetonicum* spores whilst potassium sorbate is able to inhibit the germination process. Pleasingly, neither candidate appeared to have a substantial effect on growth. The success of these molecules is most pronounced in RCM where the effect of both is clear. They were also successful in TYIR-glucose, though growth was eventually seen in the potassium sorbate condition. CGM-glucose proved an extremely inconsistent media with which to investigate germination.

The differences seen in the different media do raise the question of the effect of media composition. Table 4.1 shows the components of each media. All the three contain yeast extract, albeit at varying concentrations. The composition of yeast extract can be variable but one estimate is shown Table 4.2 and suggests a mix of amino acids, including some of our candidates, albeit in lower overall concentrations (Tomé, 2021). The fact that RCM and TYIR use the highest and lowest amounts of yeast extract are both capable of germinating without any additive suggests that the concentration of the extract is not a key determining factor in germinating capacity. Peptone and tryptone are similar, being digestions of animal products and also containing a range of amino acids at different concentrations. Neither is present in CGM suggesting they could contain germinating capacity. However, given they are absent completely from CGM, their concentration seems insufficient to explain the inconsistency displayed by CGM.

Table 4.1 Comparison of the components in RCM, TYIR and CGM

Ingredient	g/L in RCM	g/L in TYIR	g/L in CGM
Yeast extract	13	2.5	5
Peptone	10	-	-
Tryptone	-	2.5	-
$(NH_4)_2SO_4$	0.5	-	2
$FeSO_4.7H_2O$	-	0.025	0.01
Glucose	5	50	50
Soluble starch	1	-	-
NaCl	5	-	1
Cysteine hydrochloride	0.5	-	-
Agar	0.5	-	-
K_2HPO_4	-	-	0.75
KH_2PO_4	-	-	0.75
$MgSO_4$	-	-	0.4
$MnSO_4$	-	-	0.01
Asparagine	-	-	2
Sodium acetate	3	-	-

Likely, one of the compounds that is present in low concentrations in CGM, but absent from the other two media, is influencing the process. The most likely candidate based on concentration is $MnSO_4$ which is present in CGM at a concentration of 66 µM. Whilst not extensively investigated, manganese has been suggested to exert an inhibitory effect in *Geobacillus stearothermophilus* (Cheung et al., 1982) and in *Bacillus subtilis* (Nagler and Moeller, 2015), though it was shown to have a pro-germination effect in *Clostridium butyricum* (Bester and Claassens, 1970). It's possible that, with the addition of extra solutions and extensive aliquoting of the CGM stocks utilised in this study, $MnSO_4$ was not evenly distributed across all aliquots. This could account for the variability seen in the CGM based experiments. RCM contains additional L-cysteine, beyond that present in peptone and yeast extract, which perhaps accounts for its consistent performance across all conditions. Interestingly, the addition of high concentrations of glucose in RCM seemed to inhibit growth and possibly germination somewhat.

Given that germination inhibition is typically associated with either unfavourable pH conditions (Bhattacharjee et al., 2016) or assumed to be due to competitive inhibition of germination receptors, it was an oversight not to include D-cysteine in the experiments. However, potassium sorbate is significantly cheaper (£91.20 vs £433 per kg at time of writing (SLS)) and seems to work well. It has been proposed that potassium sorbate exerts its inhibitory effects by competitive inhibition of the germination receptors (Smoot and Pierson, 1981), though this was later contested (Blocher and Busta, 1985). A more recent study suggested that the mode of action of potassium sorbate does not involve competitive inhibition and is more complicated (van Melis et al., 2011). They theorised that dissociated sorbic acid is able to penetrate either the spore core or the inner membrane to interfere with the downstream signalling following the binding of germinants. This

Table 4.2 Estimated amino acid composition of yeast extract

Amino acid	Free amino acid g/100 g yeast extract	Total amino acid g/100 g yeast extract
Alanine	3.7	4.4
Arginine	1.8	2.5
Aspartic acid	1.9	4.9
Cysteine	0.3	0.4
Glutamic acid	5.2	8.1
Glycine	1.2	2.4
Histidine	0.9	1.0
Isoleucine	2.2	2.7
Leucine	3.5	3.8
Lysine	2.2	4.0
Methionine	0.7	0.9
Phenylalanine	1.8	2.3
Proline	1.0	2.0
Serine	1.8	2.3
Threonine	1.7	2.1
Tyrosine	1.2	1.4
Tryptophan	0.6	0.6
Valine	2.6	2.9

suggest that potassium sorbate might be an effective inhibitor against all germinants, except in pH conditions that cause a change in sorbic acid dissociation.

This chapter provides the groundwork for investigating germination in *C. saccharoperbutylacetonicum*. The results of this should provide practical uses in the future but there remain many unanswered questions. It would be informative to both deconvolute the effects seen here and re-convolute them again in more controlled conditions. In the first instance, obtaining pure spore preparations by improving spore purification methods would enable the use of non-growth-based assays. This would mean that individual candidate germinants could be tested without the need to also test their effect on growth. Subsequently, it would be useful to test combinations of germinants to see if faster germination is possible, i.e., re-convolute the picture. The ultimate expression of this approach would be to start to use defined media to understand the role of complex mixes of compounds.

Chapter V – Quorum sensing and the accessory gene regulator signalling in *C. saccharoperbutylacetonicum*

5.1 Introduction

Quorum sensing, the process by which bacteria sense their local population density, is important in a variety of bacteria. In fermentation, it is common to need high density, stationary phase cells to generate high yields of the desired product. For *C. saccharoperbutylacetonicum* this can be the production of key solvents during the second, solventogenic, phase of growth induced by high densities and high concentrations of acids in the media. Despite this obvious overlap between high density fermentation and the population density sensing purpose of quorum sensing, the process remains understudied in solventogenic *Clostridia*. Only three studies characterising the accessory gene regulator (agr) (Jabbari et al., 2013; Steiner et al., 2012) and RRNPP (Kotte et al., 2020) systems have been made in *C. acetobutylicum* and none *Clostridium beijerinckii* for example. Only a single study has been conducted in *C. saccharoperbutylacetonicum* which partially characterises the complex RRNPP system (Feng et al., 2020). Therefore, the need for further research into quorum sensing in *C. saccharoperbutylacetonicum* is evident.

As mentioned, some work on the RRNPP quorum sensing system has already been undertaken, however *C. saccharoperbutylacetonicum* also possesses the genes necessary for the agr system that is common across Gram positives. The agr quorum sensing system appears to utilise a relatively simple signal and response mechanism. In *Staphylococci* and in *Clostridium acetobutylicum*, the four genes directly involved, *agrA*, *agrB*, *agrC*, and *agrD*, each encode a protein: AgrA, AgrB, AgrC and the AgrD-derived auto-inducing peptide (AIP) respectively. These can be subdivided into the proteins that create the signal – AgrB and AgrD/AIP – and those that sense it – AgrA and AgrC. AgrB is a transmembrane protein that cleaves the peptide product of *agrD* into the small AIP and allows for its export out of cell (Figure 5.1). The AIP, secreted by the population, builds up outside the cell and interacts with AgrC, another transmembrane protein. AgrC and AgrA form a two-component system which work together to respond to the presence of the AIP. Upon interaction with the AIP, AgrC autophosphorylates and, in turn, phosphorylates AgrA. AgrA is a transcription factor that modulates gene expression both directly (Queck et al., 2008) and through the alteration of RNAIII expression in *Staphylococcus aureus* (reviewed in Novick and Geisinger, 2008).

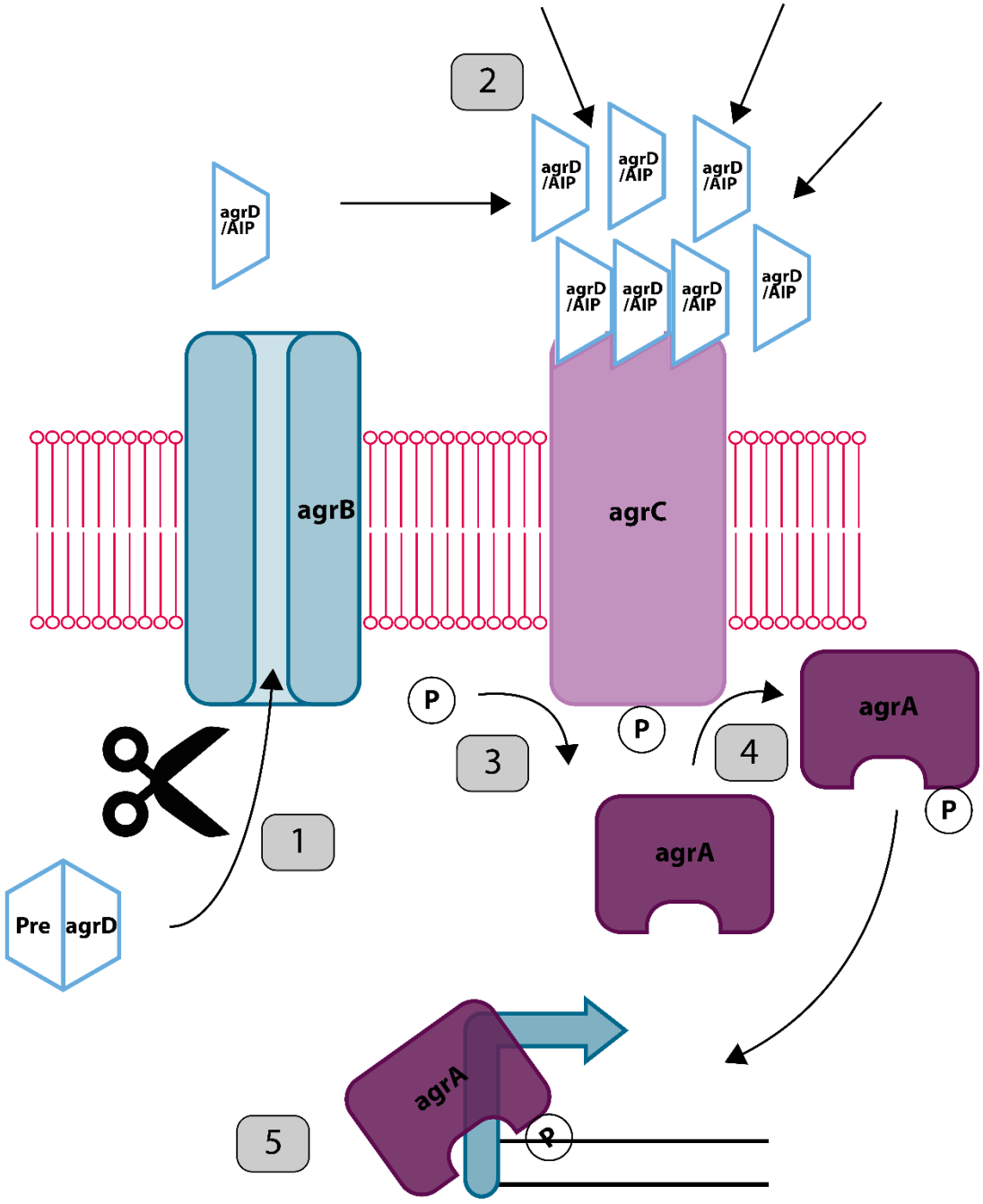

Figure 5.1 Schematic of the AGR quorum sensing mechanism. 1) Pre-agrD/AIP is cleaved into agrD/AIP and transported out of the cell by agrB. 2) AgrD/AIP secreted from the population accumulates in the environment. 3) AgrC autophosphorylates upon sensing the threshold concentration of agrD/AIP. 4) Phosphorylated agrC is then able to phosphorylate agrA. 5) Phosphorylated agrA modifies gene expression by acting as a transcription factor.

No homolog of RNAIII has been found in either *C. acetobutylicum* or *C. saccharoperbutylacetonicum* (this study) suggesting AgrA either conducts all the transcriptional regulation or a regulates a different transcription factor(s).

I decided to investigate the agr system bioinformatically through the annotation of all four genes and experimentally through the deletion and complementation of its key genes to determine their function, if any, in *C. saccharoperbutylacetonicum*. None of the agr genes were indicated as likely to be essential based on the data presented in Chapter III, suggesting the role of the agr system is not fundamental to the survival of *C. saccharoperbutylacetonicum*. However, it seems plausible that it does play a key role in the population level switch from acidogenesis to solventogenesis and also, potentially, in sporulation. All of these processes are important to almost all potential applications for *C. saccharoperbutylacetonicum*.

5.1.1 Aims and objectives

1. Establish the architecture of and gene annotations for the agr system
2. Create deletion mutants using CLEAVE™
3. Characterise candidate promoters to be used in complementation
4. Create complemented strains
5. Characterise the growth and sporulation phenotypes of mutants and complements

5.2 High density growth phenotypes cannot be induced

Typically, the preparation for large-scale fermentation utilises seed-trains with cells being grown in consecutively larger volumes of media. This helps scale up the process to the very high volumes needed to produce significant amounts of product whilst also conferring consistency on growth rates. If quorum sensing, and not simply pH, is involved in the induction of solventogenesis as has previously been implied (Kosaka et al., 2007), it is not unreasonable to theorise that maintaining cells at high density might hasten the induction of solventogenesis.

This experiment represented a simple preliminary method of testing this concept. Cells were grown until logarithmic phase (OD_{600nm} 0.3) and then centrifuged, the media removed and replaced by fresh media of either the same volume or x10 less. Growth was monitored and samples taken for acid and solvent concentration analysis. Unfortunately, the latter could not be completed in time to be included here, but the results of the former are presented below (Figure 5.2). Interestingly, the growth rate of the concentrated cells was faster than that of the control post concentration (73.4 vs 80.5 min, P=<0.0001). It is also possible that the concentrated cells reached higher ODs overall, but this would require repeats arranged so that timepoints between 600 and 1400 min could be taken. Given that is not necessarily expected that growth rate should increase with higher cell density, it is somewhat surprising that such a difference was observed. Whilst this does not indicate if solventogenesis was induced earlier in the high-density condition, the data does suggest there is a phenotypic difference caused by artificially increasing the cell density.

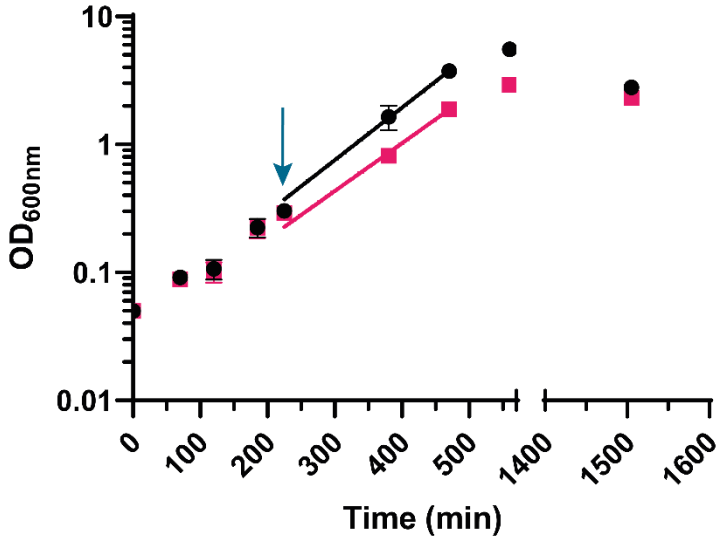

Figure 5.2 Growth curves of induced high density vs control. Artificially induced high density cultures are showing in black whilst the controls are in magenta. The teal arrow indicates the point in growth at which cells were centrifuged and resuspended. The lines shown for both conditions represent a non-linear regression analysis of the growth rate post-resuspension. Where error bars cannot be seen, error was too low to be plotted. Shown are the means and standard deviations calculated from three biological repeats and two technical repeats. A statistical Comparison of Fits conducted on the two growth rates indicated a significant difference between the two (P=<0.0001).

5.3 Identification and architecture of the *agr* genes

5.3.1 Architecture of the *agr* genes

The *agr* system in *C. saccharoperbutylacetonicum* was partially identified in previous genome annotations (Poehlein et al., 2014). Only one gene was assigned as a canonical member of the operon, *agrA* (Cspa_c18650), whilst *agrB* was labelled as a 'putative Agr*B*-like protein' (Cspa_c18660). BLAST searches of the *C. saccharoperbutylacetonicum* N1-4 HMT genome using the *Clostridium acetobutylicum* operon sequences as templates were able to identify *agrC* (Cspa_c18640) and *agrB*, both in the same orientation as for other species (Steiner et al., 2012). The order, orientated on the reverse strand, is *agrB*, *agrA* then *agrC* (Figure 5.3) which is different to that seen in *C. acetobutylicum* where agrC precedes *agrA.* However, *agrD*, the gene encoding the crucial signalling peptide, was not readily identifiable. Work to attempt to find this gene is described in 5.3.2. All the annotated genes are orientated in the same direction and located in close proximity, opening the possibility that they function as an operon. BPROM was able to identify a promoter 266 bp upstream of *agrB* which could be the promoter for all three annotated genes The a*grA* and *agrC* sequences overlap one another by 54 bp, which increases the possibility of at least those two sharing a promoter. If so, the only candidate identified by BPROM is 135 bp upstream of the *agrA* coding sequence. Alternatively, BPROM also identified two putative promoter sequences within *agrA* sequence (shown by mauve arrows in Figure 5.3), suggesting it was possible that all three genes were expressed separately, though this seems less likely for *agrA* and *agrC*. No Rho-independent terminators were found along the entirety of the *agr* region by ARNold (Gautheret and Lambert, 2001; Lesnik et al., 2001; Macke et al., 2001). This leaves open the possibility of a true operon structure with transcriptional readthrough of all genes.

In *C. saccharoperbutylacetonicum*, the *agr* genes are in close proximity to their neighbours: one encoding a putative endo-beta-mananase, downstream of *agrC*, and coded on the opposite strand; and another 969 bp sequence encoding a hypothetical protein upstream of *agrB* and coded on the same strand. The latter contains some features of CAAX proteases, though the function of these proteases is understudied in bacteria.

5.3.2 Identification of putative *agrD* locus

As mentioned, *agrD* was not annotated in previous genome assignments. Likely this is due to the enormous variability in both gene and protein sequences for *agrD*s in other species (Steiner et al., 2012). In *Clostridium acetobutylicum*, *agrD* is contiguous with *agrB* in the same manner as *agrB* and

agrC. A small cyclic peptide, agrD contains only two conserved elements across species: a ring structure containing only one, mostly conserved, cysteine in position 1 of the ring and a proline six amino acids downstream of the ring structure. Even the conservation of the cysteine is limited with 3/28 investigated agrD sequences showing serine occupying the position instead. The overall size of the peptide appears consistently to be between 44 and 57 amino acids in length. BLAST searches of these peptides against the *C. saccharoperbutylacetonicum* genome yielded no results.

The highly variable amino acid sequences seen for agrD suggest that the protein may not necessarily be identified by BLAST searches. Given the most likely location for *agrD* was in proximity to the other *agr* genes, a semi-manual search was conducted. In the first instance, the *agr* region, including up- and downstream loci was searched for the start codons, ATG, GTG and TTG using Geneious v7.1.9 'search for motifs' function. In total, 208 candidate start codons comprising of 100 ATG, 17 GTG and 91 TTG sites were identified. Once sites that were clearly inappropriate, either in the middle of other genes or leading to a short sequence (<100bp) before a stop codon, were eliminated, a total of four candidate sites remained. All four were located in the intergenic spaces between agrB and agrA, the most likely candidate region based on the genetic organisation in other species (Figure 5.3). Two of these four sites comprised the same coding region which was also in frame with agrA. Indeed, both shared a stop codon with agrA. This made them unlikely candidates as they would either need to be cleaved from the bulk of AgrA before being able to act as a signalling molecule or independently translated from a single mRNA. Additionally, whilst both share a proline at an appropriate position (amino acid 49 and 55), they do not appear to contain either a cysteine or serine associated with the cyclic ring in the correct location (amino acid 38 and 44). Together, it seemed unlikely these two start codons indicated the start of agrD.

The remaining two sites were located closer to *agrB* and overlapped significantly, sharing a stop codon. They were both an appropriate length – 141 and 129 bp – however neither showed a proline in the appropriate region. They did however contain a cysteine in a reasonable position. Whilst neither of these were considered perfect candidates, their coding sequences were targeted for knockout as the most likely candidates that could found.

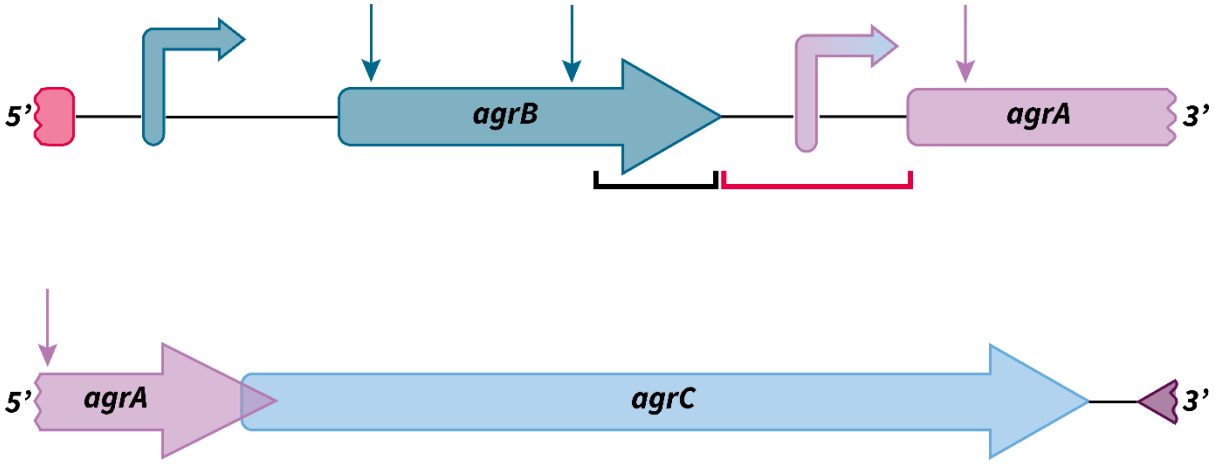

Figure 5.3 The architecture of the *agr* genes in *C. saccharoperbutylacetonicum*. The locus is split between two lines for clarity where each line represents 1,756 bp. Genes are represented by large arrows and depicted to scale relative to one another. Jagged ends are to show where the image cuts the depicted gene. *agrB* is 597 bp long, *agrA* is 732 bp long, *agrC* is 1,365 bp long. L-shaped arrows represent putative promoters if the system in *C. saccharoperbutylacetonicum* follows that of *C. acetobutylicum*. These are not to scale but are in the correct relative loci. Downward pointing arrows indicate additional candidate promoter sites identified by BPROM that could indicate each gene is expressed independently. The magenta square bracket, encompassing 287 bp in total, indicates the original site thought to contain candidate *agrD* whilst the black square bracket, encompassing 186 bp in total, indicates the site of candidate *agrD*s identified later. The small magenta rectangle with a jagged end represents the first 50 bp of a gene encoding an uncharacterised CAAX-like protease whilst the dark purple arrow represents the first 50 bp of the gene encoding an endo-beta-mannanase. Both genes are orientated oppositely to the *agr* genes.

Whilst writing this thesis, a re-analysis was conducted to obtain the figures cited above and two new candidates were discovered. In the original analysis, start codons located within other agr genes were not considered. This meant that two overlapping candidates that start within the *agrB* gene were not originally examined in detail. As with the candidates between the genes, both encompass the same coding region and also share a stop codon with each other and with *agrB*. One would produce a peptide 61 amino acids long whilst the other would produce one of 40 amino acids in length (black square bracket in Figure 5.3). Both are slightly out of the previous ranges described, but not drastically so. Both encompass a proline at an appropriate position – amino acid 45 and 24. This proline is exactly 10 amino acids along from a serine at position 35 and 14 which matches the only conserved features seen in AgrD peptides. Therefore, it seems probable that one of these two candidate coding sequences are responsible for the production of the AgrD peptide.

In order for either of these candidates to become functional signalling peptides one of three criteria must be met. Transcription and translation could occur separately from *agrB* from a promoter and ribosome binding site located within *agrB*. *agrD* could be transcribed in tandem with *agrB* but translated separately. Finally, AgrD could be cleaved from AgrB post-translation. The first two options would require a RBS of which the closest exemplar is likely 5'- ACGTCC-3' located 15 bp upstream of the start codon for the longer of the two putative peptides. No such appropriate sequence is seen in proximity to the shorter one. If a promoter is required, as for the first option, BPROM identified two putative promoters within the *agrB* sequence 44 and 342 bp upstream of the start codon to the longer peptide. Both the first two options are possible, therefore, based on the genetic sequences. The third option is less likely due to either the AgrB requiring two cleaving capacities – one to cleave the AgrD from itself and the second to cleave AgrD into its final form.

5.4 Characterisation of promoters for the *Clostridial* genetic toolbox

To aid in our objective to characterise the agr system, it was necessary to characterise and have at our disposal promoters capable of inducing a range of expression levels. This would help the complementation process necessary to ascribing phenotypes to genes by enabling us to have a wide range of expression levels available for our genes. In particular, it was necessary to characterise two inducible promoters and understand to what extent the induction of these promoters was dose-dependent and the repression of expression tight. This would, of course, allow for precise temporal control of the expression of genes under these promoters. Whilst several promoters are in regular use in the *Clostridial* field, they tend to be developed and used on an *ad*

hoc basis. No systematic attempt has been made to introduce and characterise a range of promoters (Gyulev et al., 2018). It was therefore decided to combine this *ad hoc* situation with work that may benefit the field in general by selecting 11 promoters to characterise in both *C. saccharoperbutylacetonicum* and *C. difficile*.

Seven of the eleven promoters were chosen from genes present in the *C. saccharoperbutylacetonicum* genome and from putative genes on the endogenous megaplasmid. Three are derived from the *C. difficile slpA* gene representing the core promoter with and without putative upstream promoter elements. The thiolase promoter from *C. acetobutylicum* was also selected due to its extensive use in the field (Gyulev et al., 2018). The inducible promoters used were the tetracycline inducible promoter mentioned previously and the xylose inducible promoter cloned from the *C. difficile* by Dr Joe Kirk and based on the use of the promoter in *C. perfringens* (Nariya et al., 2011). All were cloned by PCR amplification and restriction digestion into a luciferase reporter plasmid and tested for activity using luminescence as a proxy. This work was conducted in part by the author and in part by Eloise Walker as part of summer project.

5.4.1 Activity of promoters in Clostridium saccharoperbutylacetonicum

The panel of promoters produced a wide range of activity levels derived from the different promoters. The same vector backbone for all promoters was used as well as a negative control without a promoter upstream of the B*itLucOpt* gene (Pnegativecontrol). Whilst this assay is not sufficiently sensitive to comment on overall activity, it is capable of providing approximate estimates for strong, medium and weak promoters. P*hag4* showed extremely high activity, approximately 140x stronger than the next highest, it can therefore be classed as an extremely strong promoter (Figure 5.4A). The P*slpA* that includes both an UPE and additional upstream sequence demonstrates very strong activity, approximately 4x higher than the next strongest. In the next category down, P*flgB* and P*slpA* with a putative UPE also showed strong activity levels, approximately 10x more than the next category. P*hsp*, P*slpA* (short) and P*bdh* constitute the next category of promoters that showed medium activity levels, being approximately 3x more active than the next category down. Finally, P*rpoD*, P*virion structural protein* and P*atpase* form part of the lowest expressors with activity levels that are clearly above the negative controls. P*thial* does not appear to be active in this context showed no significant difference in RLU/OD_{600nm} when compared to the negative control.

Together, these results allow for the selection of promoters based on activity during logarithmic phase. Some promoters may experience changes in activity during different states of growth as part of shifts in global gene expression. In the panel tested here, it is likely that this is to be the case for the two flagella genes – P*hag4* and P*flgB* – as the cells become non-motile during stationary phase. It is also likely to be the case for P*bdh*, the butanol dehydrogenase responsible for the final step of solventogenic phase production of butanol. It is also noteworthy that promoters selected from the megaplasmid express at a measurable level suggesting those genes are active in the cell.

5.4.2 Activity of promoters in *Clostridium difficile*

Plasmids containing two of the promoters could not be conjugated into *C. difficile*: P*flgB* and P*rpoD*. These presumably had some detrimental impact on cell growth, perhaps through overexpression. The remaining promoters can again be broadly categorised by activity level. P*hag4* again displayed extremely high activity, approximately 3x higher than that of the next highest expressors (Figure 5.4B). P*slpA* with UPE and P*slpA* with UPE + additional sequence constituted the next category of very highly active promoters, active at approximately 10x that of the next highest. Interestingly, the additional sequences beyond the identified UPE reduced activity somewhat, unlike in *C. saccharoperbutylacetonicum*.

Finally, P*bdh* and P*atpase* showed the lowest level of activity in these conditions. Again, P*thial* was not active relative to the negative control. It is worth noting that, in this assay, the background luminescence was higher. This phenomenon was not thoroughly investigated but was likely due to the age of the kits when used to conduct the assay, this being the biggest difference between the assays other than the biological context. This effect was species independent and was seen also whilst developing the protocols used here.

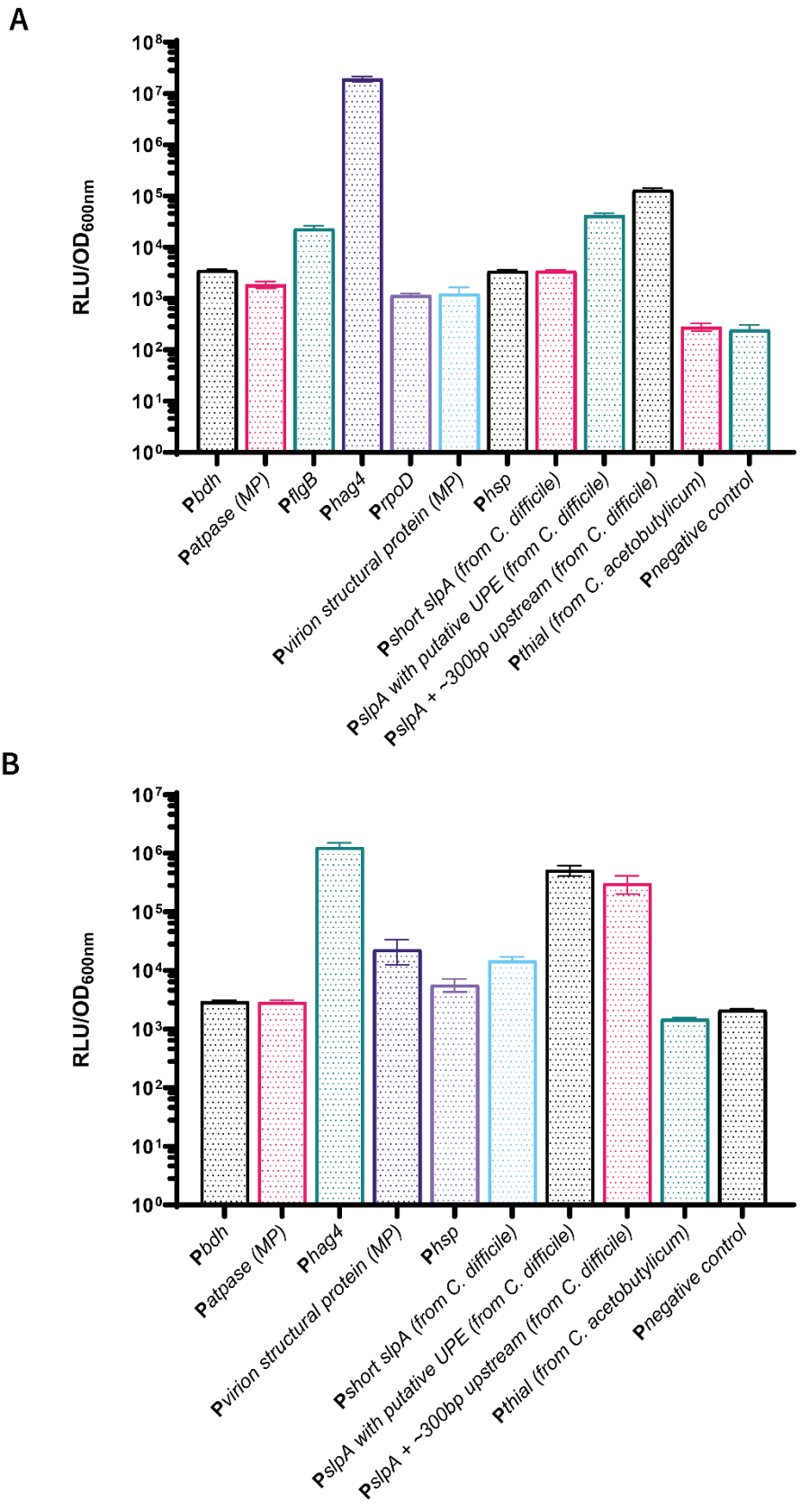

Figure 5.4 The activity of constitutive promoters as measured in relative luminescence units. A) Activity in *C. saccharoperbutylacetonicum*. A Brown-Forsythe and Welch ANOVA was

conducted comparing each promoter to the negative control. All promoters showed a significant difference when compared to the negative control (P*bdh* P= <0.0001; P*atpase* P=0.0003; P*flgB* P=0.0002 Phag4 P=<0.0001; P*rpoD* P=<0.0001; P*virion structural protein* P=0.0133; P*hsp* P=<0.0001; P*short slpA* P=<0.0001; P*slpA with putative UPE* P=<0.0001; P*slpA* + ~*300bp* P=<0.0001) except p*thial* (P=0.963) **B)** Activity in *C. difficile*. A Brown-Forsythe and Welch ANOVA was conducted comparing each promoter to the negative control. All promoters showed a significant difference when compared to the negative control (P*bdh* P= <0.0001; P*atpase* P= 0.0005; P*hag4* P= 0.0004; P*virion structural protein* P= 0.0285; P*hsp* P= 0.0104; P*short slpA* P= 0.0003; P*slpA with putative UPE* P= 0.0005; P*slpA* + ~*300bp* P= 0.0059) including P*thial* (P= <0.0001) however this was due to the pThial mean being below that of the negative control (1488 vs 2128). Pnegativecontrol utilised the same vector backbone but lacked a promoter in front of *BitLucOpt*. Samples were taken from growing cultures at OD_{600nm}0.3. Shown are the means and standard deviations of two biological repeats and three technical repeats.

5.4.3 Inducible promoters showed a dose-dependent response

Two inducible promoters were tested: P*xyl* the xylose inducible promoter (Nariya et al., 2011) and P*tet* the tetracycline inducible promoter (Corrigan and Foster, 2009; Fagan and Fairweather, 2011), in both *C. difficile* and *C. saccharoperbutylacetonicum*. A differing concentration range of inducer was chosen for both promoters as well as uninduced controls to which only the solvent of the inducer was added (Figure 5.5). Both promoters showed a dose-dependent response in both species. *C. difficile* showed a reduced rate of activity from P*tet* at the highest concentration (500 ng/mL) of anhydrotetracycline inducer, likely due to toxicity to the cells at that concentration. *C. saccharoperbutylacetonicum* is more tolerant of anhydrotetracycline and therefore continued to show an increase in activity at 500 ng/mL, though notably the level of increase is reduced.

P*xyl* was the stronger promoter in both species though the results hint at leaky expression with both 0% xylose conditions showing expression levels above that of the promoterless negative control. This was also the case for P*tet* in *C. difficile* but not in *C. saccharoperbutylacetonicum*.

Interestingly, the opposite outcome was expected given the results from Chapter III indicating uncontrolled transposase expression occurring from P*tet* controlled gene. This could partially be explained by the difference in overall expression from both promoters. Given that expression levels from P*xyl* are approximately 10x higher than those from P*tet*, the relative ability to repress expression is possibly closer than at first glance, though this would need to be probed with a more sensitive technique.

Overall, the two promoters are practical for controlling both the expression time and the expression levels in these two species. The dose-dependent responses exhibited suggest that precise control over expression levels is possible. The difference in activity between the two promoters means that a wide range of inducible promoter strengths are possible. Both promoters could make suitable candidates for use in controlling the expression of agr genes during complementation.

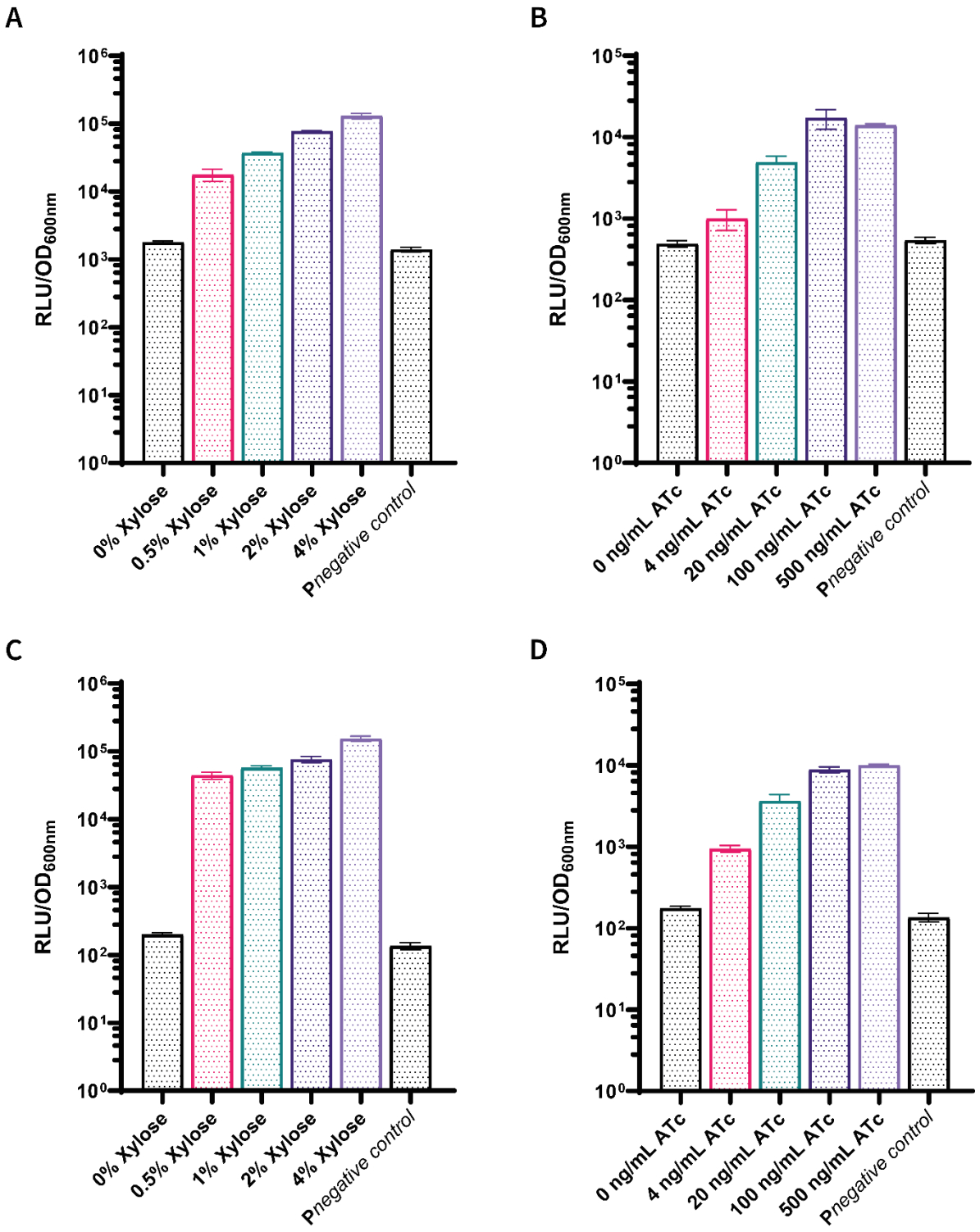

Figure 5.5 The activity of inducible promoters as measured in relative luminescence units. A) Activity of induced and uninduced P*xyl* in *C. saccharoperbutylacetonicum*. A Brown-Forsythe and Welch ANOVA was conducted comparing each promoter to the negative control. All induction concentrations showed a significant difference when compared to the negative control, including 0% xylose (0% P=0.0001; 0.5% P=0.004; 1% P=<0.0001; 2% P=<0.0001; 4% P=<0.0001) **B)** Activity of induced and uninduced P*tet* in *C. saccharoperbutylacetonicum*. A Brown-Forsythe and

Welch ANOVA was conducted comparing each promoter to the negative control. All induction concentrations showed a significant difference when compared to the negative control (4 ng/mL P=0.0457; 20 ng/mL P=0.0004; 100 ng/mL P=0.0014; 500 ng/mL P=<0.0001) except 0 ng/mL ATC (P=0.3212) **C)** Activity of induced and uninduced P*xyl* in *C. difficile*. A Brown-Forsythe and Welch ANOVA was conducted comparing each promoter to the negative control. All induction concentrations showed a significant difference when compared to the negative control, including 0% xylose (0% P=0.0002; 0.5% P=<0.0001; 1% P=<0.0001; 2% P=<0.0001; 4% P=<0.0001). **D)** Activity of induced and uninduced *Ptet* in *C. difficile*. Brown-Forsythe and Welch ANOVA was conducted comparing each promoter to the negative control. All induction concentrations showed a significant difference when compared to the negative control (4 ng/mL P=0.0457; 20 ng/mL P=0.0004; 100 ng/mL P=0.0014; 500 ng/mL P=<0.0001) including 0 ng/mL ATC (P=0.0036) The negative control is the same construct as for the tests in the constitutive promoters. Shown are the means and standard deviations of two biological repeats and three technical repeats.

5.5 Multiple sequence alignments of promoter sequences

5.5.1 All promoters

Given that all the tested promoters were derived from *Clostridial* genomes, it is plausible that they might share similar genetic features and motifs. To examine this possibility, multiple sequence alignments using MUSCLE (Madeira et al., 2022) were conducted on all the promoters tested and the results visualised using Jalview (Waterhouse et al., 2009). The two extended P*slpA* promoter sequences were excluded so as to avoid biasing the results by including several long regions of homology. Results indicated several regions of potential homology across all 11 sequences examined (Figure 5.6).

Previously, putative -10 and -35 binding regions were identified by BPROM which searches for sigma70 driven promoters (Gruber and Gross, 2003; Solovyev, 2011). In this alignment, the identified -10 and -35 regions did not align directly with one another, though were generally clustered in the 52-70bp (magenta box in Figure 5.6) and 90-107 (teal box in Figure 5.6) regions of the consensus sequence respectively. A detailed breakdown of the BPROM predicted -10 and -35 regions and other features can be seen in Appendix VII (Figure VII.1).

The only major promoter feature published concerning *Clostridial* promoters is the HU binding region in *C. difficile* (Oliveira Paiva et al., 2019). No consensus sequence was published for this binding and so promoters were searched for the known consensus recognition sequence for the *E. coli* Integration Host Factor (IHF) of 5'-ATCAANNNNTTR-3' (Hales et al., 1994). IHF has been shown to bend DNA sharply to aid in a variety of DNA manipulations such as chromatin condensation (Oppenheim et al., 1993) and to allow transcriptional machinery to bind (Dorman and Deighan, 2003; Goosen and van de Putte, 1995). Promoters were searched using the 'Find Motifs' function provided in Geneious Prime. This, in turn, is based on EMBOSS explorer's fuzznuc searcher (Rice et al., 2000) and can search nucleotide sequences for patterns including allowances for mismatches. The IHF consensus sequence was used to search both strands with up to 2 mismatches allowed. Candidate IHF sites were found in all sequences though some required more mismatch allowance than others. These tended to cluster at the 52-70 bp region of the consensus sequence with 10 of the 24 predicted motifs representing 8 of the 11 promoters clustering in this region, though not with consistent orientation or alignment (Appendix VII, Figure VII.1). Another 6 predicted IHF motifs in 6 of the promoter sequences were clustered at the 210-221 bp of the consensus sequence (dark purple box in Figure 5.6).

.

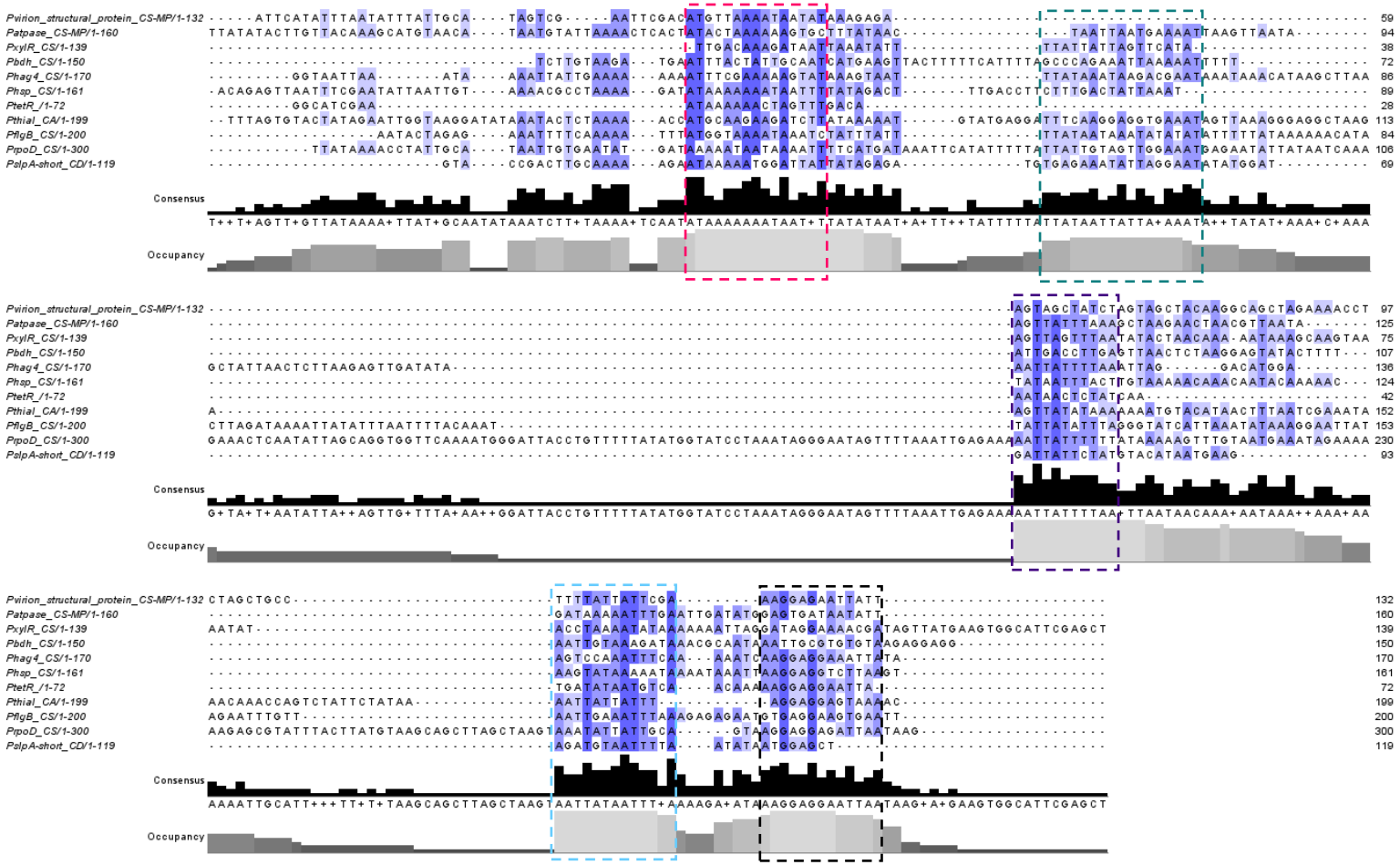

Figure 5.6 Multiple sequence alignment of all tested promoters. Tested promoter sequences as aligned by MUSCLE excluding the longer versions of the P*slpA* promoter. The darker the blue, the better consensus in that region. Consensus line indicates percentage identity at a given position. Occupancy indicates the number of different sequences represented at a given point. Three areas of high consensus are indicated: position 52-70 (magenta box), 90-107 (teal box), 211-221 (dark purple box), 286-299 (light blue box), and 308-320 (black box). All promoter sequences end prior to the gene.

Three other regions of strong consensus were also found (dark purple, light blue and black boxes in Figure 5.6). Both the dark purple and light blue consensus boxes plausibly represent alternative and -10 promoter regions. In the case of the latter, the -10 5'-TATAAT-3' consensus seems especially well conserved. This suggest BPROM was not especially accurate in identifying the -10 sequence and that this sequence is well-conserved in *Clostridia.*

5.5.2 Highly expressing promoters

To examine if there were features unique to the more highly expressing, these were aligned separately using MUSCLE. To allow for a meaningful number of sequences, all but the lowest expressors as measure in *C. saccharoperbutylacetonicum* were used for the analysis: P*bdh*, P*flgB*, P*hag4*, P*hsp*, *PslpA-short*, P*tetR*, and P*xylR*. These were once again visualised by Jalview and the regions of highest identity highlighted (Figure 5.7). Three notable areas of high identity were visible. Furthest from the gene starts was a region containing a highly conserved 5'-TTG-3' site (28-45 bp of the consensus sequence, magenta box in Figure 5.7). It is possible this represents part of the consensus recognition sequence for IHF or an IHF-like domain, though the sequence upstream of this does not seem to contain the *E. coli* consensus sequence for all promoters. This is evidenced by a lack of any clustering of the predicted IHF motifs at this site (Appendix VII Figure VII.2). In addition to the conserved TTG site, the consensus sequence also shows a highly conserved 5'-AAAA-3'. It is unclear what, if any meaning, this conserved sequence has for the promoters.

The next region of high identity is at 83-88 bp of the consensus sequence (teal box in Figure 5.7). Here this is a highly conserved 5'-AAATAA-3' sequence of unknown importance. No previously identified features appear to cluster that this point. Following this is a region containing a potential candidate for a conserved -10 region, 5'-TATAAA-3', that aligns with nearly all promoters as before. Downstream is a region of potential consensus that contains both alternative -10 regions for P*flgB* and P*slpA-short*. It also contains candidate ribosome binding sites for several of the promoters with the overall consensus sequence of 5'-TGGAGNG3'. However, for P*bdh* and P*hsp* this site is clearly located further downstream with perfect ribosome binding sites observable (5'-AGGAGG-3') (Omotajo et al., 2015). Additional small regions of consensus, mostly involving As and Ts are visible along this region, though it is hard to discern a specific sequence.

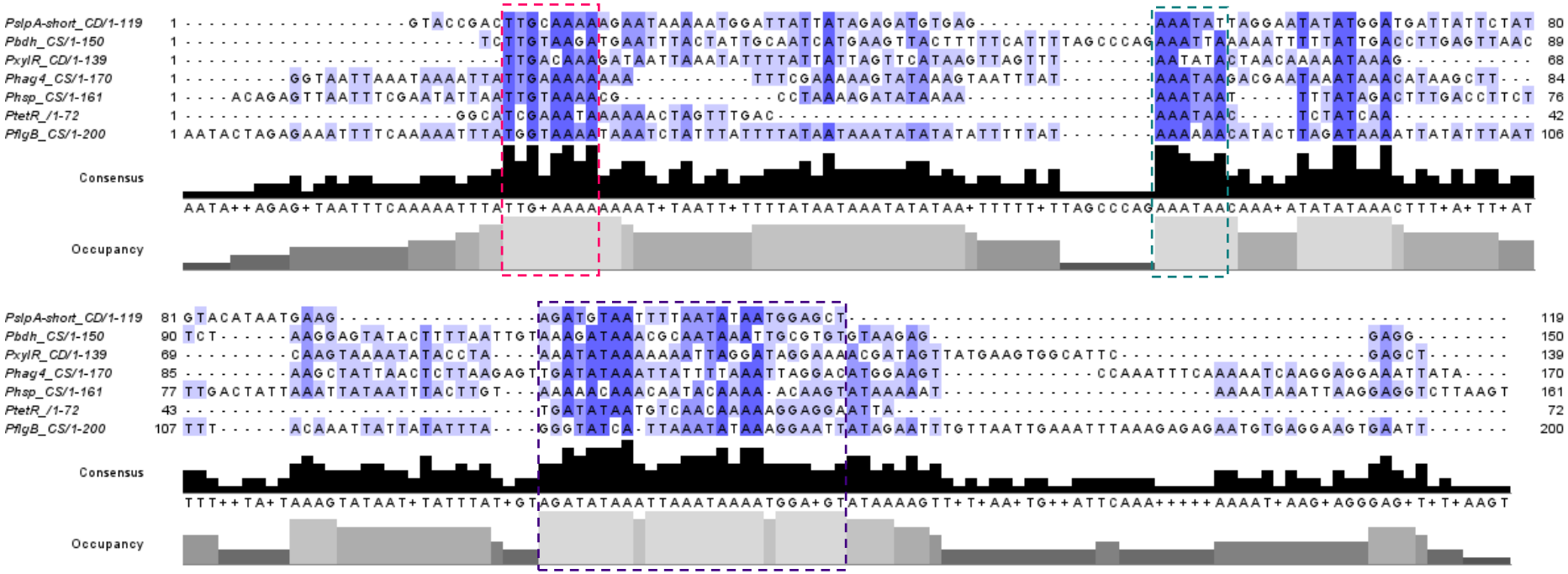

Figure 5.7 Multiple sequence alignment of highly expressing promoters. Promoters deemed to be high expressors as aligned my MUSCLE. The darker the blue, the better consensus in that region. Consensus line indicates percentage identity at a given position. Occupancy indicates the number of different sequences represented at a given point. Three areas of high consensus are indicated: position 28-35 (magenta box), 83-88 (teal box) and 145-170 (dark purple box). All promoter sequences end prior to the gene sequence.

5.5.3 Low expressing promoters

Four promoters could be broadly by classified as low expressors in *C. saccharoperbutylacetonicum*: P*virion_structural_protein*, P*atpase*, P*rpoD* and the non-expressing P*thial*. These were again aligned and viewed as before. Interestingly, these contained several regions of high identity (Figure 5.8). The first, highlighted by the magenta box, contains several distinct conserved features. The first part comprises a region of reasonably conserved As and Ts, though these not necessarily indicative in high AT genomes (70.5% in *C. saccharoperbutylacetonicum*). Of particular note is a conserved 5'-CATA-3' and a region that between 55-81 bp of the consensus sequence containing all of the BPROM predicted -10 sequences (Appendix VII Figure VII.3). The -10 regions are merely clustered and not directly aligned but likely explain the highly conserved 5'-TAAAA-3' between 62-66 bp and 5'-TAT-3' at 70-72 bp. Within the magenta box window, there is also a notably conserved 5'-AAAGAG-3' sequence at 92-98 bp of unknown importance.

The next region of consensus (teal box Figure 5.8) also contains a number of unknown sequences. These include and 5'-AAGT-3' at 11-114 bp, an 5'-ARG-3' at 133-135 bp, and a 5'-TAAA-3' at 141-145 bp. The author was unable to identify a specific importance to these sequences. Following this region is one highlighted by the dark purple box in Figure 5.8 with a consensus of 5'-AACTAACTTTWAT-3', again of unknown significance. Next, in the light blue box, contains a region well conserved in all but P*virion_structural_protein* with the sequence 5'-AAATAAAAAAANGAGTN-3' at 231-247 bp. Additionally, there is a sequence conserved in all following this, 5'-TATT-3' at 248-251 bp.

Finally, highlighted by the black box, is well conserved region containing a distinctive 5'-TATTATT-3' that could represent an alternative -10 region, however, the 5'-AGGAGG-3' RBS appears to be upstream of this sequence for the P*virion_structural_protein* and P*atpase* (both derived from the megaplasmid). For P*rpoD* and P*thial* the ribosome binding site can be clearly identified downstream of this.

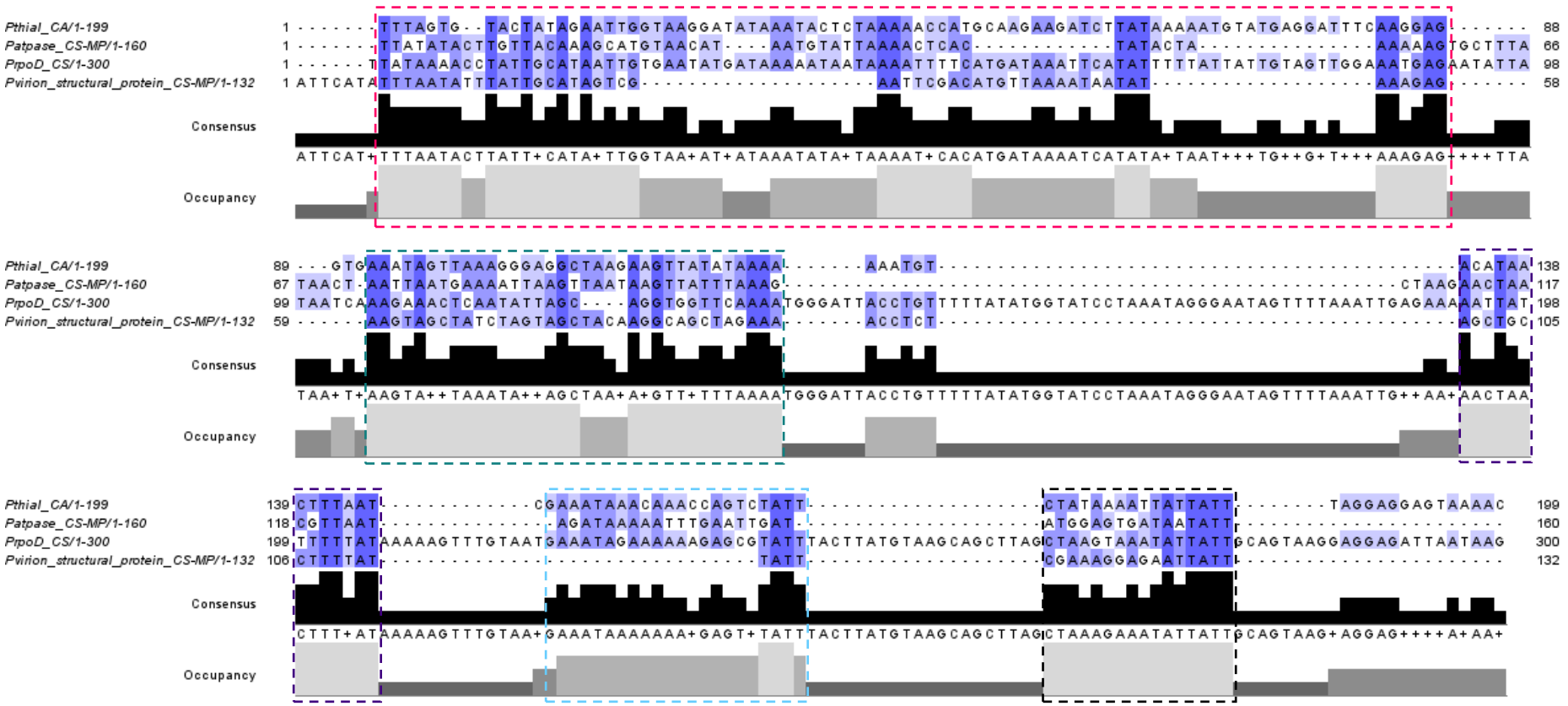

Figure 5.8 Multiple sequence alignment of low expression promoters. Promoters deemed to be low expressors as aligned my MUSCLE. The darker the blue, the better consensus in that region. Consensus line indicates percentage identity at a given position. Occupancy indicates the number of different sequences represented at a given point. Five areas of high consensus are indicated: position 8-97 (magenta box), 111-145 (teal box), 203-215 (dark purple box), 230-251 (light blue box) and 272-287 (black box).

5.5.4 C. saccharoperbutylacetonicum promoters

The previous alignments all contained promoters originating from a mix of species. It is possible, therefore, that promoter features exclusive to *C. saccharoperbutylacetonicum* were missed in the mixed analyses. Whilst there was wide variation in the observed expression levels of the different *C. saccharoperbutylacetonicum* sourced promoters, certain sequences core to the transcriptional process in *C. saccharoperbutylacetonicum* could still be present in all promoters. Therefore, all promoters originally sources from *C. saccharoperbutylacetonicum* (P*bdh*, P*flgB*, P*hag4*, P*hsp*, P*rpoD* from the genome and Pv*irion_structural_protein* and P*atpase* from the megaplasmid) were aligned together as previously described (Figure 5.9).

As before, several regions of high homology are seen across the generated consensus sequence. Furthest from the gene start site is an AT-rich sequence (magenta box in Figure 5.9). Particularly highly conserved are an 5'-AT-3' in the middle of this sequence and an 5'-AAA-3' site at the end. Interestingly, a 5'-TTG-3' is conserved within this sequence. Though the upstream sequence does not appear to match the *E. coli* consensus for the IHF, the overall sequence presented in the magenta box does contain 6 of the 16 predicted IHF motifs across all promoters representing 5 of the 7 promoters (Appendix VII Figure VII.4).

The next region of homology (teal box in Figure 5.9) shows what appears to be a consensus -10 region though, again, only 3 of the BPROM-predicted -10 regions are located in this region (Appendix VII Figure VII.4). Following this, a region with high homology in the light blue box stands. Whilst the features do not stand out as indicated a particular binding sequence, there are well-conserved 5'-TTA-3' sites at either end of the sequence. Finally, the black box highlights the ribosome binding site, which appears to be highly similar to the consensus and well conserved across a range of different genes in *C. saccharoperbutylacetonicum*.

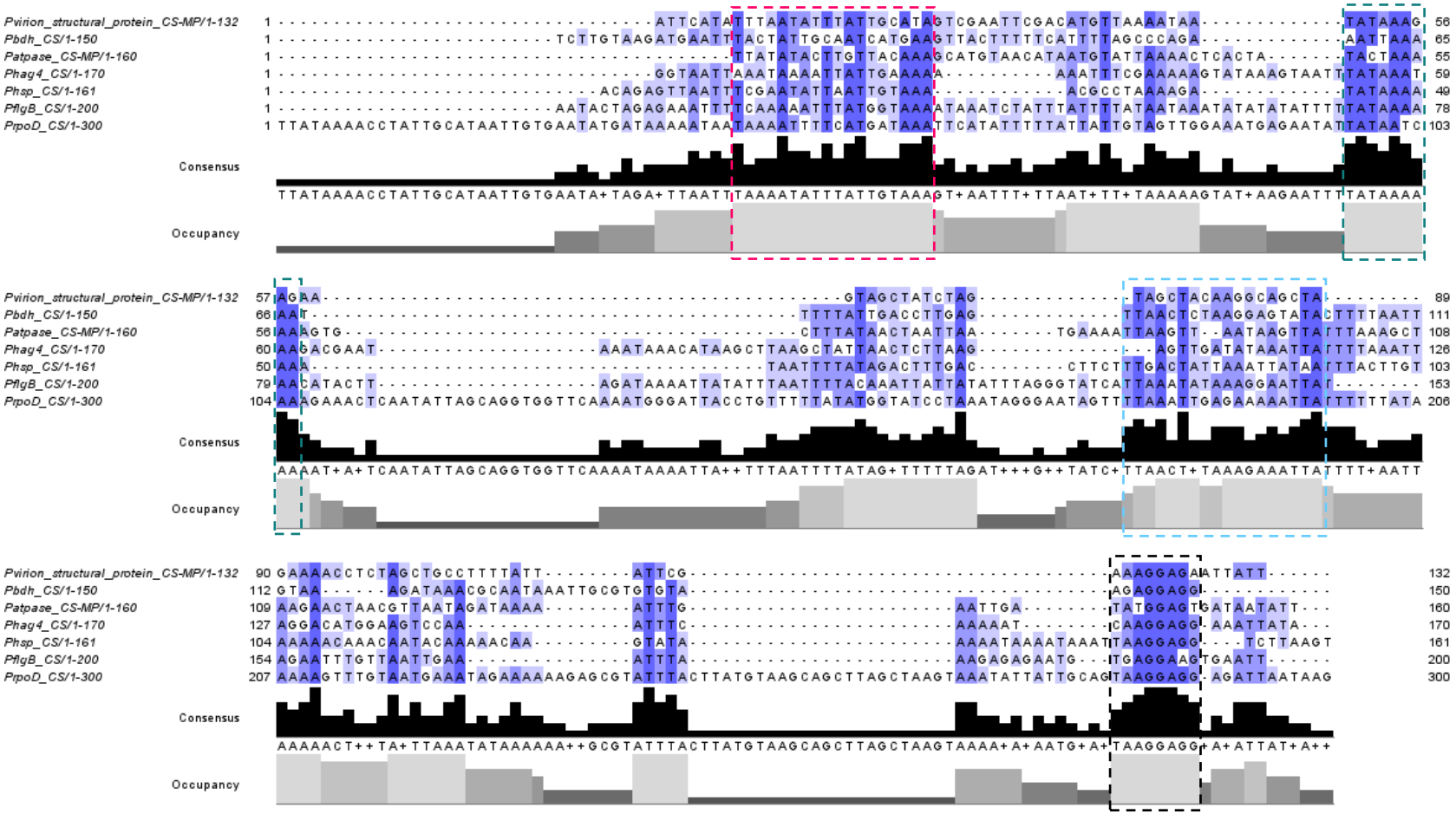

Figure 5.9 Multiple sequence alignment of promoters sourced from *C. saccharoperbutylacetonicum*. Promoters sourced from *C. saccharoperbutylacetonicum* genome as aligned my MUSCLE. The darker the blue, the better consensus in that region. Consensus line indicates percentage identity at a given position. Occupancy indicates the number of different sequences represented at a given point. Five areas of high consensus are indicated: position 42-59 (magenta box), 98-107 (teal box), 180-196 (light blue box) and 282-289 (black box).

5.6 Obtaining and characterising gene deletions in the agr system

5.6.1 Successful deletion of *agrB*, *agrC*, *agrD* and the whole system (*agrBDAC*)

Gene deletions were conducted using the CLEAVE™ genome modification system. Homology arms were designed to be between 600 and 1,200 bp with a longer length preferred where possible to increase recombination frequency whilst avoiding tricky-to-clone and synthesise repeat regions. They were also designed to retain 12-21 bp of the original gene at either end, thought this was not always possible (e.g., for the deletion of *agrB*). Homology arms were either cloned by Gibson assembly into the step 1 recombination vector or synthesised as GeneArt™ Strings™ and cloned into vector using NEBuilder. The CLEAVE™ system also requires the identification of suitable spacer sequences within the deleted regions. Appropriate spacers were found for all genes and multiple spacers cloned where possible. Spacers were synthesised by Genewiz. A full list of primers (Appendix III), homology arm plasmids (Appendix II) and spacers used (Appendix V) are listed in the appendices. Homology arms were successfully assembled for the deletion of *agrB, agrC*, *agrD* and the whole region (*agrBDAC*). The deletion process was then carried out as described in Chapter II. Successful deletion mutants, screened by PCR (Figure 5.10) and confirmed by the sequencing of PCR products, were obtained for *agrB*, *agrC*, *agrD*, and the whole region. *agrA, agrB* and *agrC* are all marked as essential by Bio-Tradis in Chapter III despite each gene containing multiple insertions. The successful deletions highlight the statistical issues with the transposon library created in that chapter.

As intimated above, complementation was also planned for and several complementation vectors using the P*Xyl* and P*rpoD* promoters were built. However, ultimately, I was unable to schedule the complementation of deletion mutants and therefore the phenotypic analyses were conducted on the deletion mutants only.

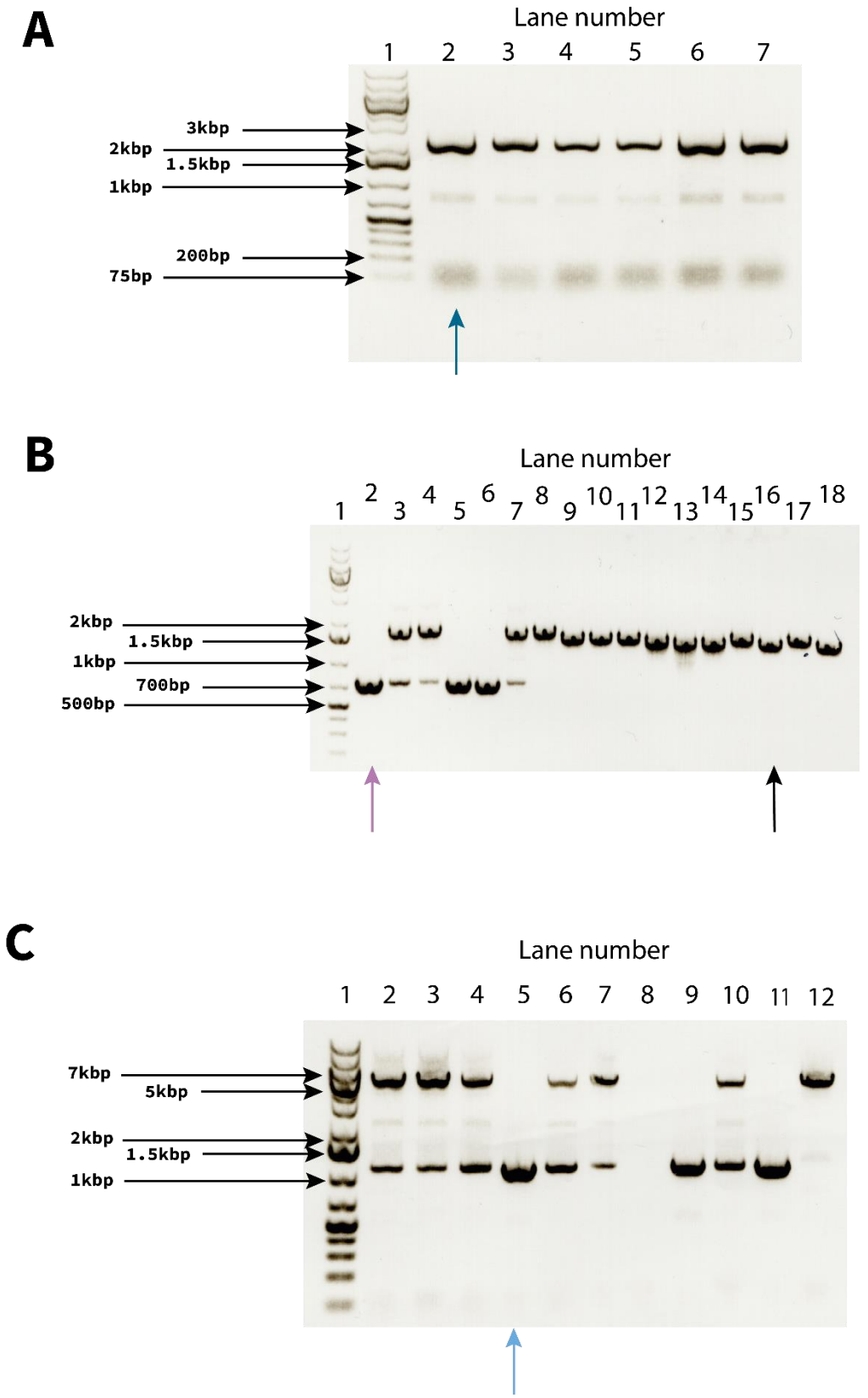

Figure 5.10 Identification of *agr* knockouts by PCR and agarose gel electrophoresis. All PCR products were run on 1% agarose gels at 100 v for 30 mins. The DNA ladder Generuler 1 kb Plus was used as sizing standard. Key band sizes are highlighted. **A)** Successful knockout of *agrC* in all cases. Wild type product would be present at 3370 bp whilst the knockout

produces the 2046 bp band seen in all of the tested clones. All clones were gel extracted and sequenced with all having their deletions confirmed. The clone in lane 2 was randomly chosen for characterisation. **B)** Successful knockout of *agrB* and *agrD*. Both knockouts utilise the same screening primers. Lane 8 contains the positive control for this PCR conducted with wild type DNA. For *agrB*, lanes 2-6, the wild type band is expected at 1645 bp (i.e. lane 3 and 4) whilst the knockout band is expected at 717 bp with three successful deletions in lanes 2, 5 and 6. All three were sequenced and had their deletions confirmed. The clone in lane 2 (lavender arrow) was randomly chosen for characterisation. For *agr'D'*, lanes 9-18, the wild type was expected at 1645 bp whilst the knockout should run to 1526 bp. With an unclear distinction between correct and incorrect clones, several were carried forward for sequencing with the clone in lane 16 one of those carrying the correct knockout. **C)** Successful knockout of the whole operon of *agrBDAC*. Wild type *agrBDAC* would produce a band at 5770 bp whilst the knockout would produce a band of 1016 bp. Lane 12 contains the positive control for this PCR conducted with wild type DNA. the Lanes 5 and 11 contained successful knockouts, both of which were confirmed by sequencing. The clone from lane 5 was randomly chosen for characterisation.

5.6.2 Phenotypic analyses

5.6.2.1 agr deletion mutants show growth defects

Conducting small-scale growth experiments represents the first steps in screening mutants for potential phenotypes. A growth screen was conducted in biological triplicate and technical duplicate for the four deletion mutants and with the background wild type strain as a control. OD_{600nm} and culture pH was monitored for 48 h, with samples taken regularly for determination of the concentration of glucose, metabolic acids and metabolic solvents. As it was not possible to decisively cure the strains of either the Step 1 homologous recombination plasmid nor the Step 3 selection plasmid, mutant strain media was supplemented with erythromycin to maintain selective pressure against reversion.

Differences in OD_{600nm} and pH could clearly be seen for all mutants when compared to the wild type (Figure 5.11A). Δ*agrC* and Δ*agrBDAC* showed the most pronounced phenotypes, displaying slower initial growth, lower final OD_{600}, and lower final pH values. Δ*agrB* showed mild differences in two of the three categories, showing initial growth rates comparable to wild type. Δ*agr'D'* also showed mild differences in all three categories. The variation between biological repeats was notably high for these latter two whilst all strains showed at least some notable variation between biological repeats. With technical repeats being consistent, it is likely that the handling prior to the screen inoculation was not even between all repeats. Nonlinear regression analysis of the initial growth rate over the first 6 h showed doubling times of 155 min for the wild type, 136 min for Δ*agrB*, 178 min for Δ*agrC*, 161 min for Δ*agr'D'*, and 502 min for Δ*agrBDAC*. These differences were significant by Comparison of Fits analysis (P=<0.0001). The extremely slow growth rate seen in the first 6 h for Δ*agrBDAC* is partly due to a long lag phase as OD_{600nm} did increase to levels comparable to Δ*agrC* after 24 h. The average final OD_{600nm} values were 11.96±2.24 for the wild type, 9.78±2.03 for Δ*agrB*, 6.00±2.66 for Δ*agrC*, 7.33±1.47 for Δ*agr'D'* and 5.07±0.56 for Δ*agrBDAC*. These differences were significant for Δ*agrC*, Δ*agr'D'* and Δ*agrBDAC*, but not for Δ*agrB*.

The media pH dropped steadily in the wild type for the first 24 h before stabilising and increasing slightly over the subsequent 24 h (Figure 5.11B). This is the expected pattern as the culture shifts from acid production to acid reabsorption and solvent production. Similar patterns are seen for Δ*agrB* and Δ*agr'D'*, though the drops are significantly lower and they recovered to a lower final pH than that seen for wild type, suggesting a slight delay in switching phases. Both Δ*agrC* and Δ*agrBDAC* showed significant difficulty in regulating culture pH with pH dropping below 5 after 24 h and being unable to recover in the subsequent 24 h. This drop suggests that neither mutant was able to shift

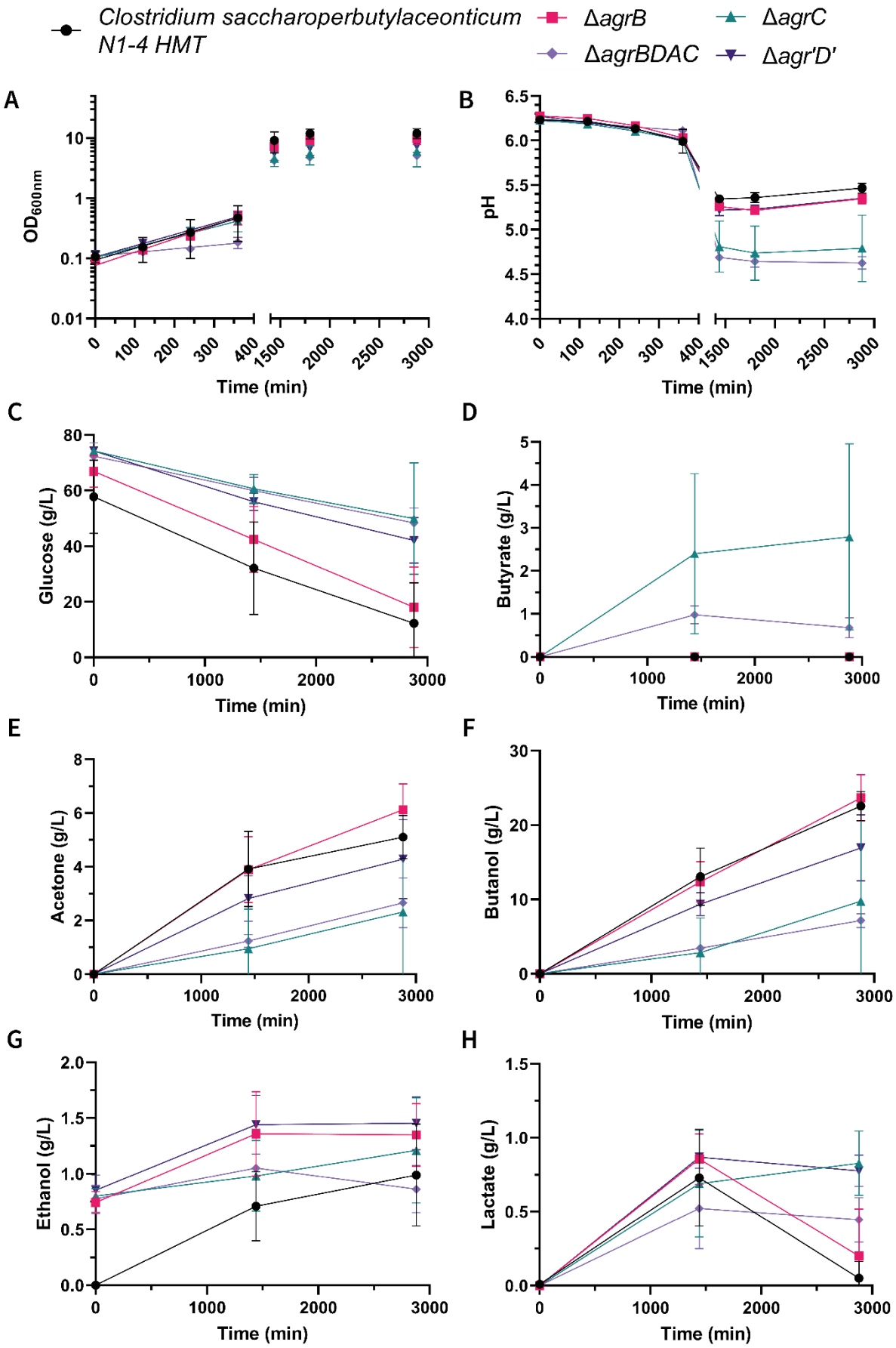

Figure 5.11 Growth profiles of *agr* gene deletions compared to wild type over 48 h.

A) OD_{600nm} over 48h. Lines represent nonlinear regression analysis of each strain. The difference in growth rates were significant by Comparison of Fits between all knockouts and the wildtype (P=<0.0001). It must be noted that Δ*agrB* grew significantly faster than the wild type. **B)** pH of the strain culture media. A Brown-Forsythe and Welch ANOVA comparing the wild type to all other conditions showed no significant difference at the 360 min timepoint but significant differences for all conditions at the final timepoint (Δ*agrB* P=0.0209; Δ*agrC* P=0.0239; Δ*agr'D'* P=0.0249; Δ*agrBDAC* P=<0.0001) **C)** Glucose concentration in g/L. A Brown-Forsythe and Welch ANOVA comparing the wild type to all other conditions showed no significant difference at the 1440 min timepoint for Δ*agrB* (P=0.6869) and Δ*agr'D'* (P=0.1030), but significant differences were seen between wild type and Δ*agrC* (P=0.0464), and Δ*agrBDAC* (P=0.0499). The same test at the 2880 min timepoint showed a significant difference between wild type and all conditions (Δ*agrC* P=0.0211; Δ*agr'D'* P=0.0225; Δ*agrBDAC* P=0.0116), except Δ*agrB* (P=0.9391). **D)** Butyrate concentration in g/L. No butyrate was detected for wild type, Δ*agrB* and Δ*agr'D'*. Due to this, statistical analysis by ANOVA was not possible. **E)** Acetone concentration in g/L. A Brown-Forsythe and Welch ANOVA comparing the wild type to all other conditions showed no significant difference at the 1440 min timepoint for Δ*agrB* (P=>0.9999) and Δ*agr'D'* (P=0.4957), but significant differences were seen between wild type and Δ*agrC* (P=0.0286), and Δ*agrBDAC* (P=0.0423). The same test at the 2880 min timepoint showed no significant difference between wild type and all conditions (Δ*agrB* P=0.2825; Δ*agrC* P=0.3367; Δ*agr'D'* P=0.6780), except Δ*agrBDAC* (P=0.0042) which was significantly lower. **F)** Butanol concentration in g/L. A Brown-Forsythe and Welch ANOVA comparing the wild type to all other conditions showed no significant difference at the 1440 min timepoint for Δ*agrB* (P=0.9939) and Δ*agr'D'* (P=0.2968), but significant differences were seen between wild type and Δ*agrC* (P=0.0192), and Δ*agrBDAC* (P=0.0166). The same test at the 2880 min timepoint showed no significant difference between wild type and all conditions (Δ*agrB* P=0.9150; Δ*agrC* P=0.2329; Δ*agr'D'* P=0.0917), except Δ*agrBDAC* (P=<0.0001) which was significantly lower **G)** Ethanol concentration in g/L. A Brown-Forsythe and Welch ANOVA comparing the wild type to all other conditions showed significant difference at the 1440 min timepoint for Δ*agrB* (P=0.0438) and Δ*agr'D'* (P=0.0141), but no significant differences were seen between wild type and Δ*agrC* (P=0.5209), and Δ*agrBDAC* (P=0.4101). The same test at the 2880 min timepoint showed no significant difference between wild type and all conditions (Δ*agrB* P=0.4785; Δ*agrC* P=0.8892; Δ*agr'D'* P=0.2596; Δ*agrBDAC* P=0.9570). **H)** Lactate concentration in g/L. A Brown-Forsythe and Welch ANOVA comparing the wild type to all other conditions showed significant difference at the 1440 min timepoint showed no significant difference between wild type and all conditions (Δ*agrB* P=0.8871; Δ*agrC* P=0.9994; Δ*agr'D'* P=0.8547; Δ*agrBDAC* P=0.7044). The same test at the 2880 min timepoint showed significant difference between wild type and all conditions (Δ*agrC* P=0.0002;

Δ*agr'D'* P=<0.0001; Δ*agrBDAC* P=0.0029) except Δ*agrB* (P=0.7340). Wild type *C. saccharoperbutylacetonicum* N1-4(HMT) is shown in black; Δ*agrB* shown in magenta; Δ*agr'D'* shown in dark purple; Δ*agrC* shown in teal; Δ*agrBDAC* shown in mauve. Data points are the means and standard deviations of three biological repeats with two technical repeats. Where error bars cannot be seen, error was too low to be plotted. Where bars are cut off by the x axis, error extended below the x axis.

from acid production to solvent production with the consequence that the environment became too acidic and the population either stopped growing, or likely, died.

5.6.2.2 Δ*agrBDAC*, Δ*agrC* and Δ*agr'D'*, but not Δ*agrB* show solvent production defects

Glucose, solvent and acid concentration measurements were conducted by Dr Victoria Green and Sasha Atmadjaja at Biocleave by high performance liquid chromatography. Due to resource restrictions, analysis could only be conducted on samples from 0 h, 24 h and 48 h. Glucose concentration started higher than anticipated averaging 71.3±4.89 g/L when 50 g/L was targeted (Figure 5.11C). Presumably this was a manual error on the part of the author, likely due to mismeasurement and loss of water during autoclaving. The high concentrations did not seem to affect the experiment where concentration dropped steadily for all five strains. In line with pattern of growth shown by OD_{600nm}, wild type and Δ*agrB* consumed glucose at the fastest rate, followed by Δ*agr'D'*, however the difference was only significant for Δ*agr'D'*. Δ*agrBDAC* and Δ*agrC* consumed significantly less glucose at the slowest rate.

Butyrate production was not observed for the wild type, Δ*agrB* and Δ*agr'D'* but was seen for Δagr*BDAC* and Δ*agrC* to final average concentrations of 0.68±.0.23 g/L and 2.79±2.16 g/L respectively (Figure 5.11D). Butyrate production likely occurred in the first ~12 h of the screen for the former three strains followed by its reabsorption by 24 h. The result somewhat reflects the pH measurements showing that these two strains acidified the culture. It should be noted that the high error seen for Δ*agrC* is again due to one biological repeat containing no butyrate – the same repeat that caused high error in OD_{600nm} and pH readings. If this is excluded, then the final concentration rises to 4.19±0.12 g/L. Interestingly, the final concentrations show a significant difference in butyrate production between Δagr*BDAC* and Δ*agrC*. This is still the case even when the faster-growing Δ*agrC* biological repeat is accounted for. This is the first substantive difference in phenotype between the two strains.

No acetate production was observed at the time points sampled hence no graph is shown. This was likely due to lower overall concentrations compared to butyrate combined with a similar production. However, it is notable that even the putative acid producing strains did not produce acetate. Final lactate levels are low for all strains though somewhat higher for Δ*agrBDAC*, Δ*agrC* and Δ*agr'D'* than wild type or Δ*agrB* (0.45±0.15, 0.83±0.22, 0.78±0.11 g/L vs 0.05±0.11 and 0.20±0.32 g/L; Figure 5.11H). In all strains, the majority of lactate was produced in the first 24 h. When all acid

production is accounted for, it is unclear what caused the drop in pH seen for Δagr*BDAC*. This may be a measurement error, a sampling error, or, less likely, a third acidic product.

Acetone was produced by all strains (Figure5.11E), although two of the three biological repeats of Δ*agrC* did not produce any acetone at all. Acetone production was significantly attenuated at the 24 h timepoint for Δ*agrBDAC* and Δ*agrC*. Final concentrations for the strains were: 5.10±0.80 g/L for wild type; 6.12±0.96 g/L for Δ*agrB*; 2.32±3.59 g/L for Δ*agrC*; 4.29±1.47 g/L for Δ*agr'D'*; and 2.66±0.93 g/L for Δ*agrBDAC*. The Δ*agrBDAC* final concentration was the only one deemed to differ significantly from the wild type.

Ethanol production did occur in all strains but was low (Figure 5.11G). It must be noted that all the deletion strains started with ethanol as erythromycin was used to maintain the selective pressure against reversion. No significant difference was seen at the final timepoint of ethanol production between the knockouts and the wild type. Finally, butanol production replicated the pattern seen in 5.6.2.1 for the different strains (Figure 5.11F). With a final concentration of 23.67±3.13 g/L, Δ*agrB* showed comparable butanol production to the wild type (22.56±1.95 g/L). As with other metrics, Δ*agr'D'* also showed reduced final butanol concentration of 16.93±4.45 g/L. Δ*agrBDAC* showed significantly attenuated butanol production, only able to produce 7.16±0.93 g/L after 48 h. Δ*agrC* again showed two distinct phenotypes between repeats with an average 9.75±14.00 g/L, which showed an insignificant difference with wild type due to high standard deviation. The one wild type-like biological repeat showed a butanol concentration of 27.81±0.77 g/L whilst the other two biological repeats barely produced butanol at all with a final concentration of 0.72±0.12 g/L.

5.6.2.3 Δ*agrBDAC* and Δ*agrC*, but not Δ*agr'D'* or Δ*agrB* show sporulation defects

A preliminary experiment was conducted to examine the ability of the mutants to sporulate in the conditions previously described. Overnight cultures of each mutant were spread on TYIR agar supplemented with 50 g/L γ-cyclodextrin. These were then left to grow and harvested as previously described after 3 days. The cells were then visualised under brightfield microscopy. As time was limited and the camera quality poor, images were not captured at this stage in the hope that better images could be taken at later date. Ultimately, this did not occur and so described here are notes on these initial observations.

The wild type displayed the classical sporulating population as seen in Chapter IV: populations of vegetative cells, sporulating cells and released spores were all observed. Spores and sporulating cells were visible for the Δ*agrB* strain. The only slight visual phenotype was an apparent smaller

average cell size and some curvature, though it must be emphasised that no objective measurements were taken. Δ*agrC* showed no sporulation or sporulating cells and clearly exhibited a smaller cell size when compared to wild type. The cells observed were motile, indicating a live population. The Δ*agr'D'* population showed wild type characteristics including sporulation and released spores. No observable differences were noted. Finally, the Δ*agrBDAC* population was also unable to sporulate, showing a similar small cell phenotype seen for Δ*agrC*. In contrast to Δ*agrC*, some enlarged cells and cell chains were seen, but these did not resemble the *Clostridial* form enlargement associated with sporulation in wild type, instead appearing longer and thinner.

Together these results implicate the agr quorum sensing system directly in sporulation. As with the growth phenotypes, the deletion of *agrB* did not result in the same phenotypes as those seen for the deletion of *agrC* and the whole system. The deletion of putative *agrD* displayed no impact whatsoever on sporulation.

5.7 Discussion

The results obtained in this chapter, whilst limited in statistical power, clearly show the importance of quorum sensing systems in the growth of *C. saccharoperbutylacetonicum*. Whilst deletion of all the genes in the *agr* system individually was not possible, intriguing phenotypes are seen for the deletions that were obtained. The whole operon is clearly of importance to the switch from acidogenesis to solventogenesis, though not essential for cell survival, despite receiving that assignation in Chapter III. Interestingly, this was not seen in *C. acetobutylicum*, where the deletion of *agrA*, *agrB*, and *agrC* had no effect on growth rate, overall OD_{600nm} or solvent production (Steiner et al., 2012). When the individual genes are analysed, *agrC* appeared to have the greatest impact on growth, solvent production and sporulation. Interestingly, whilst the OD_{600nm} and pH growth profiles were identical, Δ*agrBDAC* and Δ*agrC* showed differences in solvent and acid production. Despite these differences, *agrBDAC* and *agrC*, likely along with *agrA*, are clearly responsible for encoding the most important functions of the agr system. Canonically, AgrC is the transmembrane protein that senses the extracellular, processed AgrD-derived AIP, and autophosphorylates in response (Queck et al., 2008). The autophosphorylated AgrC is then able to phosphorylate AgrA which is a transcription factor capable of binding DNA directly (Queck et al., 2008). In *Staphylococcus aureus*, AgrA also alters the expression of RNAIII which in turn affects the expression of a variety of genes (Novick and Geisinger, 2008).

The data for Δ*agrC* is, unfortunately, confounded by the presence of one biological repeat that showed wild type growth characteristics. A repeat of this screen is essential, however, on the balance of probability, it is likely that the wild type-like biological repeat was cross-contamination from another strain, likely Δ*agrB* given the growth profile and ability to grow in the presence of erythromycin. The solvent and acid production data suggests that the loss of AgrC causes a greater defect in the ability to switch from acid to solvent production than the loss of the whole operon. This, in turn, implies that the loss of the regulation of AgrA is more detrimental to the cellular functions than the loss of AgrA. It's plausible to speculate that an uncontrolled transcriptional regulator causes more transcription dysregulation than the absence of one entirely.

In a study into the RRNPP quorum sensing system *C. saccharoperbutylacetonicum*, it was found that the RRNPP proteins were essential to regulate the switch from acidogenesis to solventogenesis in small-scale bottle screens but not in pH controlled 500 mL fermenters (Feng et al., 2020). This suggested that RRNPP quorum sensing played a key role in initiating acid reabsorption but not in initiating solvent production *per se*. Potentially, the RRNPP regulates the expression of the acetoacetyl-CoA:acetate/butyrate:CoA transferase genes *ctfA* and *ctfB* (Cspa_c56890 and Cspa_c56900 respectively), rather than the production of proteins capable of producing solvents from the core metabolism. This would fit with data suggesting the *sol* operon containing butyrylaldehyde dehydrogenase, the transferase and acetoacetate decarboxylase is transcribed in a polycistronic manner and under the control of two promoters (Kosaka et al., 2007). The production of some butanol by Δ*agrBDAC* mutant strains suggests this may also be the case for the agr operon, although it's also possible that the agr operon is the quorum sensing system that regulates the expression of other gene encoding enzymes key to solvent production.

The Δ*agr'D'* strain did show a slight defect in growth and solvent production compared to the wild type. Given the important of *agrC* and the location of *agr'D'* immediately upstream of *agrAC*, it is likely that this phenotype might be explained by the disruption of *agrAC* expression through this deletion. Whilst Δ*agr'D'* was not ultimately considered a deletion of the signalling peptide, the Δ*agrB* strain would not have been able to produce a functioning AgrD peptide if the location posited in 5.3.2 is correct. Assuming there was no reversion of *agrB* to wild type state, this implies that the role of the agr system in *C. saccharoperbutylacetonicum* has shifted from strict quorum sensing to gene regulation independent of signalling peptide concentration. Reversion to the wild type is possible using this version of CLEAVE™ without plasmid curing, especially when selection pressures are strong. However, it seems unlikely that the defects of missing AgrB were greater than that seen for missing the whole agr system. It is possible to devise a scenario where missing one gene in a

system is more detrimental then losing the whole system (e.g., toxin-antitoxin systems), but this seems relatively unlikely given that mutants were obtained in the first place. On the balance of probability, Δ*agrB* is a true deletion and the role of AgrB is not essential to the functioning of the agr system.

No homologues of *agrB* or either of the putative *agrD* sequences were found elsewhere in *C. saccharoperbutylacetonicum* by megablast, discontiguous megablast or blastn. It is possible to conceive that the processing of the signalling peptide is not essential to the ability of the peptide to function, and that the peptide might find a different route out of the cell. Therefore, perhaps, AgrB alone is not important to the agr system. However, the complicating factor is the reasonable likelihood that *agrD* was also deleted in our mutant. Whilst the auto-inducing peptide derived from AgrD can be very small – as small as 5 amino acids in *Clostridium perfringens* (Li and McClane, 2020) – the likelihood that the 7 amino acid rump peptide that could be produced from the remains *agrB* after deletion is both expressed and functional is low. It seems likely that neither AgrB nor AgrD is necessary for the role of the agr system.

There are two possible explanations for the apparent lack of a role for AgrB and AgrD: 1) that there was, or is, redundancy in the agr system that entails crosstalk from other signalling factors or 2) that the role of the agr system lies in regulating gene expression in a manner somewhat decoupled from the concept of quorum sensing. The importance of AgrC shows that the AgrAC two-component system is important to the role of the agr system and it is reasonable to conclude that a cross-talking signal sensed by AgrC on the cell surface is the more likely option of the two, otherwise it is difficult to see under what circumstances AgrC is and is not able to phosphorylate AgrA. Further work is needed to clarify the apparent unimportance of AgrB and AgrD to the agr system in *C. saccharoperbutylacetonicum*.

The stark differences in the role of the agr system between *C. saccharoperbutylacetonicum* and *C. acetobutylicum* again highlight how different the underlying biology is between two species with ostensibly similar life cycles occupying the same niche. Recently, the appreciation of the wide-ranging differences between the multitude of solventogenic and acetogenic *Clostridia* is growing (Diallo et al., 2021). This is further highlighted in this chapter with the agr system clearly being utilised in a different way by *C. saccharoperbutylacetonicum*. Such adaptations and differences in how the agr components are deployed can even be seen between different strains of *C. perfringens* (Ma et al., 2015) and in the presence of multiple copies of parts of the system in *C. difficile* (Ahmed and Ballard, 2022).

Finally, presented here are new promoters that could be useful to the expand the options available to those working with *C. saccharoperbutylacetonicum* and *C. difficile*. Further characterisation, particularly their behaviour at late stages of growth, is still necessary to understand the underlying expression dynamics over all conditions. Cursory analysis of the promoter sequences by multiple sequence alignment showed a diverse range of sequences amongst these *Clostridial* promoters. However, three stand out patterns could be seen from these sequences. Firstly, common and well-characterised promoter motifs such as the IHF domain are harder to identify in *Clostridia*. A few promoters did appear to have part of the *E. coli* consensus sequence. It is likely that a different sequence is the recognition consensus for key DNA-bending proteins. Whilst these sequences are likely to be at least partially palindromic, they might be rendered difficult to spot due to findings suggesting that one half of the palindrome is more important than the other in the binding of transcription factors (Leuze et al., 2012).

Secondly, it seemed that BPROM is unable to identify the -10 regions accurately in these *Clostridial* genomes. It is unclear why this should be the case, but multiple sequence alignment was able to show that the canonical 5'-TATAA-3' sequence was well conserved. Though the bioinformatics behind BPROM are not this author's area of expertise, it's possible that BPROM gives a certain weight to the sequence composition of the -35 region which may be highly divergent in *Clostridia*. Thirdly, the ribosome binding site is clearly well-conserved across all species in this analysis. Further analysis of these promoters by those with greater experience in the binding of transcription factors and the identification of recognition sequences is required to elucidate further information from these, and other, *Clostridial* promoters.

Chapter VI – Discussion

6.1 Results and future perspectives

As each results chapter concerned diverse topics, this discussion will aim to summarise the key results of each chapter, discuss their utility and implications, and suggest productive avenues for future work. Chapter III was focused on establishing a method of transposon mutagenesis in *C. saccharoperbutylacetonicum* with which to create large transposon mutant libraries for insertion site mapping. Multiple facets of the pipeline had to be optimised and adapted for use in *C. saccharoperbutylacetonicum,* particularly in the mutagenesis and gDNA pre-sequencing processes. Ultimately, the work was unsuccessful, able only to generate a list of 677 genes without insertions and without the statistical power necessary to make calls on essentiality. This, in turn, impacted the downstream applications of both the method and the biological transposon libraries. However, several experiments conducted to enlighten key metrics or improve on the methods showed interesting and practical results. Firstly, lists of insertion sites was generated that can be used for practical purposes by the field, even if they lack statistical power. In addition, I was able to determine the *in vivo* copy number of the megaplasmid and pRPF215, demonstrate the utility of glass beads for spread cultures on solid agar and establish the recovery rate of cells following cryopreservation at -80°C. Finally, the work presented highlights clear paths for future work to follow.

Most published studies utilising high-throughput transposon insertion site mapping include some process whereby a generated transposon library is exposed to a condition of interest to identify conditionally essential mutants. Additionally, it is common to scour the data generated for interesting observations, such as to determine the importance of genes that appear homologous to key genes in other species (e.g., Chaudhuri et al., 2013; Dembek et al., 2015; Moule et al., 2014). In *C. saccharoperbutylacetonicum*, such a process would've been invaluable to determine the nature of apparent redundancy in genome where some genes (e.g., lactate dehydrogenase) appear to have multiple homologues.

The libraries can be used as a reference with which to validate reverse genetics studies. For example, Feng, et al., were unable to delete the *qssR4* (accession number Cspa_c29260) putative RRNPP response regulator using endogenous CRISPR-Cas. They suggested that the proximity (within 100 bp) of *qssR4* to putative fatty acid biosynthetic genes (accession numbers Cspa_c29240 and Cspa_c29230) meant that QssR4 likely regulated the expression of these key genes. Our insertion

site data suggests that both hypothetical fatty biosynthetic genes were likely essential with neither containing any insertions. The picture for the *qssR4* is more complex with Bio-Tradis assigning the 1302 bp gene as essential despite the presence of 13 insertions. The presence of insertions does suggest that this gene can be disrupted without compromising overall cell viability and therefore that difficulties encountered by Feng, et al. are due to the nature of the alterations at the *qssR4* locus rather than the essentiality of the gene itself. The proximity of these genes in the same orientation suggests that genetic features that regulate the expression of the essential fatty acid biosynthesis genes may be present in or influenced by the *qssR4* sequence. Alternatively, only one domain of the QssR4 protein maybe be essential to its function. The ability of Feng, et al. to knockdown the expression of *qssR4* somewhat supports the former idea as the knockdown would be less likely to influence the expression of the downstream genes.

The above example illustrates the potential utility of a dense, statistically significant transposon insertion site library to the study of *C. saccharoperbutylacetonicum* but also demonstrates that the data generated can still have some practical application. Within this informal reference list, we suggest that the list of ORFs without insertions generated can be of particular utility. Primarily, it could be used to assess the difficulty of interfering with a particular ORF or loci or to add to the likelihood of observed phenotypes being due to the importance of a given ORF. Whilst it would be inappropriate to draw significant conclusions based on these lists, they can still serve a purpose, especially in a community in which practical outcomes are somewhat more important than the underpinning phenomena.

It is clear that the field could derive significant benefit from a meaningful library, especially when combined with experiments to examine conditional essentiality in important growth contexts such as in fermentation and high butanol environments. To do so would require improvements to the current pRPF215 transposon delivery system. To this end, I suggest either replacing the Gram positive origin of replication with one that yields a lower overall plasmid copy number or randomly mutating the current origin of replication to yield the desired phenotype. I would also recommend the implementation of a systematic documentation of origin copy numbers. This could be done following bulk DNA measures such as those employed here or in Lee et al., 2006. However, this could present an opportunity to employ single colour droplet digital PCR as has been demonstrated previously (Jahn et al., 2016; Plotka et al., 2017). Such a method can calculate the concentration of a given sequence without the need for standard curves and the technical reliance of qPCR. It could also be interesting to determine the variability of copy number between cells utilising fluorescent reporters and cell sorting (Shao et al., 2021).

It would also be useful to continue work to create a more suitable inducible promoter system. Our initial attempts at creating one based on the P*xyl* (characterised in Chapter V) were unsuccessful, apparently being unable to induce significant transposition. This is at odds with the characterisation in Chapter V which suggested that the promoter showed high activity levels. It is plausible that P*xyl* is too strong for the purpose, causing high rates of transposition in each cell leading to guaranteed disruption of essential genes and cell death. Since it was observed that P*xyl* demonstrates a dose-dependent response, it may be that lower concentrations of xylose will result in the correct balance between induction of transposition at a meaningful rate and over induction.

Chapter IV explored sporulation and germination in *C. saccharoperbutylacetonicum* with a view to offering both general overviews and information practical to the manipulation of the species. Sporulation was investigated at a morphological and ultrastructural level with the aim of establishing the physical transformations of *C. saccharoperbutylacetonicum* cells during the sporulation process. In particular, the research was focused on establishing the differences between the sporulation process when conducted with solid or liquid media, as previous studies pointed to distinct phenotypes. We also observed the previously noted differences and clarified them through both the microscopy and heat resistance assays as well as offering a hint as to why these differences are seen. The results should prove helpful to the field as a reference point for the sporulation process, as well as to add to the pool of imaged *Clostridial* spores.

Interestingly, morphological differences were seen between both the cells and spores harvested from the two sporulation methods. This was particularly evident at a population level where liquid media sporulation showed nearly all cells in the population as appearing to be in the pre-sporulation *Clostridial* form. Meanwhile, solid media sporulation showed two distinct populations corresponding to *Clostridial* form sporulating cells and typical vegetative cells. Differences were also seen at an ultrastructural level, particularly when visualising released spores. Coat assembly appeared consistently less complete for spores from liquid. Heat resistance assays suggested that either there were fewer spores, or the spores were less heat-resistant (or both) for liquid media generated spores. Together, these results paint a picture of cells struggling to sporulate and/or reach full spore maturity when in liquid media. Naturally, this indicates that anyone wishing to utilise *C. saccharoperbutylacetonicum* spores for practical purposes should aim to generate those spores on solid media.

The heat resistance assays offered a glimpse into general persistence and hardiness of the *C. saccharoperbutylacetonicum* spores. Interestingly, the results obtained did not strictly match those

of work conducted at Biocleave (resistance to 70°C; personal communication, Sasha Atmadjaja) nor that reported in the literature (resistance to 80°C; Feng et al., 2020). It's possible these discrepancies are explained by differences in handling for the former and in spore preparation for the latter (7 days in mashed potato-based media). The differences could also be explained by the different sampling times in the latter case and it would be insightful to repeat the experiment over a longer time course. Compared to other solventogenic *Clostridia*, *C. saccharoperbutylacetonicum* spores are somewhat less resistant to heat with *C. acetobutylicum* and *C. beijerinckii* both able to survive heating to 85°C for 10 min (Jamroskovic et al., 2016). It would be interesting to understand spore resistance to other insults such as UV, prolonged extreme cold and extremes of pH. Little work has been done to identify these in any solventogenic *Clostridia*. From a practical viewpoint, it would be interesting to conduct temporal studies to understand how long spores remain viable in typical storage conditions (in H_2O at 4°C for this study) and whether there is a significant drop in viability after a certain time. Overall, the sporulation studies here provide the groundwork for further investigation into all areas of sporulation and a point of comparison for those working on *C. saccharoperbutylacetonicum* and other solventogenic *Clostridia*.

In contrast to the sporulation studies, analysis of germination focused primarily on attaining practical, directly usable results. To our knowledge, no previous studies have been undertaken into *C. saccharoperbutylacetonicum* germination, therefore our work constituted the first attempt to characterise germination in this species. Both an effective germinant in L-cysteine and germination inhibitor in potassium sorbate were also identified. Our method of testing germinants was flawed, relying on outgrowth as a proxy for germination rate, a limitation forced by technical challenges around spore purification. However, coupled with growth assays of vegetative cells, they proved surprisingly indicative. Apart from the being a step removed from the germination process, the design of these assays also meant that germination typically took at least 17 h to be measurable meaning that the rate of germination could not be well-characterised.

Several lines of investigation present themselves for further studies in germination. Firstly, if spore purification protocols could be improved then spores could be purified in significant quantities. This in turn, would be invaluable to study the rate of germination. The germination rate could have practical applications as a method of precisely controlling when a culture becomes vegetative, something that might be particularly valuable in co-cultures. One interesting line of experimentation could be to co-culture spores and vegetative cells and induce germination as the vegetative cells start to decline in productivity, potentially extending the length of productive fermentation, albeit with many caveats and unknowns.

The discovery of germination inhibitors also opens similar possibilities in terms of the precise control of germination. If spores were in regular use, a cheap and easy to use inhibitor such as potassium sorbate could prove a useful additive in spore storage. Equally, it would be interesting to see if the addition of potassium sorbate could increase the sporulation rate. It is currently unknown if spores generated early in the sporulation process characterised in Chapter IV are liable to germinate when left over 7 days. Certainly, fewer spores appeared to be recovered after this time period based on the extensive microscopic evidence. As previously discussed, this could have been due to spore adhesion, either to plasticware or the agar, leading to spores that are difficult to harvest at later time points. Spore adhesion has been demonstrated for *C. difficile* spores in the context of adhesion to human Caco-2 cells (Paredes-Sabja and Sarker, 2012). Something hinting at adhesion was seen for spores generated on solid media as they tended to group together when viewed under phase contrast microscopy. However, it may just be that spores simply germinated over time. The conditions do not appear to be inhibitory to growth and likely contain L-cysteine with TYIR being the base media for sporulation. If this is the case, the addition of potassium sorbate could increase spore yields by blocking any germination.

Chapter V examined the role of the agr quorum sensing system in *C. saccharoperbutylacetonicum* and also introduced novel promoters for use in both *C. saccharoperbutylacetonicum* and *C. difficile*. This chapter was hampered by time constraints which meant that the results are incomplete and lacking rigour. Nonetheless, they should prove valuable to both academic and industrial interests in the species. A role in solventogenesis and sporulation for the agr system in *C. saccharoperbutylacetonicum* was identified for the first time.

This role of the agr system differed significantly from that seen for *C. acetobutylicum* where the agr system was only involved in granulose formation and sporulation (Steiner et al., 2012). Interestingly, our data suggested that the quorum signalling derived from the agr system identified was not important to the functioning of the system. This suggested there was either crosstalk with other quorum signals or that the role of the agr system in *C. saccharoperbutylacetonicum* is involved with general gene expression and responds in an indirect manner to the changing environment. The former is somewhat supported by a throw-away statement in a previous phylogenetic study which mentions the presence of four putative agr systems in *C. saccharoperbutylacetonicum* (Poehlein et al., 2017). However, no further details are given, and I was unable to find any homologues for the agr genes based on simple BLAST searches.

C. difficile harbours three loci containing a total of nine agr genes, most of which do not have high homology to the *S. aureus* model genes (Ahmed and Ballard, 2022). One loci contains the full 4 gene system whilst the other two contain only *agrBD* and *agrCBD*. Strikingly, the presence of a separate *agrBD*, which has been shown to play a role in virulence and pathogenesis (Darkoh et al., 2016), is notable for the results presented here. Indeed, this is not the only instance of isolated *agrBD* loci as the same has been documented for *C. botulinum* (Cooksley et al., 2010). It's possible, therefore, that something similar occurs in *C. saccharoperbutylacetonicum*, although BLAST searches of the genome utilising the *agrBD* sequence from *C. difficile* R20291 did not yield any significant results. Even if other agr systems exist and provide crosstalk, it is clear that the system investigated here plays a significant role in regulating the transcriptional response given that the deletion of *agrC* and the whole operon largely inhibited solventogenesis and abolished sporulation.

The role of the agr system in solventogenesis and sporulation demonstrates its relevance and suggests that further investigation of this system could result in the development of methods to precisely control the onset of solventogenesis. Better understanding of the operon could help identify promoter sequences that are precisely regulated by AgrA (or downstream regulators) which could offer routes to precise temporal control of gene expression without the need for the addition of exogenous compounds. To achieve these, a number of further experiments present themselves. It is necessary to characterise these agr deletion strains, with the addition an Δ*agrA* strain, in more depth. Repeats of the small-scale bottle growth assays with complementation would be essential. Furthermore, testing the mutants at the 500+ mL scale with precise pH control would help further characterise the strain phenotypes, as it did for Feng, et al. Finally, examining global transcription levels in these strains could indicate which genes are under the regulatory control of the agr system.

Chapter V also saw the characterisation of multiple novel promoters for use in both *C. saccharoperbutylacetonicum* and *C. difficile*. These should provide a direct benefits field, particularly when looking for very high or very low expression and when deciding what concentration of inducer to use for the two inducible promoters studied. Although it would be unwise to draw any biological conclusions based on these promoter assays, it was notable that promoters randomly chosen from the *C. saccharoperbutylacetonicum* megaplasmid do show reasonable and consistent activity. Naturally, this suggests that genes on the megaplasmid are expressed under normal conditions and thus may have a role in the cell. Moving forward, it would be ideal if these promoters could be characterised further, particularly in *C. saccharoperbutylacetonicum*, to account for expression during stationary phase.

6.2 Concluding remarks

This thesis took a broad approach to broad questions regarding the biological of *C. saccharoperbutylacetonicum*. The species, along with other solventogenic *Clostridia*, is developing into an important industrial powerhouse for a broad range of applications. It is for this reason that the overarching aim throughout was to provide both meaningful biological knowledge about and practical tools and methods for *C. saccharoperbutylacetonicum*. It is the author's hope that the work described herein provides both that knowledge and useful hooks for future studies.

Bibliography

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. Nat. Commun. 2018 91 *9*, 1–13.

Ahmed, U.K.B., and Ballard, J.D. (2022). Autoinducing peptide-based quorum signaling systems in *Clostridioides difficile*. Curr. Opin. Microbiol. *65*, 81–86.

Ahmed, U.K.B., Shadid, T.M., Larabee, J.L., and Ballard, J.D. (2020). Combined and distinct roles of agr proteins in *Clostridioides difficile* 630 sporulation, motility, and toxin production. MBio *11*, 1–17.

Akerley, B.J., Rubin, E.J., Camilli, A., Lampe, D.J., Robertson, H.M., and Mekalanos, J.J. (1998). Systematic identification of essential genes by *in vitro* mariner mutagenesis. Proc. Natl. Acad. Sci. U. S. A. *95*, 8927–8932.

Al-Hinai, M.A., Jones, S.W., and Papoutsakis, E.T. (2014). σK of *Clostridium acetobutylicum* is the first known sporulation-specific sigma factor with two developmentally separated roles, one early and one late in sporulation. J. Bacteriol. *196*, 287.

Al-Hinai, M.A., Jones, S.W., and Papoutsakis, E.T. (2015). The *Clostridium* sporulation programs: diversity and preservation of endospore differentiation. Microbiol. Mol. Biol. Rev. *79*, 19–37.

Al-Shorgani, N.K.N., Kalil, M.S., and Yusoff, W.M.W. (2011). The effect of different carbon sources on biobutanol production using *Clostridium saccharoperbutylacetonicum* N1-4. Biotechnology *10*, 280–285.

Al-Shorgani, N.K.N., Ali, E., Kalil, M.S., and Yusoff, W.M.W. (2012a). Bioconversion of butyric acid to butanol by *Clostridium saccharoperbutylacetonicum* N1-4 (ATCC 13564) in a limited nutrient medium. Bioenergy Res. *5*, 287–293.

Al-Shorgani, N.K.N., Kalil, M.S., and Yusoff, W.M.W. (2012b). Biobutanol production from rice bran and de-oiled rice bran by *Clostridium saccharoperbutylacetonicum* N1-4. Bioprocess Biosyst. Eng. *35*, 817–826.

Al-Shorgani, N.K.N., Tibin, E.M., Ali, E., Hamid, A.A., Yusoff, W.M.W., and Kalil, M.S. (2014). Biohydrogen production from agroindustrial wastes via *Clostridium saccharoperbutylacetonicum* N1-4 (ATCC 13564). Clean Technol. Environ. Policy *16*, 11–21.

Alsaker, K. V., and Papoutsakis, E.T. (2005). Transcriptional program of early sporulation and stationary-phase events in *Clostridium acetobutylicum*. J. Bacteriol. *187*, 7103–7118.

American Type Culture Collection (2021). Bacteriology culture guide. 28.

Anjuwon-Foster, B.R., and Tamayo, R. (2017). A genetic switch controls the production of flagella and toxins in *Clostridium difficile*. PLOS Genet. *13*, e1006701.

Atmadjaja, A.N., Holby, V., Harding, A.J., Krabben, P., Smith, H.K., and Jenkinson, E.R. (2019). CRISPR-Cas, a highly effective tool for genome editing in *Clostridium saccharoperbutylacetonicum* N1-4(HMT). FEMS Microbiol. Lett. *366*.

Atrih, A., Bacher, G., Allmaier, G., Williamson, M.P., and Foster, S.J. (1999). Analysis of peptidoglycan structure from vegetative cells of *Bacillus subtilis* 168 and role of PBP 5 in peptidoglycan maturation. J. Bacteriol. *181*, 3956–3966.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Bao Diep, D., Sigve Havarstein, L., Nissen-meyer, J., Nes, I.F., Nissen-Meyer, J., Larsen, A.G., Sletten, K., Daeschel, M., Nes, I.F., and Gen Microbiol, J. (1994). The gene encoding plantaricin A, a bacteriocin from *Lactobacillus plantarum* C11, is located on the same transcription unit as an agr-like regulatory system. Appl. Environ. Microbiol. *60*, 160.

Barquist, L., Langridge, G.C., Turner, D.J., Phan, M.D., Turner, A.K., Bateman, A., Parkhill, J., Wain, J., and Gardner, P.P. (2013). A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. Nucleic Acids Res. *41*, 4549–4564.

Barquist, L., Mayho, M., Cummins, C., Cain, A.K., Boinett, C.J., Page, A.J., Langridge, G.C., Quail, M.A., Keane, J.A., and Parkhill, J. (2016). The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. Bioinformatics *32*, 1109–1111.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-.). *315*, 1709–1712.

Bassler, B.L., and Losick, R. (2006). Bacterially Speaking. Cell *125*, 237–246.

Basu, A., Xin, F., Lim, T.K., Lin, Q., Yang, K.L., and He, J. (2017). Quantitative proteome profiles help reveal efficient xylose utilization mechanisms in solventogenic *Clostridium* sp. strain BOH3. Biotechnol. Bioeng. *114*, 1959–1969.

Bennett, S. (2004). Solexa Ltd. Pharmacogenomics *5*, 433–438.

Berezina, O. V., Zakharova, N. V., Yarotsky, C. V., and Zverlov, V. V. (2012). Microbial producers of butanol. Appl. Biochem. Microbiol. 2012 487 *48*, 625–638.

Berg, D.E., Schmandt, M.A., and Lowe, J.B. (1983). Specificity of transposon Tn5 insertion. Genetics *105*, 813–828.

Beskrovnaya, P., Sexton, D.L., Golmohammadzadeh, M., Hashimi, A., and Tocheva, E.I. (2021). Structural, metabolic and evolutionary comparison of bacterial endospore and exospore formation. Front. Microbiol. *12*.

Bester, B.H., and Claassens, J.W. (1970). The effect of various factors on the spore germination and growth of *Clostridium butyricum* and *Clostridium tyrobutyricum* various carbohydrates and buffers, age, and preheating of spores. Pbytopbylactica *2*, 237–242.

Bhattacharjee, D., McAllister, K.N., and Sorg, J.A. (2016). Germinants and their receptors in *Clostridia*. J. Bacteriol. *198*, 2767–2775.

Bi, C., Jones, S.W., Hess, D.R., Tracy, M.B.P., and Papoutsakis, E.T. (2011). SpoIIE is necessary for asymmetric division, sporulation, and expression of σF, σE, and σG but does not control solvent production in *Clostridium acetobutylicum* ATCC 824. J. Bacteriol. *193*, 5130.

Blocher, J.C., and Busta, F.F. (1985). Multiple modes of inhibition of spore germination and outgrowth by reduced pH and sorbate. J. Appl. Bacteriol. *59*, 469–478.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491–500.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. 2018 191 *19*, 1–12.

Bradley, B., Green, E., and Heeg, D. (2021). Compositions and uses thereof for treating inflammatory diseases and probiotic compositions (UK Intellectual Property Office).

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E. V., and Van Der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (80-.). *321*, 960–964.

Browning, D.F., and Busby, S.J.W. (2004). The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2004 21 *2*, 57–65.

Bruder, M.R., Pyne, M.E., Moo-Young, M., Chung, D.A., and Chou, C.P. (2016). Extending CRISPR-Cas9 technology from genome editing to transcriptional engineering in the genus *Clostridium*. Appl. Environ. Microbiol. *82*, 6109–6119.

Brunt, J., Plowman, J., Gaskin, D.J.H., Itchner, M., Carter, A.T., and Peck, M.W. (2014). Functional characterisation of germinant receptors in *Clostridium botulinum* and *Clostridium sporogenes* presents novel insights into spore germination systems. PLOS Pathog. *10*, e1004382.

Cain, A.K., Barquist, L., Goodman, A.L., Paulsen, I.T., Parkhill, J., and van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. Nat. Rev. Genet. 2020 219 *21*, 526–540.

Carter, G.P., Purdy, D., Williams, P., and Minton, N.P. (2005). Quorum sensing in *Clostridium difficile*: analysis of a luxS-type signalling system. J. Med. Microbiol. *54*, 119–127.

Cartman, S.T., Kelly, M.L., Heeg, D., Heap, J.T., and Minton, N.P. (2012). Precise manipulation of the *Clostridium difficile* chromosome reveals a lack of association between the tcdC genotype and toxin production. Appl. Environ. Microbiol. *78*, 4683.

Chambers, S.P., Prior, S.E., Barstow, D.A., and Minton, N.P. (1988). The pMTL nic− cloning vectors. I. Improved pUC polylinker regions to facilitate the use of sonicated DNA for nucleotide sequencing. Gene *68*, 139–149.

Chaudhuri, R.R., Morgan, E., Peters, S.E., Pleasance, S.J., Hudson, D.L., Davies, H.M., Wang, J., van Diemen, P.M., Buckley, A.M., Bowen, A.J., et al. (2013). Comprehensive assignment of roles for *Salmonella Typhimurium* genes in intestinal colonization of food-producing animals. PLoS Genet. *9*, e1003456.

Cheung, H.Y., Vitkovic, L., and Brown, M.R.W. (1982). Dependence of *Bacillus stearothermophilus* spore germination on nutrient depletion and manganese. J. Gen. Microbiol. *128*, 2403–2409.

Clegg, M.T., and Durbin, M.L. (2003). Tracing floral adaptations from ecology to molecules. Nat. Rev. Genet. 2003 43 *4*, 206–215.

Cook, L.C., and Federle, M.J. (2014). Peptide pheromone signaling in *Streptococcus* and *Enterococcus*. FEMS Microbiol. Rev. *38*, 473–492.

Cooksley, C.M., Davis, I.J., Winzer, K., Chan, W.C., Peck, M.W., and Minton, N.P. (2010). Regulation of neurotoxin production and sporulation by a putative AgrBD signaling system in proteolytic *Clostridium botulinum*. Appl. Environ. Microbiol. *76*, 4448 LP – 4460.

Corrigan, R.M., and Foster, T.J. (2009). An improved tetracycline-inducible expression vector for *Staphylococcus aureus*. Plasmid *61*, 126–129.

Croux, C., Nguyen, N.P.T., Lee, J., Raynaud, C., Saint-Prix, F., Gonzalez-Pajuelo, M., Meynial-Salles, I., and Soucaille, P. (2016). Construction of a restriction-less, marker-less mutant useful for functional genomic and metabolic engineering of the biofuel producer *Clostridium acetobutylicum*. Biotechnol. Biofuels *9*, 1–13.

Cui, L., and Bikard, D. (2016). Consequences of Cas9 cleavage in the chromosome of *Escherichia coli*. Nucleic Acids Res. *44*, 4243–4251.

Darkoh, C., Odo, C., and Dupont, H.L. (2016). Accessory gene regulator-1 locus is essential for virulence and pathogenesis of *Clostridium difficile*. MBio *7*.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. *97*, 6640–6645.

Davis, T. (1999). Regulation of botulinum toxin complex formation in *Clostridium botulinum* Type a NCTC 2916.

Deakin, L.J., Clare, S., Fagan, R.P., Dawson, L.F., Pickard, D.J., West, M.R., Wren, B.W., Fairweather, N.F., Dougan, G., and Lawley, T.D. (2012). The *Clostridium difficile* spo0A gene is a persistence and transmission factor. Infect. Immun. *80*, 2704–2711.

Declerck, N., Bouillaut, L., Chaix, D., Rugani, N., Slamti, L., Hoh, F., Lereclus, D., and Arold, S.T. (2007). Structure of PlcR: insights into virulence regulation and evolution of quorum sensing in Gram-positive bacteria. Proc. Natl. Acad. Sci. *104*, 18490–18495.

Dembek, M., Barquist, L., Boinett, C.J., Cain, A.K., Mayho, M., Lawley, T.D., Fairweather, N.F., and Fagan, R.P. (2015). High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. MBio *6*.

Diallo, M., Kengen, S.W.M., and López-Contreras, A.M. (2021). Sporulation in solventogenic and acetogenic *clostridia*. Appl. Microbiol. Biotechnol. 2021 1059 *105*, 3533–3557.

Dorman, C.J., and Deighan, P. (2003). Regulation of gene expression by histone-like proteins in bacteria. Curr. Opin. Genet. Dev. *13*, 179–184.

Dorman, M.J., Feltwell, T., Goulding, D.A., Parkhill, J., and Short, F.L. (2018). The capsule regulatory network of *Klebsiella pneumoniae* defined by density-TraDISort. MBio *9*.

Drepper, T., Eggert, T., Circolone, F., Heck, A., Krauß, U., Guterl, J.K., Wendorff, M., Losi, A., Gärtner, W., and Jaeger, K.E. (2007). Reporter proteins for *in vivo* fluorescence without oxygen. Nat. Biotechnol. *25*, 443–445.

Driks, A. (2003). The dynamic spore. Proc. Natl. Acad. Sci. *100*, 3007–3009.

Dupuy, B., and Sonenshein, A.L. (1998). Regulated transcription of *Clostridium difficile* toxin genes. Mol. Microbiol. *27*, 107–120.

Dürre, P. (2005). Handbook on *Clostridia* (CRC Press).

Dürre, P. (2014). Physiology and sporulation in *Clostridium*. Microbiol. Spectr. *2*.

EASAC (2012). The current status of biofuels in the European Union, their environmental impacts and future prospects.

Ellis, J.T., Hengge, N.N., Sims, R.C., and Miller, C.D. (2012). Acetone, butanol, and ethanol production from wastewater algae. Bioresour. Technol. *111*, 491–495.

Emerson, J.E., Reynolds, C.B., Fagan, R.P., Shaw, H.A., Goulding, D., and Fairweather, N.F. (2009). A novel genetic switch controls phase variable expression of CwpV, a *Clostridium difficile* cell wall protein. Mol. Microbiol. *74*, 541–556.

Engebrecht, J.A., and Silverman, M. (1984). Identification of genes and gene products necessary for bacterial bioluminescence. Proc. Natl. Acad. Sci. *81*, 4154–4158.

Estrem, S.T., Gaal, T., Ross, W., and Gourse, R.L. (1998). Identification of an UP element consensus sequence for bacterial promoters. Proc. Natl. Acad. Sci. *95*, 9761–9766.

European Parliament and Council (2018). Directive (EU) 2015/1513 of the European Parliament and of the Council amending Directive 98/70/EC relating to the quality of petrol and diesel fuels and amending Directive 2009/28/EC on the promotion of the use of energy from renewable sources (European Union).

Ezeji, T.C., Qureshi, N., and Blaschek, H.P. (2007). Bioproduction of butanol from biomass: from genes to bioreactors. Curr. Opin. Biotechnol. *18*, 220–227.

Fagan, R.P., and Fairweather, N.F. (2011). *Clostridium difficile* has two parallel and essential sec secretion systems. J. Biol. Chem. *286*, 27483–27493.

Farmanbordar, S., Amiri, H., and Karimi, K. (2018). Simultaneous organosolv pretreatment and detoxification of municipal solid waste for efficient biobutanol production. Bioresour. Technol. *270*, 236–244.

Feng, J., Zong, W., Wang, P., Zhang, Z.-T., Gu, Y., Dougherty, M., Borovok, I., and Wang, Y. (2020). RRNPP-type quorum-sensing systems regulate solvent formation, sporulation and cell motility in *Clostridium saccharoperbutylacetonicum*. Biotechnol. Biofuels *13*.

Fenton, A.K., Mortaji, L. El, Lau, D.T.C., Rudner, D.Z., and Bernhardt, T.G. (2016). CozE is a member of the MreCD complex that directs cell elongation in *Streptococcus pneumoniae*. Nat. Microbiol. 2016 23 *2*, 1–10.

Feschotte, C., and Pritham, E.J. (2007). DNA Transposons and the evolution of eukaryotic genomes. Http://Dx.Doi.Org/10.1146/Annurev.Genet.40.110405.090448 *41*, 331–368.

Fimlaid, K.A., Bond, J.P., Schutz, K.C., Putnam, E.E., Leung, J.M., Lawley, T.D., and Shen, A. (2013). Global analysis of the sporulation pathway of *Clostridium difficile*. PLOS Genet. *9*, e1003660.

Fontaine, L., Wahl, A., Fléchard, M., Mignolet, J., and Hols, P. (2015). Regulation of competence for natural transformation in *Streptococci*. Infect. Genet. Evol. *33*, 343–360.

Francis, M.B., Allen, C.A., and Sorg, J.A. (2015). Spore cortex hydrolysis precedes dipicolinic acid release during *Clostridium difficile* spore germination. J. Bacteriol. *197*, 2276–2283.

Gautheret, D., and Lambert, A. (2001). Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J. Mol. Biol. *313*, 1003–1011.

Gawronski, J.D., Wong, S.M.S., Giannoukos, G., Ward, D. V., and Akerley, B.J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. Proc. Natl. Acad. Sci. *106*, 16422–16427.

Ge, S.X., Jung, D., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics *36*, 2628–2629.

Girbal, L., Mortier-Barrière, I., Raynaud, F., Rouanet, C., Croux, C., and Soucaille, P. (2003). Development of a sensitive gene expression reporter system and an inducible promoter-repressor system for *Clostridium acetobutylicum*. Appl. Environ. Microbiol. *69*, 4985–4988.

Goodall, E.C.A., Robinson, A., Johnston, I.G., Jabbari, S., Turner, K.A., Cunningham, A.F., Lund, P.A., Cole, J.A., and Henderson, I.R. (2018). The essential genome of *Escherichia coli* K-12. MBio *9*, e02096-17.

Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R., and Gordon, J.I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe *6*, 279–289.

Goosen, N., and van de Putte, P. (1995). The regulation of transcription initiation by integration host factor. Mol. Microbiol. *16*, 1–7.

Grainge, I. (2008). Sporulation: SpoIIIE is the key to cell differentiation. Curr. Biol. *18*, R871–R872.

Gray, B., Hall, P., and Gresham, H. (2013). Targeting agr- and agr-Like quorum sensing systems for development of common therapeutics to treat multiple Gram-positive bacterial infections. Sensors 2013, Vol. 13, Pages 5130-5166 *13*, 5130–5166.

Greenblatt, I.M., and Brink, R.A. (1963). Transpositions of modulator in maize into divided and undivided chromosome segments. Nat. 1963 1974865 *197*, 412–413.

Grenha, R., Slamti, L., Nicaise, M., Refes, Y., Lereclus, D., and Nessler, S. (2013). Structural basis for the activation mechanism of the PlcR virulence regulator by the quorum-sensing signal peptide PapR. Proc. Natl. Acad. Sci. U. S. A. *110*, 1047–1052.

Grosse-Honebrink, A., Schwarz, K.M., Wang, H., Minton, N.P., and Zhang, Y. (2017). Improving gene transfer in *Clostridium pasteurianum* through the isolation of rare hypertransformable variants. Anaerobe *48*, 203–205.

Gruber, T.M., and Gross, C.A. (2003). Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. Http://Dx.Doi.Org/10.1146/Annurev.Micro.57.030502.090913 *57*, 441–466.

Gu, Y., Feng, J., Zhang, Z.T., Wang, S., Guo, L., Wang, Y., and Wang, Y. (2019). Curing the endogenous megaplasmid in *Clostridium saccharoperbutylacetonicum* N1-4 (HMT) using CRISPR-Cas9 and preliminary investigation of the role of the plasmid for the strain metabolism. Fuel *236*, 1559–1566.

Gyulev, I.S., Willson, B.J., Hennessy, R.C., Krabben, P., Jenkinson, E.R., and Thomas, G.H. (2018). Part by part: synthetic biology parts used in solventogenic *Clostridia*. ACS Synth. Biol. *7*, 311–327.

Hales, L.M., Gumport, R.I., and Gardner, J.F. (1994). Determining the DNA sequence elements required for binding integration host factor to two different target sites. J. Bacteriol. *176*, 2999.

Hamer, L., DeZwaan, T.M., Montenegro-Chamorro, M.V., Frank, S.A., and Hamer, J.E. (2001). Recent advances in large-scale transposon mutagenesis. Curr. Opin. Chem. Biol. *5*, 67–73.

Hare, R.S., Walker, S.S., Dorman, T.E., Greene, J.R., Guzman, L.M., Kenney, T.J., Sulavik, M.C., Baradaran, K., Houseweart, C., Yu, H., et al. (2001). Genetic footprinting in bacteria. J. Bacteriol. *183*, 1694–1706.

Harris, L., van Zyl, L.J., Kirby-McCullough, B.M., Damelin, L.H., Tiemessen, C.T., and Trindade, M. (2018). Identification and sequence analysis of two novel cryptic plasmids isolated from the vaginal mucosa of South African women. Plasmid *98*, 56–62.

Harry, K.H., Zhou, R., Kroos, L., and Melville, S.B. (2009). Sporulation and enterotoxin (CPE) synthesis are controlled by the sporulation-specific sigma factors SigE and SigK in *Clostridium perfringens*. J. Bacteriol. *191*, 2728–2742.

Hassan, K.A., Cain, A.K., Huang, T., Liu, Q., Elbourne, L.D.H., Boinett, C.J., Brzoska, A.J., Li, L., Ostrowski, M., Nhu, N.T.K., et al. (2016). Fluorescence-based flow sorting in parallel with transposon insertion site sequencing sdentifies multidrug efflux systems in *Acinetobacter baumannii*. MBio *7*, e01200-16.

Hasunuma, K. (2009). Genetics and molecular biology (EOLSS).

Heap, J.T., Pennington, O.J., Cartman, S.T., Carter, G.P., and Minton, N.P. (2007). The ClosTron: a universal gene knock-out system for the genus *Clostridium*. J. Microbiol. Methods *70*, 452–464.

Heap, J.T., Pennington, O.J., Cartman, S.T., and Minton, N.P. (2009). A modular system for *Clostridium* shuttle plasmids. J. Microbiol. Methods *78*, 79–85.

Heap, J.T., Kuehne, S.A., Ehsaan, M., Cartman, S.T., Cooksley, C.M., Scott, J.C., and Minton, N.P. (2010). The ClosTron: mutagenesis in *Clostridium* refined and streamlined. J. Microbiol. Methods *80*, 49–55.

Heap, J.T., Ehsaan, M., Cooksley, C.M., Ng, Y.-K., Cartman, S.T., Winzer, K., and Minton, N.P. (2012). Integration of DNA into bacterial chromosomes from plasmids without a counter-selection marker. Nucleic Acids Res. *40*, e59–e59.

Helmann, J.D., and Chamberlin, M.J. (2003). Structure and function of bacterial sigma factors. Https://Doi.Org/10.1146/Annurev.Bi.57.070188.004203 *57*, 839–872.

Henriques, A.O., and Moran, C.P. (2000). Structure and assembly of the bacterial endospore coat. Methods *20*, 95–110.

Henriques, A.O., and Moran, C.P. (2007). Structure, assembly, and function of the spore surface layers. Http://Dx.Doi.Org/10.1146/Annurev.Micro.61.080706.093224 *61*, 555–588.

Hensel, M., Shea, J.E., Gleeson, C., Jones, M.D., Dalton, E., and Holden, D.W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. Science *269*, 400–403.

Herman, N.A., Kim, S.J., Li, J.S., Cai, W., Koshino, H., and Zhang, W. (2017). The industrial anaerobe *Clostridium acetobutylicum* uses polyketides to regulate cellular differentiation. Nat. Commun. *8*.

Hitzman, D.O., Halvorson, H.O., and Ukita, T. (1957). Requirements for production and germination of spores of anaerobic bacteria. J. Bacteriol. *74*, 1–7.

Hongo, M. (1960). Process for producing butanol by fermentation (United States Patent Office).

Hongo, M., Murata, A., Kono, K., and Kato, F. (1968). Lysogeny and bacteriocinogeny in strains of *Clostridium* species. Agric. Biol. Chem. *32*, 27–33.

Huang, H., Chai, C., Li, N., Rowe, P., Minton, N.P., Yang, S., Jiang, W., and Gu, Y. (2016). CRISPR/Cas9-based efficient genome editing in *Clostridium ljungdahlii*, an autotrophic gas-fermenting bacterium. ACS Synth. Biol. *5*, 1355–1361.

Hunt, M., Silva, N. De, Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. *16*, 1–10.

Jabbari, S., Steiner, E., Heap, J.T., Winzer, K., Minton, N.P., and King, J.R. (2013). The putative influence of the agr operon upon survival mechanisms used by *Clostridium acetobutylicum*. Math. Biosci. *243*, 223–239.

Jahn, M., Vorpahl, C., Hübschmann, T., Harms, H., and Müller, S. (2016). Copy number variability of expression plasmids determined by cell sorting and droplet digital PCR. Microb. Cell Fact. *15*, 1–12.

Jamroskovic, J., Chromikova, Z., List, C., Bartova, B., Barak, I., and Bernier-Latmani, R. (2016). Variability in DPA and calcium content in the spores of *Clostridium* species. Front. Microbiol. *7*, 1791.

Jana, B., Cain, A.K., Doerrler, W.T., Boinett, C.J., Fookes, M.C., Parkhill, J., and Guardabassi, L. (2017). The secondary resistome of multidrug-resistant *Klebsiella pneumoniae*. Sci. Rep. *7*, 42483.

Jenkinson, E., and Krabben, P. (2015). Targeted mutations.

Jenkinson, E., Harding, A.J., Davies, T.E., and Atmadjaja, A.N. (2019). Processes involving *Clostridium saccharoperbutylacetonicum* (World Intellectural Property Organization).

Ji, G., Beavis, R., and Novick, R.P. (1997). Bacterial interference caused by autoinducing peptide variants. Science (80-.). *276*, 2027–2030.

Jiménez-Bonilla, P., Feng, J., Wang, S., Zhang, J., Wang, Y., Blersch, D., De-Bashan, L.E., Gaillard, P., Guo, L., and Wanga, Y. (2021). Identification and investigation of autolysin genes in *Clostridium saccharoperbutylacetonicum* strain N1-4 for enhanced biobutanol production. Appl. Environ. Microbiol. *87*, 1–2.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-.). *337*, 816–821.

Jones, D.T. (2001). Applied acetone–butanol fermentation. In *Clostridia*, (John Wiley & Sons, Ltd), pp. 125–168.

Jones, D.T., and Woods, D.R. (1986). Acetone-butanol fermentation revisited. Microbiol. Rev. *50*, 484–524.

Jones, S.W., Tracy, B.P., Gaida, S.M., and Papoutsakis, E.T. (2011). Inactivation of σF in *Clostridium acetobutylicum* ATCC 824 blocks sporulation prior to asymmetric division and abolishes σEand σG protein expression but does not block solvent formation. J. Bacteriol. *193*, 2429–2440.

Joseph, R.C., Kim, N.M., and Sandoval, N.R. (2018). Recent developments of the synthetic biology toolkit for *Clostridium*. Front. Microbiol. *9*, 154.

Kakkanat, A., Phan, M.-D., Lo, A.W., Beatson, S.A., and Schembri, M.A. (2017). Novel genes associated with enhanced motility of *Escherichia coli* ST131. PLoS One *12*, e0176290.

Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J., and Lambowitz, A.M. (2001). Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. Nat. Biotechnol. 2001 1912 *19*, 1162–1167.

Kashket, E.R., and Cao, Z.Y. (1993). Isolation of a degeneration-resistant mutant of *Clostridium acetobutylicum* NCIMB 8052. Appl. Environ. Microbiol. *59*, 4198–4202.

Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988). Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. Nature *332*, 164–166.

Keis, S., Bennett, C.F., Ward, V.K., and Jones, D.T. (1995). Taxonomy and phylogeny of industrial solvent-producing *Clostridia*. Int. J. Syst. Bacteriol. *45*, 693–705.

Keis, S., Shaheen, R., and Jones, D.T. (2001). Emended descriptions of *Clostridium acetobutylicum* and *Clostridium beijerinckii*, and descriptions of *Clostridium saccharoperbutylacetonicum* sp. nov. and *Clostridium saccharobutylicum* sp. nov. Int. J. Syst. Evol. Microbiol. *51*, 2095–2103.

Kirk, J.A., and Fagan, R.P. (2016). Heat shock increases conjugation efficiency in *Clostridium difficile*. Anaerobe *42*, 1–5.

Kleckner, N., Roth, J., and Botstein, D. (1977). Genetic engineering *in vivo* using translocatable drug-resistance elements: new methods in bacterial genetics. J. Mol. Biol. *116*, 125–159.

Kleerebezem, M., Quadri, L.E.N., Kuipers, O.P., and De Vos, W.M. (1997). Quorum sensing by peptide pheromones and two-component signal-transduction systems in Gram-positive bacteria. Mol. Microbiol. *24*, 895–904.

Kosaka, T., Nakayama, S., Nakaya, K., Yoshino, S., and Furukawa, K. (2007). Characterization of the sol operon in butanol-hyperproducing *Clostridium saccharoperbutylacetonicum* strain N1-4 and its degeneration mechanism. Biosci. Biotechnol. Biochem. *71*, 58–68.

Kotte, A.-K., Severn, O., Bean, Z., Schwarz, K., Minton, N.P., and Winzer, K. (2020). RRNPP-type quorum sensing affects solvent formation and sporulation in *Clostridium acetobutylicum*. Microbiology mic000916.

Kubyshkin, V., and Budisa, N. (2019). Anticipating alien cells with alternative genetic codes: away from the alanine world! Curr. Opin. Biotechnol. *60*, 242–249.

Kuit, W., Minton, N.P., López-Contreras, A.M., and Eggink, G. (2012). Disruption of the acetate kinase (ack) gene of *Clostridium acetobutylicum* results in delayed acetate production. Appl. Microbiol. Biotechnol. *94*, 729–741.

Lampe, D.J., Akerley, B.J., Rubin, E.J., Mekalanos, J.J., and Robertson, H.M. (1999). Hyperactive transposase mutants of the Himar1 mariner transposon. Proc. Natl. Acad. Sci. U. S. A. *96*, 11428–11433.

Langridge, G.C., Phan, M.-D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., et al. (2009). Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. Genome Res. *19*, 2308–2316.

Lederberg, J. (1946). Studies in bacterial genetics. J. Bacteriol. *52*, 503.

Lederberg, J., and Tatum, E.L. (1946). Gene recombination in *Escherichia Coli*. Nat. 1946 1584016 *158*, 558–558.

Lee, A.S.Y., and Song, K.P. (2005). LuxS/autoinducer-2 quorum sensing molecule regulates transcriptional virulence gene expression in *Clostridium difficile*. Biochem. Biophys. Res. Commun. *335*, 659–666.

Lee, H.J., and Lee, S.J. (2021). Advances in accurate microbial genome-editing CRISPR technologies. J. Microbiol. Biotechnol. *31*, 903–911.

Lee, C., Kim, J., Shin, S.G., and Hwang, S. (2006). Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. J. Biotechnol. *123*, 273–280.

Lenoir, T., and Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. J. Biomed. Discov. Collab. *1*, 11.

Leonhardt, H. (1990). Identification of a low-copy-number mutation within the pUB110 replicon and its effect on plasmid stability in *Bacillus subtilis*. Gene *94*, 121–124.

Lesiak, J.M., Liebl, W., and Ehrenreich, A. (2014). Development of an *in vivo* methylation system for the solventogen *Clostridium saccharobutylicum* NCP 262 and analysis of two endonuclease mutants. J. Biotechnol. *188*, 97–99.

Lesnik, E.A., Sampath, R., Levene, H.B., Henderson, T.J., McNeil, J.A., and Ecker, D.J. (2001). Prediction of rho-independent transcriptional terminators in *Escherichia coli*. Nucleic Acids Res. *29*, 3583–3594.

Leuze, M.R., Karpinets, T. V., Syed, M.H., Beliaev, A.S., and Uberbacher, E.C. (2012). Binding Motifs in Bacterial Gene Promoters Modulate Transcriptional Effects of Global Regulators CRP and ArcA. Gene Regul. Syst. Bio. *6*, 93.

Li, J., and McClane, B.A. (2020). Evidence that virs is a receptor for the signaling peptide of the *Clostridium perfringens* agr-like quorum sensing system. MBio *11*, 1–20.

Li, J.S., Barber, C.C., Herman, N.A., Cai, W., Zafrir, E., Du, Y., Zhu, X., Skyrud, W., and Zhang, W. (2020). Investigation of secondary metabolism in the industrial butanol hyper-producer *Clostridium saccharoperbutylacetonicum* N1-4. J. Ind. Microbiol. Biotechnol. *47*, 319–328.

Li, Q., Chen, J., Minton, N.P., Zhang, Y., Wen, Z., Liu, J., Yang, H., Zeng, Z., Ren, X., Yang, J., et al. (2016). CRISPR-based genome editing and expression control systems in *Clostridium acetobutylicum* and *Clostridium beijerinckii*. Biotechnol. J. *11*, 961–972.

Li, Y., Davis, A., Korza, G., Zhang, P., Li, Y.Q., Setlow, B., Setlow, P., and Hao, B. (2012). Role of a SpoVA Protein in Dipicolinic Acid Uptake into Developing Spores of *Bacillus subtilis*. J. Bacteriol. *194*, 1875.

Liao, Z., Zhang, Y., Luo, S., Suo, Y., Zhang, S., and Wang, J. (2017). Improving cellular robustness and butanol titers of *Clostridium acetobutylicum* ATCC824 by introducing heat shock proteins from an extremophilic bacterium. J. Biotechnol. *252*, 1–10.

Liao, Z., Suo, Y., Xue, C., Fu, H., and Wang, J. (2018). Improving the fermentation performance of *Clostridium acetobutylicum* ATCC 824 by strengthening the VB1 biosynthesis pathway. Appl. Microbiol. Biotechnol. 1–13.

Liu, Z., Qiao, K., Tian, L., Zhang, Q., Liu, Z.Y., and Li, F.L. (2015). Spontaneous large-scale autolysis in *Clostridium acetobutylicum* contributes to generation of more spores. Front. Microbiol. *6*, 950.

Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R., and Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. *10*, 599–606.

Long, S., Jones, D.T., and Woods, D.R. (1983). Sporulation of *Clostridium acetobutylicum* P262 in a defined medium. Appl. Environ. Microbiol. *45*, 1389–1393.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. PLoS Comput. Biol. *13*.

Ma, M., Li, J., and McClane, B.A. (2015). Structure-function analysis of peptide signaling in the *Clostridium perfringens* Agr-like quorum sensing system. J. Bacteriol. *197*, 1807–1818.

Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A., and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. *29*, 4724–4735.

Mackey, B., and Morris, J.G. (1972). Calcium dipicolinate-provoked germination and outgrowth of spores of *Clostridium pasteurianum*. Jourtinl Gen. Microbiol. *73*, 3–5.

Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. *50*, W276–W279.

Magge, A., Setlow, B., Cowan, A.E., and Setlow, P. (2009). Analysis of dye binding by and membrane potential in spores of *Bacillus* species. J. Appl. Microbiol. *106*, 814–824.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. Nat. Rev. Microbiol. 2011 96 *9*, 467–477.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in *Staphylococci* by targeting DNA. Science (80-.). *322*, 1843–1845.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. *36*, 344–355.

van Melis, C.C.J., Nierop Groot, M.N., and Abee, T. (2011). Impact of sorbic acid on germinant receptor-dependent and -independent germination pathways in *Bacillus cereus*. Appl. Environ. Microbiol. *77*, 2552.

Mermelstein, L.D., Papoutsakis, E.T., Mermelstein, N.D.E., Welker, G.N., Bennett, E.T., and Papoutsakis, B./ (1993). *In vivo* methylation in *Escherichia coli* by the *Bacillus subtilis* phage phi 3T I methyltransferase to protect plasmids from restriction upon transformation of *Clostridium acetobutylicum* ATCC 824. Appl. Environ. Microbiol. *59*, 1077.

Minton, N., and Morris, J.G. (1981). Isolation and partial characterization of three cryptic plasmids from strains of *Clostridium butyricum*. J. Gen. Microbiol. *127*, 325–331.

Minton, N.P., Mauchline, M.L., Lemmon, M.J., Brehm, J.K., Fox, M., Michael, N.P., Giaccia, A., and Brown, J.M. (1995). Chemotherapeutic tumour targeting using *Clostridial* spores. FEMS Microbiol. Rev. *17*, 357–364.

Mitchell, W.J., Albasheri, K.A., and Yazdanian, M. (1995). Factors affecting utilization of carbohydrates by *Clostridia*. FEMS Microbiol. Rev. *17*, 317–329.

Moberg, L.J. (1985). Fluorogenic assay for rapid detection of *Escherichia coli* in food. Appl. Environ. Microbiol. *50*, 1383–1387.

Monaghan, T.I. (2019). Optimising solvent production in *Clostridium saccharoperbutylacetonicum* N1-4(HMT).

Monaghan, T.I., Baker, J.A., Krabben, P., Davies, E.T., Jenkinson, E.R., Goodhead, I.B., Robinson, G.K., and Shepherd, M. (2021a). Deletion of glyceraldehyde-3-phosphate dehydrogenase (gapN) in *Clostridium saccharoperbutylacetonicum* N1-4(HMT) using CLEAVE™ increases the ATP pool and accelerates solvent production. Microb. Biotechnol.

Monaghan, T.I., Baker, J.A., Robinson, G.K., and Shepherd, M. (2021b). Parallel bioreactor system for accessible and reproducible anaerobic culture. Access Microbiol. *3*.

Moule, M.G., Hemsley, C.M., Seet, Q., Guerra-Assunção, J.A., Lim, J., Sarkar-Tyson, M., Clark, T.G., Tan, P.B.O., Titball, R.W., Cuccui, J., et al. (2014). Genome-wide saturation mutagenesis of *Burkholderia pseudomallei* K96243 predicts essential genes and novel targets for antimicrobial development. MBio *5*, e00926-13.

Nagaraju, S., Davies, N.K., Walker, D.J.F., Köpke, M., and Simpson, S.D. (2016). Genome editing of *Clostridium autoethanogenum* using CRISPR/Cas9. Biotechnol. Biofuels *9*, 1–8.

Nagler, K., and Moeller, R. (2015). Systematic investigation of germination responses of *Bacillus subtilis* spores in different high-salinity environments. FEMS Microbiol. Ecol. *91*, 23.

Nakayama, S., Kosaka, T., Hirakawa, H., Matsuura, K., Yoshino, S., and Furukawa, K. (2008). Metabolic engineering for solvent productivity by downregulation of the hydrogenase gene cluster *hupCBA* in *Clostridium saccharoperbutylacetonicum* strain N1-4. Appl. Microbiol. Biotechnol. *78*, 483–493.

Nariya, H., Miyata, S., Kuwahara, T., and Okabe, A. (2011). Development and characterization of a xylose-inducible gene expression system for *Clostridium perfringens*. Appl. Environ. Microbiol. *77*, 8439–8441.

Nealson, K.H., Platt, T., and Hastings, J.W. (1970). Cellular control of the synthesis and activity of the bacterial luminescent system. J. Bacteriol. *104*, 313–322.

Neiditch, M.B., Capodagli, G.C., Prehna, G., and Federle, M.J. (2017). Genetic and structural analyses of RRNPP intercellular peptide signaling of Gram-positive bacteria. Https://Doi.Org/10.1146/Annurev-Genet-120116-023507 *51*, 311–333.

Nerandzic, M.M., and Donskey, C.J. (2013). Activate to eradicate: inhibition of *Clostridium difficile* spore outgrowth by the synergistic effects of osmotic activation and nisin. PLoS One *8*, e54740.

Noguchi, T., Tashiro, Y., Yoshida, T., Zheng, J., Sakai, K., and Sonomoto, K. (2013). Efficient butanol production without carbon catabolite repression from mixed sugars with *Clostridium saccharoperbutylacetonicum* N1-4. J. Biosci. Bioeng. *116*, 716–721.

Nölling, J., Breton, G., Omelchenko, M. V., Makarova, K.S., Zeng, Q., Gibson, G., Hong Mei Lee, Dubois, J., Qiu, D., Hitti, J., et al. (2001). Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J. Bacteriol. *183*, 4823–4838.

Novick, R.P., and Geisinger, E. (2008). Quorum sensing in *Staphylococci*. Http://Dx.Doi.Org/10.1146/Annurev.Genet.42.110807.091640 *42*, 541–564.

O'Connor, J.R., Lyras, D., Farrow, K.A., Adams, V., Powell, D.R., Hinds, J., Cheung, J.K., and Rood, J.I. (2006). Construction and analysis of chromosomal *Clostridium difficile* mutants. Mol. Microbiol. *61*, 1335–1351.

O'Donnell, M., Langston, L., and Stillman, B. (2013). Principles and concepts of DNA replication in bacteria, archaea, and eukarya. Cold Spring Harb. Perspect. Biol. *5*.

Oliveira Paiva, A.M., Friggen, A.H., Hossein-Javaheri, S., and Smits, W.K. (2016). The signal sequence of the abundant extracellular metalloprotease PPEP-1 can be used to decrete dynthetic reporter proteins in *Clostridium difficile*. ACS Synth. Biol. *5*, 1376–1382.

Oliveira Paiva, A.M., Friggen, A.H., Qin, L., Douwes, R., Dame, R.T., and Smits, W.K. (2019). The bacterial chromatin protein HupA can remodel DNA and associates with the nucleoid in *Clostridium difficile*. J. Mol. Biol. *431*, 653–672.

Oliveira Paiva, A.M., Friggen, A.H., Douwes, R., Wittekoek, B., and Smits, W.K. (2022). Practical observations on the use of fluorescent reporter systems in *Clostridioides difficile*. Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol. *115*, 297–323.

Omotajo, D., Tate, T., Cho, H., and Choudhary, M. (2015). Distribution and diversity of ribosome binding sites in prokaryotic genomes. BMC Genomics *16*.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods *6*, 767–772.

Oppenheim, A.B., Rudd, K.E., Mendelson, I., and Teff, D. (1993). Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in Escherichia coli. Mol. Microbiol. *10*, 113–122.

Oshiro, M., Hanada, K., Tashiro, Y., and Sonomoto, K. (2010). Efficient conversion of lactic acid to butanol with pH-stat continuous lactic acid and glucose feeding method by *Clostridium saccharoperbutylacetonicum*. Appl. Microbiol. Biotechnol. *87*, 1177–1185.

Paredes-Sabja, D., and Sarker, M.R. (2012). Adherence of *Clostridium difficile* spores to Caco-2 cells in culture. J. Med. Microbiol. *61*, 1208–1218.

Paredes-Sabja, D., Torres, J.A., Setlow, P., and Sarker, M.R. (2008). *Clostridium perfringens* spore germination: characterization of germinants and their receptors. J. Bacteriol. *190*, 1190–1201.

Paredes-Sabja, D., Shen, A., and Sorg, J.A. (2014). *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. Trends Microbiol. *22*, 406–416.

Paul, C., Filippidou, S., Jamil, I., Kooli, W., House, G.L., Estoppey, A., Hayoz, M., Junier, T., Palmieri, F., Wunderlin, T., et al. (2019). Bacterial spores, from ecology to biotechnology. Adv. Appl. Microbiol. *106*, 79–111.

Perego, M., Higgins, C.F., Pearce, S.R., Gallagher, M.P., and Hoch, J.A. (1991). The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. Mol. Microbiol. *5*, 173–185.

Pérez-Arellano, I., Zúñiga, M., and Pérez-Martínez, G. (2001). Construction of compatible wide-host-range shuttle vectors for lactic acid bacteria and *Escherichia coli*. Plasmid *46*, 106–116.

Perez-Pascual, D., Monnet, V., and Gardan, R. (2016). Bacterial cell-cell communication in the host via RRNPP peptide-binding regulators. Front. Microbiol. *7*, 706.

Piatek, P., Humphreys, C., Raut, M.P., Wright, P.C., Simpson, S., Köpke, M., Minton, N.P., and Winzer, K. (2022). Agr quorum sensing influences the Wood-Ljungdahl pathway in *Clostridium autoethanogenum*. Sci. Reports 2022 121 *12*, 1–15.

Piewngam, P., Chiou, J., Chatterjee, P., and Otto, M. (2020). Alternative approaches to treat bacterial infections: targeting quorum-sensing. Https://Doi.Org/10.1080/14787210.2020.1750951 *18*, 499–510.

Piggot, P.J., and Hilbert, D.W. (2004). Sporulation of *Bacillus subtilis*. Curr. Opin. Microbiol. *7*, 579–586.

Pizarro-Guajardo, M., Calderón-Romero, P., Castro-Córdova, P., Mora-Uribe, P., and Paredes-Sabja, D. (2016). Ultrastructural variability of the exosporium layer of *Clostridium difficile* spores. Appl. Environ. Microbiol. *82*, 2202–2209.

Plasmidvectors.com pMTL8000 plasmids.

Plotka, M., Wozniak, M., and Kaczorowski, T. (2017). Quantification of plasmid copy number with single colour droplet digital PCR. PLoS One *12*, e0169846.

Poehlein, A., Hartwich, K., Krabben, P., Ehrenreich, A., Liebl, W., Dürre, P., Gottschalk, G., and Daniel, R. (2013). Complete genome sequence of the solvent producer *Clostridium saccharobutylicum* NCP262 (DSM 13864). Genome Announc. *1*.

Poehlein, A., Krabben, P., Dürre, P., and Daniel, R. (2014). Complete genome sequence of the solvent producer *Clostridium saccharoperbutylacetonicum* strain DSM 14923. Genome Announc. *2*.

Poehlein, A., Solano, J.D.M., Flitsch, S.K., Krabben, P., Winzer, K., Reid, S.J., Jones, D.T., Green, E., Minton, N.P., Daniel, R., et al. (2017). Microbial solvent formation revisited by comparative genome analysis. Biotechnol. Biofuels *10*, 58.

Pomerantsev, A.P., Pomerantseva, O.M., Camp, A.S., Mukkamala, R., Goldman, S., and Leppla, S.H. (2009). PapR peptide maturation: role of the NprB protease in *Bacillus cereus* 569 PlcR/PapR global gene regulation. FEMS Immunol. Med. Microbiol. *55*, 361.

Purdy, D., O'Keeffe, T.A.T., Elmore, M., Herbert, M., McLeod, A., Bokori-Brown, M., Ostrowski, A., and Minton, N.P. (2002). Conjugative transfer of *Clostridial* shuttle vectors from *Escherichia coli* to *Clostridium difficile* through circumvention of the restriction barrier. Mol. Microbiol. *46*, 439–452.

Pyne, M.E., Moo-Young, M., Chung, D.A., and Chou, C.P. (2013). Development of an electrotransformation protocol for genetic manipulation of *Clostridium pasteurianum*. Biotechnol. Biofuels *6*, 1–20.

Pyne, M.E., Bruder, M.R., Moo-Young, M., Chung, D.A., and Chou, C.P. (2016). Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in *Clostridium*. Sci. Reports 2016 61 *6*, 1–15.

Qin, X., Singh, K. V., Weinstock, G.M., and Murray, B.E. (2000). Effects of *Enterococcus faecalis* fsr genes on production of gelatinase and a serine protease and virulence. Infect. Immun. *68*, 2579–2586.

Queck, S.Y., Jameson-Lee, M., Villaruz, A.E., Bach, T.H.L., Khan, B.A., Sturdevant, D.E., Ricklefs, S.M., Li, M., and Otto, M. (2008). RNAIII-independent target gene control by the agr quorum-sensing system: insight into the evolution of virulence regulation in *Staphylococcus aureus*. Mol. Cell *32*, 150–158.

Ransom, E.M., Ellermeier, C.D., and Weiss, D.S. (2015). Use of mCherry Red fluorescent protein for studies of protein localization and gene expression in *Clostridium difficile*. Appl. Environ. Microbiol. *81*, 1652–1660.

Ravagnani, A., Jennert, K.C.B., Steiner, E., Grünberg, R., Jefferies, J.R., Wilkinson, S.R., Young, D.I., Tidswell, E.C., Brown, D.P., Youngman, P., et al. (2000). Spo0A directly controls the switch from acid to solvent production in solvent-forming *Clostridia*. Mol. Microbiol. *37*, 1172–1185.

Ravindran, S. (2012). Barbara McClintock and the discovery of jumping genes. Proc. Natl. Acad. Sci. *109*, 20198–20199.

Rice, P., Longden, L., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Robson, R.L., Robson, R.M., and Morris, J.G. (1974). The biosynthesis of granulose by *Clostridium pasteurianum*. Biochem. J. *144*, 503–511.

Rojo, F. (2001). Mechanisms of transcriptional repression. Curr. Opin. Microbiol. *4*, 145–151.

Rubin, G.M., Kidwell, M.G., and Bingham, P.M. (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell *29*, 987–994.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics *16*, 944–945.

Sanchez-Salas, J.L., Setlow, B., Zhang, P., Li, Y.Q., and Setlow, P. (2011). Maturation of released spores is necessary for acquisition of full spore heat resistance during *Bacillus subtilis* sporulation. Appl. Environ. Microbiol. *77*, 6746–6754.

Sanger, F., Brownlee, G.G., and Barrell, B.G. (1965). A two-dimensional fractionation procedure for radioactive nucleotides. J. Mol. Biol. *13*, 373-IN4.

Santa Maria, J.P., Sadaka, A., Moussa, S.H., Brown, S., Zhang, Y.J., Rubin, E.J., Gilmore, M.S., and Walker, S. (2014). Compound-gene interaction mapping reveals distinct roles for *Staphylococcus aureus* teichoic acids. Proc. Natl. Acad. Sci. U. S. A. *111*, 12510–12515.

Santangelo IY, J.D., Kuhn, A., Treuner-Lange, A., and Du ë rre, P. (1998). Sporulation and time course expression of sigma-factor homologous genes in *Clostridium acetobutylicum*. FEMS Microbiol. Lett. *161*, 157–164.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2001). Comprehensive identification of conditionally essential genes in *Mycobacteria*. Proc. Natl. Acad. Sci. U. S. A. *98*, 12712–12717.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2003). Genes required for *Mycobacterial* growth defined by high density mutagenesis. Mol. Microbiol. *48*, 77–84.

Sauer, U., Santangelo, J.D., Treuner, A., Buchholz, M., and Dürre, P. (1995). Sigma factor and sporulation genes in *Clostridium*. FEMS Microbiol. Rev. *17*, 331–340.

Schultz, D., Wolynes, P.G., Jacob, E. Ben, and Onuchic, J.N. (2009). Deciding fate in adverse times: Sporulation and competence in *Bacillus subtilis*. Proc. Natl. Acad. Sci. *106*, 21027–21034.

Scotcher, M.C., and Bennett, G.N. (2005). SpoIIE regulates sporulation but does not directly affect solventogenesis in *Clostridium acetobutylicum* ATCC 824. J. Bacteriol. *187*, 1930–1936.

Sekulovic, O., Mathias Garrett, E., Bourgeois, J., Tamayo, R., Shen, A., and Camilli, A. (2018). Genome-wide detection of conservative site-specific recombination in bacteria. PLOS Genet. *14*, e1007332.

Setlow, P. (2014). Germination of spores of *Bacillus* species: what we know and do not know. J. Bacteriol. *196*, 1297.

Setlow, P., Wang, S., and Li, Y.Q. (2017). Germination of spores of the orders *Bacillales* and *Clostridiales*. Https://Doi.Org/10.1146/Annurev-Micro-090816-093558 *71*, 459–477.

Shaheen, R., Shirley, M., and Jones, D.T. (2000). Comparative fermentation studies of industrial strains belonging to four species of solvent-producing *Clostridia*. J. Mol. Microbiol. Biotechnol. *2*, 115–124.

Shaner, N.C., Steinbach, P.A., and Tsien, R.Y. (2005). A guide to choosing fluorescent proteins. Nat. Methods 2005 212 *2*, 905–909.

Shao, B., Rammohan, J., Anderson, D.A., Alperovich, N., Ross, D., and Voigt, C.A. (2021). Single-cell measurement of plasmid copy number and promoter activity. Nat. Commun. 2021 121 *12*, 1–9.

Shen, A., Edwards, A.N., Sarker, M.R., and Paredes-Sabja, D. (2019). Sporulation and germination in *Clostridial* pathogens. Microbiol. Spectr. *7*.

Singh, R.P., Desouky, S.E., and Nakayama, J. (2016). Quorum quenching strategy targeting Gram-positive pathogenic bacteria. Adv. Exp. Med. Biol. *901*, 109–130.

Sinoquet, C., Demey, S., and Braun, F. (2008). Large-scale computational and statistical analyses of high transcription potentialities in 32 prokaryotic genomes. Nucleic Acids Res. *36*, 3332–3340.

SLS Potassium sorbate, purum p.a., | 85520-1KG | SIG001 | SLS.

Smith, V., Botstein, D., and Brown, P.O. (1995). Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. Proc. Natl. Acad. Sci. U. S. A. *92*, 6479.

Smoot, L.A., and Pierson, M.D. (1981). Mechanisms of sorbate inhibition of *Bacillus cereus* T and *Clostridium botulinum* 62A spore germination. Appl. Environ. Microbiol. *42*, 477.

Solovyev, V. (2011). Automatic annotation of microbial genomes and metagenomic sequences. In metagenomics and its applications in agriculture, biomedicine and environmental studies. pp. 61–78.

Sorg, J.A., and Sonenshein, A.L. (2008). Bile salts and glycine as cogerminants for *Clostridium difficile* spores. J. Bacteriol. *190*, 2505–2512.

Southern, E.M. (2001). DNA Microarrays. Methods Mol. Biol. *170*, 1–15.

Steiner, E., Dago, A.E., Young, D.I., Heap, J.T., Minton, N.P., Hoch, J.A., and Young, M. (2011). Multiple orphan histidine kinases interact directly with Spo0A to control the initiation of endospore formation in *Clostridium acetobutylicum*. Mol. Microbiol. *80*, 641–654.

Steiner, E., Scott, J., Minton, N.P., and Winzer, K. (2012). An agr quorum sensing system that regulates granulose formation and sporulation in *Clostridium acetobutylicum*. Appl. Environ. Microbiol. *78*, 1113–1122.

Subashchandrabose, S., Smith, S., DeOrnellas, V., Crepin, S., Kole, M., Zahdeh, C., and Mobley, H.L.T. (2016). *Acinetobacter baumannii* genes required for bacterial survival during bloodstream infection. MSphere *1*, e00013-15.

Tan, I.S., and Ramamurthi, K.S. (2014). Spore formation in *Bacillus subtilis*. Environ. Microbiol. Rep. *6*, 212.

Tanaka, S., Tashiro, Y., Kobayashi, G., Ikegami, T., Negishi, H., and Sakaki, K. (2012). Membrane-assisted extractive butanol fermentation by *Clostridium saccharoperbutylacetonicum* N1-4 with 1-dodecanol as the extractant. Bioresour. Technol. *116*, 448–452.

Tashiro, Y., Shinto, H., Hayashi, M., Baba, S., Kobayashi, G., and Sonomoto, K. (2007). Novel high-efficient butanol production from butyrate by non-growing *Clostridium saccharoperbutylacetonicum* N1-4 (ATCC 13564) with methyl viologen. J. Biosci. Bioeng. *104*, 238–240.

Thauer, R.K., Jungermann, K., and Decker, K. (1977). Energy conservation in chemotrophic anaerobic bacteria. Bacteriol. Rev. *41*, 100.

Thibault, D., Jensen, P.A., Wood, S., Qabar, C., Clark, S., Shainheit, M.G., Isberg, R.R., and van Opijnen, T. (2019). Droplet Tn-Seq combines microfluidics with Tn-Seq for identifying complex single-cell phenotypes. Nat. Commun. 2019 101 *10*, 1–13.

Tomé, D. (2021). Yeast extracts: nutritional and flavoring food ingredients. ACS Food Sci. Technol. *1*, 487–494.

Tong, X., Campbell, J.W., Balázsi, G., Kay, K.A., Wanner, B.L., Gerdes, S.Y., and Oltvai, Z.N. (2004). Genome-scale identification of conditionally essential genes in *E. coli* by DNA microarrays. Biochem. Biophys. Res. Commun. *322*, 347–354.

Tracy, B.P., Gaida, S.M., and Papoutsakis, E.T. (2008). Development and application of flow-cytometric techniques for analyzing and sorting endospore-forming *Clostridia*. Appl. Environ. Microbiol. *74*, 7497–7506.

Tracy, B.P., Jones, S.W., and Papoutsakis, E.T. (2011). Inactivation of σEand σGin *Clostridium acetobutylicum* illuminates their roles in *Clostridial*-cell-form biogenesis, granulose synthesis, solventogenesis, and spore morphogenesis. J. Bacteriol. *193*, 1414–1426.

Tripathi, S.A., Olson, D.G., Argyros, D.A., Miller, B.B., Barrett, T.F., Murphy, D.M., McCool, J.D., Warner, A.K., Rajgarhia, V.B., Lynd, L.R., et al. (2010). Development of pyrF-Based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a *pta* mutant. Appl. Environ. Microbiol. *76*, 6591–6599.

Truffaut, N., Hubert, J., and Reysset, G. (1989). Construction of shuttle vectors useful for transforming *Clostridium acetobutylicum*. FEMS Microbiol. Lett. *58*, 15–19.

Tummala, S.B., Welker, N.E., and Papoutsakis, E.T. (1999). Development and characterization of a gene expression reporter system for *Clostridium acetobutylicum* ATCC 824. Appl. Environ. Microbiol. *65*, 3793–3799.

Typas, A., Nichols, R.J., Siegele, D.A., Shales, M., Collins, S.R., Lim, B., Braberg, H., Yamamoto, N., Takeuchi, R., Wanner, B.L., et al. (2008). A tool-kit for high-throughput, quantitative analyses of genetic interactions in *E. coli*. Nat. Methods *5*, 781.

Udompijitkul, P., Alnoman, M., Banawas, S., Paredes-Sabja, D., and Sarker, M.R. (2014). New amino acid germinants for spores of the enterotoxigenic *Clostridium perfringens* type A isolates. Food Microbiol. *44*, 24–33.

Veening, J.W., Smits, W.K., and Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. Http://Dx.Doi.Org/10.1146/Annurev.Micro.62.081307.163002 *62*, 193–210.

Velásquez, J., Schuurman-Wolters, G., Birkner, J.P., Abee, T., and Poolman, B. (2014). *Bacillus subtilis* spore protein SpoVAC functions as a mechanosensitive channel. Mol. Microbiol. *92*, 813–823.

Vepachedu, V.R., and Setlow, P. (2007). Role of SpoVA proteins in release of dipicolinic acid during germination of *Bacillus subtilis* spores triggered by bodecylamine or lysozyme. J. Bacteriol. *189*, 1565–1572.

Verbeke, T.J., Giannone, R.J., Klingeman, D.M., Engle, N.L., Rydzak, T., Guss, A.M., Tschaplinski, T.J., Brown, S.D., Hettich, R.L., and Elkins, J.G. (2017). Pentose sugars inhibit metabolism and increase expression of an AgrD-type cyclic pentapeptide in *Clostridium thermocellum*. Sci. Rep. *7*, 1–11.

Vigouroux, A., and Bikard, D. (2020). CRISPR tools to control gene expression in bacteria. Microbiol. Mol. Biol. Rev. *84*.

de Vrije, T., Budde, M., van der Wal, H., Claassen, P.A.M., and López-Contreras, A.M. (2013). "*In situ*" removal of isopropanol, butanol and ethanol from fermentation broth by gas stripping. Bioresour. Technol. *137*, 153–159.

Wang, G., Zhang, P., Paredes-Sabja, D., Green, C., Setlow, P., Sarker, M.R., and Li, Y.Q. (2011). Analysis of the germination of individual *Clostridium perfringens* spores and its heterogeneity. J. Appl. Microbiol. *111*, 1212–1223.

Wang, S., Shen, A., Setlow, P., and Li, Y.Q. (2015a). Characterization of the dynamic germination of individual *Clostridium difficile* spores using Raman spectroscopy and differential interference contrast microscopy. J. Bacteriol. *197*, 2361–2373.

Wang, S., Dong, S., Wang, P., Tao, Y., and Wang, Y. (2017). Genome editing in *Clostridium saccharoperbutylacetonicum* N1-4 with the CRISPR-Cas9 system. Appl. Environ. Microbiol. *83*, 1–44.

Wang, Y., Zhang, Z.T., Seo, S.O., Choi, K., Lu, T., Jin, Y.S., and Blaschek, H.P. (2015b). Markerless chromosomal gene deletion in *Clostridium beijerinckii* using CRISPR/Cas9 system. J. Biotechnol. *200*, 1–5.

Wasels, F., Jean-Marie, J., Collas, F., López-Contreras, A.M., and Lopes Ferreira, N. (2017). A two-plasmid inducible CRISPR/Cas9 genome editing tool for *Clostridium acetobutylicum*. J. Microbiol. Methods *140*, 5–11.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Weiland-Bräuer, N. (2021). Friends or foes—microbial interactions in nature. Biol. 2021, Vol. 10, Page 496 *10*, 496.

Weizmann, C. (1919). Improvements in the bacterial fermentation of carbohydrates and in bacterial cultures for the same.

Westphal, A.J., Price, P.B., Leighton, T.J., and Wheeler, K.E. (2003). Kinetics of size changes of individual *Bacillus thuringiensis* spores in response to changes in relative humidity. Proc. Natl. Acad. Sci. *100*, 3461–3466.

Wollman, E.L., and Jacob, F. (1955). Mechanism of the transfer of genetic material during recombination in *Escherichia coli* K12. C. R. Hebd. Seances Acad. Sci. *240*, 2449–2451.

Wuster, A., and Babu, M.M. (2008). Conservation and evolutionary dynamics of the agr cell-to-cell communication system across *Firmicutes*. J. Bacteriol. *190*, 743–746.

Yamanè, T., and Shimizu, S. (1984). Fed-batch techniques in microbial processes. Bioprocess Param. Control 147–194.

Yu, M., Du, Y., Jiang, W., Chang, W.L., Yang, S.T., and Tang, I.C. (2012). Effects of different replicons in conjugative plasmids on transformation efficiency, plasmid stability, gene expression and n-butanol biosynthesis in *Clostridium tyrobutyricum*. Appl. Microbiol. Biotechnol. *93*, 881–889.

Zeytuni, N., and Zarivach, R. (2012). Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module. Structure *20*, 397–405.

Zhang, R., Peng, S., and Qin, Z. (2010). Two internal origins of replication in *Streptomyces* linear plasmid pFRL1. Appl. Environ. Microbiol. *76*, 5676–5683.

Zheng, J., Tashiro, Y., Yoshida, T., Gao, M., Wang, Q., and Sonomoto, K. (2013). Continuous butanol fermentation from xylose with high cell density by cell recycling system. Bioresour. Technol. *129*, 360–365.

Appendix I – Strains

Table A1 List of strains used in this study

Clostridium saccharoperbutylacetonicum		
Strain	**Description**	**Source**
N1-4(HMT)	One of two characterised wild type strains	Green Biologics/Biocleave
N1-4(HMT) Δ*agrB*	N1-4(HMT) containing CLEAVE™ knockout of *agrB* (Cspa_c18660)	This study
N1-4(HMT) Δ*agr'D'*	N1-4(HMT) containing CLEAVE™ knockout of *agr'D'* identified in chapter V	This study
N1-4(HMT) Δ*agrC*	N1-4(HMT) containing CLEAVE™ knockout of *agrC* (Cspa_c18640)	This study
N1-4(HMT) Δ*agrB-C*	N1-4(HMT) containing CLEAVE™ knockout of whole agr system (Cspa_c18640, Cspa_c18650 and Cspa_c18660)	This study
Escherichia coli		
NEB5α	Chemically competent cells used for all cloning at The University of Sheffiled	New England Biolabs
CA434	Strain HB101 carrying the IncPβ conjugative plasmid, R702. Used as a conjugation donor.	(Purdy et al., 2002)
TOP10	Chemically competent cells used for plasmid propogation	Lab stocks
DH10β	Chemically competent cells used for routine cloning and plasmid propagation at Biocleave	Thermo Fisher
Clostridiodes difficile		
R20291	Ribotype 027 strain responsible for the 2004 Stoke-Mandeville outbreak	(Stabler et al., 2009)

Appendix II – Plasmids

Table A2 List of plasmids used in this study

Plasmid name	Description	Source
pRPF185	pMTL960-P*tet*(SacI)-gusA-2x-terminator 12 Terminator from *C. pasteurianum fdx* gene cloned into NheI/KpnI sites of pRPF177 as 2 annealed oligos (1469/1470) - Cd630	(Fagan and Fairweather, 2011)
pRPF215	pMTL960-P*tet-himar1*(Tnase) + *ermB* transposon. Transposon mutagenesis plasmid.	(Dembek et al., 2015)
pJAK112	pMTLSC7215 *slpA* KO (SacI and BamHI cloning sites added)	(Fuchs et al., 2021)
pAF259	pRPF185 derived plasmid encoding an anhydrotetracyclin inducible *BitlucOpt*	(Oliveira Paiva et al., 2019)
pJAK175	pAF259 with *tetR* and P*tet* removed and P*xyl* and *xylR* added to create a xylose inducible *BitlucOpt*	Dr Joseph Kirk
pLLK007	pAF259 with *tetR* + P*tet* removed and a multiple cloning site inserted	This study
pLLK009	pAP259 with t*etR* + P*tet* removed and P*thial* from *C. acetobutylicum* (Ca_c2873)	This study
pLLK010	pAP259 with shortest p*SlpA* from pJAK106 (from *C. difficile* R20291)	This study
pLLK011	pAP259 with p*SlpA* containing putative UPE (from *C. difficile* R20291) from pJAK107	This study
pLLK012	pAP259 with p*SlpA* + ~ 300bp upstream of promoter from (from *C. difficile* R20291) pJAK127	This study
pEW001	pAF259 with *bdh* (Cspa_c56790) promoter controlling *BitlucOpt*	This study
pEW002	pAF259 with *chromosomal ATPase* (from megaplasmid) promoter controlling *BitlucOpt*	This study
pEW003	pAF259 with *flgB* (Cspa_c45250) promoter controlling *BitlucOpt*	This study
pEW004	pAF259 with *hag4* (Cspa_c45710) promoter controlling *BitlucOpt*	This study
pEW005	pAF259 with *rpoD* (CSPA_c43810) promoter controlling *BitlucOpt*	This study
pEW007	pAF259 with *virion structural protein* (Cspa_135p00700 from *C. saccharoperbutylacetonicum* megaplasmid) promoter controlling *BitlucOpt*	This study
pEW008	pAF259 with *heat-shock protein* (CSPA_c16680) promoter controlling *BitlucOpt*	This study
pMTL82151	Backbone for homologous recombination vectors	(Heap et al., 2009)

pMTL83251	Backbone for CRISPR-Cas targeting vectors	(Heap et al., 2009)
pMTL83251_ldr	Backbone for CRISPR-Cas targeting vectors including leader sequence	(Atmadjaja et al., 2019)
pLLK018	pMTL82151_homologyarms_agrB_knockout	This study
pLLK019	pMTL82151_homologyarms_agr'D'_knockout	This study
pLLK020	pMTL82151_homologyrms_agrC_knockout	This study
pLLK021	pMTL82151_homologyarms_agrB-C_knockout	This study
pLLK022	pMTL82151_homologyarms_agrA_knockout	This study
pLLK023	pMTL83251_ldr_*agrA*spacers	This study
pLLK024	pMTL83251_ldr_*agrB*spacers	This study
pLLK025	pMTL83251_ldr_*agrC*spacers	This study
pLLK026	pMTL83251_ldr_*agr'D*'spacers	This study
pLLK027	pMTL83251_ldr_*agrB-Cspacers*	This study

Appendix III – Primers

Table A3 List of primers used in this study

Name	Sequence 5'-3'	Purpose
RF1000	GATCGAGCTCTTCTTTTCCTCCTCTTACACAC	Amplifies a 187bp fragment across the bdh promter from the c.saccharoperbutylacetonicum N1-4 genome. Pairs with RF1001. Gives expected product
RF1001	GATCGGTACCGTTCAACTAGATTTATGTGCAAG	Amplifies a 187bp fragment across the bdh promter from the c.saccharoperbutylacetonicum N1-4 genome. Pairs with RF1000. Gives expected product
RF1520	GAAAGTTACACGTTACTAAAGGCATAAAAATAAGAAGCCTGCAAATGC	Forward Primer 1st PCR (Specific for Erm Transposon)
RF1521	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTACGGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Fw Primer 2nd PCR Index 6.1
RF1522	GACTGGAGTTCAGACGTGTGCTCTTCCGATC	Reverse Primer 1st PCR (NEB Adapter)
RF1523	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAGCTGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Fw Primer 2nd PCR Index 7.2
RF1524	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCATGCGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Fw Primer 2nd PCR Index 8.1
RF1525	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGATCGATGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Fw Primer 2nd PCR Index 9.3
RF1526	GACTGGAGTTCAGACGTGTGCTCTTCCGATC	Reverse Primer 1st PCR (NEB Adapter)
RF1629	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTACGTAGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Transposon library Primer 2nd PCR Index 6.3
RF1630	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTAGGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Transposon library Primer 2nd PCR Index 7.4
RF1631	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGCATGGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Transposon library Primer 2nd PCR Index 8.3
RF1632	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATCGATCGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Transposon library Primer 2nd PCR Index 9.4
RF1640	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Primer for amplifying TraDIS libraries for Genewiz QC (contains fw seq site - transposon binding site)

RF1662	GAAATCTACTTCAACTTATAACCTTGG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (12 oclock) p145-172
RF1663	ATTCTATGCAGCTTGGAATC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (12 oclock) p356-376
RF1664	TCATTTCAAGCAAGACCTG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (6 oclock) p68087-68106
RF1665	CCAACTCTTACATCAGAAGC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (6 oclock) p68294-68313
RF1666	CTGCTAATTTAGTAGCCTCAAG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (3 oclock) p33970-33991
RF1667	GCTAATTCATATGCAAAGTTAGC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (3 oclock) p34182-34204
RF1668	GTTACAACTGAAACCTTTAATGC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (9 oclock) p101845-101967
RF1669	TTCAGGGCCAGCATATAG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Megaplasmid (9 oclock) p102050-102072
RF1670	AATTGTGCTAGGTTTGGTAC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (12 oclock) p6530150-6530169
RF1671	CGGCATCCATTCTTATTCC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (12 oclock) p0000092-0000110
RF1672	CAGCTGAAATATTACTTTTAGGAAG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (6 oclock) p3265086-3265110
RF1673	TACACGTTCATTTGTAGCTG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (6 oclock) p3265257-3265276
RF1674	ATGAAATTTCTTCGTTACCAGG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (3 oclock) p1632451-1632472
RF1675	CTCACACATGCAAACATCAG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (3 oclock) p1632667-1632686
RF1676	TTTTCCTATTAGCAGCTATTATAAGTAC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (9 oclock) p4897294-4897321
RF1677	GAAGACTGACAAAGGCTATG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (9 oclock) p4897481-4897500
RF1678	CTATATTAAGCACTGATTAGTACTATAACTC	qPCR primer for copy number determination pRPF215 (ori) p9144-9174
RF1679	AGCTAGAATCCTAATTAGTAGGTG	qPCR primer for copy number determination pRPF215 (ori) p91-114
RF1680	GACCAAAAGAAGTAGTAACTGATG	qPCR primer for copy number determination pRPF215 Opposite Ori p4296-4319

RF1681	AAACATCTTTCAGAATCATCTACTC	qPCR primer for copy number determination pRPF215 Opposite Ori p4503-4527
RF1714	GAATGCCTTTTACAGAGGTATC	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (6 oclock) p3265020-3265041
RF1715	GAATATGTTTCCGAAAGAAATGG	qPCR primer for copy number determination *C. saccharoperbutylacetonicum* Genome (6 oclock) p3265286-3265308
RF1748	GATCGGTACCGTAACAAATATGGTGACAAC	clone promoter ATPase chromosome (*C. saccharoperbutylacetonicum* megaplasmid)
RF1749	GATCGAGCTCAGCCATAATATTATCACTCC	clone promoter ATPase chromosome (*C. saccharoperbutylacetonicum* megaplasmid)
RF1750	GATCGGTACCTTTCGTGCAGAAGATTCTAG	clone promoter flgB (*C. saccharoperbutylacetonicum*)
RF1751	GATCGAGCTCCCATTAATTCACTTCCTCAC	clone promoter flgB (*C. saccharoperbutylacetonicum*)
RF1752	GATCGGTACCCTAGATGAAGAACTTGTAAGG	clone promoter hag4 (*C. saccharoperbutylacetonicum*)
RF1753	GATCGAGCTCATTATCATTATAATTTCCTCCTTG	clone promoter hag4 (*C. saccharoperbutylacetonicum*)
RF1754	GATCGGTACCTAGCTATACAGAGTTAATTTCG	clone promoter hsp (*C. saccharoperbutylacetonicum*)
RF1755	GATCGAGCTCCAAACATACTTAAGACCTCC	clone promoter hsp (*C. saccharoperbutylacetonicum*)
RF1756	GATCGGTACCACTAACCCTGTGAAGTTGGC	clone promoter rpoD (*C. saccharoperbutylacetonicum*)
RF1757	GATCGAGCTCGGCTCCACCTTATTAATCTCC	clone promoter rpoD (*C. saccharoperbutylacetonicum*)
RF1758	GATCGGTACCGTGTATCGGCTGGTGATATC	clone promoter rrfB (*C. saccharoperbutylacetonicum*)
RF1759	GATCGAGCTCTGAGCCAGGATCAAACTCTC	clone promoter rrfB (*C. saccharoperbutylacetonicum*)
RF1760	GATCGGTACCGGTGAAATGAGAACTGGCTG	clone promoter virion structural protein (*C. saccharoperbutylacetonicum* megaplasmid)
RF1761	GATCGAGCTCCTTTCGAATAATAAAAGGCAGCTAG	clone promoter virion structural protein (*C. saccharoperbutylacetonicum* megaplasmid)
RF1777	CTGCAGTAAAGGAGAAAATTTTG	Binds upstream of bitluc in pAP259
RF1778	GGCTTCTTATTTTTATGTGTTAGACCGGGGACTTATCAGC	Binds across ITR and FDX terminator of pRPF215 for inverse PCR amplification (pairs with RF1779)
RF1779	CAACCTGTTACCAGTGTGCTGGCGGCCGCCCCTTCAATATTC	Binds across the ITR and EcoRI restriction site with the aim for replacing EcoRI with NotI through inverse pcr (pairs with RF1778)
RF1780	GCTAGCGCGGCCGCGAGCTCCTGCAGTAAAGGAGAAAATTTTG	Binds just upstream of bitlucopt (in RBS) to inverse pcr pAP259 to remove the TetR/Promoter system and leave a MCS (Pairs with RF1780)
RF1781	CTCGAGGGATCCGGTACCGATGCAGAATTCGCCCTTAAG	Binds just after the FDX terminator on pAP259 to remove the TetR/Promoter system and leave a MCS (Pairs with RF1781) ny inverse PCR
RF2181	CTGTATCCATATGACCATGATTACGAATTCTCACTTCTTAAACAGAACTAACAC	Gibson primer for homology arm construction for AgrC KO for plasmid pLLK015
RF2182	GAAATAGTACTTAATCTATTAATACGCAATAAATAAAAATTTACAG	Gibson primer for homology arm construction for AgrC KO for plasmid pLLK015
RF2183	ATTGCGTATTAATAGATTAAGTACTATTTCAAGCATTTCTTC	Gibson primer for homology arm construction for AgrC KO for plasmid pLLK015

RF2184	TGCCAAGCTTGCATGTCTGCAGGCCTCGAGCAAAATGTATTTATACAAATTTTAAGCTTTG	Gibson primer for homology arm construction for AgrC KO for plasmid pLLK015
RF2312	GTTTTAGGAATAAGTTGTTATTACTG	For screening for knockouts in *C. saccharoperbutylacetonicum* AgrC Fw
RF2313	GTATATAAAGGGGTTGTTGAAG	For screening for knockouts in *C. saccharoperbutylacetonicum* AgC Rev
RF2314	CTCAAGCATGATAATTAACGG	For screening for knockouts *C. saccharoperbutylacetonicum* AgrA Fw
RF2315	GAATGACTATAAGAACTACTTTG	For screening for knockouts in *C. saccharoperbutylacetonicum* AgrA Rev
RF2316	CAGTAAGTGTGCTATTAGCATTG	For screening for knockouts in *C. saccharoperbutylacetonicum* AgrB Fw
RF2317	CAAGCATTATATAATCACCTCCATT	For screening for knockouts in *C. saccharoperbutylacetonicum* AgrB Rev
RF2318	CTTTGATCTAGATTATCTTCACAG	For screening for knockouts in *C. saccharoperbutylacetonicum* AgrD Fw
RF2319	CTTACTGAAAGGTAGGGCC	For screening for knockouts in *C. saccharoperbutylacetonicum* AgrD Rev
RF2320	GGCTGGCAATCATGATGC	For screening for knockouts in *C. saccharoperbutylacetonicum* agrB-C_FW pairs with RF2317
RF2321	GTACAGCATTAAGGCAATCAG	For screening for knockouts in *C. saccharoperbutylacetonicum* agrB_chk_FW pairs with RF2317
M13 rev	CAGGAAACAGCTATGACC	Screening in pMTL vectors
M13 fw	TGTAAAACGACGGCCAGT	Screening in pMTL vectors
NF793	CACCTCCTTTTTGACTTTAAGCCTACGAATACC	Screening in pRPF and pJAK vectors
NF794	CACCGACGAGCAAGGCAAGACCG	Screening in pRPF and pJAK vectors

Appendix IV – *C. saccharoperbutylacetonicum* ORFs without insertions

Table A4 List of ORFs without insertions in *C. saccharoperbutylacetonicum*

locus_tag	gene_name	start	end	strand	read_count	ins_index	gene_length	ins_count	Function
Cspa_c00010	dnaA	101	1456	1	0	0	1356	0	chromosomal replication initiator protein DnaA
Cspa_c00020	dnaN	1718	2818	1	0	0	1101	0	DNA polymerase III subunit beta
Cspa_c00040	recF	3133	4215	1	0	0	1083	0	DNA replication and repair protein RecF
Cspa_c00060	gyrB1	4543	6450	1	0	0	1908	0	DNA gyrase subunit B
Cspa_c00080	rrfA	9446	10946	1	0	0	1501	0	
Cspa_c00220	addB	18652	22128	1	0	0	3477	0	ATP-dependent helicase/deoxyribonuclease subunit B
Cspa_c00230	addA	22210	25941	1	0	0	3732	0	ATP-dependent helicase/nuclease subunit A
Cspa_c00240	Cspa_c00240	26028	27203	-1	0	0	1176	0	putative L,D-transpeptidase
Cspa_c00570	ileS	61509	64622	1	0	0	3114	0	isoleucine--tRNA ligase IleS
Cspa_c01010	murC	109459	110838	-1	0	0	1380	0	UDP-N-acetylmuramate--L-alanine ligase MurC
Cspa_c01340	lysS	141661	143166	1	0	0	1506	0	lysine--tRNA ligase LysS
Cspa_c01400	murD	146144	147520	1	0	0	1377	0	UDP-N-acetylmuramoylalanine--D-glutamate ligase MurD
Cspa_c01420	rrfB	149180	150680	1	0	0	1501	0	
Cspa_c01490	Cspa_c01490	154469	155632	-1	0	0	1164	0	metal-dependent amidase/aminoacylase/carboxypeptidase
Cspa_c01810	rpoB	181197	184910	1	0	0	3714	0	DNA-directed RNA polymerase subunit beta
Cspa_c01850	fusA1	189775	191841	1	0	0	2067	0	elongation factor G
Cspa_c02080	secY	202865	204151	1	0	0	1287	0	protein translocase subunit SecY
Cspa_c02530	buk1	247241	248308	1	0	0	1068	0	butyrate kinase Buk
Cspa_c02760	leuC2	279239	280498	1	0	0	1260	0	3-isopropylmalate dehydratase large subunit LeuC
Cspa_c03300	Cspa_c03300	344824	347433	1	0	0	2610	0	hypothetical protein
Cspa_c03510	glcD1	371401	372804	1	0	0	1404	0	glycolate oxidase subunit GlcD
Cspa_c03610	Cspa_c03610	383719	384888	1	0	0	1170	0	glycosyl transferase group 1

Cspa_c03650	luxQ	386757	388268	-1	0	0	1512	0	autoinducer 2 sensor kinase/phosphatase LuxQ
Cspa_c03720	Cspa_c03720	400250	401632	-1	0	0	1383	0	MATE efflux family protein
Cspa_c03760	baeE	432093	433181	1	0	0	1089	0	polyketide biosynthesis protein BaeE
Cspa_c04340	bcd1	502498	503637	1	0	0	1140	0	acyl-CoA dehydrogenase, short-chain specific
Cspa_c04360	etfA3	504453	505460	1	0	0	1008	0	electron-transferring flavoprotein alpha-subunit EtfA
Cspa_c04520	pgi	521898	523247	1	0	0	1350	0	glucose-6-phosphate isomerase Pgi
Cspa_c04630	pcrA	532720	536058	1	0	0	3339	0	ATP-dependent DNA helicase PcrA
Cspa_c04640	ligA	536099	538090	1	0	0	1992	0	DNA ligase LigA
Cspa_c05050	rrfC	560682	562182	1	0	0	1501	0	
Cspa_c05100	tldD	566869	568254	1	0	0	1386	0	protein TldD
Cspa_c05120	Cspa_c05120	569844	570953	1	0	0	1110	0	signal transduction histidine kinase
Cspa_c05300	Cspa_c05300	589027	590043	1	0	0	1017	0	spore coat protein, CotS family
Cspa_c05330	yfmM	592043	593509	-1	0	0	1467	0	ABC transporter ATP-binding protein YfmM
Cspa_c05630	Cspa_c05630	627798	629381	1	0	0	1584	0	putative AAA-ATPase
Cspa_c05860	Cspa_c05860	650555	651664	1	0	0	1110	0	Ig-like domain-containing surface protein
Cspa_c05930	abgA1	657787	659202	1	0	0	1416	0	6-phospho-beta-glucosidase AbgA
Cspa_c06050	pyrG	670216	671826	1	0	0	1611	0	CTP synthase PyrG
Cspa_c06070	ugtP	673786	674901	-1	0	0	1116	0	processive diacylglycerol glucosyltransferase UgtP
Cspa_c06240	atpA	688895	690409	1	0	0	1515	0	ATP synthase subunit alpha
Cspa_c06780	Cspa_c06780	743210	744220	1	0	0	1011	0	amidohydrolase 2
Cspa_c06810	rnfC	746699	748018	1	0	0	1320	0	electron transport complex protein RnfC
Cspa_c06900	Cspa_c06900	758438	760711	1	0	0	2274	0	aldehyde dehydrogenase, iron-sulfur subunit
Cspa_c07120	mcp41	782361	784142	1	0	0	1782	0	methyl-accepting chemotaxis protein 4
Cspa_c07240	eutB	795555	796922	1	0	0	1368	0	ethanolamine ammonia-lyase heavy chain
Cspa_c07270	Cspa_c07270	798522	799589	-1	0	0	1068	0	2-nitropropane dioxygenase
Cspa_c07340	rrfD	802658	804158	1	0	0	1501	0	
Cspa_c07380	Cspa_c07380	809639	811081	1	0	0	1443	0	multidrug resistance efflux pump
Cspa_c07790	xkdK	846578	847885	1	0	0	1308	0	phage-like element PBSX protein XkdK
Cspa_c07880	Cspa_c07880	853529	854572	1	0	0	1044	0	phage Mu protein-like protein gp47

Cspa_c07960	argD	862284	863468	-1	0	0	1185	0	acetylornithine aminotransferase ArgD
Cspa_c07980	argJ	864623	865846	-1	0	0	1224	0	arginine biosynthesis bifunctional protein ArgJ
Cspa_c07990	argC	865882	866916	-1	0	0	1035	0	N-acetyl-gamma-glutamyl-phosphate reductase ArgC
Cspa_c08010	argG	868663	869886	-1	0	0	1224	0	argininosuccinate synthase ArgG
Cspa_c08080	glpA	877596	879023	1	0	0	1428	0	FAD dependend glycerol-3-phosphate dehydrogenase GlpA
Cspa_c08140	Cspa_c08140	882747	884021	1	0	0	1275	0	hypothetical protein
Cspa_c08280	pbg	904430	906514	1	0	0	2085	0	beta-galactosidase Pbg
Cspa_c08300	Cspa_c08300	907784	909505	1	0	0	1722	0	methyl-accepting chemotaxis protein
Cspa_c08350	apbE1	914214	915266	1	0	0	1053	0	thiamine biosynthesis lipoprotein ApbE
Cspa_c08370	Cspa_c08370	916055	917437	1	0	0	1383	0	hypothetical protein
Cspa_c08740	nadB	964828	966135	1	0	0	1308	0	L-aspartate oxidase NadB
Cspa_c08790	trpS	969077	970096	1	0	0	1020	0	tryptophan--tRNA ligase TrpS
Cspa_c08990	spoVAD1	994703	995722	1	0	0	1020	0	stage V sporulation protein AD
Cspa_c09080	kdpB	1007102	1009165	1	0	0	2064	0	potassium-transporting ATPase B chain
Cspa_c09290	yqfd	1029549	1030664	1	0	0	1116	0	sporulation protein YqfD
Cspa_c09300	Cspa_c09300	1030729	1032816	1	0	0	2088	0	metal dependent phosphohydrolase
Cspa_c09390	Cspa_c09390	1041673	1042770	1	0	0	1098	0	hypothetical protein
Cspa_c09420	mdtG	1045058	1046251	1	0	0	1194	0	multidrug resistance protein MdtG
Cspa_c09470	Cspa_c09470	1051597	1052856	1	0	0	1260	0	extracellular solute-binding protein, family 1
Cspa_c09510	Cspa_c09510	1056866	1058854	1	0	0	1989	0	hypothetical protein
Cspa_c09550	bglH1	1063321	1064754	1	0	0	1434	0	aryl-phospho-beta-D-glucosidase BglH
Cspa_c09570	cotS2	1066248	1067312	-1	0	0	1065	0	spore coat protein CotS
Cspa_c09580	Cspa_c09580	1067546	1068571	1	0	0	1026	0	dGTP triphosphohydrolase
Cspa_c09590	dnaG	1068783	1070564	1	0	0	1782	0	DNA primase DnaG

Cspa_c09600	rpoD2	1070570	1071706	1	0	0	1137	0	RNA polymerase sigma factor RpoD
Cspa_c09770	Cspa_c09770	1089933	1091651	1	0	0	1719	0	subtilase family
Cspa_c09810	glgP1	1095965	1098325	1	0	0	2361	0	glycogen phosphorylase GlgP
Cspa_c09860	nrgA1	1105814	1107046	1	0	0	1233	0	ammonium transporter NrgA
Cspa_c09880	Cspa_c09880	1109539	1111152	-1	0	0	1614	0	DNA invertase Pin-like site-specific recombinase
Cspa_c10150	Cspa_c10150	1128187	1129341	1	0	0	1155	0	prophage lambdaBa04, portal protein
Cspa_c10170	Cspa_c10170	1130095	1131348	1	0	0	1254	0	phage protein
Cspa_c10450	murE1	1155093	1156451	-1	0	0	1359	0	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate-2, 6-diaminopimelate ligase MurE
Cspa_c10620	hcp3	1179359	1181005	-1	0	0	1647	0	hydroxylamine reductase Hcp
Cspa_c10680	pgcA	1185229	1186956	1	0	0	1728	0	phosphoglucomutase PgcA
Cspa_c10690	Cspa_c10690	1187115	1188215	1	0	0	1101	0	TPR repeat-containing protein
Cspa_c10710	mviN	1189019	1190548	1	0	0	1530	0	integral membrane protein MviN
Cspa_c10730	Cspa_c10730	1191847	1193010	1	0	0	1164	0	glycosyl transferase group 1
Cspa_c10780	Cspa_c10780	1196724	1197944	-1	0	0	1221	0	transposase, mutator type
Cspa_c10940	pfl1	1215529	1217757	1	0	0	2229	0	formate acetyltransferase Pfl
Cspa_c10990	Cspa_c10990	1221714	1222760	1	0	0	1047	0	exopolysaccharide biosynthesis protein
Cspa_c11210	asnO	1245520	1247364	1	0	0	1845	0	asparagine synthetase [glutamine-hydrolyzing] 3
Cspa_c11310	nadE	1255191	1257089	-1	0	0	1899	0	glutamine-dependent NAD(+) synthetase NadE
Cspa_c11350	Cspa_c11350	1258915	1260450	1	0	0	1536	0	ATPase, histidine kinase-, DNA gyrase B-, and HSP90-like domain protein
Cspa_c11440	clcA	1266885	1268483	1	0	0	1599	0	H(+)/Cl(-) exchange transporter ClcA

Cspa_c11520	purH	1277979	1279487	1	0	0	1509	0	bifunctional purine biosynthesis protein PurH
Cspa_c11550	Cspa_c11550	1282419	1283480	1	0	0	1062	0	cell wall binding repeat-containing protein
Cspa_c11660	fabF1	1293790	1295025	1	0	0	1236	0	3-oxoacyl-[acyl-carrier-protein] synthase 2
Cspa_c11690	accC1	1296106	1297458	1	0	0	1353	0	biotin carboxylase AccC
Cspa_c12030	Cspa_c12030	1329409	1330755	1	0	0	1347	0	UDP-glucose 6-dehydrogenase
Cspa_c12070	Cspa_c12070	1334271	1335641	1	0	0	1371	0	glycosyltransferase
Cspa_c12130	rfbB1	1340074	1341123	1	0	0	1050	0	dTDP-glucose 4,6-dehydratase RfbB
Cspa_c12180	Cspa_c12180	1346675	1347694	1	0	0	1020	0	fucose 4-O-acetylase
Cspa_c12240	macB4	1351596	1353575	1	0	0	1980	0	macrolide export ATP-binding/permease protein MacB
Cspa_c12440	alaS	1375107	1377746	1	0	0	2640	0	alanine--tRNA ligase AlaS
Cspa_c12570	ftsZ	1390385	1391515	1	0	0	1131	0	cell division protein FtsZ
Cspa_c12710	Cspa_c12710	1402904	1404241	1	0	0	1338	0	NifB/MoaA family Fe-S oxidoreductase
Cspa_c12720	engA	1404241	1405557	1	0	0	1317	0	GTP-binding protein EngA
Cspa_c12800	coaBC	1411384	1412568	1	0	0	1185	0	putative coenzyme A biosynthesis bifunctional protein CoaBC
Cspa_c12810	priA	1412613	1414808	1	0	0	2196	0	primosomal protein N'
Cspa_c12850	rsmB	1416944	1418260	1	0	0	1317	0	ribosomal RNA small subunit methyltransferase B
Cspa_c12860	rlmN1	1418275	1419327	1	0	0	1053	0	putative dual-specificity RNA methyltransferase RlmN
Cspa_c12880	sps	1420042	1421940	1	0	0	1899	0	serine/threonine-protein kinase Sps
Cspa_c12950	recG	1427324	1429357	1	0	0	2034	0	ATP-dependent DNA helicase RecG
Cspa_c13000	Cspa_c13000	1432308	1433510	-1	0	0	1203	0	hypothetical protein UPF0348

Cspa_c13100	smc	1440913	1444476	1	0	0	3564	0	chromosome partition protein Smc
Cspa_c13130	ffh	1446066	1447418	1	0	0	1353	0	signal recognition particle protein Ffh
Cspa_c13220	Cspa_c13220	1452318	1453838	1	0	0	1521	0	Mg chelatase, subunit ChlI
Cspa_c13330	dxr	1464573	1465736	1	0	0	1164	0	1-deoxy-D-xylulose 5-phosphate reductoisomerase Dxr
Cspa_c13340	rasP	1465751	1466773	1	0	0	1023	0	RIP metalloprotease RasP
Cspa_c13350	ispG	1466848	1467897	1	0	0	1050	0	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase IspG
Cspa_c13370	nusA	1468711	1469877	1	0	0	1167	0	transcription elongation protein NusA
Cspa_c13400	infB	1470488	1472566	1	0	0	2079	0	translation initiation factor IF-2
Cspa_c13550	rny	1488407	1489948	1	0	0	1542	0	ribonuclease Y
Cspa_c13640	ribBA	1499339	1500532	1	0	0	1194	0	riboflavin biosynthesis protein RibBA
Cspa_c13680	kup2	1505465	1507486	-1	0	0	2022	0	putative potassium transport system protein Kup
Cspa_c13700	Cspa_c13700	1508124	1509530	1	0	0	1407	0	hypothetical protein
Cspa_c14050	cpsA	1539176	1540354	1	0	0	1179	0	thermostable carboxypeptidase 1
Cspa_c14120	nifE1	1545090	1546538	1	0	0	1449	0	nitrogenase iron-molybdenum cofactor biosynthesis protein NifE
Cspa_c14130	nifK1	1546543	1547811	1	0	0	1269	0	nitrogenase molybdenum-iron protein beta subunit NifK
Cspa_c14180	ssuA	1552909	1553955	1	0	0	1047	0	nitrate/sulfonate/bicarbonate ABC transport system substrate binding protein SsuA
Cspa_c14210	gldA	1555674	1556762	1	0	0	1089	0	glycerol dehydrogenase GldA
Cspa_c14320	metC2	1569224	1570384	1	0	0	1161	0	cystathionine beta-lyase MetC
Cspa_c14420	Cspa_c14420	1579700	1581127	-1	0	0	1428	0	drug resistance transporter, EmrB/QacA subfamily

Cspa_c14540	licR1	1593143	1594633	1	0	0	1491	0	putative licABCH operon regulator
Cspa_c14620	cobT	1602208	1603302	1	0	0	1095	0	nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase CobT
Cspa_c14690	cobD2	1607435	1608508	1	0	0	1074	0	threonine-phosphate decarboxylase CobD
Cspa_c14890	hemA	1627947	1629152	1	0	0	1206	0	glutamyl-tRNA reductase HemA
Cspa_c14970	xynD1	1636797	1638374	1	0	0	1578	0	arabinoxylan arabinofuranohydrolase XynD
Cspa_c15090	engD2	1651031	1652167	-1	0	0	1137	0	endoglucanase D
Cspa_c15320	hisZ	1679260	1680390	1	0	0	1131	0	ATP phosphoribosyltransferase regulatory subunit HisZ
Cspa_c15340	hisD	1681136	1682434	1	0	0	1299	0	histidinol dehydrogenase HisD
Cspa_c15350	hisC2	1682421	1683458	1	0	0	1038	0	histidinol-phosphate aminotransferase HisC
Cspa_c15580	pncB	1702096	1703574	-1	0	0	1479	0	nicotinate phosphoribosyltransferase PncB
Cspa_c15640	Cspa_c15640	1710972	1712153	1	0	0	1182	0	amidohydrolase
Cspa_c15850	Cspa_c15850	1738451	1739899	1	0	0	1449	0	phage late control gene D protein GPD
Cspa_c15860	Cspa_c15860	1739998	1741101	1	0	0	1104	0	pentapeptide repeat protein
Cspa_c15960	Cspa_c15960	1754189	1755565	1	0	0	1377	0	putative ATPase
Cspa_c15970	Cspa_c15970	1755700	1757445	1	0	0	1746	0	AAA-ATPase-like protein
Cspa_c16120	pel	1776119	1777924	1	0	0	1806	0	pectate trisaccharide-lyase Pel
Cspa_c16150	glnK	1781061	1782323	1	0	0	1263	0	sensor histidine kinase GlnK
Cspa_c16170	alsS1	1783623	1785302	1	0	0	1680	0	acetolactate synthase AlsS
Cspa_c16350	Cspa_c16350	1806132	1807493	1	0	0	1362	0	radical SAM domain-containing protein

Cspa_c16360	Cspa_c16360	1808161	1809399	1	0	0	1239	0	putative Zn-dependent peptidase
Cspa_c16430	Cspa_c16430	1816504	1817676	1	0	0	1173	0	aspartate/tyrosine/aromatic aminotransferase
Cspa_c16600	pulA	1834349	1836322	1	0	0	1974	0	pullulanase PulA
Cspa_c16670	rlmI	1844972	1846159	1	0	0	1188	0	rRNA (guanine-N(2)-)-methyltransferase RlmI
Cspa_c16750	Cspa_c16750	1855213	1856598	1	0	0	1386	0	amino acid/polyamine/organocation transporter, APC superfamily
Cspa_c17220	Cspa_c17220	1896573	1897976	1	0	0	1404	0	two-component system signal transduction histidine kinase
Cspa_c17310	thlA2	1905661	1906839	1	0	0	1179	0	acetyl-CoA acetyltransferase ThlA
Cspa_c17350	Cspa_c17350	1910509	1911990	1	0	0	1482	0	succinate dehydrogenase/fumarate reductase, flavoprotein subunit
Cspa_c17410	Cspa_c17410	1918284	1919552	1	0	0	1269	0	PTS system, lactose/cellobiose family IIC component
Cspa_c17450	Cspa_c17450	1924125	1925435	1	0	0	1311	0	PTS system, lactose/cellobiose family IIC component
Cspa_c17480	bglA2	1926387	1927814	1	0	0	1428	0	6-phospho-beta-glucosidase BglA
Cspa_c17490	malL1	1927950	1929632	1	0	0	1683	0	oligo-1,6-glucosidase MalL
Cspa_c17500	Cspa_c17500	1930037	1931596	1	0	0	1560	0	hypothetical protein DUF1703
Cspa_c17610	celB1	1945008	1946261	1	0	0	1254	0	cellobiose permease IIC component CelB
Cspa_c17630	Cspa_c17630	1946697	1948139	1	0	0	1443	0	beta-glucosidase/6-phospho-beta- glucosidase/beta-galactosidase
Cspa_c17650	Cspa_c17650	1950342	1951916	-1	0	0	1575	0	cysteine synthase
Cspa_c17680	Cspa_c17680	1954738	1956057	1	0	0	1320	0	ABC-type sugar transport system, periplasmic component
Cspa_c17700	Cspa_c17700	1956959	1958725	1	0	0	1767	0	integral membrane sensor signal transduction histidine kinase
Cspa_c17720	Cspa_c17720	1960431	1961996	-1	0	0	1566	0	cellobiose phosphorylase

Cspa_c17760	uxaC1	1965652	1967052	-1	0	0	1401	0	uronate isomerase UxaC
Cspa_c17940	dxs1	1986336	1988093	1	0	0	1758	0	1-deoxy-D-xylulose-5-phosphate synthase 1
Cspa_c18160	Cspa_c18160	2002284	2003498	1	0	0	1215	0	glycosyltransferase family 28
Cspa_c18330	Cspa_c18330	2017664	2019211	1	0	0	1548	0	drug resistance transporter, EmrB/QacA subfamily
Cspa_c18430	metG3	2029514	2031145	1	0	0	1632	0	methionine--tRNA ligase 1
Cspa_c18480	mepA2	2034137	2035498	1	0	0	1362	0	multidrug export protein MepA
Cspa_c18510	msbA1	2038092	2039807	1	0	0	1716	0	lipid A export ATP-binding/permease protein MsbA
Cspa_c18530	Cspa_c18530	2041486	2042601	1	0	0	1116	0	peptide maturation system protein, TIGR04066 family
Cspa_c18570	cbpA	2047416	2050181	1	0	0	2766	0	cellulose-binding protein A
Cspa_c18590	celK1	2053442	2056057	1	0	0	2616	0	cellulose 1,4-beta-cellobiosidase CelK
Cspa_c18620	celA2	2058821	2060647	1	0	0	1827	0	endoglucanase A
Cspa_c18630	Cspa_c18630	2060676	2062094	1	0	0	1419	0	endo-beta-mannanase
Cspa_c18640	Cspa_c18640	2062164	2063528	-1	0	0	1365	0	signal transduction histidine kinase regulating citrate/malate metabolism
Cspa_c18690	celH	2068223	2069362	1	0	0	1140	0	endoglucanase H
Cspa_c19030	Cspa_c19030	2102309	2104024	1	0	0	1716	0	methyl-accepting chemotaxis protein
Cspa_c19110	Cspa_c19110	2112626	2113810	1	0	0	1185	0	1,2-diacylglycerol 3-glucosyltransferase
Cspa_c19200	hcp4	2121483	2123123	1	0	0	1641	0	hydroxylamine reductase Hcp
Cspa_c19220	pgaC	2123760	2124983	-1	0	0	1224	0	poly-beta-1,6-N-acetyl-D-glucosamine synthase PgaC
Cspa_c19250	Cspa_c19250	2127385	2128631	1	0	0	1247	0	pseudogene
Cspa_c19280	Cspa_c19280	2129980	2131203	1	0	0	1224	0	arabinose efflux permease

Cspa_c19300	Cspa_c19300	2132123	2133268	1	0	0	1146	0	hypothetical protein
Cspa_c19360	ybdL	2136108	2137276	1	0	0	1169	0	pseudogene
Cspa_c19390	metN2	2138865	2139893	1	0	0	1029	0	methionine import ATP-binding protein MetN 1
Cspa_c19420	Cspa_c19420	2142240	2143469	1	0	0	1230	0	hypothetical protein
Cspa_c19610	eryC	2159502	2160596	-1	0	0	1095	0	erythromycin biosynthesis sensory transduction protein EryC
Cspa_c19630	Cspa_c19630	2161039	2162166	-1	0	0	1128	0	hypothetical protein DUF563
Cspa_c19720	Cspa_c19720	2176104	2177105	-1	0	0	1002	0	hypothetical protein
Cspa_c19780	wbpA1	2183050	2184366	1	0	0	1317	0	UDP-N-acetyl-D-glucosamine 6-dehydrogenase WbpA
Cspa_c19820	Cspa_c19820	2188938	2192225	-1	0	0	3288	0	collagen triple helix repeat protein
Cspa_c19830	Cspa_c19830	2192556	2193974	-1	0	0	1419	0	glycosyl transferase family 2
Cspa_c19990	Cspa_c19990	2212827	2213831	1	0	0	1005	0	alpha/beta hydrolase family
Cspa_c20070	Cspa_c20070	2218425	2220101	1	0	0	1677	0	hypothetical protein DUF1703
Cspa_c20290	Cspa_c20290	2243173	2244378	-1	0	0	1206	0	alpha/beta hydrolase family
Cspa_c20380	Cspa_c20380	2252587	2253612	1	0	0	1026	0	nucleoside-diphosphate-sugar epimerase
Cspa_c20520	thlA3	2265175	2266353	1	0	0	1179	0	acetyl-CoA acetyltransferase ThlA
Cspa_c20530	Cspa_c20530	2266379	2267410	1	0	0	1032	0	hypothetical protein containing Fe-S binding domain
Cspa_c20800	mglB2	2291418	2292485	1	0	0	1068	0	D-galactose-binding periplasmic protein MglB
Cspa_c20840	Cspa_c20840	2294109	2295833	1	0	0	1725	0	PAS domain-containing S-box protein
Cspa_c20890	thlA4	2301013	2302197	1	0	0	1185	0	acetyl-CoA acetyltransferase ThlA
Cspa_c20900	bcd3	2302413	2303552	1	0	0	1140	0	acyl-CoA dehydrogenase, short-chain specific

Cspa_c20920	etfA4	2304441	2305448	1	0	0	1008	0	electron transfer flavoprotein alpha subunit EtfA
Cspa_c20970	Cspa_c20970	2312284	2313444	-1	0	0	1161	0	putative transposase
Cspa_c21050	rlmN2	2321661	2322695	1	0	0	1035	0	ribosomal RNA large subunit methyltransferase Cfr
Cspa_c21080	xynD2	2324412	2325809	1	0	0	1398	0	arabinoxylan arabinofuranohydrolase XynD
Cspa_c21450	eutE	2362580	2363911	1	0	0	1332	0	ethanolamine utilization protein EutE
Cspa_c21480	mtlA	2366916	2368328	1	0	0	1413	0	PTS system mannitol-specific EIICB component MtlA
Cspa_c21580	gutB	2379036	2380052	1	0	0	1017	0	sorbitol dehydrogenase GutB
Cspa_c21590	Cspa_c21590	2380168	2382105	1	0	0	1938	0	alpha amylase
Cspa_c21670	Cspa_c21670	2389214	2390215	1	0	0	1002	0	GGGtGRT protein
Cspa_c21720	Cspa_c21720	2393128	2394459	-1	0	0	1332	0	hypothetical protein
Cspa_c21910	Cspa_c21910	2404771	2405964	1	0	0	1194	0	alpha/beta fold family hydrolase
Cspa_c22070	Cspa_c22070	2418622	2419707	1	0	0	1086	0	theronine dehydrogenase-like Zn-dependent dehydrogenase
Cspa_c22110	Cspa_c22110	2431403	2432770	-1	0	0	1368	0	major facilitator superfamily MFS_1
Cspa_c22170	Cspa_c22170	2436899	2438170	1	0	0	1272	0	hypothetical protein
Cspa_c22180	Cspa_c22180	2438256	2439467	1	0	0	1212	0	RND family efflux transporter, MFP subunit
Cspa_c22270	Cspa_c22270	2449673	2452153	1	0	0	2481	0	hypothetical protein
Cspa_c22390	ilvD2	2467281	2469017	1	0	0	1737	0	dihydroxy-acid dehydratase IlvD
Cspa_c22400	xynB3	2469122	2470777	1	0	0	1656	0	beta-xylosidase XynB
Cspa_c22440	Cspa_c22440	2474575	2475750	-1	0	0	1176	0	hypothetical protein
Cspa_c22490	resE6	2479468	2481450	1	0	0	1983	0	sensor histidine kinase ResE

Cspa_c22750	Cspa_c22750	2507205	2508221	1	0	0	1017	0	NADH-flavin oxidoreductase/NADH oxidase NADH
Cspa_c22810	Cspa_c22810	2514241	2516067	1	0	0	1827	0	enterochelin esterase-like enzyme
Cspa_c22940	ycxD1	2525876	2527219	-1	0	0	1344	0	putative HTH-type transcriptional regulator YcxD
Cspa_c23020	ycxD2	2532813	2534147	1	0	0	1335	0	putative HTH-type transcriptional regulator YcxD
Cspa_c23050	Cspa_c23050	2536750	2538042	1	0	0	1293	0	carbohydrate ABC transporter substrate-binding protein, CUT1 family
Cspa_c23090	Cspa_c23090	2542418	2543716	1	0	0	1299	0	glycosyl hydrolases family 39
Cspa_c23350	bglC2	2566692	2568098	1	0	0	1407	0	aryl-phospho-beta-D-glucosidase BglC
Cspa_c23400	dmsA	2573042	2575414	1	0	0	2373	0	dimethyl sulfoxide reductase DmsA
Cspa_c23480	Cspa_c23480	2584812	2586149	1	0	0	1338	0	two component transcriptional regulator, AraC family
Cspa_c23510	xylG	2588697	2590274	1	0	0	1578	0	xylose import ATP-binding protein XylG
Cspa_c23520	xylH1	2590276	2591457	1	0	0	1182	0	xylose transport system permease protein XylH
Cspa_c23600	Cspa_c23600	2603035	2604519	1	0	0	1485	0	L-fucose isomerase
Cspa_c23640	tetA	2606490	2607725	1	0	0	1236	0	tetracycline resistance protein, class C
Cspa_c23760	Cspa_c23760	2616995	2618104	1	0	0	1110	0	putative choloylglycine hydrolase
Cspa_c23810	Cspa_c23810	2622082	2623383	1	0	0	1302	0	methyl-accepting chemotaxis protein
Cspa_c23840	ascB	2624589	2626058	1	0	0	1470	0	6-phospho-beta-glucosidase AscB
Cspa_c24010	metH3	2638104	2641745	1	0	0	3642	0	5-methyltetrahydrofolate--homocysteine methyltransferase MetH
Cspa_c24110	apbE2	2651839	2652888	1	0	0	1050	0	thiamine biosynthesis lipoprotein ApbE
Cspa_c24250	Cspa_c24250	2668525	2669778	-1	0	0	1254	0	arabinose efflux permease
Cspa_c24270	Cspa_c24270	2670700	2672463	1	0	0	1764	0	ABC-type multidrug transport system, ATPase and permease component

Cspa_c24300	Cspa_c24300	2675488	2678328	1	0	0	2841	0	multi-sensor hybrid histidine kinase
Cspa_c24420	Cspa_c24420	2688651	2689841	1	0	0	1191	0	putative signal transduction protein
Cspa_c24430	Cspa_c24430	2689865	2691007	1	0	0	1143	0	hypothetical protein
Cspa_c24440	Cspa_c24440	2691009	2692049	1	0	0	1041	0	histidine kinase
Cspa_c24530	yjbG	2701149	2703050	1	0	0	1902	0	oligoendopeptidase F
Cspa_c24570	Cspa_c24570	2708618	2710246	1	0	0	1629	0	PAS domain S-box
Cspa_c24660	Cspa_c24660	2717607	2718719	1	0	0	1113	0	hypothetical protein DUF3810
Cspa_c24790	Cspa_c24790	2733399	2734748	1	0	0	1350	0	putative efflux protein, MATE family
Cspa_c24820	Cspa_c24820	2737096	2738109	1	0	0	1014	0	hypothetical protein
Cspa_c24930	Cspa_c24930	2745985	2747415	1	0	0	1431	0	sensory transduction histidine kinase
Cspa_c24960	hscC2	2752124	2753824	1	0	0	1701	0	chaperone protein HscC
Cspa_c25230	cysS2	2780962	2782374	1	0	0	1413	0	cysteine--tRNA ligase CysS
Cspa_c25260	rpfG6	2785376	2786527	1	0	0	1152	0	cyclic di-GMP phosphodiesterase response regulator RpfG
Cspa_c25420	Cspa_c25420	2796870	2797907	-1	0	0	1038	0	hypothetical protein
Cspa_c25570	ruvB	2809470	2810510	1	0	0	1041	0	holliday junction ATP-dependent DNA helicase RuvB
Cspa_c25760	Cspa_c25760	2830945	2832111	1	0	0	1167	0	polysaccharide pyruvyl transferase
Cspa_c25770	Cspa_c25770	2832126	2833325	1	0	0	1200	0	coenzyme F420-reducing hydrogenase, beta subunit
Cspa_c25780	Cspa_c25780	2833391	2834578	1	0	0	1188	0	glycosyl transferase group 1
Cspa_c25890	ptk	2844528	2845823	1	0	0	1296	0	tyrosine-protein kinase Ptk
Cspa_c25920	Cspa_c25920	2848559	2849659	1	0	0	1101	0	glycosyltransferase

Cspa_c25940	Cspa_c25940	2850620	2851702	1	0	0	1083	0	putative acyltransferase
Cspa_c25960	Cspa_c25960	2852954	2854021	1	0	0	1068	0	lipopolysaccharide biosynthesis protein
Cspa_c26190	pheS	2873093	2874112	1	0	0	1020	0	phenylalanine--tRNA ligase alpha subunit PheS
Cspa_c26200	pheT	2874324	2876702	1	0	0	2379	0	phenylalanine--tRNA ligase beta subunit PheT
Cspa_c26250	spoVD2	2881394	2883658	1	0	0	2265	0	stage V sporulation protein D
Cspa_c26260	murE3	2883810	2885273	1	0	0	1464	0	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diaminopimelate ligase MurE
Cspa_c26270	murF	2885417	2886787	1	0	0	1371	0	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase MurF
Cspa_c26290	ftsW	2887828	2888964	1	0	0	1137	0	lipid II flippase FtsW
Cspa_c26870	Cspa_c26870	2929834	2931117	1	0	0	1284	0	phage portal protein, SPP1 Gp6-like protein
Cspa_c26910	Cspa_c26910	2933572	2934678	1	0	0	1107	0	P22 coat protein
Cspa_c26970	Cspa_c26970	2936575	2937657	1	0	0	1083	0	phage tail sheath protein
Cspa_c27010	Cspa_c27010	2938818	2940770	1	0	0	1953	0	hypothetical protein
Cspa_c27040	Cspa_c27040	2942207	2943673	1	0	0	1467	0	NlpC/P60 family protein
Cspa_c27070	xkdT	2944505	2945572	1	0	0	1068	0	phage-like element PBSX protein XkdT
Cspa_c27120	Cspa_c27120	2948402	2949484	1	0	0	1083	0	bacterial surface protein containing Ig-like domain
Cspa_c27130	Cspa_c27130	2949877	2951049	1	0	0	1173	0	hypothetical protein
Cspa_c27170	Cspa_c27170	2952831	2954063	1	0	0	1233	0	hypothetical protein
Cspa_c27380	spoIIIAE1	2966433	2967692	1	0	0	1260	0	stage III sporulation protein AE
Cspa_c27480	dxs2	2973709	2975565	1	0	0	1857	0	1-deoxy-D-xylulose-5-phosphate synthase Dxs
Cspa_c27530	spoIIIAH	2979778	2980968	1	0	0	1191	0	SpoIVB peptidase

Cspa_c27690	Cspa_c27690	2991086	2992177	1	0	0	1092	0	bacterial surface protein containing Ig-like domain
Cspa_c27780	Cspa_c27780	3001425	3002432	1	0	0	1008	0	putative phage-associated protein
Cspa_c27900	Cspa_c27900	3017226	3019484	1	0	0	2259	0	glycosyl transferase family 39
Cspa_c27910	Cspa_c27910	3019606	3021261	1	0	0	1656	0	hypothetical protein
Cspa_c28030	Cspa_c28030	3034181	3035494	1	0	0	1314	0	putative efflux protein, MATE family
Cspa_c28160	Cspa_c28160	3049546	3050595	1	0	0	1050	0	multidrug resistance efflux pump
Cspa_c28170	Cspa_c28170	3050576	3051799	1	0	0	1224	0	ABC-type multidrug transport system, permease component
Cspa_c28180	Cspa_c28180	3051774	3052991	1	0	0	1218	0	ABC-2 family transporter protein
Cspa_c28190	Cspa_c28190	3053143	3054447	1	0	0	1305	0	hypothetical protein
Cspa_c28280	Cspa_c28280	3062238	3063395	1	0	0	1158	0	hypothetical protein DUF4367
Cspa_c28310	dacF	3065329	3066519	1	0	0	1191	0	D-alanyl-D-alanine carboxypeptidase DacF
Cspa_c28380	Cspa_c28380	3071803	3073080	1	0	0	1278	0	glucose-inhibited division protein A
Cspa_c28470	Cspa_c28470	3080855	3081865	1	0	0	1011	0	ABC-type nitrate/sulfonate/bicarbonate transport system, periplasmic component
Cspa_c28510	Cspa_c28510	3084887	3086164	-1	0	0	1278	0	hypothetical protein
Cspa_c28530	Cspa_c28530	3087579	3088856	-1	0	0	1278	0	hypothetical protein
Cspa_c28540	Cspa_c28540	3088928	3090202	-1	0	0	1275	0	hypothetical protein DUF4179
Cspa_c28570	Cspa_c28570	3092981	3093997	1	0	0	1017	0	histidine kinase
Cspa_c28720	Cspa_c28720	3112147	3113148	1	0	0	1002	0	acyltransferase 3
Cspa_c28750	Cspa_c28750	3114710	3115711	1	0	0	1002	0	hypothetical protein

Cspa_c28830	Cspa_c28830	3124886	3126136	1	0	0	1251	0	hypothetical protein
Cspa_c28850	Cspa_c28850	3127680	3129509	-1	0	0	1830	0	hypothetical protein
Cspa_c28950	potD	3135877	3136971	1	0	0	1095	0	spermidine/putrescine-binding periplasmic protein PotD
Cspa_c29000	gbsA	3141298	3142779	1	0	0	1482	0	betaine aldehyde dehydrogenase GbsA
Cspa_c29030	patB1	3146333	3147511	1	0	0	1179	0	cystathionine beta-lyase PatB
Cspa_c29060	ramA1	3150744	3153629	1	0	0	2886	0	alfa-L-rhamnosidase RamA
Cspa_c29080	licR3	3154879	3156843	1	0	0	1965	0	putative licABCH operon regulator
Cspa_c29130	xynB4	3159599	3160606	1	0	0	1008	0	endo-1,4-beta-xylanase B
Cspa_c29200	Cspa_c29200	3170286	3171437	1	0	0	1152	0	arabinose efflux permease family protein
Cspa_c29210	rpfG7	3171920	3173680	1	0	0	1761	0	cyclic di-GMP phosphodiesterase response regulator RpfG
Cspa_c29220	Cspa_c29220	3174123	3176096	1	0	0	1974	0	methyl-accepting chemotaxis sensory transducer with cache sensor
Cspa_c29290	fabF2	3180951	3182186	1	0	0	1236	0	3-oxoacyl-[acyl-carrier-protein] synthase 2
Cspa_c29310	accC2	3182867	3184219	1	0	0	1353	0	biotin carboxylase AccC
Cspa_c29340	Cspa_c29340	3186122	3192157	1	0	0	6036	0	amino acid adenylation domain protein
Cspa_c29350	Cspa_c29350	3192174	3193337	1	0	0	1164	0	major facilitator superfamily MFS_1
Cspa_c29400	patA	3205253	3206608	1	0	0	1356	0	putrescine aminotransferase PatA
Cspa_c29440	Cspa_c29440	3210088	3211224	-1	0	0	1137	0	esterase/lipase
Cspa_c29470	Cspa_c29470	3212874	3215537	1	0	0	2664	0	DEAD/DEAH box helicase domain-containing protein
Cspa_c29480	rep	3215547	3217793	1	0	0	2247	0	ATP-dependent DNA helicase Rep
Cspa_c29640	mglB5	3231752	3232804	-1	0	0	1053	0	D-galactose-binding periplasmic protein MglB

Cspa_c29700	sbp	3236843	3237880	1	0	0	1038	0	sulfate-binding protein Sbp
Cspa_c29730	cysA	3239706	3240767	1	0	0	1062	0	sulfate/thiosulfate import ATP-binding protein CysA
Cspa_c29740	Cspa_c29740	3240789	3242462	1	0	0	1674	0	succinate dehydrogenase/fumarate reductase flavoprotein subunit
Cspa_c29860	Cspa_c29860	3253352	3254770	1	0	0	1419	0	beta-glucosidase/6-phospho-beta- glucosidase/beta-galactosidase
Cspa_c29920	Cspa_c29920	3260708	3261997	1	0	0	1290	0	PTS system, lactose/cellobiose family IIC subunit
Cspa_c29940	xynB5	3263506	3266019	-1	0	0	2514	0	beta-xylosidase XynB
Cspa_c30000	Cspa_c30000	3272475	3273491	1	0	0	1017	0	oxidoreductase domain protein
Cspa_c30030	Cspa_c30030	3275716	3277503	1	0	0	1788	0	diguanylate cyclase/phosphodiesterase
Cspa_c30040	ackA2	3277742	3278977	1	0	0	1236	0	acetate kinase AckA
Cspa_c30050	polC	3279150	3283538	1	0	0	4389	0	DNA polymerase III PolC-type
Cspa_c30080	Cspa_c30080	3286651	3287952	-1	0	0	1302	0	major facilitator superfamily MFS_1
Cspa_c30100	asrA	3288991	3289998	1	0	0	1008	0	anaerobic sulfite reductase subunit A
Cspa_c30130	narB	3291929	3294013	1	0	0	2085	0	nitrate reductase NarB
Cspa_c30150	padH	3294481	3295728	1	0	0	1248	0	NADH-dependent phenylglyoxylate dehydrogenase subunit epsilon
Cspa_c30170	moeA2	3297145	3298170	1	0	0	1026	0	molybdopterin biosynthesis protein MoeA
Cspa_c30180	moeA3	3298174	3299379	1	0	0	1206	0	molybdopterin molybdenumtransferase MoeA
Cspa_c30290	Cspa_c30290	3307328	3308701	1	0	0	1374	0	TIGR00299 family protein
Cspa_c30350	Cspa_c30350	3313416	3315074	1	0	0	1659	0	ABC transporter ATP-binding protein
Cspa_c30390	Cspa_c30390	3319334	3320509	1	0	0	1176	0	amino acid/amide ABC transporter substrate-binding protein, HAAT family

Cspa_c30450	Cspa_c30450	3325371	3326636	1	0	0	1266	0	putative ferric reductase
Cspa_c30470	Cspa_c30470	3327910	3329019	-1	0	0	1110	0	superfamily II DNA and RNA helicase
Cspa_c30490	trpE	3330695	3332086	1	0	0	1392	0	anthranilate synthase component 1
Cspa_c30510	trpD	3332692	3333702	1	0	0	1011	0	anthranilate phosphoribosyltransferase TrpD
Cspa_c30640	Cspa_c30640	3344528	3346087	1	0	0	1560	0	hypothetical protein
Cspa_c30660	fbpC	3346855	3347901	-1	0	0	1047	0	Fe(3+) ions import ATP-binding protein FbpC 2
Cspa_c30690	Cspa_c30690	3349890	3351410	1	0	0	1521	0	[NiFe]-hydrogenase/urease maturation factor Ni(2+)-binding GTPase
Cspa_c30790	nifK3	3359866	3361230	1	0	0	1365	0	nitrogenase molybdenum-iron protein beta chain
Cspa_c30810	nifB2	3362947	3365646	1	0	0	2700	0	FeMo cofactor biosynthesis protein NifB
Cspa_c30880	anfD	3370574	3372142	1	0	0	1569	0	nitrogenase iron-iron protein alpha chain
Cspa_c30900	anfK	3372523	3373926	1	0	0	1404	0	nitrogenase iron-iron protein beta chain
Cspa_c31070	Cspa_c31070	3386947	3388668	-1	0	0	1722	0	DNA invertase Pin-like site-specific recombinase
Cspa_c31080	Cspa_c31080	3388967	3391246	1	0	0	2280	0	TPR repeat-containing protein
Cspa_c31090	Cspa_c31090	3391512	3394376	1	0	0	2865	0	NTPase
Cspa_c31110	Cspa_c31110	3395556	3397808	1	0	0	2253	0	hypothetical protein
Cspa_c31160	Cspa_c31160	3402676	3404184	1	0	0	1509	0	hypothetical protein
Cspa_c31170	Cspa_c31170	3404497	3405885	1	0	0	1389	0	hypothetical protein
Cspa_c31730	Cspa_c31730	3446778	3448154	-1	0	0	1377	0	hypothetical protein
Cspa_c31820	Cspa_c31820	3457487	3458845	1	0	0	1359	0	FAD/FMN-containing dehydrogenase
Cspa_c31930	pldB	3468066	3469163	1	0	0	1098	0	lysophospholipase L2

Cspa_c31940	Cspa_c31940	3469418	3470848	1	0	0	1431	0	drug resistance transporter, EmrB/QacA subfamily
Cspa_c31960	Cspa_c31960	3472023	3473132	1	0	0	1110	0	putative amidase domain-containing protein
Cspa_c32040	ilvB2	3480211	3481893	-1	0	0	1683	0	acetolactate synthase isozyme 1 large subunit IlvB
Cspa_c32070	Cspa_c32070	3483337	3485274	1	0	0	1938	0	ABC transporter ATP-binding protein
Cspa_c32090	Cspa_c32090	3485491	3488757	1	0	0	3267	0	superfamily II DNA/RNA helicase, SNF2 family
Cspa_c32200	Cspa_c32200	3496724	3497779	-1	0	0	1056	0	transcriptional regulator, AraC family
Cspa_c32210	Cspa_c32210	3498196	3499569	1	0	0	1374	0	sugar (glycoside-pentoside-hexuronide) transporter
Cspa_c32220	ramA2	3499816	3502518	1	0	0	2703	0	alpha-L-rhamnosidase RamA
Cspa_c32270	gapN	3506965	3508425	1	0	0	1461	0	NADP-dependent glyceraldehyde-3-phosphate dehydrogenase GapN
Cspa_c32290	Cspa_c32290	3511669	3512892	1	0	0	1224	0	FliB family protein
Cspa_c32310	Cspa_c32310	3513616	3514839	1	0	0	1224	0	SAM dependent methyltransferase
Cspa_c32330	mutL	3515222	3517198	1	0	0	1977	0	DNA mismatch repair protein MutL
Cspa_c32360	Cspa_c32360	3518567	3519856	1	0	0	1290	0	aluminum resistance family protein
Cspa_c32490	mglB7	3529242	3530285	1	0	0	1044	0	D-galactose-binding periplasmic protein MglB
Cspa_c32570	Cspa_c32570	3536722	3537780	1	0	0	1059	0	transposase
Cspa_c32580	ftsH3	3538298	3540109	-1	0	0	1812	0	ATP-dependent zinc metalloprotease FtsH
Cspa_c32590	fold2	3540230	3541390	-1	0	0	1161	0	putative Zn-dependent hydrolases of the beta-lactamase Fold protein
Cspa_c32670	Cspa_c32670	3548373	3550916	-1	0	0	2544	0	glycosyl hydrolases family 2, sugar binding domain protein
Cspa_c32920	dltS	3578065	3579315	1	0	0	1251	0	sensor protein DltS
Cspa_c32940	Cspa_c32940	3579611	3580930	-1	0	0	1320	0	putative ATPase

Cspa_c33090	Cspa_c33090	3596492	3597562	-1	0	0	1071	0	hypothetical protein
Cspa_c33110	Cspa_c33110	3598771	3599850	-1	0	0	1080	0	L-Ala-D/L-Glu epimerase
Cspa_c33160	Cspa_c33160	3605818	3607167	1	0	0	1350	0	iron only hydrogenase large subunit, C-terminal domain protein
Cspa_c33260	Cspa_c33260	3616030	3617763	-1	0	0	1734	0	subtilase family
Cspa_c33390	Cspa_c33390	3635618	3637159	-1	0	0	1542	0	Na:galactoside symporter family permease
Cspa_c33410	uxaA	3638971	3640458	-1	0	0	1488	0	altronate hydrolase UxaA
Cspa_c33420	uxaB	3640729	3642177	-1	0	0	1449	0	altronate oxidoreductase UxaB
Cspa_c33440	uxaC2	3643305	3644705	-1	0	0	1401	0	uronate isomerase UxaC
Cspa_c33450	Cspa_c33450	3646847	3646259	-1	0	0	1413	0	sugar (glycoside-pentoside-hexuronide) transporter
Cspa_c33550	recQ2	3654367	3656823	-1	0	0	2457	0	putative ATP-dependent DNA helicase RecQ
Cspa_c33570	dbpA	3658005	3659447	-1	0	0	1443	0	ATP-dependent RNA helicase DbpA
Cspa_c33580	Cspa_c33580	3659804	3661027	1	0	0	1224	0	superfamily II DNA and RNA helicase
Cspa_c33760	leuA3	3676800	3678470	-1	0	0	1671	0	2-isopropylmalate synthase LeuA
Cspa_c33790	Cspa_c33790	3681215	3682660	-1	0	0	1446	0	ABC-type antimicrobial peptide transport system, permease component
Cspa_c33810	Cspa_c33810	3683408	3684865	-1	0	0	1458	0	multidrug resistance efflux pump
Cspa_c33820	Cspa_c33820	3684868	3686043	-1	0	0	1176	0	hypothetical protein
Cspa_c34060	Cspa_c34060	3703731	3704771	-1	0	0	1041	0	hypothetical protein
Cspa_c34070	Cspa_c34070	3705376	3706668	-1	0	0	1293	0	major facilitator superfamily
Cspa_c34350	hemN2	3732251	3734083	-1	0	0	1833	0	oxygen-independent coproporphyrinogen-III oxidase HemN
Cspa_c34360	Cspa_c34360	3734192	3735193	-1	0	0	1002	0	hypothetical protein

Cspa_c34550	Cspa_c34550	3747813	3748955	-1	0	0	1143	0	hypothetical protein
Cspa_c34600	Cspa_c34600	3751462	3752574	-1	0	0	1113	0	hypothetical protein
Cspa_c34680	pbpX	3757179	3758195	-1	0	0	1017	0	putative penicillin-binding protein PbpX
Cspa_c34860	Cspa_c34860	3771931	3774231	-1	0	0	2301	0	acetyl-CoA carboxylase, biotin carboxylase
Cspa_c35010	Cspa_c35010	3786846	3788006	1	0	0	1161	0	putative transposase
Cspa_c35060	Cspa_c35060	3791368	3792678	-1	0	0	1311	0	flagellar capping protein
Cspa_c35130	Cspa_c35130	3798978	3800120	-1	0	0	1143	0	aspartate/tyrosine/aromatic aminotransferase
Cspa_c35210	folC	3807176	3808492	-1	0	0	1317	0	folylpolyglutamate synthase FolC
Cspa_c35220	Cspa_c35220	3808871	3810064	-1	0	0	1194	0	aspartate/tyrosine/aromatic aminotransferase
Cspa_c35350	Cspa_c35350	3824089	3825285	-1	0	0	1197	0	hypothetical protein DUF2974
Cspa_c35430	Cspa_c35430	3832977	3834035	-1	0	0	1059	0	amino acid/amide ABC transporter membrane protein 2, HAAT family
Cspa_c35450	Cspa_c35450	3834956	3836119	-1	0	0	1164	0	amino acid/amide ABC transporter substrate-binding protein, HAAT family
Cspa_c35630	Cspa_c35630	3850138	3851391	-1	0	0	1254	0	major facilitator superfamily MFS_1
Cspa_c35680	Cspa_c35680	3855773	3857020	1	0	0	1248	0	hypothetical protein
Cspa_c35730	Cspa_c35730	3861821	3863500	-1	0	0	1680	0	oligoendopeptidase, M3 family
Cspa_c35790	Cspa_c35790	3868338	3869573	-1	0	0	1236	0	hypothetical protein
Cspa_c35800	Cspa_c35800	3869702	3871114	-1	0	0	1413	0	signal transduction histidine kinase
Cspa_c35860	Cspa_c35860	3876644	3878134	-1	0	0	1491	0	hypothetical protein
Cspa_c35910	fabF3	3884095	3885330	-1	0	0	1236	0	3-oxoacyl-[acyl-carrier-protein] synthase 2
Cspa_c36040	Cspa_c36040	3897395	3898567	-1	0	0	1173	0	malate dehydrogenase

Cspa_c36090	aruS	3904703	3906598	-1	0	0	1896	0	sensor histidine kinase AruS
Cspa_c36110	Cspa_c36110	3908777	3910999	-1	0	0	2223	0	methyl-accepting chemotaxis protein
Cspa_c36170	rsgA2	3918950	3920023	-1	0	0	1074	0	putative ribosome biogenesis GTPase RsgA 1
Cspa_c36350	Cspa_c36350	3935154	3936551	-1	0	0	1398	0	hypothetical protein
Cspa_c36430	Cspa_c36430	3947688	3948782	-1	0	0	1095	0	SEFIR domain-containing protein
Cspa_c36660	Cspa_c36660	3967171	3968616	-1	0	0	1446	0	hypothetical protein
Cspa_c36680	Cspa_c36680	3969056	3970474	-1	0	0	1419	0	phage late control gene D protein
Cspa_c36750	Cspa_c36750	3973084	3974244	-1	0	0	1161	0	phage replisome organizer, putative, N-terminal region
Cspa_c36800	Cspa_c36800	3976101	3977279	1	0	0	1179	0	tyrosine recombinase XerC-like protein
Cspa_c36840	Cspa_c36840	3985841	3986986	-1	0	0	1146	0	hypothetical protein
Cspa_c37070	Cspa_c37070	4003462	4005129	1	0	0	1668	0	zinc metalloprotease
Cspa_c37110	Cspa_c37110	4008212	4009606	-1	0	0	1395	0	hypothetical protein
Cspa_c37170	Cspa_c37170	4018886	4019932	1	0	0	1047	0	hypothetical protein
Cspa_c37330	Cspa_c37330	4036150	4037391	-1	0	0	1242	0	NADPH-dependent glutamate synthase
Cspa_c37340	pckA1	4037494	4039062	-1	0	0	1569	0	phosphoenolpyruvate carboxykinase [ATP] 1
Cspa_c37350	Cspa_c37350	4039412	4040806	-1	0	0	1395	0	putative permease
Cspa_c37360	Cspa_c37360	4040866	4041975	-1	0	0	1110	0	putative 3-methylitaconate isomerase
Cspa_c37370	leuC3	4042200	4044134	-1	0	0	1935	0	3-isopropylmalate dehydratase large subunit 1
Cspa_c37420	citC1	4047557	4048669	-1	0	0	1113	0	[citrate [pro-3S]-lyase] ligase CitC
Cspa_c37470	alsS2	4054416	4056086	-1	0	0	1671	0	acetolactate synthase AlsS

Cspa_c37480	mglB8	4056632	4057684	1	0	0	1053	0	D-galactose-binding periplasmic protein MglB
Cspa_c37500	Cspa_c37500	4058229	4060046	-1	0	0	1818	0	RNA-directed DNA polymerase
Cspa_c37590	Cspa_c37590	4066965	4068356	-1	0	0	1392	0	4Fe-4S ferredoxin, iron-sulfur binding domain protein
Cspa_c37670	apbE3	4075117	4076178	-1	0	0	1062	0	thiamine biosynthesis lipoprotein ApbE
Cspa_c37720	mco	4078766	4080250	1	0	0	1485	0	multicopper oxidase Mco
Cspa_c37750	Cspa_c37750	4085123	4086496	-1	0	0	1374	0	signal transduction histidine kinase
Cspa_c37770	Cspa_c37770	4087622	4089217	1	0	0	1596	0	NHL repeat containing protein
Cspa_c37830	Cspa_c37830	4096102	4097460	-1	0	0	1359	0	hypothetical protein
Cspa_c37900	mrdB	4101543	4102658	1	0	0	1116	0	Rod shape-determining protein RodA
Cspa_c38070	bglP5	4116736	4118631	-1	0	0	1896	0	PTS system beta-glucoside-specific EIIBCA component BglP
Cspa_c38110	hydF	4120676	4121911	-1	0	0	1236	0	iron-only hydrogenase maturation protein HydF
Cspa_c38120	aspA	4122287	4123684	1	0	0	1398	0	aspartate ammonia-lyase AspA
Cspa_c38130	mglB9	4123859	4124902	-1	0	0	1044	0	D-galactose-binding periplasmic protein MglB
Cspa_c38150	fucO	4126193	4127347	-1	0	0	1155	0	lactaldehyde reductase FucO
Cspa_c38170	rhaA	4128484	4129740	-1	0	0	1257	0	L-rhamnose isomerase RhaA
Cspa_c38180	rhaB	4129818	4131215	-1	0	0	1398	0	rhamnulokinase RhaB
Cspa_c38200	Cspa_c38200	4132885	4134273	-1	0	0	1389	0	putative peptidoglycan-binding domain-containing protein
Cspa_c38220	Cspa_c38220	4135283	4136302	-1	0	0	1020	0	putative cation transporter
Cspa_c38240	Cspa_c38240	4137159	4138745	1	0	0	1587	0	ammonium transporter
Cspa_c38420	Cspa_c38420	4155325	4156398	-1	0	0	1074	0	hypothetical protein

Cspa_c38430	Cspa_c38430	4156802	4157950	-1	0	0	1149	0	FAD-dependent pyridine nucleotide-disulfide oxidoreductase
Cspa_c38440	cydC1	4158205	4159929	-1	0	0	1725	0	ATP-binding/permease protein CydC
Cspa_c38450	cydC2	4159926	4161641	-1	0	0	1716	0	ATP-binding/permease protein CydC
Cspa_c38460	cydB	4161659	4162663	-1	0	0	1005	0	cytochrome d ubiquinol oxidase subunit 2
Cspa_c38510	Cspa_c38510	4167167	4168198	-1	0	0	1032	0	hypothetical protein
Cspa_c38520	hndC1	4169007	4170110	-1	0	0	1104	0	NADP-reducing hydrogenase subunit HndC
Cspa_c38530	Cspa_c38530	4170125	4172215	-1	0	0	2091	0	ferredoxin
Cspa_c38610	Cspa_c38610	4181943	4183664	-1	0	0	1722	0	methyl-accepting chemotaxis protein
Cspa_c38830	Cspa_c38830	4203582	4204829	-1	0	0	1248	0	ammonium transporter
Cspa_c38900	Cspa_c38900	4211670	4213100	-1	0	0	1431	0	glycoside hydrolase family 25
Cspa_c38970	Cspa_c38970	4218846	4220003	-1	0	0	1158	0	YibE/F family protein
Cspa_c38990	Cspa_c38990	4221948	4223729	-1	0	0	1782	0	ABC transporter permease/ATP-binding protein
Cspa_c39240	ubiD	4250553	4252043	-1	0	0	1491	0	3-octaprenyl-4-hydroxybenzoate carboxy-lyase UbiD
Cspa_c39330	Cspa_c39330	4262685	4263779	-1	0	0	1095	0	response regulator receiver protein
Cspa_c39390	tauA	4268423	4269463	-1	0	0	1041	0	taurine-binding periplasmic protein TauA
Cspa_c39560	Cspa_c39560	4283650	4284861	-1	0	0	1212	0	hypothetical protein
Cspa_c39810	Cspa_c39810	4311484	4312680	-1	0	0	1197	0	putative neuraminidase
Cspa_c39820	Cspa_c39820	4313018	4314430	-1	0	0	1413	0	Na+/proline symporter
Cspa_c39850	Cspa_c39850	4316845	4318005	-1	0	0	1161	0	alcohol dehydrogenase, class IV
Cspa_c39970	fruA2	4332525	4333601	-1	0	0	1077	0	PTS system fructose-specific EIIBBC component FruA

Cspa_c39990	manR	4334016	4336415	-1	0	0	2400	0	putative transcriptional regulator ManR
Cspa_c40050	kipA	4344111	4345112	-1	0	0	1002	0	KipI antagonist
Cspa_c40160	braC	4354618	4355799	-1	0	0	1182	0	leucine-, isoleucine-, valine-, threonine-, and alanine-binding protein BraC
Cspa_c40180	Cspa_c40180	4357425	4358624	-1	0	0	1200	0	methylase involved in ubiquinone/menaquinone biosynthesis
Cspa_c40680	hypF	4401346	4403628	-1	0	0	2283	0	carbamoyltransferase HypF2
Cspa_c40700	hypD	4404951	4406033	-1	0	0	1083	0	hydrogenase isoenzymes formation protein HypD
Cspa_c40820	Cspa_c40820	4413814	4414914	-1	0	0	1101	0	plasmid pRiA4b ORF-3-like protein
Cspa_c41010	Cspa_c41010	4435224	4436486	-1	0	0	1263	0	hypothetical protein
Cspa_c41030	Cspa_c41030	4438316	4439605	-1	0	0	1290	0	hypothetical protein
Cspa_c41060	Cspa_c41060	4440191	4441498	1	0	0	1308	0	alpha-galactosidase
Cspa_c41050	Cspa_c41050	4441482	4443494	-1	0	0	2013	0	alpha-glucosidase, family 31 of glycosyl hydrolase
Cspa_c41220	Cspa_c41220	4458963	4460012	-1	0	0	1050	0	putative transcriptional regulator
Cspa_c41280	Cspa_c41280	4464846	4466207	1	0	0	1362	0	hypothetical protein
Cspa_c41320	Cspa_c41320	4470385	4471608	-1	0	0	1224	0	amidase, hydantoinase/carbamoylase family
Cspa_c41330	preA	4471758	4473260	-1	0	0	1503	0	NAD-dependent dihydropyrimidine dehydrogenase subunit PreA
Cspa_c41340	Cspa_c41340	4473351	4474718	-1	0	0	1368	0	D-hydantoinase
Cspa_c41350	Cspa_c41350	4474789	4475862	-1	0	0	1074	0	ABC-type nitrate/sulfonate/bicarbonate transport system, periplasmic component
Cspa_c41390	Cspa_c41390	4478645	4479799	-1	0	0	1155	0	alcohol dehydrogenase, class IV
Cspa_c41430	Cspa_c41430	4483982	4485415	-1	0	0	1434	0	FAD/FMN-containing dehydrogenase

Cspa_c41440	lctP2	4485508	4487043	-1	0	0	1536	0	L-lactate permease LctP
Cspa_c41500	Cspa_c41500	4492649	4494565	1	0	0	1917	0	Na+/H+ antiporter
Cspa_c41590	ftsH4	4502467	4504200	1	0	0	1734	0	ATP-dependent zinc metalloprotease FtsH 1
Cspa_c41640	Cspa_c41640	4512038	4513084	-1	0	0	1047	0	hypothetical protein
Cspa_c41680	ppk	4518763	4520817	-1	0	0	2055	0	polyphosphate kinase Ppk
Cspa_c41800	Cspa_c41800	4537478	4538719	-1	0	0	1242	0	putative zinc-dependent protease
Cspa_c41980	Cspa_c41980	4553791	4555044	-1	0	0	1254	0	periplasmic protease
Cspa_c42160	Cspa_c42160	4567581	4568597	-1	0	0	1017	0	putative beta-xylosidase
Cspa_c42220	araN	4575920	4577269	-1	0	0	1350	0	putative arabinose-binding protein AraN
Cspa_c42330	ansB2	4588553	4589983	1	0	0	1431	0	aspartate ammonia-lyase AnsB
Cspa_c42400	Cspa_c42400	4594722	4596401	1	0	0	1680	0	methyl-accepting chemotaxis sensory transducer
Cspa_c42410	Cspa_c42410	4596440	4597615	1	0	0	1176	0	ABC transporter, periplasmic substrate-binding protein
Cspa_c42420	Cspa_c42420	4597794	4599131	-1	0	0	1338	0	xanthine/uracil/vitamin C permease
Cspa_c42450	Cspa_c42450	4600434	4601789	-1	0	0	1356	0	ABC-type sugar transport system, periplasmic component
Cspa_c42480	Cspa_c42480	4603600	4604778	-1	0	0	1179	0	ABC-type sugar transport system, ATPase component
Cspa_c42590	Cspa_c42590	4611744	4612802	-1	0	0	1059	0	exopolysaccharide biosynthesis protein
Cspa_c42880	Cspa_c42880	4641355	4642359	-1	0	0	1005	0	sugar phosphate isomerases/epimerase
Cspa_c42970	Cspa_c42970	4656329	4658320	-1	0	0	1992	0	methyl-accepting chemotaxis sensory transducer with cache sensor
Cspa_c43040	Cspa_c43040	4664219	4665328	-1	0	0	1110	0	hypothetical protein
Cspa_c43080	Cspa_c43080	4669912	4671618	-1	0	0	1707	0	hypothetical protein

Cspa_c43100	xylB2	4674980	4676617	-1	0	0	1638	0	xylosidase/arabinosidase XylB
Cspa_c43120	Cspa_c43120	4679264	4680661	-1	0	0	1398	0	sugar (glycoside-pentoside-hexuronide) transporter
Cspa_c43130	Cspa_c43130	4681485	4683080	-1	0	0	1596	0	response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain
Cspa_c43150	Cspa_c43150	4684981	4686507	-1	0	0	1527	0	ABC-type sugar transport system, periplasmic component
Cspa_c43220	Cspa_c43220	4697138	4698187	-1	0	0	1050	0	oxidoreductase
Cspa_c43260	Cspa_c43260	4700689	4701792	-1	0	0	1104	0	putative dehydrogenase
Cspa_c43330	spoVAD2	4707914	4708915	1	0	0	1002	0	stage V sporulation protein AD
Cspa_c43490	Cspa_c43490	4722845	4723873	-1	0	0	1029	0	hypothetical protein
Cspa_c43550	egsA	4729083	4730144	-1	0	0	1062	0	glycerol-1-phosphate dehydrogenase [NAD(P)+]
Cspa_c43600	Cspa_c43600	4733180	4734370	-1	0	0	1191	0	PTS system, lactose/cellobiose family IIC component
Cspa_c43650	Cspa_c43650	4738523	4739758	-1	0	0	1236	0	arabinose efflux permease
Cspa_c43680	Cspa_c43680	4742656	4744062	-1	0	0	1407	0	signal transduction histidine kinase
Cspa_c43770	alr	4753709	4754872	1	0	0	1164	0	alanine racemase Alr
Cspa_c43910	Cspa_c43910	4772171	4773433	-1	0	0	1263	0	Mn2+ and Fe2+ transporters of the NRAMP family
Cspa_c44000	cotA	4780094	4781896	1	0	0	1803	0	spore coat protein A
Cspa_c44130	Cspa_c44130	4795183	4796196	1	0	0	1014	0	Zn-dependent hydrolase
Cspa_c44140	leuC4	4796284	4798221	-1	0	0	1938	0	3-isopropylmalate dehydratase large subunit 1
Cspa_c44150	citN	4798367	4799677	-1	0	0	1311	0	citrate transporter CitN
Cspa_c44360	Cspa_c44360	4838808	4841018	-1	0	0	2211	0	ABC-type transport system, involved in lipoprotein release, permease component

Cspa_c44450	Cspa_c44450	4853587	4854876	-1	0	0	1290	0	hypothetical protein
Cspa_c44490	tapC	4857398	4858600	-1	0	0	1203	0	type IV pilus assembly protein TapC
Cspa_c44520	pilT	4861469	4862506	-1	0	0	1038	0	twitching mobility protein PilT
Cspa_c44710	Cspa_c44710	4882562	4883992	-1	0	0	1431	0	transcriptional regulator, GntR family with aminotransferase domain containing protein
Cspa_c44740	hflX	4885896	4887686	-1	0	0	1791	0	GTPase HflX
Cspa_c44960	hag2	4909831	4911564	-1	0	0	1734	0	flagellin
Cspa_c45320	pseI	4946847	4947899	-1	0	0	1053	0	pseudaminic acid synthase PseI
Cspa_c46010	murG	5012639	5013712	-1	0	0	1074	0	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase MurG
Cspa_c46020	recJ2	5013909	5015675	-1	0	0	1767	0	single-stranded-DNA-specific exonuclease RecJ
Cspa_c46170	phnW	5032818	5033951	-1	0	0	1134	0	2-aminoethylphosphonate--pyruvate transaminase PhnW
Cspa_c46190	ppm	5035189	5036487	-1	0	0	1299	0	phosphoenolpyruvate phosphomutase Ppm
Cspa_c46300	Cspa_c46300	5047326	5049344	-1	0	0	2019	0	methyl-accepting chemotaxis sensory transducer with cache sensor
Cspa_c46410	splB	5061334	5062380	-1	0	0	1047	0	spore photoproduct lyase SplB
Cspa_c46430	nspC	5063015	5064157	-1	0	0	1143	0	carboxynorspermidine/carboxyspermidine decarboxylase NspC
Cspa_c46590	Cspa_c46590	5081448	5082866	1	0	0	1419	0	putative DNA modification/repair radical SAM protein
Cspa_c46670	mglA	5091483	5092982	-1	0	0	1500	0	galactose/methyl galactoside import ATP-binding protein MglA
Cspa_c46680	mglB11	5093138	5094208	-1	0	0	1071	0	D-galactose-binding periplasmic protein MglB
Cspa_c46860	xylH2	5114592	5115806	-1	0	0	1215	0	xylose transport system permease protein XylH

Cspa_c46870	Cspa_c46870	5115808	5117358	-1	0	0	1551	0	ABC-type sugar transport system, ATPase component
Cspa_c46880	Cspa_c46880	5117428	5118558	-1	0	0	1131	0	ABC-type xylose transport system, periplasmic component
Cspa_c46940	araA	5124203	5125669	1	0	0	1467	0	L-arabinose isomerase AraA
Cspa_c46980	Cspa_c46980	5129838	5131358	-1	0	0	1521	0	monosaccharide ABC transporter ATP-binding protein, CUT2 family
Cspa_c47020	mro	5134973	5136028	1	0	0	1056	0	aldose 1-epimerase Mro
Cspa_c47040	Cspa_c47040	5137737	5139053	-1	0	0	1317	0	PTS system, lactose/cellobiose family IIC component
Cspa_c47200	argS2	5157993	5159756	-1	0	0	1764	0	arginine--tRNA ligase ArgS
Cspa_c47450	Cspa_c47450	5188837	5189898	1	0	0	1062	0	oxidoreductase
Cspa_c47480	fccA2	5192160	5193608	1	0	0	1449	0	fumarate reductase flavoprotein subunit FccA
Cspa_c47560	nuoG	5201629	5203758	-1	0	0	2130	0	NADH-quinone oxidoreductase subunit G 2
Cspa_c47580	gyaR	5205390	5206439	-1	0	0	1050	0	glyoxylate reductase GyaR
Cspa_c47590	xylB3	5206452	5207999	-1	0	0	1548	0	D-xylulose kinase XylB
Cspa_c47760	bglA5	5229310	5232288	-1	0	0	2979	0	beta-glucosidase A
Cspa_c47780	Cspa_c47780	5233827	5235023	-1	0	0	1197	0	hypothetical protein
Cspa_c47840	Cspa_c47840	5243199	5244539	-1	0	0	1341	0	hypothetical protein
Cspa_c47890	Cspa_c47890	5251167	5252495	-1	0	0	1329	0	putative peptidase
Cspa_c48000	Cspa_c48000	5263951	5264991	1	0	0	1041	0	ABC-type hemin transport system, periplasmic component
Cspa_c48010	Cspa_c48010	5265370	5267172	1	0	0	1803	0	hypothetical protein
Cspa_c48070	Cspa_c48070	5274439	5275551	-1	0	0	1113	0	methionine synthase, vitamin-B12 independent
Cspa_c48120	cas3	5282852	5285068	-1	0	0	2217	0	CRISPR-associated nuclease/helicase Cas3

Cspa_c48240	glgD1	5299522	5300667	-1	0	0	1146	0	glycogen biosynthesis protein GlgD
Cspa_c48260	Cspa_c48260	5302605	5303870	-1	0	0	1266	0	PTS system, lactose/cellobiose family IIC component
Cspa_c48300	Cspa_c48300	5308842	5310137	-1	0	0	1296	0	hypothetical protein DUF4038
Cspa_c48310	Cspa_c48310	5310418	5311842	-1	0	0	1425	0	L-fucose isomerase
Cspa_c48370	Cspa_c48370	5319534	5321501	-1	0	0	1968	0	hypothetical protein
Cspa_c48440	Cspa_c48440	5329273	5331018	1	0	0	1746	0	integral membrane sensor signal transduction histidine kinase
Cspa_c48500	xsa1	5337649	5339127	-1	0	0	1479	0	alpha-N-arabinofuranosidase 2
Cspa_c48520	yesM2	5340814	5342628	-1	0	0	1815	0	sensor histidine kinase YesM
Cspa_c48540	xylB4	5344689	5346236	-1	0	0	1548	0	xylosidase/arabinosidase XylB
Cspa_c48550	Cspa_c48550	5346249	5347385	-1	0	0	1137	0	endoglucanase Y
Cspa_c48580	Cspa_c48580	5349339	5351009	-1	0	0	1671	0	extracellular solute-binding protein family 1
Cspa_c48680	Cspa_c48680	5367746	5369053	-1	0	0	1308	0	carbohydrate ABC transporter substrate-binding protein, CUT1 family
Cspa_c48850	Cspa_c48850	5390534	5392036	-1	0	0	1503	0	major facilitator superfamily MFS_1
Cspa_c48860	uxaC3	5392085	5393488	-1	0	0	1404	0	uronate isomerase UxaC
Cspa_c48880	uxuA3	5395503	5396564	-1	0	0	1062	0	mannonate dehydratase 1
Cspa_c48950	Cspa_c48950	5405270	5406640	-1	0	0	1371	0	signal transduction histidine kinase
Cspa_c48970	Cspa_c48970	5407344	5408612	-1	0	0	1269	0	hypothetical protein
Cspa_c49190	Cspa_c49190	5430856	5432523	1	0	0	1668	0	drug resistance transporter, EmrB/QacA subfamily
Cspa_c49200	Cspa_c49200	5432686	5434401	-1	0	0	1716	0	methyl-accepting chemotaxis protein
Cspa_c49220	Cspa_c49220	5436260	5438224	1	0	0	1965	0	alpha-L-arabinofuranosidase

Cspa_c49250	mtbA	5441780	5442823	-1	0	0	1044	0	methylcobamide:CoM methyltransferase MtbA
Cspa_c49400	Cspa_c49400	5459625	5461079	-1	0	0	1455	0	drug resistance transporter, EmrB/QacA subfamily
Cspa_c49450	uxuA4	5467072	5468133	-1	0	0	1062	0	mannonate dehydratase 1
Cspa_c49570	bglH4	5478781	5480187	-1	0	0	1407	0	aryl-phospho-beta-D-glucosidase BglH
Cspa_c49580	bglP8	5480190	5482046	-1	0	0	1857	0	PTS system beta-glucoside-specific EIIBCA component BglP
Cspa_c49620	malH2	5484584	5485909	-1	0	0	1326	0	maltose-6'-phosphate glucosidase MalH
Cspa_c49760	Cspa_c49760	5501840	5503210	-1	0	0	1371	0	putative efflux protein, MATE family
Cspa_c49990	Cspa_c49990	5525175	5526314	-1	0	0	1140	0	amidohydrolase
Cspa_c50270	Cspa_c50270	5551087	5552268	-1	0	0	1182	0	amidohydrolase 3
Cspa_c50350	Cspa_c50350	5559027	5560055	-1	0	0	1029	0	D-alanine-D-alanine ligase
Cspa_c50400	Cspa_c50400	5562453	5564075	-1	0	0	1623	0	signal transduction histidine kinase regulating citrate/malate metabolism
Cspa_c50420	maeB	5565449	5566621	-1	0	0	1173	0	NADP-dependent malic enzyme MaeB
Cspa_c50480	msbA2	5571461	5573212	-1	0	0	1752	0	lipid A export ATP-binding/permease protein MsbA
Cspa_c50490	bioA	5573645	5574985	-1	0	0	1341	0	adenosylmethionine-8-amino-7-oxononanoate aminotransferase BioA
Cspa_c50620	Cspa_c50620	5587376	5588704	-1	0	0	1329	0	6-phospho-alpha-glucosidase 2
Cspa_c50630	Cspa_c50630	5588726	5590321	-1	0	0	1596	0	PTS system alpha-glucoside-specific EIICB component
Cspa_c51060	eno2	5642615	5643907	-1	0	0	1293	0	enolase Eno
Cspa_c51120	gap	5651670	5652671	-1	0	0	1002	0	glyceraldehyde-3-phosphate dehydrogenase Gap
Cspa_c51130	cggR	5652749	5653798	-1	0	0	1050	0	central glycolytic genes regulator
Cspa_c51390	yhbH	5682823	5683995	-1	0	0	1173	0	sporulation protein YhbH

Cspa_c51500	Cspa_c51500	5693624	5695477	1	0	0	1854	0	ABC-type multidrug transport system, ATPase and permease components
Cspa_c51740	Cspa_c51740	5723905	5725116	-1	0	0	1212	0	carbohydrate ABC transporter substrate-binding protein, CUT1 family
Cspa_c52080	rrfF	5763217	5764717	-1	0	0	1501	0	
Cspa_c52250	rrfG	5788910	5790410	-1	0	0	1501	0	
Cspa_c52380	mreB2	5804017	5805033	-1	0	0	1017	0	Rod shape-determining protein MreB
Cspa_c52540	mppE	5823847	5824980	-1	0	0	1134	0	putative metallophosphoesterase MppE
Cspa_c52660	Cspa_c52660	5834555	5835775	-1	0	0	1221	0	transposase, mutator type
Cspa_c52780	Cspa_c52780	5850607	5851623	-1	0	0	1017	0	putative glycosylase
Cspa_c52840	Cspa_c52840	5857328	5859073	-1	0	0	1746	0	polygalacturonase
Cspa_c52900	dgoT	5864887	5866191	-1	0	0	1305	0	D-galactonate transporter DgoT
Cspa_c52930	Cspa_c52930	5868823	5870745	-1	0	0	1923	0	NADH-flavin oxidoreductase, Old yellow enzyme family
Cspa_c52970	Cspa_c52970	5873939	5875948	-1	0	0	2010	0	NADH-flavin oxidoreductase, Old yellow enzyme family
Cspa_c53000	Cspa_c53000	5878173	5879672	-1	0	0	1500	0	3-octaprenyl-4-hydroxybenzoate carboxy-lyase
Cspa_c53150	leuS	5895345	5897795	1	0	0	2451	0	leucine--tRNA ligase LeuS
Cspa_c53500	Cspa_c53500	5928473	5929729	-1	0	0	1257	0	glycosyl transferase, group 1
Cspa_c53510	galE3	5929804	5930817	-1	0	0	1014	0	UDP-glucose 4-epimerase GalE
Cspa_c53520	Cspa_c53520	5931138	5932202	1	0	0	1065	0	UDP-N-acetylglucosamine 2-epimerase
Cspa_c53550	epsC	5934549	5936411	-1	0	0	1863	0	putative polysaccharide biosynthesis protein EpsC
Cspa_c53660	Cspa_c53660	5950384	5951724	-1	0	0	1341	0	PTS system, lactose/cellobiose family IIC component
Cspa_c53710	bgxA2	5956806	5959223	-1	0	0	2418	0	periplasmic beta-glucosidase/beta-xylosidase BgxA

Cspa_c53720	Cspa_c53720	5959714	5961174	-1	0	0	1461	0	beta-glucosidase/6-phospho-beta- glucosidase/beta-galactosidase
Cspa_c53740	Cspa_c53740	5962242	5963606	-1	0	0	1365	0	alpha-galactosidase/6-phospho-beta-glucosidase, family 4 of glycosyl hydrolase
Cspa_c53860	Cspa_c53860	5978940	5980790	-1	0	0	1851	0	choline/ethanolamine kinase
Cspa_c54020	Cspa_c54020	6013703	6014830	1	0	0	1128	0	hypothetical protein
Cspa_c54050	Cspa_c54050	6018269	6019951	-1	0	0	1683	0	hypothetical protein
Cspa_c54060	Cspa_c54060	6019985	6021685	-1	0	0	1701	0	hypothetical protein DUF1703
Cspa_c54540	Cspa_c54540	6089741	6090916	1	0	0	1176	0	cation diffusion facilitator family transporter
Cspa_c54600	helD3	6096607	6098778	-1	0	0	2172	0	helicase IV
Cspa_c54680	alaXL	6107381	6108580	1	0	0	1200	0	alanyl-tRNA editing protein AlaX-L
Cspa_c54800	Cspa_c54800	6123544	6125094	-1	0	0	1551	0	methyl-accepting chemotaxis sensory transducer
Cspa_c54850	cheB2	6129421	6130449	-1	0	0	1029	0	chemotaxis response regulator protein-glutamate methylesterase 2
Cspa_c54870	Cspa_c54870	6131381	6133081	-1	0	0	1701	0	methyl-accepting chemotaxis sensory transducer
Cspa_c54990	Cspa_c54990	6144029	6145117	-1	0	0	1089	0	peptidoglycan-binding domain 1 protein
Cspa_c55050	pyk2	6155760	6157181	-1	0	0	1422	0	pyruvate kinase Pyk
Cspa_c55130	uvrC2	6167373	6169247	-1	0	0	1875	0	UvrABC system protein C
Cspa_c55390	degT2	6196433	6197614	-1	0	0	1182	0	pleiotropic regulatory protein DegT
Cspa_c55560	cinA	6212994	6214235	-1	0	0	1242	0	putative competence-damage inducible protein CinA
Cspa_c55680	rrfH	6226852	6228352	-1	0	0	1501	0	
Cspa_c55780	glgD2	6238007	6239107	-1	0	0	1101	0	glycogen biosynthesis protein GlgD
Cspa_c55790	glgC2	6239134	6240294	-1	0	0	1161	0	glucose-1-phosphate adenylyltransferase GlgC

Cspa_c55830	glgB	6246375	6248894	-1	0	0	2520	0	1,4-alpha-glucan branching enzyme GlgB 1
Cspa_c55860	rrfJ	6252854	6254354	-1	0	0	1501	0	
Cspa_c56230	dnaX2	6294903	6296531	-1	0	0	1629	0	DNA polymerase III subunit gamma/tau
Cspa_c56490	malX	6330514	6332073	1	0	0	1560	0	PTS system maltose- and glucose-specific EIICB component MalX
Cspa_c57320	dnaC1	6410464	6411774	-1	0	0	1311	0	replicative DNA helicase DnaC
Cspa_c57550	rrfK	6437713	6439213	-1	0	0	1501	0	
Cspa_c57650	hppA	6447819	6449954	1	0	0	2136	0	K(+)-insensitive pyrophosphate-energized proton pump HppA
Cspa_c57850	gtfA2	6475833	6477302	1	0	0	1470	0	sucrose phosphorylase GtfA
Cspa_c135p00040	Cspa_c135p00040	2002	3462	-1	0	0	1461	0	hypothetical protein
Cspa_c135p00470	Cspa_c135p00470	69787	71181	1	0	0	1395	0	glycine rich protein
Cspa_c135p00950	Cspa_c135p00950	128152	129486	-1	0	0	1335	0	MoxR-like ATPase

Appendix V – Targeting spacer sequences

Table A5 List of targeting spacer sequences used in this study

Spacer sequence 5'-3'	Target
TTATATCTAATGAGACTAAAAAATTCAATTGTAAAAT	*agrA*
AAGGGATATATTATATTTTTAACTGCTCATGCTGAAT	*agrA*
TTACTGCTTTTCTAGTTTTTGCAGCTCTTAGAATTTT	*agrB*
ATGTATTATTATAAATTACATACTCTTTTTCGGAAAT	*agrB*
GCTCTAAACTTTACTTTTTTTGTAAAGCCTAATAAAA	*agrC*
ATATAGTTTAAAAGATTTTGTAGTAACATGGAATAAA	*agrC*
ATAATTATATGGGCACTTTTTCATTTTATGTAGAAAT	*agr'D'*
AAATGTATTTATACAAATTTTAAGCTTTGTTTTGGTG	*agr'D'*
TTATATCTAATGAGACTAAAAAATTCAATTGTAAAAT	*agrB-C* (whole operon)
TTACTGCTTTTCTAGTTTTTGCAGCTCTTAGAATTTT	*agrB-C* (whole operon)
ATATAGTTTAAAAGATTTTGTAGTAACATGGAATAAA	*agrB-C* (whole operon)

Appendix VI – All pathways identified by ShinyGO gene enrichment analysis

Table A6 List of all pathways identified by ShinyGO

FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway	Genes
1.44E-08	101	450	1.934691	Nucleotide-binding	dnaA recF gyrB1 addB addA ileS murC lysS murD fusA1 buk1 pcrA yfmM pyrG atpA argG trpS kdpB asnO nadE AGF54911.1 accC1 macB4 alaS ftsZ engA priA recG tmcAL smc ffh infB ribBA glnK AGF55492.1 metG3 msbA1 metN2 AGF55850.1 resE6 xylG AGF56192.1 AGF56195.1 AGF56209.1 AGF56258.1 hscC2 cysS2 ruvB pheS pheT murE3 murF AGF56620.1 accC2 AGF56708.1 rep cysA ackA2 AGF56796.1 AGF56808.1 fbpC AGF56830.1 nifK3 AGF56869.1 AGF56968.1 AGF56970.1 ftsH3 dltS recQ2 dbpA AGF57119.1 folC AGF57341.1 rsgA2 pckA1 citC1 AGF57535.1 rhaB cydC1 cydC2 AGF57659.1 ftsH4 ppk AGF58001.1 AGF58121.1 hflX mglA AGF58440.1 AGF58451.1 argS2 cas3 AGF58648.1 AGF58788.1 msbA2 gap AGF58901.1 leuS helD3 glgC2 dnaX2 dnaC1
3.60E-08	92	405	1.958103	ATP-binding	dnaA recF gyrB1 addB addA ileS murC lysS murD buk1 pcrA yfmM pyrG atpA argG trpS kdpB asnO nadE AGF54911.1 accC1 macB4 alaS priA recG tmcAL smc glnK AGF55492.1 metG3 msbA1 metN2 AGF55850.1 resE6 xylG AGF56192.1 AGF56195.1 AGF56209.1 AGF56258.1 hscC2 cysS2 ruvB pheS pheT murE3 murF AGF56620.1 accC2 AGF56708.1 rep cysA ackA2 AGF56796.1 AGF56808.1 fbpC AGF56830.1 nifK3 AGF56869.1 AGF56968.1 AGF56970.1 ftsH3 dltS recQ2 dbpA AGF57119.1 folC AGF57341.1 pckA1 citC1 AGF57535.1 rhaB cydC1 cydC2 AGF57659.1 ftsH4 ppk AGF58001.1 AGF58121.1 mglA AGF58440.1 AGF58451.1 argS2 cas3 AGF58648.1 AGF58788.1 msbA2 AGF58901.1 leuS helD3 glgC2 dnaX2 dnaC1
2.14E-05	22	54	3.511815	MFS transporter superfamily	mdtG AGF55212.1 AGF55603.1 AGF55696.1 AGF55976.1 tetA AGF56190.1 AGF56681.1 AGF56696.1 AGF56769.1 AGF56955.1 AGF56982.1 AGF57100.1 AGF57106.1 AGF57168.1

					AGF57324.1 AGF58065.1 AGF58118.1 AGF58638.1 AGF58672.1 AGF58693.1 dgoT
2.14E-05	79	375	1.815928	Signal	AGF54723.1 murE1 AGF54931.1 ssuA xynD1 engD2 pel AGF55505.1 AGF55538.1 cbpA celK1 celA2 Endo-beta-mannanase celH AGF55708.1 xynD2 AGF55982.1 AGF55983.1 AGF56009.1 AGF56046.1 AGF56070.1 AGF56074.1 apbE2 yjbG AGF56247.1 AGF56477.1 AGF56478.1 AGF56532.1 AGF56579.1 AGF56582.1 AGF56591.1 dacF AGF56610.1 AGF56614.1 AGF56616.1 AGF56617.1 AGF56648.1 potD AGF56705.1 sbp AGF56877.1 AGF56878.1 pldB AGF56957.1 mglB7 AGF57143.1 AGF57197.1 AGF57221.1 AGF57306.1 AGF57396.1 AGF57467.1 AGF57471.1 AGF57477.1 mglB8 mco AGF57543.1 AGF57584.1 AGF57650.1 braC AGF57888.1 AGF57951.1 araN AGF57994.1 AGF57998.1 AGF58061.1 AGF58068.1 mglB11 AGF58441.1 bglA5 AGF58531.1 AGF58542.1 AGF58553.1 AGF58554.1 AGF58611.1 AGF58621.1 Alpha-L-arabinofuranosidase AGF58925.1 bgxA2 AGF59147.1
2.21E-05	16	31	4.448986	Helicase	addB addA pcrA priA recG ruvB AGF56708.1 rep AGF56808.1 AGF56970.1 recQ2 dbpA AGF57119.1 cas3 helD3 dnaC1
3.36E-05	38	135	2.426345	Glycosidase	abgA1 pbg bglH1 xynD1 engD2 pulA bglA2 malL1 AGF55533.1 AGF55542.1 celK1 celA2 Endo-beta-mannanase celH xynD2 xynB3 bglC2 ascB ramA1 xynB4 AGF56747.1 xynB5 ramA2 AGF57650.1 AGF57858.1 Alpha-galactosidase AGF57969.1 xylB2 bglA5 xsa1 xylB4 AGF58608.1 bglH4 malH2 AGF58813.1 Polygalacturonase bgxA2 Alpha-galactosidase/6-phospho-beta-glucosidase
7.36E-05	20	50	3.447964	Major Facilitator Superfamily	mdtG AGF55212.1 AGF55603.1 AGF55696.1 AGF55976.1 tetA AGF56190.1 AGF56681.1 AGF56696.1 AGF56769.1 AGF56955.1 AGF57100.1 AGF57106.1 AGF57168.1 AGF57324.1 AGF58118.1 AGF58638.1 AGF58672.1 AGF58693.1 dgoT
8.12E-05	29	93	2.687929	Glycoside hydrolase superfamily	abgA1 pbg bglH1 engD2 pulA bglA2 malL1 AGF55533.1 celA2 Endo-beta-mannanase celH AGF55925.1 AGF56074.1 bglC2 ascB AGF56747.1 xynB5 AGF57650.1 AGF57858.1 Alpha-galactosidase bglA5 AGF58583.1 xsa1

					Alpha-L-arabinofuranosidase bglH4 bgxA2 AGF59117.1 glgB gtfA2
0.000254	17	41	3.574109	Major facilitator superfamily	mdtG AGF55212.1 AGF55603.1 AGF55696.1 AGF55976.1 tetA AGF56190.1 AGF56681.1 AGF56696.1 AGF56769.1 AGF56955.1 AGF57168.1 AGF57324.1 AGF58118.1 AGF58672.1 AGF58693.1 dgoT
0.000278	135	803	1.449175	Coiled coil	addA lysS rpoB fusA1 buk1 AGF54148.1 AGF54324.1 yfmM AGF54511.1 AGF54607.1 apbE1 nadB AGF54723.1 dnaG rpoD2 glgP1 AGF54764.1 AGF54793.1 AGF54849.1 pfl1 AGF54911.1 alaS AGF55043.1 priA rsmB smc nusA infB rny licR1 hisZ hisD Amidohydrolase AGF55356.1 AGF55405.1 AGF55406.1 AGF55413.1 AGF55492.1 malL1 metG3 AGF55673.1 eryC AGF55765.1 AGF55850.1 mtlA AGF55937.1 AGF55982.1 AGF55983.1 resE6 xylG metH3 AGF56195.1 AGF56209.1 yjbG AGF56222.1 cysS2 AGF56343.1 pheS spoIIIAE1 AGF56579.1 AGF56582.1 AGF56620.1 AGF56648.1 patB1 rpfG7 AGF56695.1 patA AGF56708.1 rep AGF56735.1 AGF56796.1 AGF56808.1 AGF56825.1 AGF56868.1 AGF56869.1 NTPase AGF56872.1 AGF56968.1 AGF56970.1 mutL dltS recQ2 AGF57142.1 AGF57143.1 AGF57167.1 AGF57267.1 AGF57274.1 AGF57296.1 fabF3 aruS AGF57372.1 AGF57435.1 AGF57440.1 AGF57495.1 AGF57537.1 AGF57543.1 aspA AGF57582.1 AGF57602.1 hndC1 AGF57621.1 AGF57659.1 fruA2 AGF57892.1 ftsH4 AGF57969.1 araN AGF57998.1 AGF58066.1 AGF58121.1 AGF58198.1 hflX hag2 murG recJ2 AGF58383.1 AGF58451.1 gyaR bglA5 AGF58553.1 cas3 yesM2 AGF58621.1 bioA AGF58814.1 AGF59042.1 leuS galE3 epsC AGF59131.1 helD3 alaXL AGF59225.1 uvrC2 degT2
0.000433	108	614	1.516206	Hydrolase	addB addA AGF53968.1 luxQ pcrA yfmM abgA1 atpA AGF54463.1 pbg kdpB AGF54706.1 bglH1 AGF54734.1 purH macB4 priA recG rasP rny ribBA cpsA xynD1 engD2 Amidohydrolase pulA bglA2 malL1 AGF55533.1 AGF55542.1 msbA1 celK1 celA2 Endo-beta-mannanase celH metN2 AGF55765.1 AGF55795.1 xynD2 AGF55956.1 xynB3 AGF56074.1 bglC2 xylG AGF56141.1 ascB yjbG rpfG6 ruvB spoIIIAH dacF

					ramA1 xynB4 rpfG7 AGF56708.1 rep cysA AGF56747.1 xynB5 polC AGF56808.1 fbpC pldB AGF56970.1 ramA2 ftsH3 fold2 AGF57028.1 uxaA recQ2 dbpA AGF57119.1 hemN2 rsgA2 AGF57467.1 AGF57650.1 hypF AGF57858.1 Alpha-galactosidase AGF57885.1 D-hydantoinase ftsH4 AGF57933.1 AGF57951.1 AGF57969.1 xylB2 AGF58166.1 recJ2 mglA AGF58451.1 bglA5 cas3 xsa1 xylB4 AGF58608.1 bglH4 malH2 AGF58752.1 AGF58780.1 msbA2 AGF58813.1 Polygalacturonase bgxA2 Alpha-galactosidase/6-phospho-beta-glucosidase helD3 cheB2 dnaC1 hppA
0.000436	14	31	3.892862	Phosphotransferase system, EIIC	AGF55511.1 AGF55515.1 celB1 mtlA AGF56753.1 bglP5 fruA2 AGF58113.1 AGF58457.1 AGF58579.1 bglP8 AGF58814.1 AGF59111.1 malX
0.000436	28	97	2.488221	Ligase	ileS murC lysS murD ligA pyrG argG trpS murE1 asnO nadE accC1 alaS coaBC tmcAL pncB metG3 cysS2 pheS pheT murE3 murF accC2 folC citC1 argS2 AGF58788.1 leuS
0.000436	14	31	3.892862	Phosphotransferase system, EIIC	AGF55511.1 AGF55515.1 celB1 mtlA AGF56753.1 bglP5 fruA2 AGF58113.1 AGF58457.1 AGF58579.1 bglP8 AGF58814.1 AGF59111.1 malX
0.000524	65	323	1.734657	Metal-binding	gyrB1 addB ileS lysS leuC2 ligA pyrG rnfC pbg apbE1 kdpB dnaG hcp3 pgcA alaS priA rlmN1 dxr rasP ispG ribBA gldA hisD AGF55405.1 dxs1 metG3 hcp4 rlmN2 AGF55972.1 dmsA metH3 apbE2 yjbG cysS2 pheS pheT dxs2 ackA2 narB moeA2 moeA3 trpE trpD nifK3 nifB2 anfD ilvB2 AGF56970.1 ftsH3 AGF57072.1 AGF57365.1 rsgA2 pckA1 apbE3 rhaA ubiD AGF57885.1 D-hydantoinase ftsH4 ppk egsA araA AGF58560.1 maeB eno2
0.000708	37	151	2.112163	Mixed, incl. glycosidase, and xylose isomerase-like superfamily	AGF54727.1 xynD1 engD2 pel AGF55542.1 celK1 celA2 celH xynD2 xynB3 AGF56009.1 AGF56046.1 AGF56074.1 ramA1 xynB4 xynB5 AGF56761.1 ramA2 AGF57028.1 AGF57247.1 rhaA rhaB AGF57854.1 AGF57858.1 AGF57969.1 AGF58041.1 xylB2 AGF58075.1 AGF58079.1 AGF58498.1 AGF58583.1 AGF58584.1 xsa1 xylB4 AGF58608.1 Alpha-L-arabinofuranosidase Polygalacturonase
0.000708	21	64	2.828408	Sugar transport	AGF55511.1 AGF55515.1 celB1 AGF55538.1 mtlA xylG AGF56753.1

					bglP5 fruA2 AGF57998.1 AGF58001.1 AGF58068.1 AGF58113.1 AGF58440.1 AGF58457.1 AGF58579.1 bglP8 AGF58814.1 AGF58925.1 AGF59111.1 malX
0.000991	16	42	3.283775	Major facilitator superfamily domain	mdtG AGF55212.1 AGF55603.1 AGF55696.1 AGF55976.1 tetA AGF56190.1 AGF56681.1 AGF56696.1 AGF56769.1 AGF56955.1 AGF56982.1 AGF58118.1 AGF58672.1 AGF58693.1 dgoT
0.001054	17	47	3.11784	Pyridoxal phosphate-dependent transferase, major domain	argD metC2 cobD2 hisC2 AGF55413.1 eryC ycxD1 ycxD2 patB1 patA AGF56997.1 AGF57274.1 AGF57283.1 AGF58224.1 phnW bioA degT2
0.001054	17	47	3.11784	Pyridoxal phosphate-dependent transferase	argD metC2 cobD2 hisC2 AGF55413.1 eryC ycxD1 ycxD2 patB1 patA AGF56997.1 AGF57274.1 AGF57283.1 AGF58224.1 phnW bioA degT2
0.001058	8	12	5.746606	Transmembrane secretion effector	mdtG AGF55212.1 AGF56190.1 AGF57168.1 AGF57324.1 AGF58118.1 AGF58672.1 AGF58693.1
0.00174	37	159	2.005891	Mixed, incl. butanoate metabolism, and pyruvate metabolism	buk1 leuC2 glcD1 bcd1 etfA3 pfl1 alsS1 thlA2 AGF55505.1 thlA3 thlA4 bcd3 etfA4 eutE AGF55972.1 ilvD2 gbsA patA ackA2 ilvB2 leuA3 AGF57365.1 pckA1 AGF57496.1 leuC3 alsS2 AGF57892.1 AGF57896.1 lctP2 leuC4 phnW ppm fccA2 gyaR maeB eno2 pyk2
0.00174	9	16	4.848699	Aminotransferase, class I/classII	cobD2 hisC2 AGF55413.1 ycxD1 ycxD2 patB1 AGF57274.1 AGF57283.1 AGF58224.1
0.00174	5	5	8.61991	PD-(D/E)XK nuclease superfamily 9	AGF55367.1 AGF55520.1 AGF55773.1 AGF59150.1 AGF59151.1
0.00174	5	5	8.61991	PD-(D/E)XK nuclease superfamily	AGF55367.1 AGF55520.1 AGF55773.1 AGF59150.1 AGF59151.1
0.001853	8	13	5.30456	Phosphotransferase system, EIIC component, type 3	AGF55511.1 AGF55515.1 celB1 AGF56753.1 AGF58113.1 AGF58457.1 AGF58579.1 AGF59111.1
0.001853	8	13	5.30456	Phosphotransferase system, cellobiose-type IIC component	AGF55511.1 AGF55515.1 celB1 AGF56753.1 AGF58113.1 AGF58457.1 AGF58579.1 AGF59111.1
0.002082	15	41	3.153625	Pyridoxal phosphate-	argD metC2 cobD2 hisC2 AGF55413.1 eryC ycxD2 patB1 patA AGF57274.1

				dependent transferase domain 1	AGF57283.1 AGF58224.1 phnW bioA degT2
0.002334	24	87	2.377906	Histidine kinase-like ATPases	gyrB1 luxQ AGF54306.1 AGF54911.1 glnK AGF55492.1 AGF55540.1 AGF55634.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 mutL dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58597.1 yesM2 AGF58648.1 AGF58793.1
0.002351	20	66	2.612094	Glycosidase, and Pectin lyase fold	AGF54727.1 xynD1 engD2 pel celK1 celA2 xynD2 xynB3 AGF56074.1 xynB5 ramA2 AGF57247.1 AGF57858.1 AGF57969.1 xylB2 xsa1 xylB4 AGF58608.1 Alpha-L-arabinofuranosidase Polygalacturonase
0.002376	9	17	4.563482	Helicase, C-terminal	priA recG AGF56708.1 AGF56808.1 AGF56970.1 recQ2 dbpA AGF57119.1 cas3
0.002376	9	17	4.563482	Glycosyl hydrolases family 1, N-terminal conserved site	abgA1 bglH1 bglA2 AGF55533.1 bglC2 ascB AGF56747.1 bglH4 AGF59117.1
0.002376	9	17	4.563482	DEAD-like helicases superfamily	priA recG AGF56708.1 AGF56808.1 AGF56970.1 recQ2 dbpA AGF57119.1 cas3
0.002403	30	122	2.11965	Butanoate metabolism, and Pyruvate metabolism	buk1 leuC2 bcd1 etfA3 pfl1 alsS1 thlA2 AGF55505.1 thlA3 thlA4 bcd3 etfA4 eutE AGF55972.1 ilvD2 gbsA ackA2 ilvB2 leuA3 AGF57365.1 pckA1 AGF57496.1 leuC3 alsS2 AGF57892.1 leuC4 fccA2 maeB eno2 pyk2
0.002403	24	88	2.350884	Histidine kinase/HSP90-like ATPase superfamily	gyrB1 luxQ AGF54306.1 AGF54911.1 glnK AGF55492.1 AGF55540.1 AGF55634.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 mutL dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58597.1 yesM2 AGF58648.1 AGF58793.1
0.002596	10	21	4.104719	GHKL domain	AGF54911.1 glnK AGF55634.1 AGF56222.1 AGF56258.1 dltS AGF57341.1 AGF58121.1 AGF58648.1 AGF58793.1
0.002705	62	328	1.629373	P-loop containing nucleoside triphosphate hydrolase	dnaA recF addB addA fusA1 pcrA yfmM AGF54357.1 pyrG atpA macB4 engA priA recG smc ffh AGF55094.1 infB AGF55366.1 AGF55520.1 msbA1 AGF55623.1 metN2 AGF55850.1 xylG AGF56192.1 ruvB ptk AGF56708.1 rep cysA AGF56796.1 AGF56808.1 fbpC AGF56830.1 NTPase AGF56872.1 AGF56968.1 AGF56970.1 ftsH3

					AGF57055.1 recQ2 dbpA AGF57119.1 rsgA2 hydF cydC1 cydC2 AGF57659.1 ftsH4 AGF58001.1 pilT hflX mglA AGF58440.1 AGF58451.1 cas3 msbA2 AGF58901.1 helD3 dnaX2 dnaC1
0.00279	8	14	4.925663	DEAD/DEAH box helicase	priA recG AGF56708.1 AGF56808.1 recQ2 dbpA AGF57119.1 AGF57659.1
0.002824	6	8	6.464932	Mur ligase, central	murC murD murE1 murE3 murF folC
0.002824	6	8	6.464932	Mur-like, catalytic domain superfamily	murC murD murE1 murE3 murF folC
0.002824	6	8	6.464932	Mur ligase middle domain	murC murD murE1 murE3 murF folC
0.003218	25	96	2.244768	Phosphotransferase system (PTS), and Glycosyl hydrolase family 1	AGF55511.1 AGF55515.1 bglA2 celB1 AGF55533.1 mtlA bglC2 licR3 AGF56747.1 AGF56753.1 bglP5 fruA2 manR AGF58113.1 AGF58457.1 AGF58579.1 bglH4 bglP8 malH2 AGF58813.1 AGF58814.1 AGF59111.1 AGF59117.1 Alpha-galactosidase/6-phospho-beta-glucosidase malX
0.003224	9	18	4.309955	Helicase superfamily 1/2, ATP-binding domain	priA recG AGF56708.1 AGF56808.1 AGF56970.1 recQ2 dbpA AGF57119.1 cas3
0.003224	9	18	4.309955	Helicase conserved C-terminal domain	priA recG AGF56708.1 AGF56808.1 AGF56970.1 recQ2 dbpA AGF57119.1 cas3
0.003403	22	80	2.370475	Histidine kinase/HSP90-like ATPase	gyrB1 luxQ AGF54911.1 glnK AGF55492.1 AGF55540.1 AGF55634.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58597.1 yesM2 AGF58648.1 AGF58793.1
0.004366	44	215	1.764075	Oxidoreductase	glcD1 bcd1 AGF54511.1 argC glpA nadB hcp3 AGF54977.1 dxr ispG nifK1 gldA hemA hisD AGF55505.1 hcp4 wbpA1 bcd3 gutB dmsA gbsA AGF56695.1 narB padH nifK3 anfD anfK gapN uxaB hemN2 AGF57365.1 AGF57493.1 mco fucO cydB hndC1 preA AGF57896.1 egsA fccA2 nuoG gyaR maeB gap
0.004366	8	15	4.597285	Helicase superfamily c-terminal domain	priA recG AGF56708.1 AGF56808.1 AGF56970.1 recQ2 dbpA AGF57119.1
0.00505	9	19	4.083115	DNA replication	dnaA dnaN recF ligA dnaG priA polC dnaX2 dnaC1

0.005389	7	12	5.028281	Mixed, incl. uronate isomerase, and mannonate dehydratase	uxaC1 uxaA uxaB uxaC2 uxaC3 uxuA3 uxuA4
0.005835	21	78	2.320745	Pyruvate metabolism, and Valine, leucine and isoleucine biosynthesis	leuC2 pfl1 alsS1 AGF55505.1 eutE AGF55972.1 ilvD2 gbsA ackA2 ilvB2 leuA3 pckA1 AGF57496.1 leuC3 alsS2 AGF57892.1 leuC4 fccA2 maeB eno2 pyk2
0.005835	6	9	5.746606	DEAD/DEAH box helicase domain	recG AGF56708.1 AGF56808.1 recQ2 dbpA AGF57119.1
0.005835	6	9	5.746606	AAA-ATPase-like domain	AGF54357.1 AGF55367.1 AGF55520.1 AGF55773.1 AGF59150.1 AGF59151.1
0.006145	22	84	2.257595	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	gyrB1 luxQ AGF54911.1 glnK AGF55492.1 AGF55540.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 mutL dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58597.1 yesM2 AGF58648.1 AGF58793.1
0.006496	8	16	4.309955	Glycoside hydrolase family 1, active site	abgA1 bglH1 bglA2 AGF55533.1 bglC2 ascB AGF56747.1 bglH4
0.006734	9	20	3.878959	Glycoside hydrolase family 1	abgA1 bglH1 bglA2 AGF55533.1 bglC2 ascB AGF56747.1 bglH4 AGF59117.1
0.006734	9	20	3.878959	Glycosyl hydrolase family 1	abgA1 bglH1 bglA2 AGF55533.1 bglC2 ascB AGF56747.1 bglH4 AGF59117.1
0.010321	8	17	4.056428	Aminoacyl-tRNA synthetase	ileS lysS trpS metG3 pheS pheT argS2 leuS
0.010321	8	17	4.056428	Mixed, incl. rhamnose metabolism, and alpha-l-rhamnosidase, six-hairpin glycosidase domain	AGF55542.1 ramA1 xynB4 AGF57028.1 rhaA rhaB AGF58583.1 AGF58584.1
0.011134	6	10	5.171946	Nitrogenase /oxidoreductase, component 1	nifE1 nifK1 nifK3 nifB2 anfD anfK
0.011134	6	10	5.171946	Nitrogenase component 1 type	nifE1 nifK1 nifK3 nifB2 anfD anfK

				Oxidoreductase	
0.011134	6	10	5.171946	Predicted AAA-ATPase	AGF54357.1 AGF55367.1 AGF55520.1 AGF55773.1 AGF59150.1 AGF59151.1
0.011767	5	7	6.157078	Mur ligase, C-terminal domain superfamily	murC murD murE3 murF folC
0.012551	10	26	3.31535	Bacterial extracellular solute-binding protein	AGF54723.1 AGF55538.1 AGF56070.1 potD araN AGF57994.1 AGF57998.1 AGF58068.1 AGF58621.1 AGF58925.1
0.013429	12	36	2.873303	Peptidoglycan biosynthesis, and Cell cycle protein	murC murD murE1 spoVD2 murE3 murF ftsW dacF mrdB alr murG AGF58788.1
0.013429	11	31	3.058678	Mixed, incl. alpha amylase, catalytic domain, and amino acid permease	glgP1 pgcA pulA AGF55445.1 malL1 AGF55925.1 glgD1 glgD2 glgC2 glgB gtfA2
0.013429	8	18	3.831071	Six-hairpin glycosidase superfamily	AGF54727.1 AGF55542.1 celK1 ramA1 ramA2 AGF57247.1 AGF57854.1 AGF58608.1
0.013429	7	14	4.309955	Invasin/intimin cell-adhesion fragments	AGF55708.1 AGF56477.1 AGF56478.1 AGF56532.1 AGF56877.1 AGF56878.1 AGF57396.1
0.013429	14	46	2.623451	FAD/NAD(P)-binding domain superfamily	AGF54475.1 glpA nadB AGF55505.1 AGF56601.1 AGF56735.1 padH AGF57493.1 AGF57603.1 AGF57613.1 fccA2 nuoG AGF59038.1 AGF59042.1
0.013429	12	36	2.873303	NAD	ligA nadE AGF54977.1 gldA hisD uxaB egsA malH2 AGF58813.1 gap galE3 Alpha-galactosidase/6-phospho-beta-glucosidase
0.013429	22	90	2.107089	Zinc	ileS ligA pbg dnaG alaS priA ribBA gldA hisD metG3 AGF55972.1 metH3 yjbG cysS2 AGF56970.1 ftsH3 rsgA2 AGF57467.1 AGF57885.1 ftsH4 egsA AGF58560.1
0.013429	12	36	2.873303	Aminotransferase class I and II	argD metC2 cobD2 hisC2 AGF55413.1 eryC ycxD1 ycxD2 patB1 AGF57274.1 AGF57283.1 AGF58224.1
0.013429	15	51	2.535268	His Kinase A (phospho-acceptor) domain	luxQ AGF54911.1 AGF55492.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58648.1

0.013429	7	14	4.309955	Type III restriction enzyme, res subunit	priA recG AGF56708.1 AGF56808.1 recQ2 dbpA AGF57119.1
0.013488	13	41	2.733142	Homologous recombination, and Helicase	dnaN addB addA pcrA recG AGF56708.1 rep polC AGF56970.1 recQ2 recJ2 mppE dnaX2
0.016261	17	63	2.326007	Glycoside hydrolase superfamily, and Amino acid permease	pbg glgP1 pgcA pulA AGF55445.1 malL1 AGF55925.1 Alpha-galactosidase mro bglA5 glgD1 AGF59011.1 bgxA2 glgD2 glgC2 glgB gtfA2
0.016524	21	86	2.104862	ABC transporters	AGF54723.1 AGF55538.1 mglB2 AGF56070.1 xylG xylH1 mglB5 mglB8 mglB9 araN AGF57998.1 AGF58001.1 mglA mglB11 xylH2 AGF58440.1 AGF58441.1 AGF58451.1 AGF58611.1 AGF58621.1 AGF58925.1
0.016703	6	11	4.701769	Mixed, incl. glycogen biosynthesis, and glucose-1-phosphate adenylyltransferase, glgd subunit	pgcA pulA glgD1 glgD2 glgC2 glgB
0.016703	6	11	4.701769	ABC transporter, permease, and Periplasmic binding protein	xylG xylH1 xylH2 AGF58440.1 AGF58441.1 AGF58451.1
0.018269	7	15	4.022624	Cell shape, and Diaminopimelate epimerase	murC murD spoVD2 murE3 murF ftsW murG
0.018269	22	93	2.039118	ABC transporters, and Bacterial extracellular solute-binding protein	AGF54723.1 AGF55538.1 mglB2 AGF55992.1 AGF56070.1 xylG xylH1 mglB5 mglB8 mglB9 araN AGF57998.1 AGF58001.1 mglA mglB11 xylH2 AGF58440.1 AGF58441.1 AGF58451.1 AGF58611.1 AGF58621.1 AGF58925.1
0.018269	7	15	4.022624	Bacterial Ig-like, group 2	AGF55708.1 AGF56477.1 AGF56478.1 AGF56532.1 AGF56877.1 AGF56878.1 AGF57396.1

0.018269	7	15	4.022624	Bacterial Ig-like domain (group 2)	AGF55708.1 AGF56477.1 AGF56478.1 AGF56532.1 AGF56877.1 AGF56878.1 AGF57396.1
0.018269	7	15	4.022624	Bacterial Ig-like domain 2	AGF55708.1 AGF56477.1 AGF56478.1 AGF56532.1 AGF56877.1 AGF56878.1 AGF57396.1
0.01873	16	59	2.337603	Mixed, incl. dna repair, and mismatch repair	dnaN addB addA pcrA recG ruvB AGF56708.1 rep polC AGF56970.1 mutL recQ2 recJ2 mppE uvrC2 dnaX2
0.01873	4	5	6.895928	ABC-2 family transporter protein, and RND efflux pump, membrane fusion protein, barrel-sandwich domain	AGF55603.1 AGF56579.1 AGF56580.1 AGF56581.1
0.01873	4	5	6.895928	Aminotransferases, class-I, pyridoxal-phosphate-binding site	cobD2 AGF55413.1 AGF57274.1 AGF57283.1
0.01873	4	5	6.895928	SsuA/THI5-like	ssuA AGF56610.1 tauA AGF57888.1
0.019513	5	8	5.387443	DNA helicase, UvrD/REP type	addB addA pcrA rep helD3
0.020032	15	54	2.394419	Signal transduction histidine kinase, dimerisation/phosphoacceptor domain	luxQ AGF54911.1 AGF55492.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58648.1
0.020899	9	24	3.232466	Periplasmic binding protein, and ABC transporter, permease	mglB2 xylG xylH1 mglA mglB11 xylH2 AGF58440.1 AGF58441.1 AGF58451.1
0.021354	19	77	2.126991	Mixed, incl. helicase, and homologous recombination	dnaN addB addA pcrA dnaG priA recG ruvB AGF56708.1 rep polC AGF56970.1 mutL recQ2 recJ2 mppE uvrC2 dnaX2 dnaC1

0.021477	34	171	1.7139	ABC transporters, and ABC transporter transmembrane region	AGF54723.1 clcA AGF55538.1 msbA1 metN2 mglB2 AGF55992.1 AGF56070.1 xylG xylH1 AGF56192.1 potD mglB5 mglB8 mglB9 cydC1 cydC2 cydB AGF57659.1 AGF57903.1 araN AGF57998.1 AGF58001.1 mglA mglB11 xylH2 AGF58440.1 AGF58441.1 AGF58451.1 AGF58553.1 AGF58611.1 AGF58621.1 AGF58901.1 AGF58925.1
0.022052	38	198	1.654326	Kinase	buk1 luxQ AGF54306.1 AGF54911.1 sps glnK AGF55492.1 AGF55540.1 AGF55634.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 ptk AGF56620.1 licR3 ackA2 dltS AGF57341.1 aruS pckA1 AGF57535.1 bglP5 rhaB manR ppk AGF58121.1 xylB3 AGF58597.1 yesM2 AGF58648.1 bglP8 AGF58793.1 AGF58814.1 AGF59131.1 pyk2 malX
0.022337	8	20	3.447964	Alpha amylase, catalytic domain, and Glucose-1-phosphate adenylyltransferase	glgP1 pgcA pulA malL1 glgD1 glgD2 glgC2 glgB
0.022337	8	20	3.447964	Thiolase-like	spoVAD1 fabF1 thlA2 thlA3 thlA4 fabF2 fabF3 spoVAD2
0.022337	15	55	2.350884	Signal transduction histidine kinase, dimerisation/phosphoacceptor domain superfamily	luxQ AGF54911.1 AGF55492.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58648.1
0.022337	15	55	2.350884	His Kinase A (phosphoacceptor) domain	luxQ AGF54911.1 AGF55492.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58648.1
0.0231	27	127	1.832579	Magnesium	gyrB1 lysS ligA pyrG apbE1 kdpB dnaG pgcA ribBA dxs1 apbE2 pheS pheT murE3 dxs2 ackA2 moeA2 moeA3 trpE trpD ilvB2 AGF57072.1 apbE3 ppk eno2 pyk2 hppA
0.02326	14	50	2.413575	Starch and sucrose metabolism	AGF55511.1 AGF55515.1 bglA2 celB1 AGF55533.1 bglC2 AGF56747.1 AGF56753.1 AGF58113.1 AGF58457.1 AGF58579.1 bglH4 AGF59111.1 AGF59117.1
0.02326	14	50	2.413575	Mixed, incl. xylose	AGF55542.1 ramA1 xynB4 AGF56761.1 AGF57028.1 rhaA rhaB AGF57854.1

				isomerase-like superfamily, and oxidoreductase, n-terminal	AGF58041.1 AGF58075.1 AGF58079.1 AGF58498.1 AGF58583.1 AGF58584.1
0.02326	6	12	4.309955	Mixed, incl. rhamnose metabolism, and glycosyl hydrolases family 2, sugar binding domain	AGF55542.1 AGF57028.1 rhaA rhaB AGF58583.1 AGF58584.1
0.02335	7	16	3.77121	Glycosyl hydrolases family 43, and Alpha-L-arabinofuranosidase, C-terminal	xynD1 xynD2 AGF57247.1 xsa1 xylB4 AGF58608.1 Alpha-L-arabinofuranosidase
0.02412	17	67	2.187141	Histidine kinase domain	luxQ AGF54911.1 glnK AGF55492.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58648.1 AGF58793.1
0.02521	15	56	2.308904	Signal transduction histidine kinase-related protein, C-terminal	AGF54911.1 glnK AGF55492.1 resE6 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 dltS AGF57341.1 aruS AGF57535.1 AGF58121.1 AGF58648.1
0.02538	21	91	1.98921	HAMP (Histidine kinases, Adenylyl cyclases, Methyl binding proteins, Phosphatases) domain	mcp41 AGF54607.1 AGF54911.1 AGF55492.1 AGF55540.1 AGF55673.1 AGF56195.1 AGF56258.1 AGF56683.1 AGF57341.1 AGF57372.1 AGF57535.1 AGF57621.1 AGF57993.1 AGF58050.1 AGF58121.1 yesM2 AGF58648.1 AGF58673.1 AGF59225.1 AGF59232.1
0.02559	12	40	2.585973	Signal transduction histidine kinase, dimerisation/phosphoacceptor domain, and OmpR/PhoB-type DNA-	luxQ AGF54911.1 resE6 AGF56195.1 AGF56222.1 AGF56258.1 AGF56620.1 AGF56764.1 dltS AGF57341.1 AGF57535.1 AGF57690.1

				binding domain	
0.02619	11	35	2.709114	Nitrogen metabolism, and Nitrogen regulatory protein PII	nrgA1 hcp3 nifE1 nifK1 hcp4 nifK3 nifB2 anfD anfK AGF57584.1 AGF57643.1
0.027044	14	51	2.36625	Mixed, incl. pentose and glucuronate interconversions, and lactose permease-like	uxaC1 AGF56982.1 AGF57100.1 uxaA uxaB uxaC2 AGF57106.1 AGF58065.1 AGF58118.1 AGF58638.1 uxaC3 uxuA3 uxuA4 dgoT
0.027095	51	292	1.505532	Cytoplasm	dnaA dnaN recF gyrB1 ileS murC lysS murD fusA1 buk1 pgi argD argJ argC argG trpS rpoD2 hcp3 alaS ftsZ rsmB rlmN1 tmcAL smc ffh nusA infB hisZ rlmI metG3 hcp4 rlmN2 cysS2 pheS pheT murE3 murF ackA2 polC dbpA rsgA2 pckA1 rhaA hflX argS2 bioA eno2 leuS alaXL cheB2 uvrC2
0.027095	51	292	1.505532	Cell membrane	secY atpA rnfC kdpB mviN AGF54911.1 macB4 rny kup2 AGF55212.1 AGF55492.1 AGF55511.1 AGF55515.1 celB1 AGF55603.1 pgaC metN2 mtlA xylG xylH1 AGF56258.1 ftsW cysA AGF56753.1 AGF56955.1 ftsH3 AGF57140.1 AGF57304.1 AGF57341.1 AGF57535.1 bglP5 fruA2 lctP2 AGF57903.1 ftsH4 AGF58113.1 AGF58121.1 AGF58144.1 murG xylH2 AGF58457.1 AGF58579.1 AGF58648.1 AGF58672.1 AGF58693.1 bglP8 AGF58814.1 AGF58925.1 AGF59111.1 malX hppA
0.027145	3	3	8.61991	Beta-ketoacyl synthase	fabF1 fabF2 fabF3
0.027145	3	3	8.61991	L-fucose isomerase, N-terminal/central domain superfamily	AGF56125.1 araA AGF58584.1
0.027145	3	3	8.61991	3-oxoacyl-[acyl-carrier-protein] synthase 2	fabF1 fabF2 fabF3
0.031283	21	93	1.946431	HAMP domain	mcp41 AGF54607.1 AGF54911.1 AGF55492.1 AGF55540.1 AGF55673.1

					AGF56195.1 AGF56258.1 AGF56683.1 AGF57341.1 AGF57372.1 AGF57535.1 AGF57621.1 AGF57993.1 AGF58050.1 AGF58121.1 yesM2 AGF58648.1 AGF58673.1 AGF59225.1 AGF59232.1
0.031496	7	17	3.549375	Glycosyl hydrolase, five-bladed beta-propellor domain superfamily	xynD1 xynD2 xynB3 AGF57969.1 xylB2 xylB4 AGF59023.1
0.031496	7	17	3.549375	Glycosyl transferases group 1	AGF54179.1 AGF54849.1 AGF54981.1 AGF55681.1 AGF56343.1 AGF56357.1 AGF59095.1
0.032771	33	171	1.663491	AAA+ ATPase domain	dnaA yfmM macB4 engA ffh AGF55094.1 AGF55366.1 msbA1 metN2 AGF55850.1 xylG AGF56192.1 ruvB cysA AGF56796.1 fbpC AGF56830.1 AGF56968.1 ftsH3 AGF57055.1 cydC1 cydC2 AGF57659.1 ftsH4 AGF58001.1 pilT mglA AGF58440.1 AGF58451.1 msbA2 AGF58901.1 dnaX2 dnaC1
0.032771	6	13	3.97842	Alpha-helical ferredoxin	AGF54475.1 asrA AGF57493.1 AGF57613.1 preA nuoG
0.032771	6	13	3.97842	Cellulase (glycosyl hydrolase family 5)	abgA1 bglH1 engD2 AGF55533.1 Endo-beta-mannanase AGF56074.1
0.032869	13	47	2.38423	Mixed, incl. nitrogen metabolism, and nitrogen regulatory protein pii	nrgA1 hcp3 nifE1 nifK1 hcp4 AGF56825.1 AGF56830.1 nifK3 nifB2 anfD anfK AGF57584.1 AGF57643.1
0.034136	4	6	5.746606	Arginine biosynthesis	argD argJ argC argG
0.034136	4	6	5.746606	Nitrogenase component 1, conserved site	nifK3 nifB2 anfD anfK
0.034136	4	6	5.746606	Mur ligase, C-terminal	murC murE3 murF folC
0.034136	4	6	5.746606	Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain superfamily	nadB AGF55505.1 AGF56735.1 fccA2
0.034136	4	6	5.746606	Capsid protein	AGF54324.1 cotS2 AGF56456.1 cotA

0.034136	4	6	5.746606	Virion	AGF54324.1 cotS2 AGF56456.1 cotA
0.034136	4	6	5.746606	Mur ligase family, glutamate ligase domain	murC murE3 murF folC
0.034136	33	172	1.65382	ATPases associated with a variety of cellular activities	dnaA yfmM macB4 engA ffh AGF55094.1 AGF55366.1 msbA1 metN2 AGF55850.1 xylG AGF56192.1 ruvB cysA AGF56796.1 fbpC AGF56830.1 AGF56968.1 ftsH3 AGF57055.1 cydC1 cydC2 AGF57659.1 ftsH4 AGF58001.1 pilT mglA AGF58440.1 AGF58451.1 msbA2 AGF58901.1 dnaX2 dnaC1
0.03443	8	22	3.134513	Translocase	atpA rnfC kdpB macB4 metN2 xylG cysA hppA
0.036509	9	27	2.873303	Aminotransferase	argD hisC2 AGF55413.1 patA AGF57274.1 AGF57283.1 AGF58224.1 phnW bioA
0.040638	7	18	3.352187	Mixed, incl. pas domain, and helix_turn_helix, lux regulon	luxQ AGF54911.1 AGF56195.1 AGF56620.1 AGF56764.1 AGF57341.1 AGF57690.1
0.040638	7	18	3.352187	Cell shape	murC murD mviN murE3 murF ftsW murG
0.043069	8	23	2.998229	Nitrogen metabolism, and Nitrogenase cofactor biosynthesis protein NifB	hcp3 nifE1 nifK1 hcp4 nifK3 nifB2 anfD anfK
0.043069	5	10	4.309955	Aminoacyl-tRNA synthetase, class Ia, anticodon-binding, and Aminoacyl-tRNA synthetase, class II (D/K/N)	ileS lysS metG3 argS2 leuS
0.043069	5	10	4.309955	Lactose permease-like	AGF56982.1 AGF57100.1 AGF57106.1 AGF58065.1 AGF58638.1
0.043069	5	10	4.309955	Mixed, incl. hlyd membrane-fusion protein of t1ss, and	AGF55603.1 AGF56579.1 AGF56580.1 AGF56581.1 AGF56955.1

				drug resistance transporter emrb-like	
0.043069	5	10	4.309955	Drug resistance transporter EmrB-like	AGF55212.1 AGF55603.1 AGF56955.1 AGF58672.1 AGF58693.1
0.043069	6	14	3.694247	Glycoside hydrolase, family 43	xynD1 xynD2 xynB3 AGF57969.1 xylB2 xylB4
0.043069	5	10	4.309955	Aminoacyl-tRNA synthetase, class Ia, anticodon-binding	ileS metG3 cysS2 argS2 leuS
0.043069	5	10	4.309955	Thiamine pyrophosphate	alsS1 dxs1 dxs2 ilvB2 alsS2
0.043069	115	786	1.261183	Transferase	dnaN rpoB buk1 AGF54179.1 luxQ baeE AGF54306.1 ugtP argD argJ apbE1 dnaG glgP1 AGF54849.1 pfl1 AGF54911.1 purH fabF1 AGF54981.1 rsmB rlmN1 sps licR1 cobT hisZ hisC2 pncB glnK alsS1 AGF55413.1 rlmI AGF55492.1 thlA2 AGF55535.1 AGF55540.1 dxs1 AGF55586.1 AGF55634.1 AGF55681.1 pgaC AGF55749.1 thlA3 thlA4 rlmN2 mtlA resE6 metH3 apbE2 AGF56195.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56341.1 AGF56343.1 ptk AGF56357.1 AGF56359.1 dxs2 AGF56553.1 AGF56620.1 AGF56635.1 licR3 fabF2 patA AGF56753.1 ackA2 polC moeA2 moeA3 trpD AGF56869.1 ilvB2 AGF56992.1 dltS leuA3 AGF57274.1 AGF57283.1 AGF57341.1 fabF3 aruS pckA1 alsS2 AGF57510.1 apbE3 AGF57535.1 bglP5 rhaB fruA2 manR AGF57775.1 hypF ppk AGF58121.1 AGF58224.1 pseI murG phnW xylB3 AGF58560.1 AGF58597.1 yesM2 AGF58648.1 mtbA bglP8 AGF58793.1 bioA AGF58814.1 AGF59095.1 AGF59131.1 pyk2 glgC2 glgB dnaX2 malX gtfA2
0.043069	6	14	3.694247	FAD dependent oxidoreductase	glpA nadB AGF55505.1 AGF57613.1 fccA2 nuoG
0.043069	6	14	3.694247	Glycosyl hydrolases family 43	xynD1 xynD2 xynB3 AGF57969.1 xylB2 xylB4

0.043069	8	23	2.998229	MFS/sugar transport protein	mdtG tetA AGF56982.1 AGF57100.1 AGF57106.1 AGF58065.1 AGF58638.1 AGF58672.1
0.043595	9	28	2.770685	Glycosyl hydrolase, all-beta	pbg pulA malL1 celA2 AGF57858.1 Alpha-galactosidase xsa1 Alpha-L-arabinofuranosidase glgB
0.052343	7	19	3.175756	Metal-dependent hydrolase	AGF54463.1 uxaC1 uxaC2 D-hydantoinase uxaC3 AGF58752.1 AGF58780.1
0.055472	9	29	2.675144	FAD binding domain	nadB AGF55505.1 AGF56601.1 AGF56735.1 AGF57613.1 fccA2 nuoG AGF59038.1 AGF59042.1
0.058079	6	15	3.447964	Nitrogen metabolism	hcp3 nifE1 nifK1 hcp4 anfD anfK
0.058079	6	15	3.447964	Mixed, incl. drug resistance transporter emrb-like, and hlyd membrane-fusion protein of t1ss	AGF55603.1 AGF56579.1 AGF56580.1 AGF56581.1 AGF56955.1 AGF58693.1
0.058079	6	15	3.447964	Glycosyl transferase, family 1	AGF54179.1 AGF54849.1 AGF54981.1 AGF55681.1 AGF56343.1 AGF59095.1
0.058079	4	7	4.925663	Acyltransferase 3	AGF54992.1 AGF56359.1 AGF56635.1 AGF56638.1
0.058079	4	7	4.925663	FAD-dependent oxidoreductase 2, FAD binding domain	nadB AGF55505.1 AGF56735.1 fccA2
0.058079	4	7	4.925663	UvrD-like helicase, ATP-binding domain	addA pcrA rep helD3
0.058079	4	7	4.925663	Immunoglobulin E-set	pulA cbpA celK1 glgB
0.058079	4	7	4.925663	UvrD/REP helicase N-terminal domain	addA pcrA rep helD3
0.058079	6	15	3.447964	MatE	AGF54190.1 mviN mepA2 AGF56244.1 AGF56566.1 AGF58729.1
0.058079	4	7	4.925663	Acyltransferase family	AGF54992.1 AGF56359.1 AGF56635.1 AGF56638.1
0.058474	7	20	3.016968	Benzoate degradation, and Electron transfer	bcd1 etfA3 thlA2 thlA3 thlA4 bcd3 etfA4

				flavoprotein domain	
0.058474	7	20	3.016968	Pentose and glucuronate interconversions, and KDPG/KHG aldolase	uxaC1 uxaA uxaB uxaC2 uxaC3 uxuA3 uxuA4
0.058474	5	11	3.918141	Mixed, incl. cellulose binding, type iv, and alpha-l-arabinofuranosidase, c-terminal	xynD1 xynD2 AGF57247.1 xsa1 Alpha-L-arabinofuranosidase
0.058474	3	4	6.464932	Mur ligase, N-terminal catalytic domain	murC murE3 murF
0.058474	3	4	6.464932	Aconitase/3-isopropylmalate dehydratase large subunit, alpha/beta/alpha domain	leuC2 leuC3 leuC4
0.058474	3	4	6.464932	Ammonium transporter	nrgA1 AGF57584.1 AGF57643.1
0.058474	3	4	6.464932	Tetracycline resistance protein TetA/multidrug resistance protein MdtG	mdtG tetA AGF56190.1
0.058474	3	4	6.464932	ApbE-like superfamily	apbE1 apbE2 apbE3
0.058474	3	4	6.464932	Uronate isomerase	uxaC1 uxaC2 uxaC3
0.058474	3	4	6.464932	Glucose-1-phosphate adenylyltransferase	glgD1 glgD2 glgC2
0.058474	3	4	6.464932	RNA helicase, DEAD-box type, Q motif	AGF56808.1 dbpA AGF57119.1
0.058474	3	4	6.464932	UvrD-like DNA	addA pcrA rep

				helicase, C-terminal	
0.058474	13	52	2.154977	Rossmann-like alpha/beta/alpha sandwich fold	ileS etfA3 argG trpS nadE tmcAL metG3 etfA4 cysS2 citC1 ppm argS2 leuS
0.058474	3	4	6.464932	Aconitase/3-isopropylmalate dehydratase large subunit, alpha/beta/alpha, subdomain 1/3	leuC2 leuC3 leuC4
0.058474	3	4	6.464932	Ammonium transporter, conserved site	nrgA1 AGF57584.1 AGF57643.1
0.058474	3	4	6.464932	Aconitase family, 4Fe-4S cluster binding site	leuC2 leuC3 leuC4
0.058474	3	4	6.464932	Thiolase, active site	thlA2 thlA3 thlA4
0.058474	3	4	6.464932	Ammonium transporter AmtB-like domain	nrgA1 AGF57584.1 AGF57643.1
0.058474	3	4	6.464932	Flavin transferase ApbE	apbE1 apbE2 apbE3
0.058474	5	11	3.918141	Dihydropyrimidine dehydrogenase domain II	AGF54475.1 AGF57493.1 AGF57613.1 preA nuoG
0.058474	3	4	6.464932	Ammonium/urea transporter	nrgA1 AGF57584.1 AGF57643.1
0.058474	3	4	6.464932	Alpha-L-rhamnosidase, six-hairpin glycosidase domain	AGF55542.1 ramA1 ramA2
0.058474	3	4	6.464932	Aconitase, iron-sulfur domain	leuC2 leuC3 leuC4
0.058474	3	4	6.464932	Primosome	dnaG priA dnaC1

0.058474	3	4	6.464932	Ammonia transport	nrgA1 AGF57584.1 AGF57643.1
0.058474	3	4	6.464932	Aconitase family (aconitate hydratase)	leuC2 leuC3 leuC4
0.058474	3	4	6.464932	Ammonium Transporter Family	nrgA1 AGF57584.1 AGF57643.1
0.058474	3	4	6.464932	Mur ligase family, catalytic domain	murC murE3 murF
0.058474	3	4	6.464932	ApbE family	apbE1 apbE2 apbE3
0.058474	3	4	6.464932	Glucuronate isomerase	uxaC1 uxaC2 uxaC3
0.058474	3	4	6.464932	Bacterial alpha-L-rhamnosidase concanavalin-like domain	AGF55542.1 ramA1 ramA2
0.058474	3	4	6.464932	UvrD-like helicase C-terminal domain	addA pcrA rep
0.058474	3	4	6.464932	Bacterial alpha-L-rhamnosidase 6 hairpin glycosidase domain	AGF55542.1 ramA1 ramA2
0.061965	12	47	2.200828	Pyridoxal phosphate	argD glgP1 metC2 hisC2 AGF55535.1 eryC patA alr phnW nspC bioA degT2
0.062043	27	142	1.638997	Lyase	leuC2 eutB rfbB1 coaBC ribBA metC2 cobD2 pel ilvD2 patB1 larC trpE uxaA pckA1 leuC3 citC1 aspA ubiD ansB2 leuC4 splB nspC uxuA3 mtbA uxuA4 eno2 AGF59045.1
0.068864	6	16	3.232466	Six-hairpin glycosidase-like superfamily	AGF55542.1 celK1 ramA1 ramA2 AGF57854.1 AGF58608.1
0.068864	6	16	3.232466	Peptidoglycan synthesis	murC murD mviN murE3 murF murG
0.069964	17	78	1.878698	Mixed, incl. hlyd family secretion protein, and ftsx-like permease family	AGF54515.1 macB4 AGF55212.1 AGF55603.1 AGF55982.1 AGF55983.1 AGF56579.1 AGF56580.1 AGF56581.1 AGF56582.1 AGF56955.1 AGF57140.1 AGF57142.1 AGF57143.1 AGF57543.1 AGF58189.1 AGF58693.1

0.073554	7	21	2.873303	Periplasmic binding protein	mglB2 mglB5 mglB7 mglB8 mglB9 mglB11 AGF58441.1
0.078038	5	12	3.591629	Predicted AAA-ATPase, and AAA domain	AGF55366.1 AGF55367.1 AGF55520.1 AGF59150.1 AGF59151.1
0.078038	5	12	3.591629	Citrate cycle (TCA cycle), and PrpF protein	AGF55505.1 AGF57496.1 leuC3 leuC4 fccA2
0.078038	5	12	3.591629	Domain of unknown function DUF11	AGF55708.1 AGF56478.1 AGF56877.1 AGF56878.1 AGF57396.1
0.078038	5	12	3.591629	Multi antimicrobial extrusion protein	AGF54190.1 mepA2 AGF56244.1 AGF56566.1 AGF58729.1
0.078038	5	12	3.591629	Secreted	pel AGF57267.1 AGF57467.1 hag2 eno2
0.078038	5	12	3.591629	Thiamine pyrophosphate enzyme, C-terminal TPP binding domain	alsS1 dxs1 dxs2 ilvB2 alsS2
0.078038	5	12	3.591629	Dihydropyrimidine dehydrogenase domain II, 4Fe-4S cluster	AGF54475.1 AGF57493.1 AGF57613.1 preA nuoG
0.079476	4	8	4.309955	Dockerin type I repeat	celK1 celA2 Endo-beta-mannanase celH
0.079476	4	8	4.309955	Dockerin domain	celK1 celA2 Endo-beta-mannanase celH
0.079476	4	8	4.309955	Dockerin domain superfamily	celK1 celA2 Endo-beta-mannanase celH
0.079476	4	8	4.309955	Isoprene biosynthesis	dxr ispG dxs1 dxs2
0.079476	4	8	4.309955	Oxidoreductase family, C-terminal alpha/beta domain	AGF56761.1 AGF58075.1 AGF58079.1 AGF58498.1
0.079476	4	8	4.309955	tRNA synthetases class I (M)	ileS metG3 cysS2 leuS
0.079476	4	8	4.309955	AAA domain	addA pcrA rep helD3

0.085956	8	27	2.554047	Chemotaxis methyl-accepting receptor HlyB-like, 4HB MCP domain	AGF54607.1 AGF55673.1 AGF57372.1 AGF57621.1 AGF57993.1 AGF58673.1 AGF59225.1 AGF59232.1
0.085956	8	27	2.554047	Four helix bundle sensory module for signal transduction	AGF54607.1 AGF55673.1 AGF57372.1 AGF57621.1 AGF57993.1 AGF58673.1 AGF59225.1 AGF59232.1
0.087323	6	17	3.042321	DEAD-like helicases superfamily, and DExx box DNA helicase domain superfamily	pcrA recG AGF56708.1 rep AGF56970.1 recQ2
0.088799	7	22	2.742698	Pyridine nucleotide-disulphide oxidoreductase	AGF54475.1 AGF56601.1 padH AGF57603.1 AGF57613.1 AGF59038.1 AGF59042.1
0.0917	13	56	2.00105	Cysteine and methionine metabolism, and Diaminopimelate biosynthesis	asnO metC2 AGF55406.1 AGF55413.1 AGF55535.1 metH3 patB1 AGF57072.1 AGF57274.1 AGF57283.1 aspA ansB2 AGF58560.1
0.098684	3	5	5.171946	Mixed, incl. betaine-homocysteine s-methyltransferase, bhmt, and putative c-s lyase	metC2 patB1 AGF58560.1
0.098684	13	57	1.965944	Mixed, incl. cell cycle, and peptidoglycan biosynthesis	murC murD murE1 ftsZ spoVD2 murE3 murF ftsW dacF mrdB alr murG AGF58788.1
0.098684	3	5	5.171946	Nitrogen fixation, and Nitrogenase iron-iron,	hcp3 anfD anfK

				delta subunit	
0.098684	3	5	5.171946	Glycogen biosynthesis, and Glucose-1-phosphate adenylyltransferase, GlgD subunit	glgD2 glgC2 glgB
0.098684	3	5	5.171946	Anticodon-binding domain of tRNA, and Lysine-tRNA ligase, class II	ileS lysS leuS
0.098684	3	5	5.171946	Alpha-L-arabinofuranosidase, C-terminal, and Cellulose binding, type IV	xynD1 xsa1 Alpha-L-arabinofuranosidase
0.098684	3	5	5.171946	Glycosyl hydrolases family 2, sugar binding domain, and L-arabinose isomerase	AGF57028.1 AGF58583.1 AGF58584.1
0.098684	6	18	2.873303	O-Antigen nucleotide sugar biosynthesis, and Glycosyl transferases group 1	mviN AGF54849.1 wbpA1 AGF59095.1 AGF59097.1 degT2
0.098684	3	5	5.171946	Mixed, incl. hpt domain, and metal-dependent hydrolase hdod	AGF56195.1 AGF56620.1 AGF57690.1
0.098684	3	5	5.171946	Mixed, incl. ferritin-like domain, and fist n domain	AGF56207.1 AGF56208.1 AGF56209.1
0.098684	3	5	5.171946	Mixed, incl. globin-like	cydC1 cydC2 cydB

				superfamily, and cytochrome ubiquinol oxidase subunit 1	
0.098684	3	5	5.171946	Glycoside hydrolase, family 4	malH2 AGF58813.1 Alpha-galactosidase/6-phospho-beta-glucosidase
0.098684	3	5	5.171946	MoaB/Mog domain	moeA2 moeA3 cinA
0.098684	3	5	5.171946	Thiolase	thlA2 thlA3 thlA4
0.098684	3	5	5.171946	Glycosyl transferase, family 28, C-terminal	ugtP AGF55586.1 murG
0.098684	7	23	2.623451	ABC transporter type 1, transmembrane domain	msbA1 AGF56192.1 cydC1 cydC2 AGF57659.1 msbA2 AGF58901.1
0.098684	3	5	5.171946	Thiamine pyrophosphate enzyme, central domain	alsS1 ilvB2 alsS2
0.098684	3	5	5.171946	Methionyl/Valyl/Leucyl/Isoleucyl-tRNA synthetase, anticodon-binding	ileS metG3 leuS
0.098684	3	5	5.171946	Thiolase, conserved site	thlA2 thlA3 thlA4
0.098684	3	5	5.171946	Thiolase, acyl-enzyme intermediate active site	thlA2 thlA3 thlA4
0.098684	3	5	5.171946	Thiolase, C-terminal	thlA2 thlA3 thlA4
0.098684	3	5	5.171946	Glycosyl hydrolase, family 4, C-terminal	malH2 AGF58813.1 Alpha-galactosidase/6-phospho-beta-glucosidase
0.098684	3	5	5.171946	Leucine-binding protein domain	AGF56800.1 AGF57306.1 braC
0.098684	3	5	5.171946	PDZ superfamily	AGF55043.1 rasP spoIIIAH

0.098684	24	129	1.603704	NAD(P)-binding domain superfamily	argC AGF54977.1 rfbB1 dxr hemA wbpA1 AGF55804.1 gutB AGF55972.1 AGF56695.1 AGF56761.1 uxaB AGF57365.1 AGF58075.1 AGF58079.1 AGF58498.1 gyaR malH2 maeB AGF58813.1 gap galE3 epsC Alpha-galactosidase/6-phospho-beta-glucosidase
0.098684	3	5	5.171946	MoaB/Mog-like domain superfamily	moeA2 moeA3 cinA
0.098684	7	23	2.623451	ABC transporter type 1, transmembrane domain superfamily	msbA1 AGF56192.1 cydC1 cydC2 AGF57659.1 msbA2 AGF58901.1
0.098684	5	13	3.31535	Beta-ketoacyl synthase, N-terminal domain	fabF1 thlA2 thlA4 fabF2 fabF3
0.098684	7	23	2.623451	Aminotransferase class-V	metC2 hisC2 AGF55413.1 eryC AGF57274.1 phnW degT2
0.098684	7	23	2.623451	ABC transporter transmembrane region	msbA1 AGF56192.1 cydC1 cydC2 AGF57659.1 msbA2 AGF58901.1
0.098684	3	5	5.171946	Probable molybdopterin binding domain	moeA2 moeA3 cinA
0.098684	3	5	5.171946	Receptor family ligand binding region	AGF56800.1 AGF57306.1 braC
0.098684	3	5	5.171946	Family 4 glycosyl hydrolase	malH2 AGF58813.1 Alpha-galactosidase/6-phospho-beta-glucosidase
0.098684	3	5	5.171946	Thiolase, C-terminal domain	thlA2 thlA3 thlA4
0.098684	3	5	5.171946	Family 4 glycosyl hydrolase C-terminal domain	malH2 AGF58813.1 Alpha-galactosidase/6-phospho-beta-glucosidase
0.098684	3	5	5.171946	Alanine-glyoxylate amino-transferase	ycxD1 ycxD2 AGF58224.1
0.098684	3	5	5.171946	Periplasmic binding	AGF56800.1 AGF57306.1 braC

				protein domain	
0.098684	8	28	2.462831	NAD(P)-binding Rossmann-like domain	AGF54475.1 AGF55505.1 AGF56601.1 AGF57493.1 AGF57613.1 nuoG AGF59038.1 AGF59042.1
0.098684	5	13	3.31535	Iron-containing alcohol dehydrogenase	gldA fucO AGF57742.1 AGF57892.1 egsA
0.101847	20	103	1.673769	ABC transporter, conserved site	yfmM macB4 msbA1 metN2 xylG AGF56192.1 cysA AGF56796.1 fbpC AGF56830.1 AGF56968.1 cydC1 cydC2 AGF57659.1 AGF58001.1 mglA AGF58440.1 AGF58451.1 msbA2 AGF58901.1
0.103135	10	40	2.154977	Mixed, incl. biotin-lipoyl like, and ftsx-like permease family	AGF54515.1 macB4 AGF55982.1 AGF55983.1 AGF56582.1 AGF57140.1 AGF57142.1 AGF57143.1 AGF57543.1 AGF58189.1
0.103972	4	9	3.831071	Gfo/Idh/MocA-like oxidoreductase, C-terminal	AGF56761.1 AGF58075.1 AGF58079.1 AGF58498.1
0.103972	4	9	3.831071	Glycosyltransferase subfamily 4-like, N-terminal domain	AGF54849.1 AGF54981.1 AGF55681.1 AGF56357.1
0.103972	4	9	3.831071	Nitrogen fixation	nifK3 nifB2 anfD anfK
0.118765	7	24	2.51414	Mixed, incl. helicase, and dna topoisomerase 3-like, toprim domain	pcrA recG AGF56708.1 rep AGF56970.1 recQ2 mppE
0.118765	7	24	2.51414	Mixed, incl. drug resistance transporter emrb-like, and hlyd family secretion protein	AGF55212.1 AGF55603.1 AGF56579.1 AGF56580.1 AGF56581.1 AGF56955.1 AGF58693.1
0.119383	10	41	2.102417	Bacterial chemotaxis	mcp41 AGF56146.1 AGF56683.1 AGF57372.1 AGF57621.1 AGF58050.1

					AGF58383.1 AGF59225.1 cheB2 AGF59232.1
0.120632	20	105	1.641888	Isomerase	gyrB1 pgi pgcA dxr uxaC1 AGF56125.1 AGF56695.1 AGF57072.1 uxaC2 AGF57496.1 rhaA AGF58041.1 alr ppm araA mro AGF58584.1 uxaC3 galE3 AGF59097.1
0.1223	5	14	3.078539	DHS-like NAD/FAD-binding domain superfamily	etfA3 alsS1 etfA4 ilvB2 alsS2
0.1223	5	14	3.078539	Iron-containing alcohol dehydrogenase	gldA fucO AGF57742.1 AGF57892.1 egsA
0.122612	6	19	2.722077	Mixed, incl. abc transporter, and rna helicase, dead-box type, q motif	yfmM rsmB AGF56796.1 AGF56808.1 dbpA AGF57119.1
0.122612	6	19	2.722077	Cell wall biogenesis/ degradation	murC murD mviN murE3 murF murG
0.122612	6	19	2.722077	Sugar (and other) transporter	mdtG AGF55976.1 tetA AGF56190.1 AGF56696.1 AGF58672.1
0.124863	16	79	1.745804	Two-component regulatory system, and Signal transduction histidine kinase, dimerisation/phosphoacceptor domain	luxQ AGF54911.1 AGF55492.1 resE6 AGF56195.1 AGF56222.1 AGF56258.1 AGF56620.1 AGF56764.1 dltS AGF57341.1 aruS AGF57535.1 AGF57690.1 AGF58648.1 AGF58650.1
0.127188	8	30	2.298643	Glucose permease domain IIB, and PRD domain	licR3 bglP5 manR bglP8 malH2 AGF58813.1 AGF58814.1 malX
0.127188	8	30	2.298643	DNA damage	recF addB addA ligA recG ruvB mutL uvrC2
0.127188	8	30	2.298643	DNA repair	recF addB addA ligA recG ruvB mutL uvrC2
0.133399	9	36	2.154977	ATPase family associated with various	dnaA AGF55094.1 AGF55366.1 AGF55850.1 ruvB ftsH3 AGF57055.1 ftsH4 dnaX2

				cellular activities (AAA)	
0.141081	7	25	2.413575	Pyruvate metabolism, and Thiamine pyrophosphate enzyme, central domain	alsS1 ilvB2 pckA1 alsS2 maeB eno2 pyk2
0.141081	7	25	2.413575	Mixed, incl. benzoate degradation, and electron transfer flavoprotein domain	bcd1 etfA3 thlA2 thlA3 thlA4 bcd3 etfA4
0.141081	7	25	2.413575	Tar ligand binding domain homologue	AGF54607.1 AGF55673.1 AGF57372.1 AGF57621.1 AGF57993.1 AGF58673.1 AGF59232.1
0.144824	4	10	3.447964	Oxidoreductase, N-terminal	AGF56761.1 AGF58075.1 AGF58079.1 AGF58498.1
0.144824	4	10	3.447964	Aldehyde/histidinol dehydrogenase	hisD eutE gbsA gapN
0.144824	4	10	3.447964	Oxidoreductase family, NAD-binding Rossmann fold	AGF56761.1 AGF58075.1 AGF58079.1 AGF58498.1
0.144824	4	10	3.447964	NMT1/THI5 like	ssuA AGF56610.1 tauA AGF57888.1
0.144824	4	10	3.447964	Glycosyltransferase Family 4	AGF54849.1 AGF54981.1 AGF55681.1 AGF56357.1
0.145336	12	55	1.880708	Acyltransferase	baeE argJ pfl1 fabF1 thlA2 thlA3 thlA4 AGF56359.1 AGF56635.1 fabF2 leuA3 fabF3
0.145336	12	55	1.880708	Protein biosynthesis	ileS lysS fusA1 trpS alaS infB metG3 cysS2 pheS pheT argS2 leuS
0.147151	10	43	2.00463	Valine, leucine and isoleucine biosynthesis, and Pyruvate	leuC2 alsS1 ilvD2 ilvB2 leuA3 pckA1 alsS2 maeB eno2 pyk2
0.147151	3	6	4.309955	Aconitase, putative,	AGF57496.1 leuC3 leuC4

				and PrpF protein	
0.147151	3	6	4.309955	RNA helicase, DEAD-box type, Q motif, and GTPase, MTG1	rsmB AGF56808.1 dbpA
0.147151	3	6	4.309955	Mixed, incl. glucose-1-phosphate adenylyltransferase, and pullulanase, type i	pgcA pulA glgD1
0.147151	9	37	2.096735	Mixed, incl. pentose and glucuronate interconversions, and fcd domain	uxaC1 uxaA uxaB uxaC2 AGF58118.1 uxaC3 uxuA3 uxuA4 dgoT
0.147151	3	6	4.309955	Uronate isomerase, and Mannonate dehydratase	uxaC3 uxuA3 uxuA4
0.147151	3	6	4.309955	Class I and II aminoacyl-tRNA synthetase, tRNA-binding arm, and B3/4 domain	trpS pheS pheT
0.147151	3	6	4.309955	HRDC domain, and DExx box DNA helicase domain superfamily	pcrA recG recQ2
0.147151	3	6	4.309955	Mixed, incl. phage tail lysozyme, and copper amine oxidase-like, n-terminal	AGF54931.1 AGF56591.1 AGF59244.1
0.147151	3	6	4.309955	Mixed, incl. dna topoisomerase, type iia-like domain superfamily,	dnaA recF gyrB1

				and s4 domain	
0.147151	3	6	4.309955	DHHA1 domain	alaS recJ2 alaXL
0.147151	3	6	4.309955	Aminotransferase class-III	argD patA bioA
0.147151	3	6	4.309955	Thiamine pyrophosphate enzyme, N-terminal TPP-binding domain	alsS1 ilvB2 alsS2
0.147151	8	31	2.224493	Immunoglobulin-like fold	pulA cbpA celK1 ramA1 ramA2 bglA5 Polygalacturonase glgB
0.147151	3	6	4.309955	Thiolase, N-terminal	thlA2 thlA3 thlA4
0.147151	3	6	4.309955	Beta-xylosidase, C-terminal Concanavalin A-like domain	xynB3 xylB2 xylB4
0.147151	3	6	4.309955	DNA-directed DNA polymerase	dnaN polC dnaX2
0.147151	45	289	1.3422	Transport	secY atpA rnfC kdpB nrgA1 mviN macB4 kup2 AGF55212.1 AGF55511.1 AGF55515.1 celB1 AGF55538.1 AGF55603.1 metN2 mtlA xylG xylH1 cysA AGF56753.1 AGF56955.1 AGF57304.1 bglP5 AGF57584.1 AGF57643.1 fruA2 lctP2 AGF57903.1 AGF57998.1 AGF58001.1 AGF58068.1 AGF58113.1 xylH2 AGF58440.1 AGF58457.1 AGF58579.1 AGF58672.1 AGF58693.1 bglP8 AGF58814.1 AGF58925.1 AGF59111.1 AGF59199.1 malX hppA
0.147151	3	6	4.309955	tRNA synthetases class I (I, L, M and V)	ileS metG3 leuS
0.147151	3	6	4.309955	Flavin containing amine oxidoreductase	glpA AGF57613.1 nuoG
0.147151	3	6	4.309955	Choline/ethanolamine kinase	AGF54324.1 cotS2 AGF59131.1

0.147151	3	6	4.309955	Thiamine pyrophosphate enzyme, N-terminal TPP binding domain	alsS1 ilvB2 alsS2
0.147151	3	6	4.309955	UDP-glucose/GDP-mannose dehydrogenase family, NAD binding domain	AGF54977.1 wbpA1 galE3
0.147151	3	6	4.309955	Glycosyltransferase family 28 C-terminal domain	ugtP AGF55586.1 murG
0.147151	3	6	4.309955	Mga helix-turn-helix domain	licR1 licR3 manR
0.147151	3	6	4.309955	MFS_1 like family	mdtG AGF56190.1 AGF56696.1
0.147151	3	6	4.309955	NMT1-like family	ssuA tauA AGF57888.1
0.149593	20	109	1.581635	Mixed, incl. 4fe-4s ferredoxin-type, iron-sulphur binding domain, and sulfur metabolism	AGF54511.1 dmsA AGF56342.1 sbp cysA AGF56735.1 AGF56769.1 asrA narB padH AGF57077.1 AGF57493.1 mco AGF57603.1 hndC1 AGF57613.1 hypF hypD nuoG AGF58672.1
0.151355	23	130	1.525061	Mixed, incl. signal transduction histidine kinase, dimerisation/phosphoacceptor domain, and ompr/phob-type dna-binding domain	AGF54179.1 luxQ kdpB AGF54911.1 glnK AGF55492.1 resE6 AGF56195.1 AGF56207.1 AGF56208.1 AGF56209.1 AGF56222.1 AGF56258.1 AGF56620.1 AGF56764.1 dltS AGF57341.1 aruS AGF57535.1 AGF57690.1 AGF58121.1 AGF58648.1 AGF58650.1
0.153701	7	26	2.320745	Chemotaxis methyl-accepting receptor	AGF54607.1 AGF57372.1 AGF57621.1 AGF57993.1 AGF58673.1 AGF59225.1 AGF59232.1
0.160381	8	32	2.154977	Aminoacyl-tRNA biosynthesis	ileS lysS trpS metG3 pheS pheT argS2 leuS

0.160381	10	44	1.95907	Phosphotransferase system, and PRD domain	mtlA licR3 bglP5 fruA2 manR bglP8 malH2 AGF58813.1 AGF58814.1 malX
0.161779	9	38	2.041558	GTP-binding	fusA1 ftsZ engA ffh infB ribBA AGF56830.1 rsgA2 hflX
0.172072	4	11	3.134513	Arginine biosynthesis, and Proline biosynthesis	argD argJ argC argG
0.172072	6	22	2.350884	Nicotinate and nicotinamide metabolism, and Riboflavin metabolism	nadB nadE coaBC ribBA pncB cinA
0.172072	4	11	3.134513	Benzoate degradation, and MaoC like domain	bcd1 thlA2 thlA4 bcd3
0.172072	3	7	3.694247	Mixed, incl. multicopper oxidase, and nitronate monooxygenase	AGF54511.1 AGF56769.1 mco
0.172072	8	33	2.089675	Phosphotransferase system (PTS)	AGF55511.1 celB1 AGF55533.1 bglC2 AGF58113.1 AGF58457.1 AGF59111.1 AGF59117.1
0.172072	7	27	2.234791	Pentose and glucuronate interconversions, and FCD domain	uxaC1 uxaA uxaB uxaC2 uxaC3 uxuA3 uxuA4
0.172072	3	7	3.694247	Glycosyl hydrolases family 39, and Beta-xylosidase, C-terminal Concanavalin A-like domain	xynB3 xynB5 xylB2
0.172072	3	7	3.694247	Mixed, incl. protein of unknown function (duf1624), and glycosyltransferase wbsx	AGF56357.1 AGF56359.1 AGF56361.1

0.172072	3	7	3.694247	Type I protein exporter, and Cation/H+ exchanger	AGF56192.1 AGF57903.1 AGF58901.1
0.172072	3	7	3.694247	Mixed, incl. pd-(d/e)xk endonuclease-like domain, addab-type, and recr protein	addB addA recJ2
0.172072	2	3	5.746606	ATP-dependent RNA helicase DEAD-box, conserved site	dbpA AGF57119.1
0.172072	2	3	5.746606	Leu/Ile/Val-binding protein	AGF56800.1 braC
0.172072	2	3	5.746606	Peptidase S9, prolyl oligopeptidase, catalytic domain	xynB4 AGF58542.1
0.172072	3	7	3.694247	Aminoacyl-tRNA synthetase, class I, conserved site	ileS trpS metG3
0.172072	2	3	5.746606	Glycoside hydrolase, family 5	engD2 Endo-beta-mannanase
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogenase, N-terminal	AGF54977.1 wbpA1
0.172072	2	3	5.746606	UbiD decarboxylase family	ubiD AGF59045.1
0.172072	2	3	5.746606	RecF/RecN/SMC, N-terminal	recF smc
0.172072	3	7	3.694247	Outer membrane efflux protein	AGF56582.1 AGF57143.1 AGF57543.1

0.172072	2	3	5.746606	FAD-linked oxidase, C-terminal	glcD1 AGF57896.1
0.172072	2	3	5.746606	Carbohydrate binding module family 6	xynD1 xynD2
0.172072	2	3	5.746606	Biotin carboxylase-like, N-terminal domain	accC1 accC2
0.172072	2	3	5.746606	Biotin carboxylase, C-terminal	accC1 accC2
0.172072	3	7	3.694247	FAD linked oxidase, N-terminal	glcD1 AGF56943.1 AGF57896.1
0.172072	2	3	5.746606	Aconitase, putative	leuC3 leuC4
0.172072	2	3	5.746606	Gp5/Type VI secretion system Vgr protein, OB-fold domain	AGF55355.1 AGF57428.1
0.172072	2	3	5.746606	Cellulose binding, type IV	xynD1 xynD2
0.172072	2	3	5.746606	Baseplate protein J-like	AGF54565.1 xkdT
0.172072	2	3	5.746606	DNA helicase, DnaB-like, N-terminal	dnaG dnaC1
0.172072	2	3	5.746606	Peptidase M16, C-terminal	AGF55406.1 AGF57933.1
0.172072	2	3	5.746606	Alpha-L-rhamnosidase, concanavalin-like domain	ramA1 ramA2
0.172072	2	3	5.746606	Cupredoxin	mco cotA
0.172072	2	3	5.746606	Alanine racemase/group IV decarboxylase, C-terminal	alr nspC
0.172072	2	3	5.746606	Valyl/Leucyl/Isoleucyl-	ileS leuS

				tRNA synthetase, editing domain	
0.172072	2	3	5.746606	Aspartate decarboxylase-like domain superfamily	dmsA narB
0.172072	2	3	5.746606	Alpha-L-arabinofuranosidase, C-terminal	xsa1 Alpha-L-arabinofuranosidase
0.172072	2	3	5.746606	Biotin carboxylation domain	accC1 accC2
0.172072	2	3	5.746606	Peptidase M16, N-terminal	AGF55406.1 AGF57933.1
0.172072	2	3	5.746606	Threonyl/alanyl tRNA synthetase, SAD	alaS alaXL
0.172072	3	7	3.694247	Phosphotransferase system, EIIB component, type 2	mtlA licR3 manR
0.172072	2	3	5.746606	Phosphotransferase system, EIIC component, type 2	mtlA fruA2
0.172072	2	3	5.746606	Bacterial alpha-L-rhamnosidase N-terminal	ramA1 ramA2
0.172072	2	3	5.746606	DExx box DNA helicase domain superfamily	pcrA rep
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogenase, dimerisation	AGF54977.1 wbpA1
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogen	AGF54977.1 wbpA1

				ase, C-terminal	
0.172072	2	3	5.746606	Malic enzyme, conserved site	AGF57365.1 maeB
0.172072	2	3	5.746606	DNA helicase DnaB, N-terminal/DNA primase DnaG, C-terminal	dnaG dnaC1
0.172072	2	3	5.746606	FAD-linked oxidase-like, C-terminal	glcD1 AGF57896.1
0.172072	2	3	5.746606	Vanillyl-alcohol oxidase, C-terminal subdomain 2	glcD1 AGF57896.1
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogenase	AGF54977.1 wbpA1
0.172072	2	3	5.746606	DHBP synthase RibB-like alpha/beta domain superfamily	ribBA hypF
0.172072	2	3	5.746606	Folylpolyglutamate synthetase, conserved site	murE3 folC
0.172072	3	7	3.694247	Beta-ketoacyl synthase, active site	fabF1 fabF2 fabF3
0.172072	2	3	5.746606	Phosphodiester glycosidase	AGF54875.1 AGF58012.1
0.172072	2	3	5.746606	Tail sheath protein, C-terminal domain	xkdK AGF56462.1
0.172072	2	3	5.746606	Bacilysin exporter BacE, putative	AGF57324.1 AGF58118.1

0.172072	7	28	2.154977	FAD/NAD(P)-binding domain	padH AGF57493.1 AGF57603.1 AGF57613.1 nuoG AGF59038.1 AGF59042.1
0.172072	2	3	5.746606	Fumarase/histidase, N-terminal	aspA ansB2
0.172072	2	3	5.746606	Recombinase zinc beta ribbon domain	AGF54764.1 AGF56868.1
0.172072	10	46	1.873893	Periplasmic binding protein-like I	mglB2 mglB5 AGF56800.1 mglB7 AGF57306.1 mglB8 mglB9 braC mglB11 AGF58441.1
0.172072	2	3	5.746606	Enolase-like, N-terminal	AGF57072.1 eno2
0.172072	2	3	5.746606	Enolase C-terminal domain-like	AGF57072.1 eno2
0.172072	3	7	3.694247	ABC-transporter extension domain	yfmM AGF56796.1 AGF56968.1
0.172072	2	3	5.746606	Electron transfer flavoprotein, alpha subunit, N-terminal	etfA3 etfA4
0.172072	2	3	5.746606	Alpha-L-rhamnosidase C-terminal domain	ramA1 ramA2
0.172072	2	3	5.746606	MurE/MurF, N-terminal	murE3 murF
0.172072	2	3	5.746606	DNA helicase, DnaB-like, N-terminal domain superfamily	dnaG dnaC1
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogenase, C-terminal domain superfamily	AGF54977.1 wbpA1
0.172072	2	3	5.746606	Enolase-like, C-terminal	AGF57072.1 eno2

				domain superfamily	
0.172072	2	3	5.746606	UROD/MetE-like superfamily	AGF58560.1 mtbA
0.172072	2	3	5.746606	Glycosyltransferase RgtA/B/C/D-like	AGF56553.1 AGF56554.1
0.172072	2	3	5.746606	Molybdopterin biosynthesis protein MoeA-like	moeA2 moeA3
0.172072	2	3	5.746606	PDZ domain 6	AGF55043.1 spoIIIAH
0.172072	2	3	5.746606	Glycogen metabolism	glgC2 glgB
0.172072	11	53	1.789038	Glycosyltransferase	ugtP glgP1 cobT hisZ pncB AGF55681.1 pgaC trpD murG glgB gtfA2
0.172072	5	16	2.693722	Lipoprotein	apbE1 apbE2 apbE3 AGF58189.1 AGF58925.1
0.172072	8	33	2.089675	Manganese	ligA ribBA pckA1 rhaA ubiD araA uxuA3 uxuA4
0.172072	5	16	2.693722	Metalloprotease	rasP yjbG ftsH3 AGF57467.1 ftsH4
0.172072	2	3	5.746606	Rhamnose metabolism	rhaA rhaB
0.172072	3	7	3.694247	Thiolase, N-terminal domain	thlA2 thlA3 thlA4
0.172072	2	3	5.746606	Biotin carboxylase, N-terminal domain	accC1 accC2
0.172072	4	11	3.134513	Prolyl oligopeptidase family	AGF55795.1 xynB4 AGF56705.1 AGF58542.1
0.172072	2	3	5.746606	Insulinase (Peptidase family M16)	AGF55406.1 AGF57933.1
0.172072	2	3	5.746606	DnaB-like helicase N terminal domain	dnaG dnaC1
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogenase family, central domain	AGF54977.1 wbpA1
0.172072	5	16	2.693722	PAS fold	luxQ AGF55850.1 AGF56222.1 rpfG7 AGF58793.1

0.172072	5	16	2.693722	Cys/Met metabolism PLP-dependent enzyme	metC2 AGF55413.1 patB1 AGF56997.1 degT2
0.172072	3	7	3.694247	FAD binding domain	glcD1 AGF56943.1 AGF57896.1
0.172072	2	3	5.746606	3-octaprenyl-4-hydroxybenzoate carboxy-lyase	ubiD AGF59045.1
0.172072	2	3	5.746606	Biotin carboxylase C-terminal domain	accC1 accC2
0.172072	2	3	5.746606	FAD linked oxidases, C-terminal domain	glcD1 AGF57896.1
0.172072	2	3	5.746606	Carbohydrate binding module (family 6)	xynD1 xynD2
0.172072	2	3	5.746606	UDP-glucose/GDP-mannose dehydrogenase family, UDP binding domain	AGF54977.1 wbpA1
0.172072	2	3	5.746606	Substrate binding domain of ABC-type glycine betaine transport system	ssuA tauA
0.172072	2	3	5.746606	Baseplate J-like protein	AGF54565.1 xkdT
0.172072	2	3	5.746606	Phage tail sheath protein subtilisin-like domain	xkdK AGF56462.1
0.172072	2	3	5.746606	Peptidase M16 inactive domain	AGF55406.1 AGF57933.1
0.172072	2	3	5.746606	Alpha-L-arabinofuranosidase C-	xsa1 Alpha-L-arabinofuranosidase

				terminal domain	
0.172072	3	7	3.694247	D-ala D-ala ligase C-terminus	accC1 accC2 AGF58788.1
0.172072	2	3	5.746606	Threonyl and Alanyl tRNA synthetase second additional domain	alaS alaXL
0.172072	2	3	5.746606	Anticodon-binding domain of tRNA	ileS leuS
0.172072	2	3	5.746606	Alpha-L-rhamnosidase N-terminal domain	ramA1 ramA2
0.172072	3	7	3.694247	ABC transporter	yfmM AGF56796.1 AGF56968.1
0.172072	2	3	5.746606	Dolichyl-phosphate-mannose-protein mannosyltransferase	AGF56553.1 AGF56554.1
0.172072	2	3	5.746606	Bacterial alpha-L-rhamnosidase C-terminal domain	ramA1 ramA2
0.172072	2	3	5.746606	Phage tail sheath C-terminal domain	xkdK AGF56462.1
0.172072	2	3	5.746606	Cellulose Binding Domain Type IV	xynD1 xynD2
0.172072	2	3	5.746606	PASTA	sps spoVD2
0.172072	2	3	5.746606	UDP binding domain	AGF54977.1 wbpA1
0.175292	5	17	2.535268	Fatty acid metabolism	fabF1 accC1 fabF2 accC2 fabF3
0.175292	5	17	2.535268	Fatty acid biosynthesis	fabF1 accC1 fabF2 accC2 fabF3
0.183611	4	12	2.873303	Nicotinate and nicotinamid	nadB nadE pncB cinA

				e metabolism	
0.183611	4	12	2.873303	Mixed, incl. rna helicase, dead-box type, q motif, and abc transporter	yfmM rsmB AGF56808.1 dbpA
0.183611	4	12	2.873303	Mixed, incl. udp-n-acetylglucosamine 2-epimerase, and udp-n-acetyl-d-mannosamine/glucosamine dehydrogenase	wbpA1 AGF59095.1 AGF59097.1 degT2
0.183611	4	12	2.873303	Mostly uncharacterized, incl. phage tail fibre protein, and baseplate j-like protein	AGF56462.1 AGF56466.1 AGF56469.1 xkdT
0.183611	4	12	2.873303	Alcohol dehydrogenase, iron-type/glycerol dehydrogenase GldA	gldA fucO AGF57742.1 AGF57892.1
0.183611	4	12	2.873303	Glucose inhibited division protein A	AGF56601.1 AGF57613.1 AGF59038.1 AGF59042.1

Appendix VII – Annotated sequence features of tested promoters

Figure VII.1 All promoters with identified features and motifs.

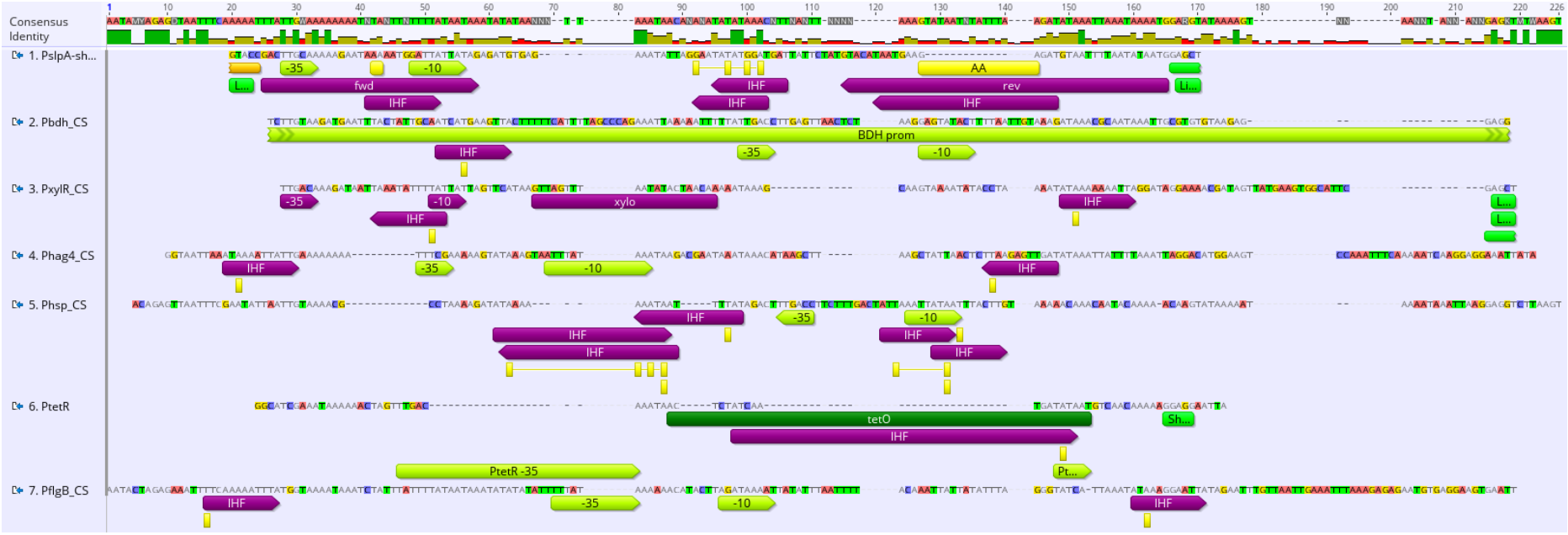

Figure VII.2 High expressing promoters with identified features and motifs.

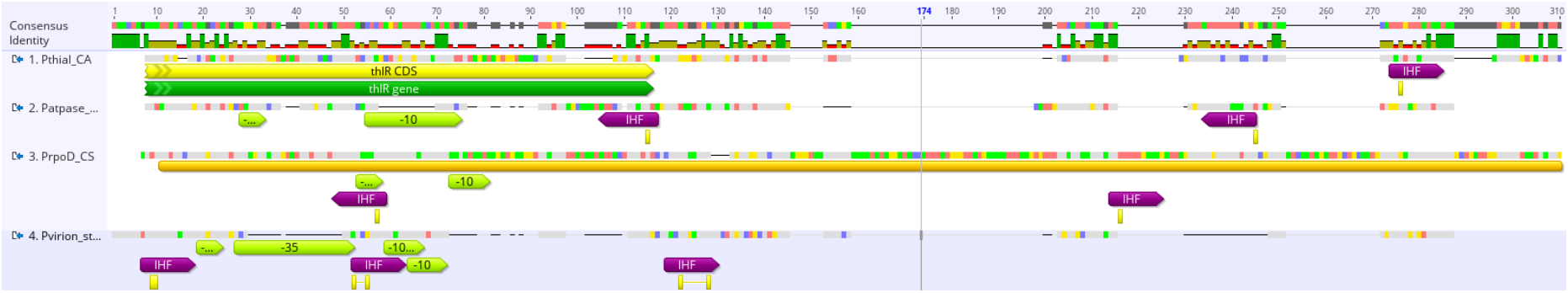

Figure VII.3 Low expressing promoters with identified features and motifs.

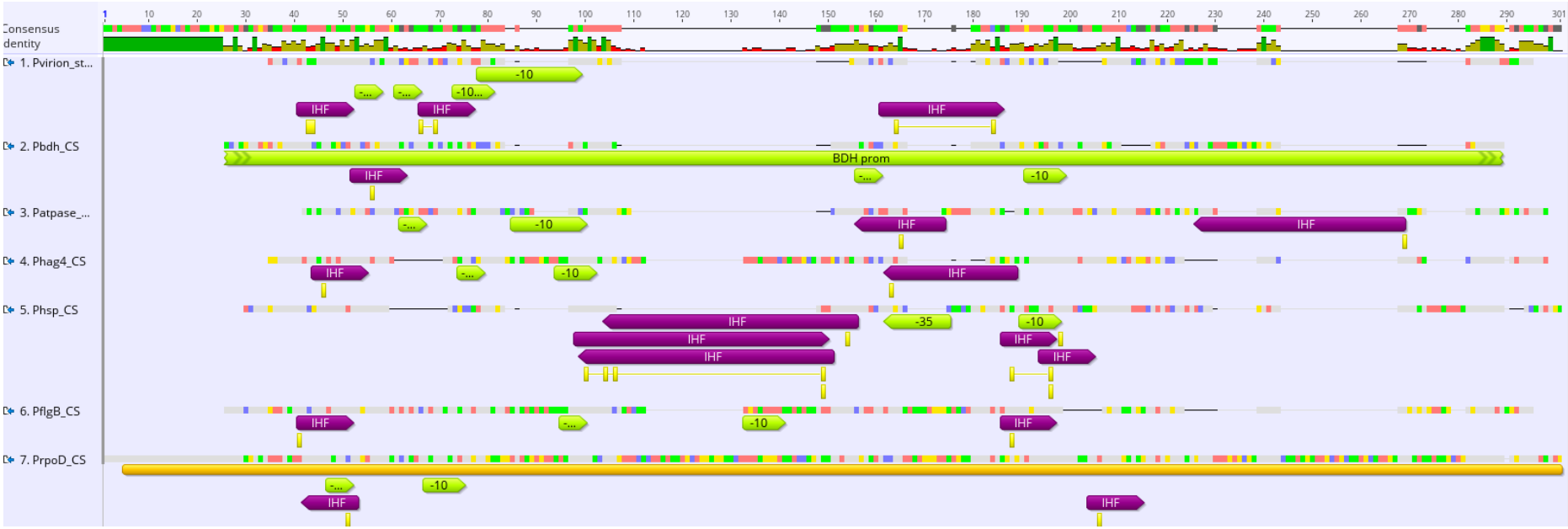

Figure VII.4 *C. saccharoperbutylacetonicum* derived promoters with identified features and motifs.