This is a repository copy of Part by part: synthetic biology parts used in solventogenic Clostridia.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/125408/

Version: Accepted Version

Article:
Gyulev, Ivan, Willson, Benjamin James, Hennessy, Rosanna Catherine et al. (3 more authors) (2018) Part by part: synthetic biology parts used in solventogenic Clostridia. ACS Synthetic Biology. pp. 311-327. ISSN 2161-5063

https://doi.org/10.1021/acssynbio.7b00327

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
Part by part: synthetic biology parts used in solventogenic Clostridia

Authors:

Ivan S. Gyulev¹*, Benjamin J. Willson¹*, Rosanna C. Hennessy², Preben Krabben³, Elizabeth R. Jenkinson⁴, Gavin H. Thomas¹§

Author affiliation:

1. Department of Biology, University of York, Wentworth Way, York, YO10 5DD, United Kingdom
2. Microbial Ecology and Biotechnology, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark
3. 9 Marsh Lane, Didcot, Oxfordshire, OX11 8FD, United Kingdom
4. Green Biologics Limited, Milton Park, Abingdon, Oxfordshire, OX14 4RU, United Kingdom

* These authors contributed equally to this work.

§ - Phone: 44-1904-328678. E-mail: gavin.thomas@york.ac.uk.
ABSTRACT: The solventogenic Clostridia are of interest to the chemical industry because of their natural ability to produce chemicals such as butanol, acetone and ethanol from diverse feedstocks. Their use as whole cell factories presents multiple metabolic engineering targets that could lead to improved sustainability and profitability of Clostridium industrial processes. However, engineering efforts have been held back by the scarcity of genetic and synthetic biology tools. Over the last decade, genetic tools to enable transformation and chromosomal modifications have been developed, but the lack of a broad palette of synthetic biology parts remains one of the last obstacles to the rapid engineered improvement of these species for bioproduction. We have systematically reviewed existing parts that have been used in the modification of solventogenic Clostridia, revealing a narrow range of empirically chosen and non-engineered parts that are in current use. The analysis uncovers elements, such as promoters, transcriptional terminators and ribosome binding sites where increased fundamental knowledge is needed for their reliable use in different applications. Together, the review provides the most comprehensive list of parts used and also presents areas where an improved toolbox is needed for full exploitation of these industrially important bacteria.

Keywords

Clostridium, ABE fermentation, solventogenesis, transcription termination, reporters, promoters
Abbreviations

ABE: acetone-butanol-ethanol
ACE: allele-coupled exchange
aSD: anti Shine-Dalgarno sequence, found at the 3’ end of 16S rRNA
CDS: coding sequence
FbFP: flavin-binding fluorescent protein
RBS: ribosome binding site
RNAP: RNA-polymerase
SD: Shine-Dalgarno sequence
UTR: untranslated region

Acknowledgements:

We acknowledge the BBSRC for support for Ivan Gyulev through an iCASE PhD studentship with Green Biologics Ltd (BB/M016331/1), Benjamin Willson from an IB Catalyst grant for Project DETOX (BB/N01040X/1) and Rosanna Hennessy for a BBSRC IUK award (BB/L011522/1). We thank Stephen Thorpe for writing a Python script that we used to compare the overlaps of the outputs of the terminator prediction algorithms.
One of the earliest industrial biotechnology processes used by man was the production of acetone by bacterial fermentation. The bacterium used was from the genus *Clostridium*, which are Gram-positive, spore forming, obligately anaerobic organisms. In addition to acetone, these processes produced butanol and ethanol, leading to the fermentation being given the name ‘ABE’ for ‘acetone-butanol-ethanol’. ABE-producing clostridia typically show a biphasic growth pattern, producing acetic and butyric acids during the early stages of growth, and then undergoing a metabolic 'switch', reassimilating the produced acids and producing solvents. Driven by the demands of industry, many different ABE-producing clostridial isolates were identified; those strains that were used in industry are primarily represented by four species, namely *Clostridium acetobutylicum*, *Clostridium beijerinckii*, *Clostridium saccharobutylicum*, and *Clostridium saccharoperbutylacetonicum*, although ABE fermentation has been observed in other clostridia. Furthermore, not all butanol-producing clostridia produce the canonical trio of solvent products; for example, *Clostridium pasteurianum* produces butanol and 1,3-propanediol (instead of acetone) from glycerol. While the ABE process had until recently fallen out of favour due to competition from the petrochemical industry, the necessity of identifying alternative fuels has renewed interest in the production of butanol as a potential biofuel candidate and sustainable commodity chemical.
Figure 1. Phylogenetic tree of 16S sequences from selected organisms from the genus Clostridium. The tree was built using Maximum Likelihood method based on the Tamura-Nei model\(^1\). The bootstrap consensus tree inferred from 1000 replicates\(^2\) is taken to represent the evolutionary history of the taxa analyzed\(^3\). Evolutionary analyses were conducted in MEGA5\(^4\). Green circles denote butanol-producing species used in industrial biotechnology (IB), red squares mark risk group 2 species\(^5\) (risk group 3 being highest risk).
A recent comparative genomics study of many industrial saccharolytic strains (those preferring sugar as a carbon source) revealed that the known solventogens fall within two sister clades: one exemplified by *C. acetobutylicum* and one by *C. beijerinckii* (Figure 1). Interestingly, another comparative genomics paper which included more genus-wide species but fewer industrial solventogens supported the split, with the genus’ type species *Clostridium butyricum* being more closely related to *C. beijerinckii* and the pathogen *Clostridium tetani* clustering closer to *C. acetobutylicum*. Altogether, these findings serve to re-iterate that complex traits within the Clostridia such as pathogenicity are paraphyletic (also see Fig. 1). On the other hand, solventogenesis (of ethanol, butanol or acetone) may be very widespread in the genus, but there has not yet been a definitive comparative study reporting the extent of its conservation to our knowledge, and species and strains certainly vary in their productivity. Topics of engineering interest have included improved characteristics such as solvent production, sugar utilisation, growth on alternative feedstocks such as lignocellulose and the production of alternative products.

**A shortage of biological parts for the solventogenic Clostridia**

To manipulate the solventogenic Clostridia using rational engineering, the development of reliable standard genetic parts for use in synthetic biology is essential. The assembly of these parts into synthetic gene cassettes and larger gene clusters means that they must work together in a consistent and predictable manner to become a useful tool for strain engineering. The basic parts of synthetic biology are the minimal sequence elements with biological function in gene expression (Figure 2), including promoters, ribosome binding sites, transcriptional terminators and other factors in the mRNA that affect stability. Also, synthetic biology tools rely on gene reporters to measure levels of gene expression, and these elements are often combined into plasmids which need their own origins of replication. Work in *Escherichia coli, Saccharomyces cerevisiae* and other model organisms has been transformed in the past decade by the rapid development, characterization and standardization of parts. However, work in these organisms benefits from a legacy of biological knowledge that is not necessarily available when working in non-model organisms, making the engineering of the latter even more
challenging.

Figure 2. Features of genetic constructs depicted using symbols from the SBOLv\textsuperscript{29}. Examples of promoters, ribosomal binding sites, reporters, terminators and replicons are described in the main text.

It is also important to note that, of the existing body of molecular biology research in this genus, the majority has been carried out in *C. acetobutylicum* ATCC 824. Work in this species has benefited from a relatively early development of a reliable electroporation protocol\textsuperscript{30} and from the publication of the whole genome sequence in 2001, making it the first published clostridial genome\textsuperscript{31}. However, *C. acetobutylicum* is only one of many industrially promising solventogenic clostridia. While the engineering of some of these species has historically been hindered by a lack of transformation protocols, these are now available for the transformation of the aforementioned five industrial species\textsuperscript{30,32-35}. In several cases, development of transformation protocols has required the circumvention of restriction systems which degrade incorrectly methylated DNA. Indeed, electroporation of *C. acetobutylicum* only became viable with the expression of the *B. subtilis* phage $\phi$3T I methylase in the *E. coli* cloning host; this methylates the sequence GCNGC, which would otherwise be cleaved by the *C. acetobutylicum* Cac824I type II restriction enzyme\textsuperscript{30}. Likewise, transformation of *C. pasteurianum* ATCC 6013 (DSM 525) requires the methylation of CGCG sequences, which has been accomplished by the use of the M.FnuDI\textsuperscript{33} or M.BepI\textsuperscript{36} methylases. Other organisms have more complex restriction systems; in *C. saccharobutylicum* NCP 262, which has two type I restriction systems, expression of the methylation and specificity domains of these
systems on a plasmid in *E. coli* was sufficient to allow transformation by conjugation. Some developments have also been made in transforming non-type strains, which may have restriction patterns which differ from those of the type strains. For example, *C. pasteurianum* NRRL B-598 is part of the *C. beijerinckii* cluster but requires the use of a *dam/dcm* strain of *E. coli* for successful transformation, suggesting that the type IV system of this strain is particularly important; conversely, *C. beijerinckii* NCIMB 8052 can be transformed with much greater efficiency even with DNA from a *dam+/dcm* host. While an analysis of such developments in the entire genus *Clostridium* would be beyond the scope of this review, recent publications by Pyne et al. and Minton et al. provide a comprehensive review of the development of *Clostridium* strains for genetic engineering. However, it is certain that the range of genetically tractable *Clostridium* species and strains will expand with future research.

The resurgent interest in solventogenic *Clostridium* species suggests that synthetic biology tools are needed and there are clear shortages of particular biological parts with characterized activity available for the engineering of these organisms. We believe this warrants a concerted effort to address the weaknesses in the toolbox. Furthermore, with the establishment of transformation protocols throughout the genus *Clostridium*, it is also important to consider the potential of adapting existing parts and tools for use in other, less well-developed solventogens. Here, we assess the currently available parts used in synthetic biology projects, starting first with reporters available to measure gene expression, then the promoters, terminators working at the level of transcription and then post-transcriptional factors including mRNA stability, ribosome binding sites, riboswitches and codon usage. Finally we review the replicons currently being used to assemble synthetic gene clusters and briefly provide an overview of methods to introduce these genetic elements onto the chromosome, which has been reviewed elsewhere. Overall, we aim to summarize and systematize the existing biological parts used in the industrial solventogenic *Clostridium* species in order to highlight areas where more research is required and accelerate progress in creating a full range of synthetic biology tools for manipulating these industrially important organisms.

---

### Reporters of gene expression
Genetically-encoded reporter systems are the major in vivo gene expression measurement techniques available and are required to measure the activity of different parts being tested. The oxygen sensitivity of obligate anaerobes such as the clostridia limits the ease of use of many reporter systems, including some popular fluorescence-based and enzymatically-based reporters, and has led to the development of some novel reporters which are slowly being adopted.

**Fluorescent Reporter Proteins**

The use of fluorescent reporter proteins is now widespread in biology. Successful use requires the correct folding and maturation of the fluorescent protein to enable detection. The level of signal for the fluorescent protein must be sufficiently high to enable accurate detection as there is no signal amplification as seen in enzymatic reporters.

**GFP-family fluorescent proteins**

The highly engineered family of Green Fluorescent Proteins (GFPs) now includes variants with improved brightness and photostability and with a range of different colours (different excitation and emission maxima). The major limitation of the GFP-like proteins for anaerobes is the requirement of molecular oxygen for chromophore maturation and fluorescence. This excludes the possibility of real-time gene expression monitoring; however, in vivo observations are still possible in some oxygen-tolerant clostridia such as *Clostridium perfringens*, where yellow fluorescent protein (YFP) was synthesized and fluorescence was developed after exposure of live cells to atmospheric oxygen.

Using a similar principle, it is possible to obtain a snapshot of the protein levels in whole cells by exposing fixed anaerobically grown bacteria to atmospheric O₂ in a process termed ‘aerobic fluorescence recovery’, enabled by the discovery that GFP-family proteins can undergo maturation in fixed cells. Two studies have demonstrated the technique in *C. difficile* using paraformaldehyde and glutaraldehyde to fix cells expressing codon-optimized variants of CFP (cyan) and mCherry (red). This approach works as the fixatives cross-link primary amines (especially lysine residues) which are absent from the GFP-family chromophores. Fixed cells offer an advantage to
the imaging of dying live cells as they more accurately represent normal protein localization (gene expression changes are also likely to occur in dying or metabolically stressed cells)\textsuperscript{46}. So far, the technique has not been used to systematically quantify gene expression but rather to label proteins and track their intracellular localization\textsuperscript{46,47}.

**Flavin-binding Fluorescent proteins**

The flavin-binding fluorescent proteins (FbFPs) are a class of alternative fluorescent reporters capable of maturation in anaerobic conditions. The FbFPs are small proteins (\textasciitilde 11-15kDa) that have oxygen-independent fluorescent properties, using a flavin-mononucleotide (FMN) cofactor as the chromophore.

A pioneering study in 2007 reported the development of three anaerobic fluorescent reporters derived from the LOV (Light-, Oxygen- or Voltage-sensing) domains of bacterial proteins: BsFbFP derived from YtvA of *Bacillus subtilis*, PpFbFP from SB2 of *Pseudomonas putida* plus an *E. coli* codon optimized variant of BsFbFP called EcFbFP\textsuperscript{42}. This domain was engineered to emit fluorescence by mutagenesis of a reactive cysteine in the FMN-binding pocket\textsuperscript{42}.

These proteins are now commercially available from Evocatal GmbH under the trademark name ‘evoglow’. Currently listed variants marketed for use in *Clostridium* species include Bs1 (monomeric BsFbFP), Bs2 (dimeric BsFbFP) and Pp1 (PpFbFP).

The reporters’ functionality in *C. acetobutylicum* was demonstrated by Schulz in 2013; the highest fluorescence levels were observed for a *C. acetobutylicum* codon-optimized Pp1 (referred to in text as “Pp2”), followed by codon-optimized Bs2 (referred to as “Bs3”). Interestingly the same constructs gave inverse results in *E. coli*\textsuperscript{48}. While Evocatal GmbH offers C-Pp1 and C-Bs2 *Clostridium* codon-optimized reporters with publicly available nucleotide sequences, it is not clear whether the nucleotide sequences (reflective of the codon optimization approach) are the same as Pp2 and Bs3. A study in 2014 reported the placement of the evoglow Pp1 reporter downstream of the *cipP* promoter to monitor growth of *Clostridium cellulolyticum* on cellulose\textsuperscript{49}. The functionality of the evoglow Bs2 variant has also been demonstrated in *C. pasteurianum*\textsuperscript{50} and in *Clostridium ljungdahlii*\textsuperscript{51}.

Another FbFP example, developed from the *Arabidopsis thaliana* LOV2 domain of the blue-light receptor protein Phot2, is the improved LOV (iLOV) FbFP\textsuperscript{52}. The iLOV FbFP has been further modified for enhanced properties, generating variants such as
photostable iLOV 2.1 (phiLOV 2.1) and others. While phiLOV 2.1 FbFP has improved photostability, it is still substantially lower than that of GFP. A recent study demonstrated the utility of a \textit{C. difficile} codon optimized phiLOV 2.1 in three \textit{Clostridium} species (expression was driven from an engineered strong \textit{xyl} promoter). Under the test conditions, the three species exhibited varying levels of fold-increase of fluorescence over their background autofluorescence level: 3.2-fold increase in \textit{C. difficile} R20291 (not a \textit{Clostridium} sensu stricto species), 5.6-fold in the pathogenic toxin-producer \textit{Clostridium sordellii} ATCC 9714 and 4.5-fold in the solventogen \textit{C. acetobutylicum} ATCC 824. While the FbFP reporters will be detectable when expression is driven from very strong promoters, further improvements in brightness and photostability will make them more generally useful for multiple applications.

One of the drawbacks of the FbFPs is that all of the available variants are of the same colour; additionally, unfolding caused by translocation has been reported to result in loss of fluorescence due to the loss of the bound flavin cofactor. These limitations prevent the use of FbFPs as the sole reporters for multi-output circuits or secretion. Engineering of FbFPs for different emission spectra has proven to be somewhat complicated; while a simulation study of a theoretical iLOV mutant predicted a red shift increase, experimental results showed that the mutation resulted in an overall brightness decrease and a blue shift of emission. Research is underway to develop more fluorescent derivatives of the LOV domains and to explore the potential for diversifying their colour spectrum. The other limitation of FbFPs is their relative dimness, which can decrease the sensitivity of a reporter system; depending on the measurement instrument used, a promoter’s expression level might be below the detection limit. Photostability improvements (as in phiLOV) have resulted in a dimmer mutant; for applications where a longer exposure time is crucial, the dimmer, stable protein is preferable, but for an automated single measurement application such as flow cytometry, the brighter variant may be more suitable. Overall, all fluorescent protein reporters provide the ability to quantify cell-to-cell heterogeneity of gene expression (when used in flow cytometry or microscopy), which can be very useful, and while there are still some limitations with FbFPs, they have the potential to be a route to reliable \textit{in vivo} real-time expression monitoring in \textit{Clostridium}. 
Enzymatic Reporter Proteins

Enzymatic reporters catalyse a (preferably) unique reaction either in vivo or in vitro and the specific activity is calculated to estimate protein levels and thus gene expression. Enzymatic reporters require the addition of substrates and cofactors to the assay reaction mixture as well as the production of cell lysate. This often means that enzymatic assays involve more preparation steps than fluorescent proteins but can have adjustable sensitivity by varying substrate levels. Also, in vitro lysate assays are, by definition, bulk population measurements.

Chloramphenicol acetyltransferase

The first reporter which has been extensively used in the Clostridium genus, both in vivo and, more quantitatively, in vitro, is the chloramphenicol acetyltransferase (CAT) reporter (encoded by catP). The system was first developed for use in C. perfringens (a medically relevant non-solventogenic bacterium)\textsuperscript{58} and has since been used in C. acetobutylicum\textsuperscript{59}. Chloramphenicol acetyltransferase is an enzyme that catalyzes the covalent attachment of an acetyl group from acetyl-CoA to chloramphenicol\textsuperscript{60}, and is the basis of the chloramphenicol resistance marker found in many bacterial vectors. Transfer of the acetyl group exposes the thiol group of CoA, allowing the progress of the reaction to be observed by addition of 5,5'-dithiobis-(2-nitrobenzoic acid) (DTNB, also known as Ellman's reagent); this compound reacts with the free thiol, releasing 5-thio-2-nitrobenzoate, which can be detected by measurement of absorbance at 412 nm\textsuperscript{60, 61}. The assay relies on a continuous spectrophotometric rate determination to calculate specific activity. Potential drawbacks include endogenous activity in chloramphenicol resistant strains (which may be resolved by disruption of the resistance gene) as well as high levels of endogenous non-specific coenzyme A transferases\textsuperscript{62} (knockouts of which would be more laborious and would likely have growth and phenotypic effects) and the cost of the substrate acetyl-CoA.

Carbohydrate hydrolases: β-glucuronidase, β-galactosidase, amylase, endoglucanase

The CAT assay's drawbacks led to the adaptation of a classical E. coli reporter in Clostridium: the β-galactosidase enzyme, encoded by lacZ (the gene was derived from Thermoaerobacter thermosulfurigenes)\textsuperscript{62}. To quantify enzyme activity spectrophotometrically, ortho-nitrophenyl-β-galactoside (ONPG) is used in an in vitro assay and an increase in absorbance at 420 nm due to the release of ortho-nitrophenol is
measured. Similarly, the GUS reporter system, which utilizes β-glucuronidase (E. coli gusA), has been used in C. acetobutylicum in a fluorimetric assay with a cell lysate\textsuperscript{63}. The fluorimetric assay measures the release of 4-methylumbelliferone (4-MU) after cleavage of 4-MU-β-galactoside or 4-MU-β-glucuronide (by a β-galactosidase or β-glucuronidase, respectively), 4-MU emits light at 460 nm when excited by 365 nm light\textsuperscript{64}. Both the β-galactosidase and β-glucuronidase reporter systems benefit from the commercially available range of fluorometric, spectrophotometric and histochemical substrates and kits, making them an improvement over the CAT assay, although they are still not inexpensive.

The endogenous \textit{amyP} gene (encoding an amylase expressed during solventogenesis) has been used in C. acetobutylicum\textsuperscript{65} as a reporter to study the phenomenon of strain degeneration (loss of solventogenesis) which is often caused by loss of the pSOL1 megaplasmid on which \textit{amyP} is located. A codon-optimized amylase (\textit{AmyE}\textsubscript{opt}) has been used successfully as a secreted reporter in \textit{C. difficile} by addition of a zinc metalloprotease PPEP-1 signal sequence\textsuperscript{66}. It is noteworthy that the strain used in the above study was not capable of degrading starch under laboratory conditions; use of amylase as a reporter in amylolytic strains (such as \textit{C. acetobutylicum} ATCC 824) may require knockout of endogenous amylases to increase signal to background ratio.

In \textit{C. beijerinckii}, a secreted endoglucanase (\textit{eglA}) was cloned from \textit{C. saccharobutylicum} NCP 262 and used as a reporter for gene expression\textsuperscript{67}; assays used were agar plate and cell lysate assays which measured substrate (carboxymethylcellulose, CMC) clearance on plates and product (p-nitrophenol from cleavage of p-nitrophenyl cellobioside) accumulation in lysate, respectively.

\textit{Lipase and Alkaline phosphatase}

The lipase encoded by \textit{tliA}, from \textit{Pseudomonas fluorescens} SIK W1, has also been used as a reporter in \textit{C. beijerinckii} NCIMB 8052\textsuperscript{68}. Lipases are enzymes which hydrolyse the ester bonds found in long-chain acylglycerols, releasing fatty acids\textsuperscript{69}. Activity can thus be assessed by an enzymatic assay measuring the cleavage of p-nitrophenyl-esters\textsuperscript{69} such as p-nitrophenyl decanoate\textsuperscript{68}. However, this reporter was only useable in \textit{C. beijerinckii} as this species has a very low endogenous lipase activity. As such, it may not be suitable in other species, such as \textit{C. acetobutylicum}, which has been observed to show inducible lipase activity\textsuperscript{70}. 
A colorimetric alkaline phosphatase assay was developed and used in *C. difficile* in 2015 based on the *phoZ* gene product from *Enterococcus faecalis*\(^71\). Activity of this gene can be determined by a colorimetric assay with *p*-nitrophenyl phosphate as the substrate. Use in solventogenic clostridia could be limited by native phosphatase activity. In order to examine the suitability of *C. difficile* as a host for this reporter, BLAST analysis was used to screen for *phoZ* homologues, and activity towards 5-bromo-4-chloro-3-indolyl phosphate was tested\(^71\). While the four main industrial strains do not have a *phoZ* homologue, homologues can be found in other species such as *C. pasteurianum* (CLPA_RS02340, with 29% identity to *phoZ*).

**Luciferase**

The luciferase (*lucB*) reporter was also used in *C. acetobutylicum* successfully, allowing luminescence detection\(^72\). However, it also requires oxygen, ATP and luciferin. Notably, the cells used in the luciferase assay were neither lysed nor fixed, but were live cells that were exposed to atmospheric conditions, washed, and kept on ice. This treatment could conceivably introduce changes in gene expression levels prior to measurement. This assay has the lowest background signal level but the requirement for live cell exposure to oxygen may introduce variability. A codon-optimized luciferase (*sLuc\(^{opt}\)*) was also successfully secreted in *C. difficile* using the aforementioned zinc metalloprotease signal peptide\(^66\).

While there are many reporter choices available to clostridial researchers, we would argue that the multiplicity of reporters used has not helped ease the comparison of data obtained by different laboratories. Altogether, a single reporter has not been established as a community standard; given the drawbacks of each particular system, it is difficult to identify one standout reporter, although our hope would be that improved fluorescence reporters with increased brightness and photostability would be the most useful and enable single cell studies in live cells. Currently we would advocate a choice of more than one reporter, enzymatic and fluorescent, and their systematic and comparative use within the genus. Such an approach could provide insight into the sources of gene expression variability between strains and provide a starting point for future engineering. To achieve this seamlessly, it is advisable to explore translational fusions between reporters; examples from *E. coli* include a FRET pair YFP-FbFP fusion\(^73\) and
Gemini (lacZa-GFP) fusion\textsuperscript{74}. If successfully applied to the \textit{Clostridium} species, such bifunctional reporters have the potential to become a single standard reporter.

Additionally, α-peptide complementation strategies using the \textit{E. coli} lacZ gene in \textit{B. subtilis}\textsuperscript{75} and \textit{Mycobacterium smegmatis}\textsuperscript{76} have been reported but a similar approach in Clostridia has not been attempted. Interestingly, the \textit{T. thermosulfurigenes}-derived LacZ protein appears to lack the α-complementation region\textsuperscript{77} and is a dimer\textsuperscript{78} (rather than a tetramer like \textit{E. coli} LacZ).

\textbf{Transcription}

Transcription is the first stage of gene expression and the main stage for regulation of gene expression. Most existing biological parts in \textit{Clostridium} fall within the transcription category and are mostly promoters (summarized in Table 1). In contrast, the termination of transcription has been studied relatively little and we present some analysis on the potential to study and improve the parts available for reliable termination.

\textbf{Transcription initiation}

Promoters are DNA elements that are capable of driving transcription by RNA polymerase (RNAP) of downstream regions. Core promoter architecture in bacteria includes the -10 region or TATAAT box (Pribnow sequence), the -35 region and a spacer (with an optimal length of 17bp in \textit{E. coli}) between the two conserved regions\textsuperscript{79}. This motif is recognized by the housekeeping sigma factor that provides sequence specificity to RNAP (called RpoD in \textit{E. coli} and SigA in \textit{B. subtilis}, \textit{C. acetobutylicum} and other solventogens)\textsuperscript{80,81}. Both the consensus sequence and sigma factors bear very significant similarities between Firmicutes and \textit{Escherichia}.

In addition, some \textit{E. coli} promoters contain AT-rich UP elements (upstream of -35 region) that are responsible for recognition by the carboxy-terminal domain of the RNAP α-subunit\textsuperscript{82}, an additional \textit{E. coli} promoter feature is the 'extended -10' region\textsuperscript{83} (upstream of the -10 region and within the spacer). Interestingly, near-consensus promoters (including ones with UP elements) have been found to be significantly more common in Firmicutes (including Clostridia) than in other bacteria\textsuperscript{84}, a feature that was not explicable through higher AT-content alone. In Clostridia additional sequences with resemblances to UP elements (termed phased A-tracts) have been described\textsuperscript{85} and a conserved extended -10 region that differs from its \textit{E. coli} analogue has also been proposed\textsuperscript{86}. The majority of promoters used in the genetic engineering of clostridia have
been identified from the transcriptional units of important metabolic genes and most have been shown to be able to drive strong constitutive expression of a gene of interest. The most commonly used constitutive promoter in *C. acetobutylicum* (Table 1) is that of the native thiolase gene (*thlA*)\(^88,89\), which has been used for the expression of genes involved in production of solvents such as butanol\(^90\) and isopropanol\(^26\), the transcriptional regulator gene *tetR*\(^91\) (used a minimal promoter variant- miniPthl), and, in a modified form, for the expression of cellulosomal scaffoldins\(^92\) and glycoside hydrolases\(^25,93\). The *C. acetobutylicum* thiolase enzyme catalyzes the condensation of acetoacetyl-CoA from acetyl-CoA which are central metabolic intermediates\(^88\). The *C. acetobutylicum* *thlA* promoter (hereinafter referred to as *thlA*) is a sigA-dependent promoter (as evidenced by its near-consensus -35 and -10 regions\(^89\)); however, in its normally chromosomal context it is also the subject of regulation by the redox-sensing transcriptional repressor Rex\(^94\). Rex inactivation was found to increase native *thlA* activity in *C. acetobutylicum* about 12-fold\(^95\). The Rex-binding site has been omitted from the core promoter in commonly used synthetic constructs\(^96\); yet this binding site is conserved in the promoters of orthologous genes from other solventogens (according to the RegPrecise database\(^97\) and our promoter region alignments), suggesting that there are additional levels of control for this promoter that could be easily added by addition of the Rex sites if desirable. Indeed, some studies using similar promoters such as *thlA* and *ptb* have reported conflicting strength/activity findings, which may well be due to the cloning of regions of varying length (that likely include regulator binding sites such as Rex) and choosing to include the native RBSs or not (such as the ones reported between promoters in the Schulz 2013\(^48\) and Girbal 2003\(^63\) studies).

Thiolase promoters have also been used for expression in other clostridia such as *C. beijerinckii*\(^98\) and *C. pasteurianum*\(^99\). However, gene expression data (RNAseq from *C. beijerinckii*) indicates that there are other genes that have higher expression levels than the ones currently used as promoter sources; many of those genes encoded are hypothetical proteins\(^100\). Recently, Yang *et al.,* 2017 constructed a sequence logo of 18 promoters previously identified in *C. acetobutylicum* which revealed a strongly conserved -10 region. Using degenerate oligos to mutagenize the core *thlA* promoter elements, the authors were able to generate a promoter library of variable strengths,
including mutations that increased the strength of the promoter, suggesting that \textit{thlA}
can be improved further using synthetic biology approaches.

The use of a constitutive promoter may not always be desirable; it may be preferable to
use an inducible promoter, allowing controlled expression of a gene of interest. So far,
the only naturally inducible promoters exemplified in a solventogenic clostridium are
the \textit{Staphylococcus xylosus xylA} promoter, which is repressed by XylR in the absence of
\textit{D-xylose} \textsuperscript{63}, and the \textit{C. perfringens bgaL} promoter, which is repressed by BglR in the
absence of lactose\textsuperscript{101}. However, constitutive promoters can be made inducible by
addition of operator sites (for transcriptional repressors) or by the addition of binding
sites for activators, and several such promoters have been developed for use in
solventogenic clostridia. LacI-repressible versions of thiolase and ferredoxin promoters
have allowed the construction of clostridial expression constructs for genes where
expression in \textit{E. coli} would be toxic\textsuperscript{24,25,93}. In \textit{C. acetobutylicum}, the \textit{fac}
(Clostridium \textit{pasteurianum} ferredoxin promoter with lac operator) is able to function as
an IPTG-inducible system in the presence of LacI, allowing approximately 10-fold
induction\textsuperscript{59}. A TetR-repressed, anhydrotetracycline-inducible promoter has also been
used in \textit{C. acetobutylicum}, generated by the fusion of the chloramphenicol
acetyltransferase promoter \textit{pcm} with the tetracycline operator \textit{tetO}\textsuperscript{91}. This promoter
could achieve up to 313-fold induction, although high levels of anhydrotetracycline were
inhibitory to growth. However, operators are known to influence the basal activity of the
promoter\textsuperscript{102}. In fact, a recent study in \textit{E. coli} found the core RpoD promoter too sensitive
to sequence context and operator insertions to be a suitable target for forward
engineering efforts and turned to extracellular sigma factors (ECFs) with T7 RNAP
whose promoter core sequences they found to be more insensitive to operator
addition\textsuperscript{103}. An alternative strategy - CRISPR-mediated repression of transcription - has
been demonstrated in several solventogenic species, namely \textit{C. acetobutylicum}\textsuperscript{104}, \textit{C.
\textit{beijerinckii}}\textsuperscript{98} and \textit{C. pasteurianum}\textsuperscript{50},

Orthogonal expression systems, i.e. promoters that require other non-native elements
for activity (commonly an alternative sigma factor or phage polymerase), first found
widespread use in the cloning of genes with toxic products. This approach allows the
total repression of genes until they are introduced into the organism of interest. A
commonly used example in *E. coli* is the T7 promoter, which requires the phage T7 polymerase for activity\textsuperscript{105}. An example from Clostridia is expression driven from the *C. difficile* *tcdB* promoter that is dependent on the native sigma factor TcdR. The *tcdB* promoter is highly active in *C. acetobutylicum*, but only when *tcdR* has been integrated into the genome\textsuperscript{59}. This enabled the high-level expression of a *mariner* transposon in *C. acetobutylicum* without negative effects on the *E. coli* cloning host. There is a strong case for designing these orthologous-type systems as simply using a clostridial housekeeping promoter (such as the *thlA* promoter) with a clostridial ribosome binding site would not provide orthogonality in commonly used cloning hosts such as *E. coli* or *B. subtilis* due to the degree of conservation between the transcription and translation initiation systems amongst these organisms; on the contrary, as a rule strong clostridial promoters and RBSs (see translation initiation section) often retain their strength in *E. coli* whereas the opposite is observed more rarely, this is likely due to the on average higher similarity to the Bacteria-wide consensus translation and transcription initiation signals in *Clostridium*\textsuperscript{86,106}. Interestingly, AT-rich DNA can be toxic to *E. coli* (clostridial genomes are very AT-rich) due to transcriptional activity from spurious intragenic promoters and RNAP titration\textsuperscript{107}. This is exacerbated by the fact that the *E. coli* extended -10 region, which is sufficient alone for transcription initiation in *E. coli*, could also be present by chance in Clostridium-derived sequences, increasing the likelihood of spurious and unpredictable transcription during cloning in *E. coli*. 
<table>
<thead>
<tr>
<th>Promoter</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constitutive</strong></td>
<td></td>
</tr>
<tr>
<td>thlA (C. acetobutylicum)</td>
<td>Widely used for constitutive gene expression in <em>C. acetobutylicum</em> for example expression on <em>C. beijerinckii</em> ald, <em>C. ljungdahlii</em> bdh(^9), <em>C. acetobutylicum</em> adc, ctfA, ctfB, (^{26}) &amp; tetR (miniPthl variant)(^9). Activity analyzed using GusA(^63) and several FbFP(^48)reporters.</td>
</tr>
<tr>
<td>thl (C. perfringens)</td>
<td>Used for expression of srtA genes from <em>C. acetobutylicum</em>, <em>L. monocytogenes</em>, and <em>B. cereus</em>(^25).</td>
</tr>
<tr>
<td>fdx (C. sporogenes)</td>
<td>Activity analyzed using CatP reporter(^59), Used in the ClosTron system for expression of the Ll.LtrB intron(^108). Used for expression of spoA integrated into the chromosome for complementation of a spoA mutant(^109).</td>
</tr>
<tr>
<td><strong>Controlled</strong></td>
<td></td>
</tr>
<tr>
<td>thl (thl with double lac operator)</td>
<td>lac-repressed version of thlA. Used to express a miniscaffoldin cipc(^19) &amp; weakened version for expression of mannanase man5K(^93).</td>
</tr>
<tr>
<td>thlOid (thl with single lac operator)</td>
<td>A lac-repressible version, used to drive chromosomal expression of the <em>C. cellulolyticum</em> glycoside hydrolase xyn10A(^25).</td>
</tr>
<tr>
<td>fac (single lac operator, derived from C. pasteurianum)</td>
<td>Activity analyzed using CatP reporter(^59,110) and Pp2 FbFP(^48), Formerly used in the ClosTron system for expression of the Ll.LtrB intron(^110), Used for expression of codA in the <em>C. acetobutylicum</em> knockout vector pMTL-SC7515(^109).</td>
</tr>
<tr>
<td>fdxOid (fdx with single lac operator)</td>
<td>Used for chromosomal expression of <em>C. cellulolyticum</em> glycoside hydrolase cel9G(^25). Has RBS from <em>C. acetobutylicum</em> thlA promoter.</td>
</tr>
<tr>
<td>facOid (fac with single lac operator)</td>
<td>Used for chromosomal expression of <em>C. cellulolyticum</em> glycoside hydrolase cel48F(^25). Has RBS from <em>C. acetobutylicum</em> thlA promoter.</td>
</tr>
<tr>
<td>tcdB (C. difficile)</td>
<td>Developed for an orthogonal expression system, requires the exogenous sigma factor TcdR for function(^59).</td>
</tr>
<tr>
<td>xylA (S. xylosus)</td>
<td>Repressed by XylR and inducible by D-xylose; assessed with β-glucuronidase reporter giving 17-fold induction with D-xylose(^63).</td>
</tr>
<tr>
<td>pcm (tetO1-containing variants)</td>
<td>Repressed by TetR; inducible in the presence of anhydrotetracycline(^91).</td>
</tr>
<tr>
<td>bgaL (C. perfringens)</td>
<td>Repressed by BgaR and inducible by lactose(^101).</td>
</tr>
<tr>
<td>xyl/tetO (B. subtilis)</td>
<td>Repressed by TetR(^111); activity in <em>C. acetobutylicum</em> shown by phiLOV 2.1 Opt FbFP expression(^55)</td>
</tr>
<tr>
<td><strong>Fermentation phase-specific</strong></td>
<td></td>
</tr>
<tr>
<td>adc (C. acetobutylicum)</td>
<td>Activity assessed using β-glucuronidase(^63), β-galactosidase(^72) and Pp2 FbFP(^48). Primarily active after onset of solventogenesis.</td>
</tr>
<tr>
<td>ptb (C. acetobutylicum)</td>
<td>Activity assessed using β-glucuronidase(^63), luciferase(^72) and Pp2 FbFP(^48). Active during acidogenesis (not solventogenesis)</td>
</tr>
<tr>
<td>ptb (C. beijerinckii)</td>
<td>Activity assessed through expression of lac(^159).</td>
</tr>
<tr>
<td>sol (C. acetobutylicum)</td>
<td>Activity assessed using β-galactosidase &amp; luciferase reporters(^72). Weak expression, primarily active during late exponential phase.</td>
</tr>
<tr>
<td>bdhA (C. acetobutylicum)</td>
<td>Analyzed using β-galactosidase reporter(^72). Comparatively weak expression, primarily active in early exponential phase.</td>
</tr>
<tr>
<td>bdhB (C. acetobutylicum)</td>
<td>Analyzed using β-galactosidase reporter(^72). Primarily active until onset of solventogenesis.</td>
</tr>
<tr>
<td>hydA (C. acetobutylicum)</td>
<td>Activity analyzed using β-glucuronidase(^63) and Pp2 FbFP(^48). High activity during acidogenesis, decreases to low after phase shift</td>
</tr>
<tr>
<td>adhe2 (C. acetobutylicum)</td>
<td>Activity analyzed using Pp2 FbFP(^48). Strong expression during solventogenesis(^48,112)</td>
</tr>
</tbody>
</table>
Transcription termination

Bacteria have two distinct mechanisms that function in transcription termination. Both types of terminators are usually located in the 3’ end of transcriptional units. Rho-dependent terminators rely on the Rho protein that recognizes a target sequence and causes RNA polymerase to fall off of the template DNA. The specific DNA recognition sites, rut sites, have been used before in genetic circuits, but not extensively, perhaps due to the relatively poor mechanistic understanding of the process. The second mechanism, which also has been reported to be more widespread, is referred to as Rho-independent or intrinsic termination and relies on the folding of a short GC rich hairpin followed by a poly-U transcribed sequence. The hairpin folding and subsequent transcription of the U-tract causes the polymerase to pause and leads to transcript and RNAP dissociation. This is also facilitated by the fact that ribo-uracil-deoxyribo-adenine is the weakest nucleotide base pair. Intrinsic terminators are often found downstream of operons; however, they are also involved in transcription attenuation when present within coding regions or downstream of promoters.

While a number of terminators have been used in the construction of clostridial expression constructs, often derived from clostridial genes such as adc, fdx, CD0164, and thlA, there have been few published analyses of terminator strengths in clostridia. The C. pasteurianum fdx terminator has been shown to be highly effective at preventing read-through inhibition of the replicon from the fac promoter in a clostridial vector, and screening of a selection of terminators in C. acetobutylicum showed that the E. coli rrnB terminator T1 loop was able to function as an efficient terminator, reducing expression of a downstream gene. However, these analyses only took into consideration the effect of the terminator on a downstream target under the influence of a single promoter. As previously mentioned, the introduction of a terminator may have an effect on expression of an upstream gene by influencing mRNA stability, which is not a desirable feature of a standard part. However, this effect can be quantified in an appropriately designed assay. Furthermore, the efficiency of termination may increase or decrease depending on promoter activity; this has been recently exploited for the development of a genetic band-pass filter in E. coli. Terminator strength has also been shown to be influenced by the hairpin’s proximity to a stop codon or when present within a coding region. A recent study demonstrated...
the gradual increase in termination efficiency as distance between the stop codon and hairpin increased. The ribosome was found to repress transcriptional termination when the stop codon and hairpin were in close proximity – termination repression was relieved when insulator sequences of approximately 30 bp were used. Expanding the range of characterized clostridial terminators is clearly necessary for the development of a clostridial synthetic biology toolbox.

One approach that can be undertaken is the use of algorithms to extract putative terminator sequences from genomes, also known as ‘part mining’. Several such bioinformatics tools exist; they rely on seed sequences, secondary RNA structure features or both. These bioinformatics tools were developed to aid genome annotations by identifying the 3’ ends of operons and may not then detect all the features required by a functional terminator. In order to evaluate the relative difference in sequences retrieved by different programs, we compared the output of three searches performed on the *C. acetobutylicum* ATCC 824 genome (Figure 3A). The three search algorithms that were chosen are as follows: TransTermHP (TTHP), RNIE and WebGeSTer (WG). TTHP is a widely used tool, while RNIE’s authors state that their tool eliminates false positives in comparison to TTHP. Finally, WebGeSTer was selected because the authors classify the results into different types of intrinsic terminators based on overall secondary structure. They discovered that canonical U-tract containing intrinsic terminators (termed ‘L-shaped’) form the majority of structures found within Firmicutes, whereas they are a minority in the *E. coli* model where they were first identified.
Figure 3. Application of terminator prediction software to C. acetobutylicum. A. Outputs from predicted terminator sets derived from the C. acetobutylicum genome by application of TransTermHP (TTHP), RNIE and WebGeSTer were compared. B. An alignment of predictions of a putative bi-directional terminator (marked by T sign) by all three algorithms (note all three produce separate predictions for plus (+) and minus (-)). C. RNA structures predicted by RNAfold for the TransTermHP sequence, drawn with VARNA; it is worth noting that the base-pairing between the A-tract and U-tract may not form until after termination. Terminator lengths: WG: maximum length (max) - 48bp, minimum (min) - 28bp, average (avg) - 35.8bp; RNIE: max - 45bp, min - 34bp, avg - 41.7bp, TTHP: max - 89bp, min - 41bp, avg - 57.2bp.
In our analysis, each algorithm was used to produce sets of putative terminator predictions, which were examined to identify overlapping predictions and are presented in Fig. 3A. Using the default settings of the different algorithms, the total number of predictions per program varied considerably, with TTHP and WG each predicting over 1000 terminators while RNIE predicted fewer than 300 (Fig. 3A). Despite this much smaller set, only about half of the predictions were shared with either TTHP and WG sets or both, producing a small set of 104 terminators shared by all three methods, an example of which is illustrated in Fig. 3B, an intergenic region separating two converging coding sequences, CA_C0323 (a sensory transduction histidine kinase) and CA_C0324 (a tetratricopeptide-repeats-containing protein). This also illustrates a class of terminators with particular utility in synthetic biology, namely bi-directional terminators, which are capable of terminating transcription from both the plus and minus strands. Our analysis finds that 37 bi-directional terminators are supported by all three algorithms, supporting the concepts that there are plentiful targets in clostridial genomes for ‘part mining’ and future experimental characterisation of these likely strong terminators. An interesting feature of the bi-directional terminators is their hypersymmetry\(^{131}\) (can be seen in Fig. 3B). This leads to the possibility to form an extended hairpin between the so-called A-tract (can act as U-tract in the reverse transcriptional orientation) and the U-tract (as shown in Fig. 3C). However, as pausing and termination occur at the U-tract\(^{117}\), the formation of an extended hairpin may not be relevant and does not appear to contribute to strength in all terminators\(^{132}\). Also, note that in this example the terminator on the minus strand (also referred to as Crick strand) is predicted to have a more stable hairpin (rG and rU form a basepair while rA and rC do not), whereas terminators from both strands are capable of forming an extended hairpin if fully transcribed as a result of base pairing between the A- and U-tracts. Combined with having a longer U-tract (Fig. 3C), the Crick strand terminator is possibly the stronger one, interestingly WebGeSTer predicts a smaller terminator downstream of the one depicted in Fig. 3A (not shown). A previous analysis of the *C. acetobutylicum* genome for terminators using TransTerm\(^{127}\), a predecessor of TTHP, revealed the presence of 66 putative bi-directional terminators\(^{133}\), supporting the rationale of combining outputs from multiple algorithms to create a small part list for experimental characterisation.
Translation initiation

Bacterial ribosome-binding sites (RBS) are short sequences located in the 5’ untranslated region of messenger RNA (mRNA) transcripts, consisting of a Shine-Dalgarno sequence (SD), polynucleotide spacer, and a translation initiation codon\(^{134}\). Commonly, native promoter-RBS combinations have been used (such as in the pMTL80000 vectors). An alternative strategy is to use a native RBS (such as that of the \(C.\) acetobutylicum \(thlA\) gene) fused to a new promoter, as in a study from 2016\(^{25}\) that generated several new hybrid promoters. Others have experimented with the length of the spacer\(^{135,136}\). Yet, in contrast to model organisms (\(E.\) coli and even \(B.\) subtilis), there are few published comparisons of modified RBSs for use in solventogenic clostridia. Hence the generation and screening of synthetic RBS libraries could be a promising route for optimising the expression of synthetic gene constructs in clostridia.

SD sequences provide sequence complementarity for the 3’ terminus of the 16S rRNA (known as the anti-Shine-Dalgarno or aSD\(^{137}\)) which acts as a guide for the ribosome complex enabling mRNA recognition by the translation machinery and translation initiation. While most of the knowledge on bacterial translation initiation comes from \(E.\) coli work, early studies indicated that the Firmicute \(B.\) subtilis requires a longer complementary region between the SD and the 16S rRNA to achieve comparable expression levels\(^{138,139}\). An early study on translation initiation revealed that Firmicutes have, on average, a higher complementarity of the predicted SD region to the 16S rRNA 3’ terminus than \(E.\) coli does\(^{106}\). Recently, a systematic analysis of SD-aSD pairings in \(B.\) subtilis and in \(E.\) coli confirmed this trend\(^{140}\). Replicating these studies in solventogens would provide useful information.

The spacer is the mRNA region between the SD and the start codon. The aligned spacer (which is the distance between the start codon and the 5’ end of the aSD, base-paired to a SD\(^{137,141,142}\) as revealed by sequence alignment) is of particular importance for translation initiation\(^{137}\). Defining the length of the aligned spacer precisely is difficult, as interpretations of the extent of the aSD region (starting from the 3’ 16S rRNA terminus) vary. The 3’ end of the 16S rRNA gene (aSDs) of \(B.\) subtilis is identical to that of \(C.\) perfringens\(^{143}\) (as well as identical to those of solventogenic clostridia, based on our sequence analysis of published genomes) but to our knowledge there is no reported
experimental validation of the clostridial mature 16S rRNA 3’ ends in vivo. Spacers in different species may have different optimal lengths; for example, spacers of Pyrococcus abyssi are, on average, roughly 3 nucleotides longer than those of E. coli\textsuperscript{144}, whereas the spacers in Bifidobacterium longum\textsuperscript{145} would be considered shorter. According to our definition of the putative clostridial aSD (5’ GAUCACCUCCUUUCU 3’), in C. acetobutylicum, the native RBS of the thlA promoter has an ‘aligned spacer’ of 4 bases.

The effect of altering the length of the thlA spacer was recently investigated in two studies conducted in C. acetobutylicum\textsuperscript{135,136}. Interestingly Yang et al., 2016 showed that a lengthened thlA RBS spacer with an extra 6 nucleotides (to a total of 14 bases-aligned spacer of 10bp), encoding a SalI restriction site, did not significantly alter reporter expression in comparison to the WT thlA RBS. Shortening the spacer below the WT length resulted in a decrease in expression, while further increases over an aligned spacer length of 10 bases (by the addition of a XbaI site) removed almost all of the expression\textsuperscript{135}.

Yang et al. (2016) demonstrated the potential benefits of utilising modified spacers in C. acetobutylicum\textsuperscript{135}. Overexpression of the biotin synthesis genes bioY, bioD, bioA, and bioB was observed to provide an improvement in growth phenotype and solvent production. The thlA promoter was used to drive expression of bioY, bioD and bioA; replacement of the thlA RBS spacer with a shortened, less 'effective' variant resulted in a further improvement in growth characteristics. These results demonstrate not only the ability of modified RBS to optimize expression of synthetic pathways, but also the necessity of considering the effects of any alterations to the spacer, e.g. the introduction of restriction sites, when generating synthetic constructs. In 2017, Yang et al. generated a library of spacers (a BamHI site preceded the start codon in all cases) by starting with an aligned spacer of 2 bases and progressively increasing its length by 2 bases. After testing 11 variants the authors found that an aligned spacer of 4 bases gave the strongest levels of expression followed by 8, 6 and 10 bases in that order. It is conceivable that not only the length but also the sequence of the spacer might influence translation initiation, for example the introduction of a secondary SD within the primary SD’s spacer region.

Computational tools to design RBSs exist, such as the RBS calculator, which use biophysical models of RBS recognition and translation initiation, including RNA base-pairing between the aSD and SD, spacer length and messenger secondary
structure\textsuperscript{146,147}. There are reports that \textit{de novo} design produces more accurate results than translation initiation rate prediction of natural sequences for Gram positives\textsuperscript{148}.

**Codon usage during translation**

While codon usage is not formally a 'biological part', it is an important feature of coding sequences, the differential frequency of synonymous codons amongst genomes, referred to as codon usage bias\textsuperscript{149,150}, has been shown to strongly influence heterologous protein expression levels\textsuperscript{151,152}. High genomic AT-content is characteristic of the Firmicutes and is reflected in the nucleotide composition of coding sequences. The \textit{Clostridium} species' codon usage differs from that of other Firmicutes as well as the Proteobacterium \textit{E. coli}\textsuperscript{53} and there are also bioinformatically observable variations within the \textit{Clostridium} genus itself\textsuperscript{154} but the significance of the latter in influencing gene expression has not been experimentally verified to our knowledge. Genetically encoded reporters have been used heterologously within the Firmicutes phylum without codon-optimization, for example \textit{Staphyloccoci} have been sources of reporters and antibiotic resistance genes for Clostridia. On several occasions researchers have successfully used native reporter genes from \textit{E. coli} (\textit{gusA})\textsuperscript{63}, \textit{T. thermosulfurigenes} (\textit{lacZ})\textsuperscript{72} and the firefly \textit{Photinus pyralis} (\textit{lucB})\textsuperscript{72} (we describe these reporters in more detail in the Enzymatic Reporter Protein section). Codon-optimized genes for \textit{C. difficile} have been used in \textit{C. acetobutylicum}\textsuperscript{55}, as well as bespoke \textit{C. acetobutylicum} codon-optimized ones\textsuperscript{24,48}. Researchers have also codon-optimized several GFP-like proteins for use in Firmicutes\textsuperscript{155} (with \textit{B. subtilis} in mind) but these have not been used in solventogens to our knowledge.

Codon optimization is not a trivial problem and codon optimization strategies vary considerably. The codon-adaptation index (CAI)\textsuperscript{150} has been the historical measure of codon usage bias in an organism while there are others such as the codon bias index and the effective number of codons\textsuperscript{156}. Interestingly, simply improving a heterologous gene’s CAI (making it more like a native gene or a highly expressed native gene) has not been found to correlate with expression levels\textsuperscript{157,158}. Efforts have been made to improve the indices describing codon usage bias and translational efficiency (which codon usage bias is thought to reflect) by studying endogenous gene expression\textsuperscript{159,160}. In addition, condition-specific usage tables have also been reported\textsuperscript{161}. In an alternative strategy, Welch and colleagues developed genetic algorithms to select partial least squares
regression models which revealed that codons predominantly read by tRNAs that are most highly charged during amino acid starvation were good predictors of expression levels. Based on these results the researchers developed proprietary codon optimization algorithms to maximize protein expression\textsuperscript{158,162} which allowed them to predict expression levels in \textit{E. coli}.

While it is clear that there is room for improvement in the heterologous protein expression strategies used in \textit{Clostridium} solventogens, codon optimization strategies themselves are still being developed and the underlying principles are not yet fully understood; progress and existing approaches in the field have been reviewed elsewhere\textsuperscript{163}. A good starting point is for researchers to report the details of the codon optimization strategy undertaken when publishing work containing codon-optimized genes. This way data from heterologous protein expression in solventogens can be compared more reliably.

**Post-transcriptional control of gene expression**

Tuning gene expression levels in \textit{Clostridium} species has been achieved using control at the RNA level – by either influencing translation or changing RNA degradation rates.

**Controlling mRNA stability**

Another potential avenue for optimisation of expression levels is the adjustment of mRNA stability. Altering the stability of an mRNA transcript influences the number of transcripts in the cell, thereby affecting the overall rate of translation. In bacteria, a number of factors are associated with mRNA stability, such as secondary structures, RNase recognition sites and polyadenylation, amongst others\textsuperscript{122}. The presence of secondary structures at the 5′ end of the mRNA has been observed to provide an increase in mRNA stability\textsuperscript{122}. In \textit{E. coli}, these structures prevent the binding of RNase E, an endonuclease which binds single-stranded RNA at the 5′ end and then scans for cleavage sites. This property has been exploited in \textit{E. coli} by the generation of libraries of synthetic hairpins for introduction into the 5′ untranslated region (UTR)\textsuperscript{164}. While \textit{C. acetobutylicum} has an RNase E homolog, RNase E/G, it is not certain if this behaves in the same way as \textit{E. coli} RNase E, due to having a different domain organisation\textsuperscript{165}. Nevertheless, \textit{C. acetobutylicum} also has a homolog of RNase Y\textsuperscript{166}, which fulfils the role of RNase E in \textit{B. subtilis}\textsuperscript{167}. Correspondingly, the utility of 5′ hairpins for protection of mRNA has been demonstrated in clostridia; the introduction of 5′ stem-loop sequences was confirmed to increase mRNA stability, reporter expression, and expression of the
genes adhE1 and adhE2 in both C. acetobutylicum and C. beijerinckii\textsuperscript{168}. This effect was much more pronounced during solventogenesis than during acidogenesis. Similarly, the introduction of a terminator hairpin in the 3' UTR can result in improved mRNA stability via inhibition of nuclease activity\textsuperscript{122}. Although not fully explored, this principle has been demonstrated in solventogenic clostridia; the expression of a cat reporter gene in C. acetobutylicum was observed to increase by approximately 36% when the downstream adc terminator was replaced by a synthetic terminator, BBe_B1010, from the iGEM registry\textsuperscript{168}. Additionally, a terminator with activity in the reverse orientation prevents the formation of antisense transcripts which are known to reduce protein expression levels\textsuperscript{169}.

A completely opposite approach is to reduce mRNA stability by introducing RNase sites into the 3' UTR. This may be useful in the case of proteins which form inclusion bodies when overexpressed. In E. coli, sequences derived from the cat gene, which contains 28 RNase E sites, have been shown to reduce mRNA stability\textsuperscript{170}. When combined with two poorly soluble heterologous enzymes, the cat-derived 3' UTRs were shown to result in an increase in soluble protein, with concomitant increases in enzyme activity. The authors noted that this improvement could not be observed simply by using a weaker promoter; it was proposed that by limiting the stability of the mRNA, the number of proteins that could be produced from a single transcript was decreased, thus limiting the local concentrations of protein during translation.

### Antisense RNAs and Riboswitches

Bacterial antisense RNAs (asRNAs) are transcripts with complementarity to another RNA. Several studies have used asRNAs were used to reduce native gene expression in C. acetobutylicum\textsuperscript{169,171} and in C. pasteurianum\textsuperscript{172}, demonstrating that multiple routes to regulate gene expression are available in the solventogenic clostridia.

Riboregulators are another class of naturally-occurring and generally trans-activating asRNA elements that respond to a signal nucleic acid by Watson-Crick base pairing\textsuperscript{173}. They have defined sensor and effector domains and have been rationally designed to repress\textsuperscript{174} and activate gene expression\textsuperscript{175} in E. coli. Toehold switches (a synthetic cis-regulatory subgroup of riboregulators) that activate gene expression in the presence of
cognate RNAs rely on sequestering the RBS and start codon\textsuperscript{176}. Like riboswitches, riboregulators are known to be present in \textit{Clostridium} genomes. Riboswitches are RNA sequences that are able to bind to a soluble ligand, influencing the properties of the RNA. In nature, riboswitches typically contain a binding domain, or ‘aptamer’, and an ‘expression platform’ which mediates the effect\textsuperscript{177}. Binding leads to a change in conformation of the RNA, leading to formation of a secondary structure which can act as a terminator or an anti-terminator. While riboswitches are typically found in the 5’ UTRs of mRNAs, some have been recently determined to control the expression of antisense RNAs or protein-sequestering small RNAs, while yet others have been shown to control access to recognition sequences such as RNase sites\textsuperscript{178}. The range of applications has been further developed by the creation of synthetic riboswitches. In bacteria, riboswitches have been developed that can influence translation initiation by inhibiting access to the RBS; ligand binding leads to a conformational change or even to self-cleavage, revealing the RBS and allowing translation\textsuperscript{179}. Riboswitches are found in all taxa and a number have been characterized in the solventogenic clostridia\textsuperscript{180–182}. However, riboswitches have not yet been used in the engineering of these organisms. As discussed earlier, only a limited range of inducible systems is available for the engineering of clostridia; the use of synthetic riboswitches could be a promising alternative for the creation of controlled promoters.
Replication

Plasmid origins of replication
Four replicons are in routine usage in solventogens: pCB102 (from *C. butyricum*), pBP1 (from *C. botulinum*), pCD6 (from *C. difficile*) and pIM13 (from *B. subtilis*). The pIM13 replication origin is thought to replicate via rolling-circle replication while there is evidence that pCD6 replicates in similar fashion to pIP404 (which is either theta or strand-displacement). The replication mechanisms of pCB102 and pBP1 are unknown. In *C. acetobutylicum*, two other replicons that have been used are the pAMβ1 replicon and the pUB110 replicon; the pUB110 replicon was found to be somewhat more stable than pIM13, whereas the pAMβ1 replicon was highly unstable. The *C. beijerinckii* filamentous phage CAK1’s origin of replication has been used in *C. beijerinckii* strains. Additionally, the development of a replicon specific for *C. saccharoperbutylacetonicum* N1-4 was reported in 2007, this replicon is identical to the origin of the endogenous plasmid from *C. saccharoperbutylacetonicum* N1-504. A thermosensitive origin pWV01ts derived from *Lactococcus lactis cremoris* has been shown to work in both *C. ljungdahlii* and *C. acetobutylicum*. Segregation and transformation frequencies are available; however, more work is needed to determine copy number and compatibility groups.

Chromosomal integration
The integration of DNA into the genome, while not a ‘part’ in itself, is an important consideration for synthetic biology projects. Genomic integration has several advantages over plasmid-based expression strategies, including increased stability, removal of the requirement for antibiotic selection, and standardisation of copy number. However, there are other factors that must be considered when using chromosomal integration. One implication of the integration position is the copy number effect – genes closer to the origin have a higher copy number than ones near the terminus in exponentially dividing cells due to the mechanism of DNA replication. There is contradictory evidence as to the effects of chromosome location and levels of expression. The cause of some of the apparent transcriptional insulation of chromosomal genes (and thus promoter independence) was found to be due to the gene silencing activity of the HU-protein in *E. coli*, a protein that is also present in...
Firmicutes. Genomic engineering in *E. coli* has been carried out extensively; lambda-
Red recombineering\(^1\)\(^9\)\(^7\) is well-established, and the utility of the newly developed
CRISPR technique has been demonstrated in this species\(^1\)\(^9\)\(^8\),\(^9\)\(^9\). However, in the
solventogenic clostridia, until recently published examples of genomic integration had
been still somewhat limited. An early enabling technology was ClosTron which adapted
the Ll.LtrB intron for use in *Clostridium*\(^1\)\(^1\)\(^0\). Another method for genomic integration is
Allele-Coupled Exchange (ACE), as demonstrated in *C. acetobutylicum*\(^2\)\(^4\),\(^2\)\(^5\),\(^1\)\(^9\)\(^0\). This is a
homologous recombination-based method, where homology arms with different lengths
are used to control the sequence of recombination events, and the second recombination
leads to the generation of a selectable phenotype; currently, this involves either the
truncation or repair of the *pyrE* gene or the activation of a promoterless antibiotic
resistance gene by integration downstream of a strong chromosomal promoter such as
*thlA*\(^1\)\(^9\)\(^0\). Thus, one potential drawback to this method is that it only allows integration
into a limited selection of loci. This drawback can be mitigated by the ability to carry out
multiple rounds of iterative ACE, thereby making further genomic integrations into the
same locus.

Nevertheless, many new developments have been made regarding the genetic
manipulation of solventogenic clostridia. A variety of different allelic exchange-based
strategies have been exemplified in *C. acetobutylicum*\(^1\)\(^0\)\(^1\),\(^1\)\(^0\)\(^9\),\(^2\)\(^0\)\(^0\),\(^2\)\(^0\)\(^1\) and *C. beijerinckii*\(^2\)\(^0\)\(^2\); while most of these studies have focused on the generation of in-frame deletions and
subsequent complementation, Al-Hinai et al.\(^1\)\(^0\)\(^1\) demonstrate the integration of a
heterologous gene through gene replacement. Furthermore, the generation of point
mutations through recombineering has been demonstrated in *C. acetobutylicum*\(^2\)\(^0\)\(^3\),
suggesting that the integration of DNA through this method may be feasible. Finally,
mutant selection via CRISPR has been established in almost all of the main solventogenic
species, with published examples in *C. beijerinckii*\(^i\)\(^2\)\(^1\),\(^2\)\(^0\)\(^4\), *C. acetobutylicum*\(^1\)\(^0\)\(^4\), *C.
*pasteurianum*\(^2\)\(^0\)\(^5\), and *C. saccharoperbutylacetonicum*\(^2\)\(^0\)\(^6\). Future developments are certain
to expand the range of genomic modifications that can be made in these organisms.

**Insulators**
A biological part’s adjacent sequences can have a profound effect on its behaviour
compared to the sequence context in which it was characterized\(^2\)\(^0\)\(^7\). This poses a
fundamental challenge to the synthetic biology principle of part creation and characterization\textsuperscript{208}. To counteract this issue in reproducibility genetic engineers have started utilizing a new class of parts called insulators\textsuperscript{209}. Several strategies can be undertaken to insulate a part from its genetic context, and these can be split into two main categories: DNA-level insulators (such as simply using flanking buffer zones of sequence without a biological function or secondary structure) and RNA-level insulators. The latter includes post-transcriptional modification of RNA as well as the commonly used flanking double terminators\textsuperscript{123} that prevent read-through transcription into synthetic gene constructs. Post-transcriptional insulators consist of inclusion of ribozyme-based insulators or using CRISPR-RNA-processing to decouple the 5' UTR from the coding sequence (CDS)\textsuperscript{208,210}.

Different parts require different types of insulation in order to achieve maximum reproducibility without compromising features such as strength. Promoters used in synthetic biology projects are often minimal (-35 to -10 region) and without characterized transcriptional start sites; the inclusion of important functional elements such as the UP element has been previously recommended\textsuperscript{211}. Accordingly, the addition of an upstream and downstream insulating sequence has increased reproducibility in different genetic contexts\textsuperscript{212}. It is worth noting that the strong Clostridium promoters that are in widespread use in the Clostridium community, fdx and thlA, are 200 and 150bp respectively. Even though this recombinant thlA is longer than a usual minimal promoter, a long 5'UTR contributes to this part's length (recombinant thlA is 59bp from its 5' end to the transcriptional start site), while the recombinant 5' end is slightly truncated to exclude a Rex NADH-dependent regulator binding site (as mentioned before). A study that reported the use of a minimal thlA promoter in C. acetobutylicum, miniPthl which has a truncated 5' UTR, did not test activity variation in different genetic contexts\textsuperscript{91}.

Work in E. coli has demonstrated the utility of 'bicistronic RBS', where a leader peptide is translationally coupled with the CDS of interest, in improving reliability and context independence (downstream gene sequence) when a particular 5' UTR is combined with a new coding sequence\textsuperscript{212}. An upstream RBS and start codon initiate translation of the leader peptide, the stop codon of which overlaps with the start of the downstream CDS, while the latter's RBS is positioned with the leader peptide. The ribosomes translating the leader peptide unfold the 5' UTR (of the downstream CDS), preventing it from
forming secondary structures with the mRNA of the downstream CDS and thus influencing translation. The same study employed standard transcriptional start site ("+1 promoter"): 5' UTR junctions to minimize (or insulate against) unforeseen effects of combining promoters with new 5' UTRs.

As mentioned above, the observation of ribosomal repression of transcription termination also necessitates the more widespread use of 'distance' insulators of a sequence without emergent function and secondary structure to separate the stop codon and the stem-loop hairpin\(^{126}\). Such strategies are yet to be implemented in \textit{Clostridium} engineering projects.

Discussion

The development of biorenewables as an alternative to petroleum-derived commodity chemicals and fuels has resulted in the emergence of new markets\(^{213}\). Increasing productivity, broadening the range of feedstocks, improving tolerance to solvents and by-products are all existing challenges to achieving higher sustainability and ensuring the economic viability of \textit{Clostridium}-derived biorenewables\(^{214}\). \textit{Clostridium} species remain important hosts for the biological production of solvents and their further development relies on the adaptation of novel methodologies such as synthetic biology and metabolic engineering. Much progress has been made in the latter with several projects improving industrially relevant strains; however, the availability of biological parts with known behaviour is one of the limiting factors for the rate and scale of research. The reasons for our anticipation are twofold: first, the physical availability of biological parts streamlines assembly of genetic pathways, and second, knowledge about parts' behaviour is crucial in predicting and analyzing the behaviour of pathways and genetic circuits. We have shown in this review that there are multiple areas where improved knowledge of parts would provide a better toolkit for synthetic biologists using these organisms and consequently improve the rate at which industrial biotechnology and bioenergy process development can improve.

References


(87) Pyne, M. E. (2014, April) Development of genetic tools for metabolic engineering of
Clostridium pasteurianum. University of Waterloo.


(108) Heap, J. T., Kuehne, S. A., Ehsaan, M., Cartman, S. T., Cooksley, C. M., Scott, J. C., and


(215) Lorenz, R., Bernhart, S. S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C.,

