Geobacillus thermoglucosidans as a Thermophile Chassis for Synthetic Biology

Benjamin Reeve
Imperial College London
Department of Bioengineering
Centre for Synthetic Biology and Innovation
Supervised by Dr. Tom Ellis

Thesis submitted for the degree Doctor of Philosophy from Imperial College London.

Declaration

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

Benjamin Reeve

Abstract

Current reliance on petrochemicals for fuel and chemical production is environmentally damaging and unsustainable. The most promising alternative is bioconversion of lignocellulosic biomass however the organisms commonly used in microbial fermentations for chemical production are not well suited to utilise this feedstock. Alternative microbes for lignocellulose utilisation have been identified and *Geobacillus thermoglucosidans* is one of the best-adapted organisms currently known. This organism has been used for production of biofuels from lignocellulose but it currently lacks the tools and genetic parts needed to produce a wider variety of products.

In this study, novel tools and genetic parts to enable synthetic biology with *G. thermoglucosidans* were developed. The thermostability of reporter proteins, superfolder GFP, mCherry and flavin-based anaerobic fluorescent proteins was tested and superfolder GFP was shown to be the best reporter protein available for *Geobacillus*. Two novel constitutive promoter libraries were then generated and characterised. In both *G. thermoglucosidans* and *E. coli*, both libraries showed over a 100-fold range of expression strength with the strongest variants comparable in strength to the strongest previously reported *Geobacillus* promoter pLdh. Predictable tuning of expression strength in *G. thermoglucosidans* was further demonstrated using translation initiation rate calculator software and the limitations of such tools were reviewed. Finally, a set of seven modular shuttle vectors was developed and characterised. The resulting *Geobacillus* toolkit allowed for the first time, attempts to produce a more complex biobased product via *G. thermoglucosidans* genetic engineering. An operon was designed and constructed for biosynthesis of hyaluronic acid, using a newly-discovered hyaluronan synthase from the moderate thermophile *Streptococcus thermophilus*. The promise of this new enzyme was shown in *E. coli* where heterologous hyaluronic acid production was demonstrated.

The parts and tools developed here for enable more sophisticated genetic engineering with *G. thermoglucosidans*, making this the first chassis for thermophile synthetic biology.

Acknowledgments

Firstly thanks go to my supervisor, Dr. Tom Ellis for amazing support and supervision. Thankyou also to all in the Ellis lab for the science knowledge, camaraderie and fun times, particular thanks go to Elena Martinez-Klimova for the *Geobacillus* advice and understanding. The financial support from The Department of Bioengineering, Imperial College London is gratefully acknowledged. I am also thankful for initial sponsorship and support from TMO Renewables Ltd. particularly Dr Steve Martin and Dr Alex Pudney.

Thank you to Room 603 labmates, particularly Catherine and Kealan who have been there when so many experiments have gone wrong. Thank you to housmates, Ollie, Dom, Alice, Pippa, Jem, Steve, and Joe and to fellow Wilkinson Hall wardens and to all of the amazing Imperial 2014 iGEM team. Thank you to Caitlin for so much help and finally to my parents, Mum and Dad, I could never thank you enough and could not have done this without your support.

Contents

List of Figures

Publications Resulting From This Work

Predicting translation initiation rates for designing synthetic biology

Reeve B, Hargest T, Gilbert C, Ellis T (2014)

Front. Bioeng. Biotechnol. 20(2):1. doi:10.3389/fbioe201400001

The Geobacillus plasmid set: a modular toolkit for thermophile engineering

Reeve B, Martinez-Klimova E, De Jonghe J, Leak D.J., Ellis T (2016)

ACS Synth. Biol., DOI: 10.1021/acssynbio.5b00298

1. Introduction

Chapter Summary

Current reliance on petroleum as the primary resource for fuel and commodity chemicals is environmentally damaging and unsustainable. The most viable alternative is biobased manufacturing using renewable biological feedstocks converted into products by microorganisms. Lignocellulosic biomass is the most abundant feedstock however its use is challenging. It is difficult to degrade and cannot be easily utilised by most production microorganisms. The thermophilic bacterium *Geobacillus thermoglucosidans* is an ideal chassis for utilising this feedstock but its use is currently limited by a lack of tools. *G. thermoglucosidans* has been engineered to produce simple products, such as ethanol and butanol, from lignocellulosic feedstock but with improved genetic tools it could be engineered to produce a wide range of renewable biobased products.

1.1. The Need for Biobased Manufacturing

Perhaps the most daunting challenge we face in the modern world is addressing our environmentally damaging and unsustainable consumption of natural resources (1). It is over use of fossil fuels which causes the most pressing problems (2) and our dependence on crude oil is particularly difficult to relieve. Oil remains the primary source for transportation fuel and commodity chemical products such as solvents, fertilizers, pesticides, plastics and pharmaceuticals (3,4).

Oil is a finite resource but its human and environmental impacts present a far more pressing problem than its limited supply. Extraction and transport is inherently environmentally damaging and risks disastrous spillages that destroy wildlife and human livelihoods. When transported correctly, oil remains inherently toxic containing volatile organic compounds such benzene that are a hazard to human and animal health. Processing crude oil into petrochemicals often involves environmentally damaging catalysts or disposal of by-products and is resource- intensive in energy and water demand. When burned, petroleum releases numerous pollutes into the air. The sulphur

dioxides cause acid rain leading to ecological damage and loss of biodiversity though direct contact with plants and acidification of lakes and oceans. The increased atmospheric CO_2 also causes ocean acidification, damaging marine biodiversity to the extent of risking ecosystem collapse in certain areas. The greatest concern however is the contribution of such emissions to accelerating disastrous climate change. Indeed, the World Economic Forum considers failure to mitigate and adapt to climate change the biggest threat currently facing humanity (2).

Beyond the health and environmental risks, crude oil dependence also fuels political and economic troubles. Many oil exporting countries are politically unstable and in some, oil extraction may exacerbate this instability (5). Dependence on crude oil can stoke global political and military tensions with certain states exploiting their positions as major oil exporters to gain leverage in political disputes (6). In addition, crude oil prices are particularly volatile compared to other commodities, fluctuating widely year on year due to slight changes in demand and perceived security of supply. This causes profound economic difficulties for businesses and for countries overly reliant on oil (6).

The argument for change is clear and strong. To simply scale back consumption however is a near impossible task and would certainly reduce economic growth. In the last three decades this growth has lifted millions of people out of poverty, raised living standards, connectivity and freedoms and this is projected to continue. The answer instead then, is in sustainable growth – decoupling progress and prosperity from dangerous, unsustainable fossil fuel consumption (1,4). Advances in biotechnology offer the most promising solution by facilitating the transition from a fossil fuel based economy to a 'biobased' economy. Here, sustainable, carbon neutral, plant or algal feedstocks are converted into fuels and commodity chemicals, avoiding the damaging effects of fossil fuels.

Beyond simply avoiding previous dangers, biobased products can offer significant advantages in more cost effective production methods and opportunities for new products to address the additional challenges we face in sustainable energy, agriculture, health and manufacturing (6,7). Traditional commodity chemical production often relies on costly conditions, such as high temperature, pressure or the presence of expensive catalysts. Products may be produced in complex mixtures with other

undesirable molecules – these may be difficult to separate particularly if unwanted enantiomers are present. Biological processes can be cheaper and more efficient, requiring milder conditions, potentially low cost feedstocks and catalysts (enzymes) and can produce products more specifically - particularly in terms of stereospecificity (8). Additionally the shift to biobased chemical production can also offer manufacturing plants increased flexibility, production capacity and the ability to produce or modify more complex molecules not previously viable at large scales (7).

Our increased understanding and exploration of the biological world through progress to biobased replacement chemicals is also driving the discovery of new useful natural products perhaps not accessible through traditional chemical synthesis (9). Further to this, with recent advances in biological engineering, particularly protein engineering of enzyme catalysts, completely novel biobased compounds, not found in nature can be produced at scale (10). The bioeconomy also offers potential socioeconomic benefits. Growing feedstocks for fuel and chemical production benefits more local, agricultural producers rather than global fossil fuel corporations. This boost to rural economies can increase employment, stem urbanisation and decrease inequality. In Brazil for example, production of sugarcane feedstocks employs over 1 million people. In a country without social security the industry provides vital basic income and has been credited with greatly reducing poverty in many of the poorest regions (6).

The challenge of progressing from fossil fuel based economy and towards a biobased economy is strongly pushed by the considerable human and environmental costs associated with fossil fuels. It is also pulled by the huge potential of a bioeconomy for improved manufacturing and innovative novel products to drive economic growth and address other pressing challenges.

1.2 Synthetic Biology and Biobased Chemicals

The transition to a bioeconomy will demand social and political changes but is crucially reliant on technological progress in many fields (11). The discipline with potential for the greatest impact however, is the comparatively young field of synthetic biology (12).

This discipline enables huge innovations in the development of industrially suitable organisms for conversion of biological feedstocks into commodity chemicals and fuels.

Synthetic biology is loosely defined as the design and engineering of novel biological parts, devices and systems as well as the redesign of existing, natural biological systems (13). It builds on advances in molecular biology, cell biology, systems biology and genetic engineering but aims to frame this knowledge within an engineering approach. The field's rapid growth has been enabled by exponential improvements in DNA sequencing and synthesis technologies, to build biological systems in addition to modelling and bioinformatics techniques to design and analyse them (14).

As an engineering discipline, efforts are ultimately focussed on building systems for useful applications more than further knowledge generation. The field is particularly inspired by engineering principles including standardisation, modularization and abstraction (15). Inspiration and metaphors for how to design and build with biology are often drawn from electronic engineering and information technology. As a caveat, it should be noted however that biology will always remain somewhat complex and messy. Biological components are extremely context-dependent, involve intrinsically stochastic processes and their combinations can display emergent properties. Systems are inherently dynamic, they change over time as cells grow and divide and are subject to Darwinian evolution. Life will never be programmable exactly like a computer or assembled from parts as easily as a circuit board, but these metaphors can help guide the discourse and direction of synthetic biology (16,17).

Standardisation – Standards underlie all of engineering allowing knowledge and parts to be shared and combined. Standards are vital firstly in measurement; for example, the power outputs of lightbulbs are given in watts so any two bulbs can be accurately compared. Secondly in construction, any bayonet cap light bulb will fit any bayonet socket. In synthetic biology standard measurements include agreed conditions – M9 Glucose media for growing *Escherichia coli* and output units, and product production measured in grams/litre/hour for example.

Modularization – parts or devices are built and characterised as discrete entities performing a particular function (ideally) regardless of context. This allows modules to be interchangeably combined to build more complex systems in a predictable manner. A synthetic biology community has grown around this goal of implementing standardized modular building blocks (18) with the aim of making complex designs easily assembled from standard "off-the-shelf" biological parts. The MIT-based Registry of Standard Biological Parts (www.partsregistry.org) lists and distributes thousands of widely-used genetic building blocks, largely formatted to match a DNA sequence standard called BioBricks that enables simple construction. Currently the goal of predictable modular construction is not limited by number of available parts but in their quality and characterisation (19).

Abstraction – Through abstraction, engineers can design and build complex systems without a detailed understanding of the underlying components. The complex, scientific details of components are simplified into abstract representations of their behaviour. Circuit diagrams in electronic engineering are the classic example of this - circuit designs can be represented simply to communicate what a circuit does without a need for knowledge about how a transistor actually functions for example. Abstract parts can then be used to design sub-systems, which are again abstracted (Figure 1.1). In biological systems, genetic parts can be combined to create devices with particular biological functions - invert a signal or communicate with another cell. Hierarchical layers of abstraction then enable comparatively simple design of complex large scale systems (15,20). The use of these standards, modules and abstractions, in combination with automation of tasks such as DNA synthesis and construction can lead to complete decoupling of high level system design from lower level specifications and parts fabrication.

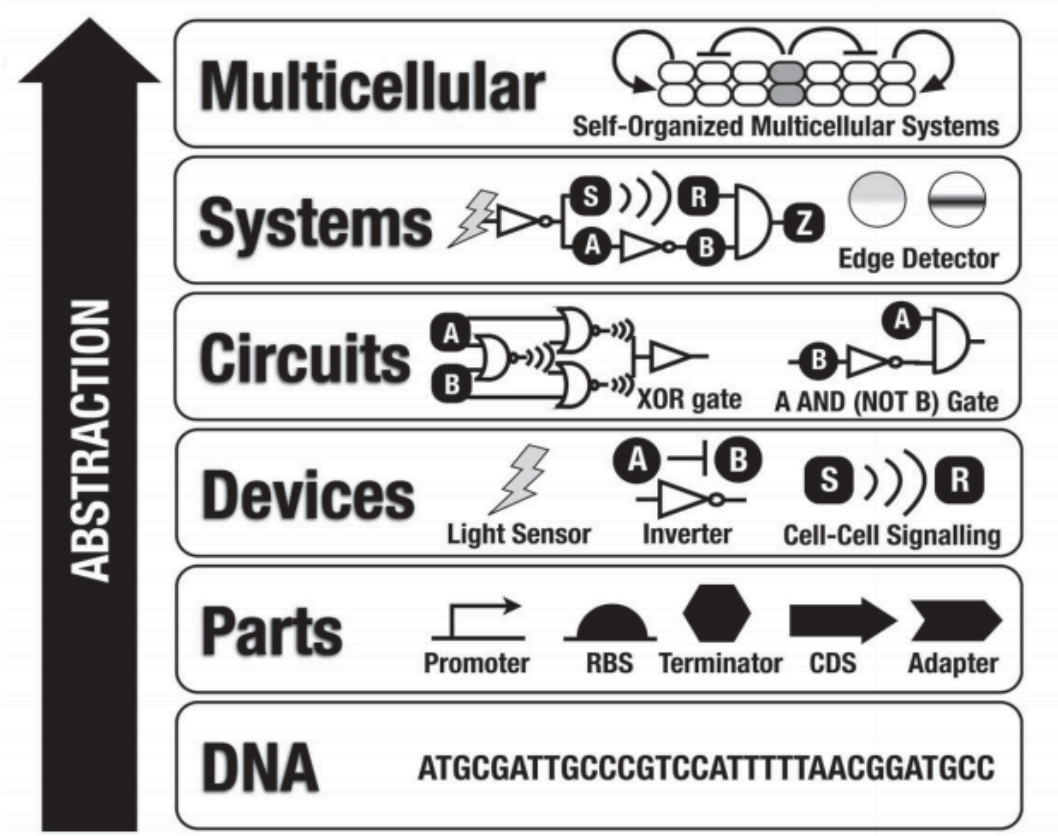


Figure 1.1. The synthetic biology abstraction hierarchy from Federici *et al*. 2013 (21)

Development of synthetic biology applications tends to follow an engineering inspired design/build/test/analyse cycle (22) (Figure 1.2). As the field advances, improvements in DNA synthesis and sequencing, and improvements in automation of DNA assembly and circuit/system testing make these cycles faster and better.

Synthetic biology also draws inspiration from elsewhere in engineering through the use or rational model-guided design. Model-guided design means that fewer experimental engineering cycles are theoretically necessary to get to the specified application, reducing the costliest part of development in synthetic biology. Genetic engineering in the 1980s and 90s focused on bespoke alterations to a handful of genetic elements that had to be individually tested and refined. With characterized, modular parts however, *in silico* modelling can guide predictable gene network design and construction. This reduces or eliminates the need for many further edits after the system is built (23). The predictable functioning of higher complexity circuits and systems remains entirely dependent on the quality and characterisation of lower order parts, however.

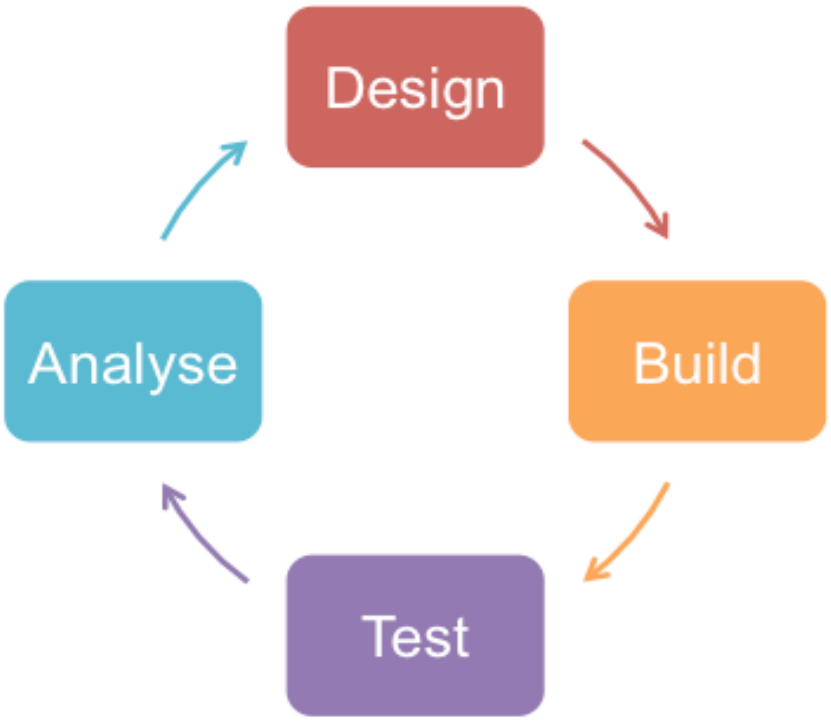


Figure 1.2. An engineering inspired design cycle. Testing and data collection informs further designs.

The above-described features tend to characterise the original, parts-based synthetic biology that emerged a decade ago. However the field, its definitions and applications are now becoming increasingly broad. Synthetic biology is increasingly merging with other disciplines such as metabolic engineering, bioprocess engineering and with areas of bioinformatics. This eroding of traditional subject boundaries along with increased collaboration between disciplines is critical to address large-scale challenges such as the transition away from fossil fuels (11). For example, a problem with unwanted inhibitor molecules in a biobased manufacturing feedstock could be solved by chemical engineers designing better pretreatments to remove them or equally by synthetic biologists improving production strain tolerance of these molecules. Only through collaborations and/or training of individuals in both disciplines can the best solutions be discovered.

Developments in many areas of engineering and biology are necessary for the transition to a bioeconomy. The US National Academies 2015 report identified the key areas for innovation (7) with synthetic biology identified as particularly relevant for building pathways and for developing new chassis strains and organisms (Figure 1.3). A chassis

in synthetic biology is defined as an "autonomous genetic and/or biochemical scaffold that functions as a dynamic platform for implanting designed biological devices" (24). More simply, the chassis is the starting organism (or cell free system) to which modifications are made and genetic constructs added.

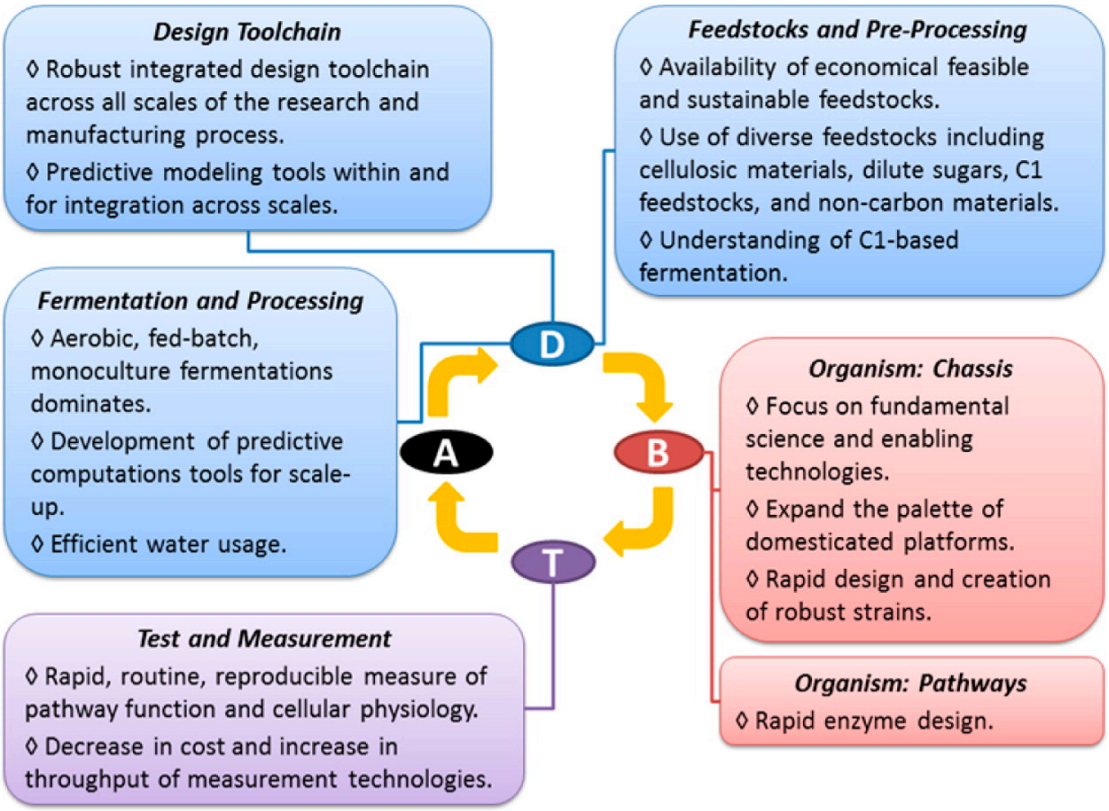


Figure 1.3. The key areas for innovation identified by the National academies report (National Research Council 2015) mapped onto the Design-Build-Test-Analyse cycle by Friedman and Ellington (Friedman & Ellington 2015). Of particular interest are the need for "diverse feedstocks including cellulosic materials" and expanding the "palette" of available chassis organisms.

1.2.1 Some Synthetic Biology Success Stories

To date synthetic biology has made considerable progress in the production of commodity chemicals from renewable feedstocks. Many products, traditionally extracted in small quantities from plant or animal sources or synthesized by organic chemistry methods from petroleum-based feedstocks now have biobased alternatives produced at scale from renewable feedstocks (Table 1.1). Historically, microbial production of chemicals was limited to nature's existing chemical repertoire: fatty acids, amino acids, alcohols, antibiotics etc. Microbes were simply mutated and selected to produce higher titres of these existing metabolites. With the advent of

genetic engineering, it became possible to produce heterologous products: molecules found in nature but not produced by the production host. In future, with the development of protein engineering and synthetic biology, novel products not previously found in nature could also be produced (10).

Compound	Institution/Company	Main applications	References
Methyl ketones	Joint BioEnergy Institute	Solvents	(25)
Farnesane	Amyris	Biodiesel and jet fuel	(26)
Propylene	Global Bioenergies	Polypropylene	(27)
Indigo	Genencor	Dyes	(28)
Vitamin C	Genencor and Eastman	Vitamin supplement	(29)
Spider silk	Bolt Threads and Spiber	Textiles	(30,31)

Table 1.1. A selection of basic and specialty chemicals now commertially produced in a one-step conversion from renewable sources to the final chemical. A synthetic biology approach was used in each case to achive production and commertially viable yields. Adapted/updated from Lopes 2015 (4).

Despite the significant successes given here, many challenges remain. Particular areas needing innovation include the utilization of more complex but cheaper, more sustainable feedstocks such as lignocellulosic biomass and expanding beyond engineering in standard laboratory strains to new ‘chassis’ strains better suited for industrial processes (Figure 1.3) (7).

1.3 Utilising Lignocellulosic Feedstocks

First generation feedstocks such as corn and sugar cane contain easily accessible sugars and can be used to feed microbial fermentations with minimal pre-treatment. These crops have been utilised on a large scale for decades in the production of bioethanol for transportation fuels and they remain preferred feedstocks for smaller scale production of high value commodity chemicals. The widespread use of first generation feedstocks for fuel production has been scaled back however due to concerns about competition

with food crops. Biofuel and chemical production is not the only (and is far from the most significant) reason for changes in crop availability and food prices over the last two decades, however biofuels has shouldered a large share of the blame (6). These concerns are understandable. Large-scale deployment of biofuels from these feedstocks does compete to some degree with food production in use of agricultural resources and could affect hunger and poverty through a complex interaction with changes in agricultural policy and investment. It may also drive land use changes such as increasing deforestation with negative environmental consequences (6).

Conversion of non-food cellulosic plant biomass, however, avoids these concerns. Lignocellulosic biomass is the most abundant renewable resource on the planet and is widely, inexpensively available. Common sources include agricultural residues such as corn stover and sugarcane bagasse. It is also available from waste streams from the food and forestry industries, or can be sourced from fast growing crops such as switchgrass, agave and poplar that can be grown on low quality land unsuitable for food production. For most lignocellulosic feedstocks, the sugar content is comparable to food feedstocks such as corn (32). However, the cost-effective conversion of these sugars to fuels and chemicals is far more challenging than with simple, starch based feedstocks.

Allternatives to using lignocellulosic biomass include so-called “third-generation” technologies. Here, bacteria or algae utilise carbon dioxide directly as the feedstock (possibly in addition to hydrogen and/or carbon monoxide). These processes involve shallow pools or tubes of photosynthetic organisms or bioreactors fed by waste gas streams. Whilst these applications are exciting and currently viable for production of certain niche chemicals they are in their relative infancy (33) and not currently ready to address the pressing, large scale need for biobased manufacturing. Third generation processes are often complex, susceptible to contamination and, whilst feedstocks are inexpensive or free, these still require other less renewable resources. Addition of

nitrogen and phosphate sources as well as trace minerals is usually essential and algal solutions have huge freshwater demands (34). Considering the scale and urgency of the need for biobased manufacturing, second generation, lignocellulosic feedstocks are the most promising despite their many challenges.

A typical process for utilising lignocellulosic feedstock is shown in Figure 1.4. The sugars in lignocellulose are locked in stable polymers of cellulose, hemicellulose and lignin. Their proportions and structure can vary considerably between different biomass sources and these polymers have evolved over millions of years to protect plants by resisting natural deconstruction. Fibres of cellulose are generally encased in a covalently linked network of lignin and hemicellulose. The cellulose fraction, (30 to 40% of biomass by dry weight) is composed of only D-glucose linked by β-1,4 glycosidic bonds while a mixture of pentoses - especially xylose and arabinose - and hexoses - galactose, glucose, and mannose - comprises the main component of hemicellulose (20 to 40% of biomass dry weight). Lignin by contrast is a complex polymer of aromatic alcohols (32). There are three major challenges to bioconversion of lignocellulose: degradation to sugars, utilization of 5-carbon sugars (particularly of xylose) and tolerance to fermentation inhibitors (3).

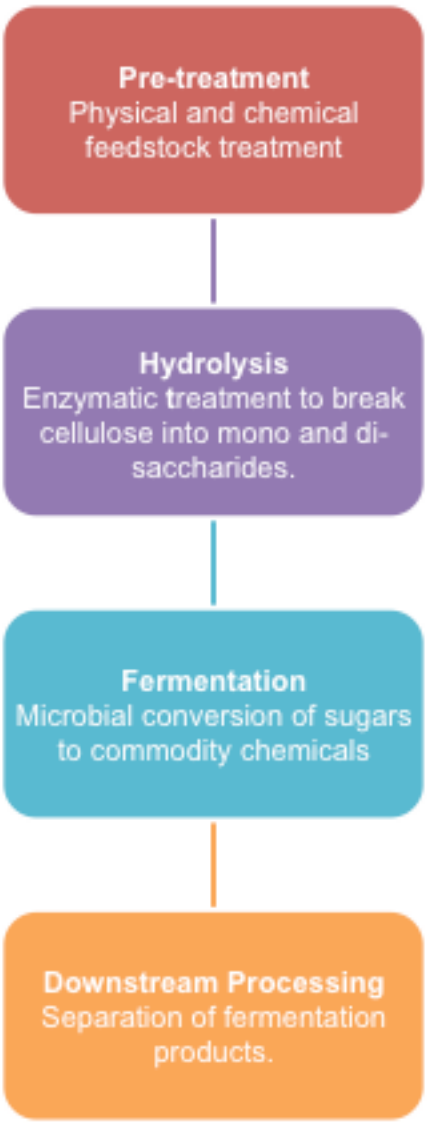


Figure 1.4. Process diagram showing commodity chemical production from second-generation cellulosic feedstocks. Adapted from Bartosiak-Jentys 2010 (153).

Degradation

In a typical process (Figure 1.4), raw feedstock is mechanically broken up and then undergoes chemical pretreatment processes. Steam pretreatment with dilute mineral acids is the most common and this is effective at degrading the pentose hemicellulose

polymers, breaking up the microscopic structure of the feedstock. The cellulose is then more accessible to cellulases in a second enzymatic digestion step (35). After pretreatment and cellulase digestion, the majority of sugars in agricultural waste will be released into the broth as monomers, dimers or short oligomers, accessible for conversion into fuels and chemicals during a microbial fermentation step. The cost of cellulase enzymes is a significant factor for the economically viable use of these feedstocks. However, ongoing efforts in synthetic biology both in academic and industrial laboratories are continually improving cellulase enzymes and enzyme complexes (36). Judicious use of cellulases can improve costs (37) but converts lower quantities of the biomass to sugar monomers and dimers, leaving higher concentrations of oligomers. One solution, and a significant trend in this research area (3) is the effort to select or engineer fermentation organisms that directly utilise oligomers such as cellodextrins and xylooligosaccharides. This approach requires strains with both transport systems that uptake sugar oligomers and appropriate enzymes to intracellularly degrade them (38).

Alternative solutions to reduce costs of the enzyme hydrolysis step include simultaneous saccharification and fermentation (SSF) and consolidated bioprocessing (CBP). With SFF approaches, the enzyme hydrolysis is combined with the microbial fermentation. This can reduce residence time, equipment costs and relieve feedback inhibition from monomeric/dimeric carbohydrates on the cellulase enzymes as the fermentation organism quickly consumes liberated sugars. SFF is challenging however because bioreactor conditions must then become a difficult compromise between acceptable enzymatic hydrolysis and fermentation efficiency (39). In CBP, saccharification of lignocellulose and fermentation are also combined into a single process step however in this case all enzymes are produced by the process organisms. Microbes or microbial consortia can potentially secrete the range of necessary enzymes for feedstock digestion in addition to converting the feedstock into product. This is hugely complex to achieve in reality, however, as optimising bioreactor conditions for a single organism is difficult and so for consortia it becomes hugely challenging. The alternative, engineering all CBP capabilities into a single organism, is again a considerable challenge (40).

Note that the processes described above are not mutually exclusive and the optimal processes could be some combination of these. Even with a separate hydrolysis step, native cellulase secretion by the production organism is immensely valuable, as is the option to engineer expression of heterologous enzymes.

5C Sugar Utilization

One of the major carbohydrates in typical lignocellulosic biomass is D-xylose, a five-carbon aldose, which is difficult for many microorganisms to metabolize. Common bioethanol producing industrial organisms such as *Saccharomyces cerevisiae* and *Zymomonas mobilis*, do not natively metabolize xylose and although some bacteria such as *Escherichia coli* have a native xylose utilisation pathway, it is not efficient and is commonly repressed by the presence of glucose (32). Production organisms would ideally be able to utilize D-xylose naturally or else require this capacity to be engineered-in by heterologous enzyme expression.

Fermentation Inhibitors

Pretreated lignocellulose hydrolysate contains a number of inhibitors that retard cell growth, slow substrate metabolism, and reduce product formation. Inhibitors may come from the plant biomass itself, such as acetic acid, or from degradation products of lignocellulose produced during the treatment process such as furfural, furans and phenolics. Furfural, a dehydration product of pentose sugars, is one of the most potent inhibitors and can completely prevent cellular growth even at low concentrations. (41). There are chemical engineering solutions to reduce inhibitor concentrations such as over-liming the feedstock immediately after acid pretreatment with $Ca(OH)_2$ or by filtering inhibitors out with active carbon. However, these increase the process complexity and operational cost (42). Instead, as knowledge about toxicity mechanisms and microbial tolerance traits emerges there is a growing interest in selecting for resistant strains and engineering increased resistance to these inhibitors (43).

1.4 Alternative Chassis – Beyond Model Organisms

"We need to move from research that is E. coli or S. cerevisiae-centric to work that utilizes organisms more suitable to fermentation and production" - Friedman & Ellington (12)

Synthetic biology continues to make astonishing advances in the degree to which organisms can be rationally redesigned, however this work remains a difficult task. The properties that have been engineered into novel strains so far are minimal compared to the possible variety of features and phenotypes observed in nature. For challenging applications such as the utilization of lignocellulose for chemical production careful selection of a chassis strain naturally well suited to the application minimises the need for further complex biological engineering. As so few model organisms have been established as chassis for synthetic biology the ideal organisms are probably "not part of the current pantheon of established production strains" (7). A huge variety of organisms have already been considered and tested for chemical production from lignocellulosic biomass.

Key attributes for strains utilising this feedstock include:

- Inhibitor tolerance – particularly to furfural and furans.
- A broad spectrum of fermentable substrates – the profile of sugars available in second generation feedstocks is broad and variable.
- Co-utilisation of sugars – using sugars simultaneously and not displaying catabolite repression.
- Native cellulase and hemicellulase activity – to aid feedstock degradation and be able to utilise longer chain substrates.

This is in addition to general requirements for industrial strains such as:

- Potentially high yields and productivity of product formation.
- Tolerance to high concentrations of the product.
- Process "hardy" – i.e. being able to tolerate fluctuations in temperature, gas solubility and pH.
- Minimal production of unwanted by products.
- Generally Recognized as Safe (GRAS) status.

- Minimal nutrient requirements – supplementing the feedstock with minerals or vitamins is not cost effective. (44)(10)

The diverse metabolic and physiological requirements for production of different compounds from different feedstocks demands a range of chassis organisms for metabolic engineering. Microorganisms with a naturally high tolerance for long-chain alcohols make more suitable as hosts for biofuel production for example, whilst strains with very low pH tolerance are advantageous for production of organic acids. *E. coli*, *S. cerevisiae*, and other model organisms are so highly used due to the extensive repertoire of genetic tools available for these hosts and a deep knowledge of their genes and metabolism. To better produce biobased products from renewable biomass, additional foundational research with organisms better suited to bioprocess engineering and production is required (10,45).

1.4.1 Candidate Industrial Microorganisms

Eukaryotes

The yeast *Saccharomyces cerevisiae* is naturally ethanologenic and ethanol tolerant and has been used as for industrial bioethanol production with first generation feedstocks since the 1970s (46). *S. cerevisiae* does not metabolise 5C sugars such as xylose however and is very sensitive to the growth inhibitors in pretreated lignocellulose feedstock. Although several recombinant strains have been developed for second generation ethanol production, industrially viable performance has been challenging (van Maris et al. 2006) (46). *S. cerevisiae* is well established as a chassis for synthetic biology however with well characterised parts and tools so is preferred for producing very complex high value products where feedstock costs are not a significant concern.

Many others fungi including filamentous *Trichoderma* and *Aspergillus* species (48,49) are better adapted to lignocellulose utilisation but are slower growing than bacterial species and are often less resistant to inhibitors and stresses.

Mesophilic Bacteria

The model organism *Escherichia coli* has been extensively engineered for commodity chemical production. This chassis has the greatest range of characterised tools and parts for synthetic biology and so is often used as a "proof of concept" chassis for ambitious metabolic engineering. This species is not particularly well suited to industrial bioreactor conditions however and does not efficiently utilise lignocellulosic feedstocks (50). A huge range of alternative Gram-negative bacteria such as *Zymomonas mobilis* and *Klebsiella oxytoca* have been engineered for renewable chemical production and may have particular biochemical advantages for production of certain products but they not readily metabolise 5C sugars and so require first generation feedstocks (44,51)(52).

The model Gram-positive bacterium, *Bacillus subtilis* has been extensively studied and is popular as an industrial production strain with many well-characterised genetic parts. This species can import and utilize 5C and cellobiose but requires significant engineering to efficiently utilise lignocellulosic feedstocks and does not grow well under anaerobic conditions, which limits its applications (53).

Thermophillic Bacteria

Thermophiles are organisms with an optimal growth temperature above 50 °C (54). As few researchers (and very few synthetic biologists and metabolic engineers) work with thermophiles, these species are often overlooked in favour of more established mesophile model organisms. However, as production strains for chemicals from lignocellulosic feedstocks, thermophilic bacteria may offer many advantages:

- Many thermophilic bacteria utilize pentoses and hexoses as well as more complex carbohydrates found in lignocellulose hydrolysates (44)

- High temperatures aid downstream processing of certain products. Volatile compounds such as ethanol can be separated from fermentation by application of a mild vacuum or by gas stripping. This greatly reduces the costs of product recovery (55)

- Thermophilic bacteria often display a high tolerance to fluctuations in pH, temperature and other stresses (56–58).

- Higher temperatures reduce the risk of contamination by common mesophilic contaminants (Sommer et al. 2004; Cripps et al. 2009), though contamination by spores of other thermophilic microbial species is possible.

- Many processes using a mesophilic production strain require cooling of the feedstock after pretreatment and subsequent heating during downstream processing steps such as distillation. With a thermophile a higher temperature can be maintained throughout the process reducing energy input (58).

- The solubility of substrates is increased at high temperatures, allowing for greater concentrations of carbohydrates to be used in feeds (58).

- Thermophiles usually have faster growth rates and potentially faster feedstock conversion than mesophilic organisms (60)

Strong candidates for thermophilic industrial applications include, *Clostridium*, *Thermoanaerobacter,* and *Geobacillus* species. These bacteria each have different advantages, which may make them the preferred chassis for production of particular products or utilisation of particular feedstocks.

***Clostridium* Species**

Lois Pasteur reported arguably the first deliberate biobased chemical production from a microbial fermentation in 1861. He a demonstrated a process that produced butanol from sugars in the absence of oxygen and the organisms responsible was later shown to be *Clostridium acetobutylicum* (61). Many clostridium species can naturally utilise lignocellulosic feedstocks and thermophilic species such as *C. thermocellum* have rapid growth and feedstock conversion. As obligate thermophiles manipulation in the laboratory is challenging but efficient tools for transformation and modification of *Clostridia* species have been developed (62). Applications are limited however as these species have comparatively poor solvent tolerance and do not efficiently ferment pentose sugars (58).

Thermoanaerobacter species

Thermoanaerobacter are a genus of thermophilic bacteria closely related to *Clostridia* that have many of the same advantages as production chassis. Several species are also better able to utilise 5C sugars such as xylose and so have been effectively used for

production of second-generation biofuel. Genetic manipulations are difficult however and tolerance to certain products and inhibitors in the feedstock is limited (63).

Geobacillus species

Geobacillus species are comparatively amenable to genetic manipulation, efficiently utilise a range of feedstocks derived from lignocellulosic biomass and have good tolerance to feedstock inhibitors as well as many potential products such as organic alcohols and acids (64,65). In a broad study of possible strains for advanced biofuel production at the U.S. department of energy's Joint BioEnergy Institute (JBEI) a *G. thermoglucosidans* strain emerged as the strongest candidate due to its efficient utilisation of a range of sugars and product tolerance (66)(67). *G. thermoglucosidans* was described as the "ideal microbe" for advanced biofuel production and with improved tools for genetic engineering, could also be an excellent chassis for many other products.

1.5 *Geobacillus* Species

The genus *Geobacillus* includes thermophilic Gram-positive endospore-forming bacteria that form a phylogenetically coherent clade within the family Bacillaceae. Species in the *Geobacillus* genus are strictly aerobic (such as *G. kaustophilus*) or facultatively anaerobic (*G. thermoglucosidans*) and are capable of growth between 40 °C and 70 °C (64).

In nature *Geobacillus*, species are obligate thermophiles occurring in environments heated by geothermal activity or by microbial activity. They have been isolated from hot springs, oil fields and gas wells, hydrothermal vents and compost heaps (68) (Figure 1.5). They are primarily decomposers of plant biomass, and in warm compost heaps *Geobacillus* species are the dominant culturable microbes. Genomic analysis shows they are highly adapted to utilise this feedstock (68). Interestingly however, many species also possess the ability to utilise more eclectic feedstocks such as long chain hydrocarbons (69).

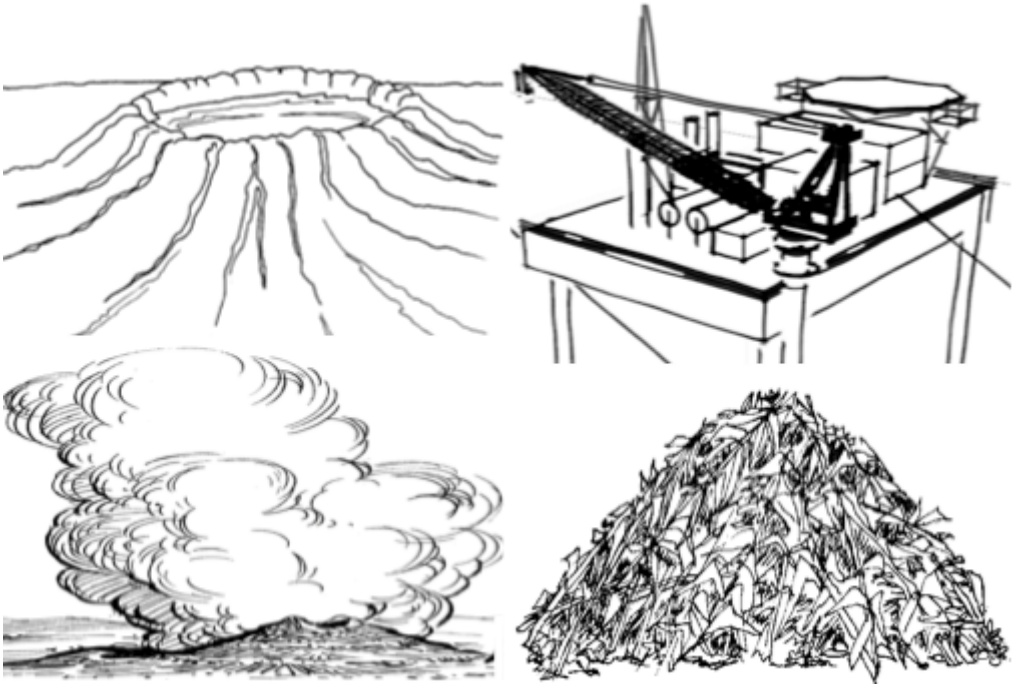


Figure 1.5. The natural habitats of *Geobacillus* species hot springs, geothermally heated oil wells, hydrothermal vents and compost heaps. Adaptations for these environments make *Geobacillus* species well suited for industrial bioreactor conditions.

The most studied *Geobacillus* species include *G. thermoglucosidans* (or *glucosidasius*) *G. kaustophilus*, *G. thermodenitrificans* and *G. strearothermophilus* (57,70–72), all shown on the cladogram below (Figure 1.6). Species in the next most closely related genus, *Anoxybacillus* are also thermophilic and, though less well studied, could have applications in bioremediation (73). Thermophilic bacilli such as *B. smithii* and *B. coagulans,* also of industrial interest (74) are only distantly related. The *Geobacillus* specie's closest well studied relative, with detailed transcriptome and proteome data (75,76), is *B. subtilis* (Figure 1.6)

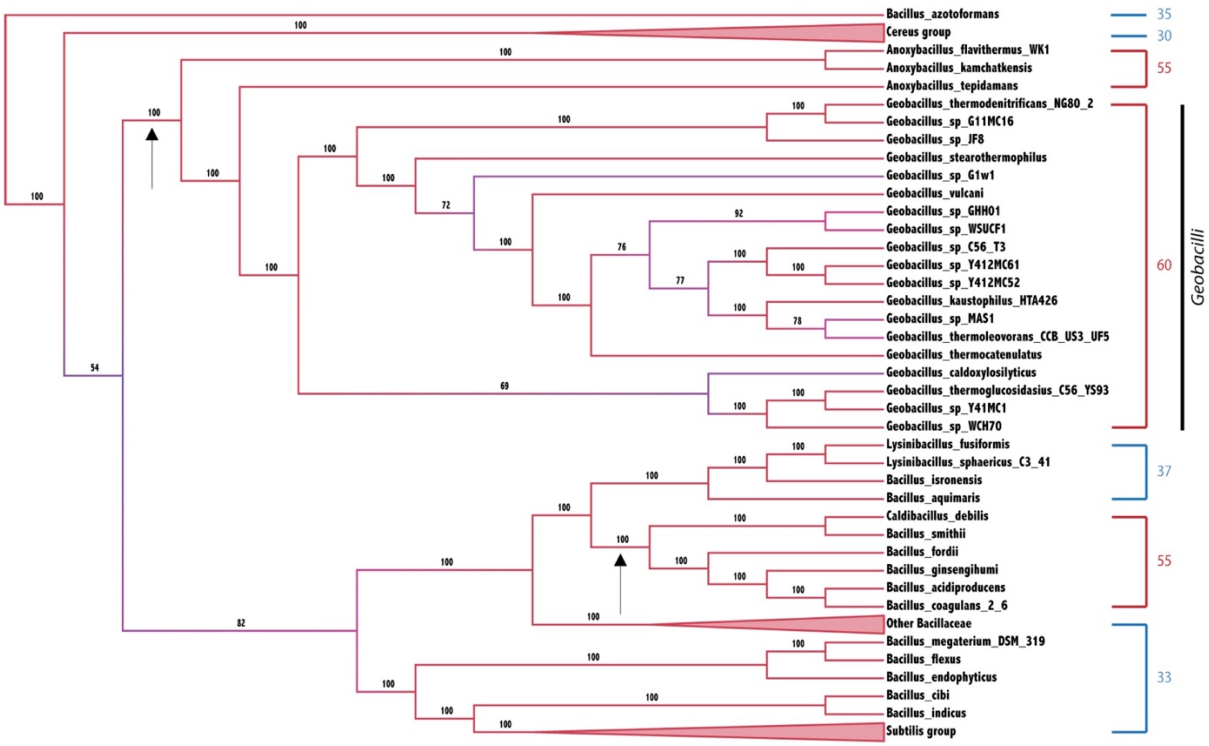


Figure 1.6. Phylogeny of *Geobacillus* species The cladogram is a pruned section of the Bacillaceae family tree generated from alignments of single copy homologous gene families universal to all Bacillaceae sampled. Branch labels show bootstrap confidence value percentages. Wedges represent strongly supported collapsed groups. Average optimum growth temperatures for the bracketed species are given on the right. Black arrows indicate suggested origins of thermophily. This figure was generated by A. Esin (Department of Life Sciences Imperial College) and reproduced here with kind permission.

Many potential applications for Geobacillus species have been developed or suggested including production of thermophilic proteins, pollution control and bioremediation (69,77–79). Production of chemicals from lignocellulosic biomass is the most promising application area however (64,80).

1.5.1 Geobacillus Engineering So Far

Despite *Geobacillus* species being so well-suited for a variety of metabolic engineering and biotechnology applications, there have been very few studies that use genetic engineering to modify these thermophiles. *Geobacillus* species were first transformed with large, naturally occurring plasmids in the early 1980s by Imanaka et al. (81). A decade later, better characterised, stably replicating shuttle vectors were then developed: pBST22 (82) and pSTE33 (83). *Geobacillus* species were then only considered as a source for thermostable proteins, however, not as potential chassis organism. Interest in biobased chemical production with these thermophiles has only

taken off within the last decade. This has reinvigorated *Geobacillus* research and now several metabolic engineering applications have been recently explored.

Metabolic Engineering

Two significant metabolic engineering achievements with *Geobacillus* species have been possible so far, both using the *G. thermoglucosidans* species as the chassis: high yield ethanol production (60) and isobutanol production (80). The genetic engineering in both studies involved gene knockouts and/or overexpression of native genes with a strong promoter.

G. thermoglucosidans naturally produces small yields of ethanol under some conditions which makes it of interest for biofuel production. Under anaerobic conditions, mixed acid fermentation occurs naturally in *G. thermoglucosidans* producing lactate, formate, acetate and ethanol. For higher yield ethanol production, Cripps *et al*. knocked out genomic copies of lactate dehydrogenase (LDH) and pyruvate formate lyase (PFL) genes to direct metabolic flux towards ethanol rather than to the usual production of lactate and formate. Pyruvate dehydrogenase PDH was upregulated by replacing its promoter on the genome with the much-stronger *G. stearothermophilus* LDH promoter (pLdh) (60) (Figure 1.7b). For carrying-out all of these genetic modifications, temperature sensitive vectors were used that included a kanamycin resistance marker flanked by 300 bp of homologous sequence that matched the required genomic region where the modification would be made. These vectors were introduced by electroporation and genomic integration was selected-for by raising growth temperature with kanamycin selection. Antibiotic selection was then removed and, after several rounds of subculturing, second recombination events excising the kanamycin marker were observed. While the genetic engineering in this study was trivial compared to that done in model organisms, it demonstrated the huge potential of *G. thermoglucosidans* as a production strain. Potential ethanol yields from the engineering *G. thermoglucosidans* proved to be high, giving 0.45 grams of ethanol per gram glucose, 90% of the theoretical maximum. The productivity, 3.2 g/l/h, was also particularly impressive, demonstrating the advantage of fast, thermophilic metabolism.

In the second study of note, an improved biofuel molecule was the target, and isobutanol is ideal in this respect as unlike ethanol, it can directly replace petrol to almost 100% in vehicle fuels. However, isobutanol production required more challenging metabolic engineering. This product is not naturally produced by *G. thermoglucosidans* and so expression of heterologous genes was needed. A previous successful strategy for isobutanol production in *E. coli* expressed an acetolactate synthase (AlsS) from *B. subtilis* and a ketoisovalerate decarboxylase (KIVD) from *L. lactis* (84), these enzymes were shown to have reasonable thermostability *in vitro* and so were used for production in *G. thermoglucosidans* (Figure 1.7b). The enzymes in the pathway that feed up to KIVD are all usually involved in valine biosynthesis and so are naturally present in most bacteria. Several putative alcohol dehydrogenase (ADH) enzymes were identified from the *G. thermoglucosidans* genome and four were shown to have activity for isobutyraldehyde reduction. Whilst heterologous expression of the KIVD enzyme from *L. lactis* alone could theoretically give isobutanol production, none was actually detected. Therefore, 8 different operon variants overexpressing other genes from the pathway were all tested. All genes had their natural ribosome binding site and were transcribed in an operon by the strong Ldh promoter of *G. thermodenitrificans* on the pNW33N plasmid. The operon giving the highest yield, 3.3 g/l, was a three gene operon containing the *L. lactis* KIVD followed by the native *G. thermoglucosidans* ketol-acid reductoisomerase (KARI) and finally the *B. subtilis* AlsS.

The yield and productivity was very low with 3.3 g/l only achieved after 36 hours growth in media with 36 g/l glucose - far worse than with *E. coli* previously (84). Whilst this may in part be due to the non-optimal growth of *G. thermoglucosidans* 50 °C (this lower than usual temperature was required as the ALS and KIVD enzymes were taken from mesophilies), theoretically yields could be considerably improved with further genetic optimisation. For example, the genes were all very strongly transcribed with no alternative promoters tested. Furthermore, no consideration to ribosome binding site (RBS) sequence and translation rate was given.

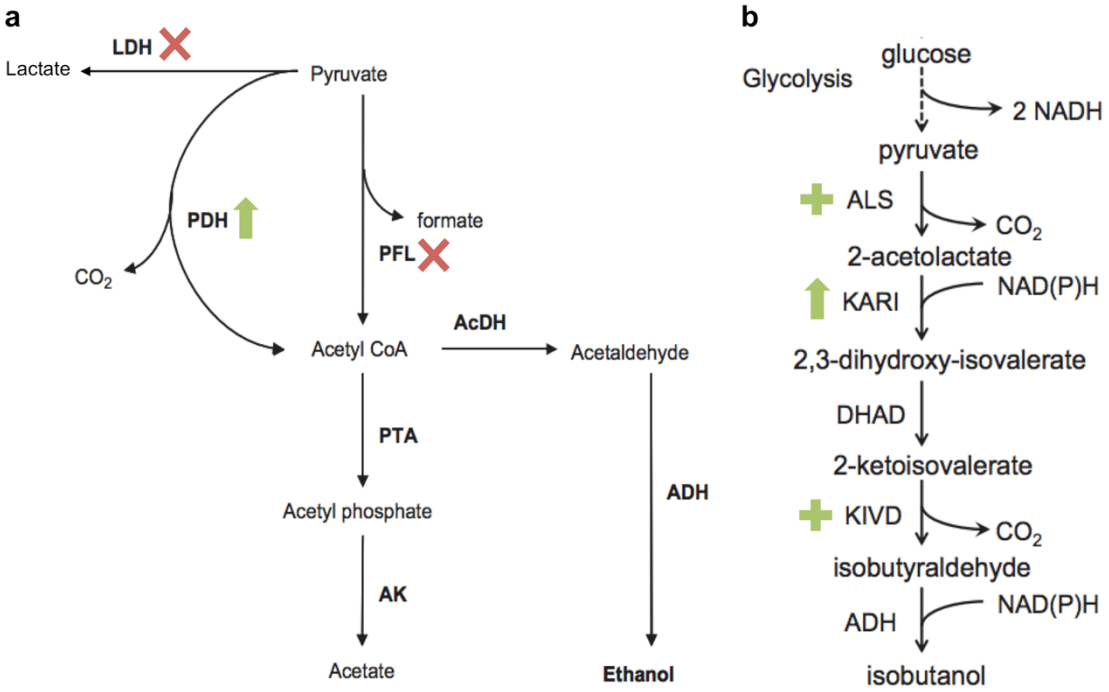


Figure 1.7 a) Metabolic engineering strategy from by Cripps *et al.* for ethanol production, enzyme abbreviations are: lactate dehydrogenase (LDH), pyruvate formate lyase (PFL) PDH, pyruvate dehydrogenase (PDH), acetaldehyde dehydrogenase (AcDH), alcohol dehydrogenase (ADH), phosphotransacetylase (PTA) and acetate kinase (AK). LDH and PFL were knocked out whilst PDH was upregulated by promoter replacement with the strong *G. stearothermophilus* LDH promoter (pLdh) (60). **b) Metabolic engineering strategy from by Lin *et al.* for isobutanol production,** enzyme abbreviations are: acetolactate synthase (ALS), ketol-acid reductoisomerase (KARI), dihydroxy acid dehydratase (DHAD), 2-ketoisovalerate decarboxylase (KIVD) and alcohol dehydrogenase (ADH). In the highest yielding strain a three gene operon was expressed from a plasmid (pNW33N) with the native *G. thermoglucosidans* KARI, an ALS from *B. subtilis* and a KIVD from *L. lactis*. Sufficient DHAD and ADH activity existed in the natural strain (80).

For these two exemplar projects with *G. thermoglucosidans* producing biofuels, both Cripps *et al.* and Lin *et al.* employed "traditional" genetic engineering strategies – either gene knockouts or strong unregulated overexpression. Expression strength was not tuned and there was not model-guided or *in silico* design. This approach may be sufficient for redirecting metabolic flux to increase the yield of a natural product (e.g. ethanol) but is not sophisticated enough to balance flux through a more complex pathway (e.g. for isobutanol). Isobutanol is still a relatively simple, low value product compared to the huge breadth of biobased products currently produced by mesophiles. Production of further products in *G. thermoglucosidans* is clearly severely restricted by the lack of tools for tuning gene expression.

Thermophilic Protein Production

Recent studies to improve protein expression systems for *G. thermoglucosidans* have involved more precise, tuneable gene expression. There are two motivations for optimising protein production, firstly for industrial production of thermophilic enzymes that may not be well expressed in a mesophilic host (79,85) and secondly to develop strains better able to degrade lignocellulose for use in CBP or SSF processes (38,86).

Bartosiak-Jentys *et al.* developed a compact plasmid, pUCG3.8 from which a native cellobiose inducible promoter (pβglu) was used to express thermophilic hydrolase enzymes for secretion (86). Thermophilic endoglucanases Cel5A, from *Thermotoga maritima* and CelA, from *Caldicellulosiruptor saccharolyticus* were expressed and secreted from *G. thermoglucosidans* however induction with cellobiose only increased expression of these enzymes 2.5-fold. This may be because the natural transcription factor regulating the promoter was titrated-out with the promoter on a multicopy plasmid or because the cellobiose inducer was being degraded. This study demonstrated the potential of *G. thermoglucosidans* as a host for production of industrially valuable cellulases or as a strain for CBP or SFF where cellulases need to be produced and secreted by cells within the fermentation. However, yields were not high compared to other more established production strains, suggesting that there is room for improvement. From this work, however, pβglu was identified and characterised. This is the first example of an inducible promoter for a *Geobacillus* species but is unfortunately not ideal as background expression is high and the old change upon induction (x 2.5) is comparatively low.

In parallel work, a protein expression system for the closely related *G. kaustophilus* has also been developed. Suzuki *et al*. screened six possible inducible promoters identified from the *G. kaustophilus* genome sequence (79). Five of these gave no expression or expression that did not vary with the addition of possible inducers. However one promoter - pGk704 - was found to be maltose-inducible showing a 12-fold induction from the genome, and a 6-fold induction from a plasmid (pSTE33) when induced. A panel of eight thermophilic enzymes were tested for expression from the plasmid with this promoter, with varying results. Reasonable yields were achieved in two cases: when expressing an amylase, AmyE from *G. stearothermophilus* and a cellulase

PH1171c from *Pyrococcus horikoshii* (a cellulose which cannot be expressed in *E. coli*). Expression of these enzymes allowed the strains to utilise insoluble starch and cellulose respectively (79). This inducible promoter is more promising for further applications with *Geobacillus* species, though fold change on induction is again low compared to inducible promoters used in model organisms.

Recently Suzuki *et al*. have further developed *G. kaustophilus* for an alternative application; for evolving more thermostable proteins (87). Four DNA repair genes were knocked-out from the genome, and this increased the natural mutation rate in these cells by up to 9,000-fold. In this new "mutator strain" the natural *pyrF* gene for uracil production was also knocked out and complemented with a mesophilic *pyrF* gene from *B. subtilis* that is not functional at 65 °C. The strain initially could not grow at 65 °C without uracil however after rounds of subculturing at 60 °C (without uracil), more thermostable mutant versions of the *B. subtilis pyrF* genes were generated (87). The same strain was then later used to evolve more thermostable chloramphenicol (88) and thiostrepton (89) resistance genes.

1.6 The Importance of Tools

"The continued development of tools using the developments in synthetic biology is the surest way to reduce the cost and time required to engineer biological systems, such as those engineered to produce pharmaceutical ingredients, fine and commodity chemicals, and fuels. While the development of biological components might be less 'sexy' than the development of solutions to important problems, those components will enable many solutions, not just the ones for which the components were developed" - Jay Keasling (10)

The synthetic biology approach relies on characterised biological parts – promoters, ribosome binding sites, expression vectors etc. – in order to rationally design and construct genetic circuits and pathways. Lack of parts characterisation and the unpredictable performance of parts under alternative conditions (such as higher temperatures or in novel chassis) present two of the greatest challenges to modern synthetic biology (19). The development of tools to control expression of genes and metabolic pathways has generally lagged behind the development of metabolic pathways and this, in turn, has lead to long development times and high costs for synthetic biology (10).

For chassis strains such as thermophiles the situation is particularly problematic. The majority of previously characterised genetic parts will not function, or their performance will be significantly altered in the new host, largely due to the much greater operating temperature of the host, which in particular affects the folding of proteins, but also because of more typical context-dependencies, such as the ability of the host's ribosome to recognise a heterologous mRNA or the recognition of new promoters by the native RNA polymerase. Developing biotechnology applications in *Geobacillus* species (metabolic engineering, protein expression and evolution of thermostable enzymes) has thus so far relied on a very limited set of tools. These existing tools and genetic parts are summarised in the table below (Table 1.2)

Part	**Notes**	**Reference**
Constitutive promoters		
G. stearothermophilus Ldh, promoter	Strong promoter used to overexpress PDC then PDH for ethanol production in *G. thermoglucosidans*. Later used to test the PheB reported gene	(60,90–92)
G. thermodenitrificans Ldh promoter	Used to express a synthetic operon for Isobutanol production in *G. thermoglucosidans*	(80)
G. kaustophilus SigA promoter	Used to express reporter genes integrated into the genome of *G. kaustophilus*	(93)
Inducible promoters		
G. thermoglucosidans pβglu, cellobiose inducible promoter	Apparently strong induction when expressing the PheB reporter but only a 2.5x increase in expression of cellulase enzymes when induced	(86)
G. kaustophilus pGK704, maltose inducible promoter	12x increase in expression of β-Gal reporter when induced on the genome, 6x when on a plasmid (pSTE33)	(79)
Modern Shuttle vectors		
pUCG3.8	Used for protein expression in *G. thermoglucosidans*. More compact version of pUCG18 (90), which was made from pUB90 (60). Kanamycin resistance marker.	(86)
pSTE33	Developed for *G. thermodenitrificans*. Also used for conjugative transfer (94) and protein expression (79)in *G. kaustophilus*. Kanamycin resistance marker	(83)
pNW33N	Used to transform *B. stearothermophilus* (72) and for isobutanol production in *G. thermoglucosidans* (80). Chloramphenicol resistance marker.	(95)

Table 1.2. Promoters and shuttle vectors previously used and published for genetic engineering applications with *Geobacillus* species

Compared to established chassis organisms such as *E. coli* and *B. subtilis* the range of previously published parts for *Geobacillus* species is incredibly limited. Established strains have a huge range of promoters to select from, many of which have characterised libraries of different strength variants in order to fine tune transcriptional strength. Alternatively, expression can tightly controlled or varied over a wide range with strongly regulated inducible promoters. These established parts have been characterised with reporter proteins and in various applications and so they can be predictably used in future designs. For building genetic constructs a range of vector backbones are then available with good characterisation data, known copy number and well-optimised transformation protocols. The lack of such genetic parts for *Geobacillus* species greatly restricts the development of applications with these species.

Aims of this study

Given the huge potential of *Geobacillus* species for production of biobased chemicals from cheap, abundant lignocellulosic feedstock, developing tools to enable synthetic biology with this organism to enable novel production strains to be engineered would be hugely valuable.

The ultimate aim of this project was to improve the parts and tools available for synthetic biology in the thermophile *G. thermoglucosidans* and to test the potential of these tools and this microbial chassis for the production a complex higher-value product. To achieve this goal the work of the thesis was broken-down into 4 specific foundational goals and a final application goal. These were as follows:

1. Test and characterise fluorescent reporter genes for use in *G. thermoglucosidans.*

Reporter proteins are vital in synthetic biology to characterise genetic parts for gene expression. Fluorescent reporters are preferred as they allow simple, non-destructive measurement and cell populations can be assessed by flow cytometry. Fluorescent proteins that function at a range of temperatures and oxygen conditions in *G. thermoglucosidans* would allow more complete characterisation of genetic parts than previously possible.

2. Generate and characterise promoter libraries for fine-tuning gene expression in this chassis.

Previous metabolic engineering with *G. thermoglucosidans* involved simple overexpression of pathway genes with a strong promoter, however, optimising yields and the production of more complex products demands tunable gene expression to balance metabolic flux. Promoter libraries are the best tools to achieve this and libraries with a wide expression range and good characterisation could be predictably reused for many applications.

3. Assess how expression can be tuned by varying translation rates with ribosome binding site design and test the applicability of existing software in design of ribosome binding site sequences for *G. thermoglucosidans.*

Tuning the translation initiation rate gives greater control over gene expression and tools exist to rationally design sequence to vary this rate in *E. coli* and related bacteria. However, these have not yet been tested for thermophiles. Predictable design via such sequence-to-output calculators would allow better optimisation with reduced *in vivo* testing.

4. Construct and characterise minimal modular shuttle vectors.

Previous *Geobacillus* plasmid vectors were generally large and poorly characterised. A compact vector set with interchangeable modules and variable copy number would be a flexible platform from which to characterise genetic parts and build constructs for future applications.

5. Using these parts and tools, design and test a strategy for the production of the high-value metabolite, in this case hyaluronic acid.

Production of a more complex product would demonstrate the utility of the tools developed in this study and the potential of this host for production of a wide variety of molecules. The valuable biopolymer hyaluronic acid, presents an excellent opportunity as a test-case in this regard.

Chapter 2: Materials And Methods

2.1 Strains, Plasmids

2.1.1 Bacterial Strains Used in This Study

Strain	Description	Source	Reference
E. coli DH10B	*DH10BF endA1 recA1 galE15 galK16 nupG rpsL ΔlacX74 F80lacZDM15 araD139, Δ(ara,leu)7697 mcrA Δ(mrr-hsdRMS-mcrBC) l*	New England Biolabs	NEB Catalogue
G. thermoglucosidans DL33	Novel isolate from the UK	David Leak Lab, University of Bath	Unpublished
G. thermoglucosidans DL44	Δ*ldh* variant of *G. thermoglucosidans* DL33	David Leak Lab, University of Bath	Cripps et al. 2009 (60)
G. thermoglucosidans TM89	Δ*ldh* variant of *G. thermoglucosidans* NCIMB 11955 (type strain)	TMO Renewables	Cripps et al. 2009 (60)
G. kaustophilus CER5420	Isolate from the Centre for Extremophile Research collection in Bath	David Leak Lab, University of Bath	Unpublished
G. thermodenitrificans K1041	WT isolate	David Leak Lab, University of Bath	Narumi et al. 1992 (96)
G. thermoleovorans DSM14791	WT isolate from sugar refinery wastewater	TMO renewables	Tai et al. 2004 (97)

Table 2.1. *E. coli* and *Geobacillus* species strains used in this study.

For *G. thermoglucosidans* work in this study, strain DL44 was used unless otherwise stated as this strain had the highest electroporation efficiency. When taking sequence from the genome, the sequence of *G. thermoglucosidans* C56-YS93 (70,98) was used as the complete sequence of strains DL33 or NCIMB 11955 has not been published.

2.1.2 Plasmid Backbones Used in This Study

Name	Description	Source	Reference
pUCG18	Amp^R, pUC18 ori, Kan^R, repBST1	David Leak Lab, Imperial College	Taylor et al. 2008 (90)
pUCG16	Amp^R, pUC18 ori, Kan^R, repBST1	David Leak Lab, Imperial College	(unpublished)
pUCG3.8	Amp^R, pUC18 ori, Kan^R, repBST1	David Leak Lab, Imperial College	(86)
pG1AK	Kan^R, Amp^R, pUC18 ori, repBSTI	This study	
pG1K	Kan^R, pUC18 ori, repBSTI	This study	
pG2K	Kan^R, pUC18 ori, repB	This study	
pG1C	Cam^R, pUC18 ori, repBSTI	This study	

Table 2.2. Plasmid shuttle vector backbones used or produced in this study.

2.2 Microbiology Methods

2.2.1 Standard Reagents

All reagent-grade chemicals were purchased from Sigma Aldrich or Fisher Scientific. Yeast extract, tryptone and soy peptone were purchased from Merck.

2.2.2 Bacterial Growth Conditions

***E. coli* media**	**Ingredients per litre of media**
LB	Tryptone 10 g, Yeast extract 5 g, NaCl 5 g, adjusted to pH7 with HCl or NaOH
LB, 2% glucose	Tryptone 10 g, Yeast extract 5 g, NaCl 5 g, Glucose 20 g, adjusted to pH7 with HCl or NaOH

Table 2.4 Media used for growth of *E. coli*

***Geobacillus* species media**	**Ingredients per litre of media**
TGP	Tryptone 17.0 g, Soy Peptone 3.0 g, NaCl 5.0 g, K_2HPO4 2.5 g, Sodium Pyruvate 4.0 g, Glycerol 4.0 ml, adjusted to pH7 with 3 M NaOH
TGP/glucose	Tryptone 17.0 g, Soy Peptone 3.0 g, NaCl 5.0 g, K_2HPO_4 2.5 g, sodium Pyruvate 4.0 g, glucose 4.0 g, adjusted to pH 7 with 3 M NaOH
BCM	Tryptone 17 g, Soy peptone 3 g, Yeast extract 5 g, NaCl 5 g, Glucose 10 g, HEPES 2 mM, adjusted to pH 7 with 5M NaOH

2TY	Tryptone 16 g, Yeast Extract 10 g, Sodium Chloride 5 g, adjusted to pH 7 with 5 M NaOH
2SPYNG	Soy Peptone 16 g, Yeast Extract 10 g, Sodium Chloride 5 g, adjusted to pH 7 with 5 M NaOH
LB	Tryptone 10 g, Yeast extract 5 g, NaCl 5 g, adjusted to pH 7 with HCl or NaOH
LB, 2% glucose	Tryptone 10 g, Yeast extract 5 g, NaCl 5 g, Glucose 20 g, adjusted to pH 7 with HCl or NaOH
LB/MW 2% glucose	Tryptone 10 g, Yeast extract 5 g, NaCl 5 g, Glucose 20 g, made up to 1 L with spring water, adjusted to pH 7 with HCl or NaOH
LB/MW	Tryptone 10 g, Yeast extract 5 g, NaCl 5 g, made up to 1 litre with spring water, adjusted to pH 7 with HCl or NaOH
Ammonium Sulphates Medium (ASM) + 1% glucose	8 mM Citric acid, 5 mM $MgSO_4$, 20 mM NaH_2PO_4, 10mM K_2SO_4, 25 mM $(NH_4)_2SO_4$, 80 µM $CaCl_2$, 1.65 µM Na_2MoO_4, 12µM Biotin, 1% w/v Glucose, Trace element solution 5ml/l Trace element solution: $ZnSO_4.7H_2O$ 1.44 g/l, $CoSO_4.6H_2O$ 0.56 g/l, $CuSO_4.5H_2O$ 0.25 g/l, $FeSO_4.6H_2O$ 5.56 g/l, $NiSO_4.6H_2O$ 0.89 g/l, $MnSO_4$ 1.69 g/l, H_3BO_3 0.08 g/l, 12M H_2SO_4 5.0 ml/l

Table 2.3 Media used for growth of *Geobacillus* species

Geobacilli were typically grown in 2SPYNG media at 55 °C with shaking at 200 rpm. Aerobic growth of Geobacilli was with 5 ml of culture in a 50 ml tube or 50 ml of culture in a 500 ml baffled flask. Microaerophilic growth was with 15 ml of culture in a filled, tightly sealed 15 ml tube. When grown on solid media, agar plates were wrapped in tin foil to reduce drying. *E. coli* growth was at 37 °C with 5 ml of media in a 15 ml tube shaking at 200 rpm.

2.2.3 Sterilization

All media and containers were sterilized before use by autoclaving at 121 ºC/103 Mpa for 15 minutes. Heat-sensitive ingredients were dissolved and filtered sterilized through a Minisart 0.2 µm syringe filter (Sartorius).

2.2.4 Antibiotic Selection

For selection based on antibiotic resistance in *E. coli*, ampicillin was used at a concentration of 100 µg/ml, kanamycin at 50 µg/ml and chloramphenicol at 12 µg/ml. For selection in *Geobacillus* species, kanamycin and chloramphenicol were used at concentrations of 12 µg/ml. Antibiotics were added to media from 1000x concentrated

stocks with ampicillin and chloramphenicol suspended in 100% ethanol and kanamycin in sterile water.

2.2.5 Storage of Bacteria

Bacterial strains on solid media were routinely stored at 4 ºC for up to 4 weeks. For longer term storage, glycerol stocks were prepared by mixing 750 µL of 50% glycerol with 750 µL of liquid culture in a microcentrifuge tube (Eppendorf UK). Stocks were stored indefinitely at -80 ºC.

2.2.6 Preparation of Chemically Competent *E. coli*

Chemically competent *E. coli* DH10B cells were prepared from glycerol stocks. 5 ml of LB media was inoculated and grown overnight at 37 ºC with shaking at 200 rpm. 500 µl of this culture was used to inoculate 50 ml of LB media in a 250 mL baffled conical flask, which was incubated at 37 ºC in a rotary shaker for 2-3 h until it reached an OD600 of 0.5. Optical density was measured using a Jenway Genova-plus spectrophometer (Bibby Scientific). The culture was cooled on ice for 10 min and the cells were harvested by centrifugation (4ºC, 4000 rpm, 5 min). The cells were resuspended in 5 ml of chilled CCMB80 buffer (80 mM $CaCl_2$, 20 mM $MnCl_2$, 10 mM $MgCl_2$, 10 mM KOac, 10% v/v glycerol) and divided into 200 µl aliquots that were stored at -80 ºC.

2.2.7 Transformation of Chemically Competent *E. coli*

10-300 ng of purified plasmid DNA was mixed with a 50 µl aliquot of chemically competent *E. coli* cells. The vial was allowed to rest on ice for 30 min, after which the cells were incubated at 42 ºC for 30 seconds and then placed back on ice for 2 minutes. Following this, the cells were diluted with 320 µl of SOC media (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose) and allowed to recover for 1 hour at 37 ºC, shaking at 200 rpm after which they were plated on an LB agar plate containing the appropriate antibiotic.

2.2.8 Preparation of Electrocompetent *Geobacillus* Strains

Electrocompetent Geobacilli were prepared and transformed a with a slightly modified protocol from that described in Taylor et al. 2008 (90). The strain was grown overnight at 55 °C on a pre-warmed 2SPYNG agar plate from which, 50 ml of pre-warmed 2SPYNG media in a 500 ml baffled flask was inoculated. The flask was incubated at 55 °C for until the OD600 reached 1.6 (approximately 3-6 hours). Cells were cooled on ice for 10 minutes and harvested by centrifugation at 4000 rpm at 4 °C for 15 minutes. The cell pellet was re-suspended in 50ml of ice cold, sterile, electroporation buffer (0.5 M sorbitol, 0.5 M mannitol and 10% v/v glycerol) and pelleted again by centrifugation. Cells were subsequently washed in 25 ml and 2x 10 ml ice-cold electroporation buffer, with centrifugation as above between each wash. The final cell pellet was re-suspended in 0.5 ml ice cold electroporation buffer and stored as 60 µl aliquots at -80 °C.

2.2.9 Electroporation of *Geobacillus* Strains

A 60 µl aliquot of electrocompetent cells was thawed on ice, mixed with ~100 ng of plasmid DNA and incubated on ice for 10 minutes. The mixture was then transferred to an ice-cold 1 mm gap electroporation cuvette (BioRad). A single exponential pulse (2500 V from a 10 µF capacitor with 600 Ω in parallel) was applied using an Xcell gene pulser (BioRad). Immediately after pulsing, 500 µl of room temperature 2SPYNG media was added to the cuvette and the contents transferred to a 2 ml tube. The cells were recovered at 55 °C for 1.5 hours with shaking at 200 rpm and then quickly plated on pre-warmed (55 °C) 2SPYNG agar plates containing appropriate selective antibiotic.

2.2.10 Mineral Nanofiber Transformation

This method was adapted from Tan *et al.* 2010 (99). A culture of *G. thermoglucosidans* was grown in 2SPYNG in a baffled flask with shaking at 55 °C to an OD600 of 1.5. This was divided into 2 ml aliquots of culture each in 2 ml microcentrifuge tubes and spun in a benchtop microcentrifuge (1 minute at >4000 g) at room temperature. The media was removed and the pelleted cells resuspended in 100 µl 2SPYNG with 0.1%

to 1% w/v sepiolite mineral (Kremer Pigmente GmbH & Co. KG (Hauptstr, Aichstetten, Germany) powdered and autoclave sterilised. 500 ng pUCG16 plasmid DNA (in triplicates plus negative controls without DNA) was then added and tubes were vortex mixed at full speed for 2 minutes. 400 µl of fresh, room temperature 2SPYNG was then added to each tube and cells were recovered at 55 °C with shaking at 200 rpm for 1.5 hours. After recovery, cells were plated on warm 2SPYNG agar plates with 12 µg/ml kanamycin and incubated for 24 hours at 55 °C.

2.2.11 Plasmid Purification

Plasmids were purified from *E. coli* and *G. thermoglucosidans* using the QIAprep Spin Miniprep Kit (Qiagen) following the manufacturer's protocol. For *Geobacillus* species, lysozyme was added to P1 buffer at 10 mg/mL and when resuspended in P1/lysozyme cells were and incubated at 37 ºC for 30 minutes with shaking at 200 rpm. A Nanodrop microvolume spectrophotometer (Thermo Fisher Scientific) was used for DNA quantification. The concentration of DNA was determined from the absorbance at 260 nm and protein contamination was determined by the ratio of the absorbance values 260:280 nm. Preparations of plasmid DNA eluted in deionised autoclaved water or elution buffer (10 mM Tris-Cl, pH 8.5) were routinely stored at -20 °C. For electroporation into *Geobacillus* species plasmid preparation were desalted with 0.25 µm pore size nitrocellulose microdialysis filter disks (Millipore). The filters were floated on the surface of a plate filled with distilled water and the droplet of eluted plasmid was applied to the filter and left for 30 minutes for salts to diffuse out before being collected into a fresh microcentrifuge tube.

2.2.12 Chelex Genomic DNA Preparation

2 ml of overnight *Geobacillus* species culture was transferred to a 2 ml microcentrifuge tube and the cells pelleted by centrifugation (8,000 rpm for 1 minute). Supernatant was removed and cells resuspended in 300 µl 5% w/v solution of Chelex-100 resin in in distilled water and vortex mixed for 1 min. Tubes were then boiled at 100 °C for 15 minutes to lyse cells. Samples were centrifuged at 8,000 rpm for 10 minutes and the

supernatant transferred to a fresh tube. This supernatant was suitable for use as template for PCR reactions to amplify genomic sequence.

2.3 Molecular Biology Methods

2.3.1 Polymerase Chain Reaction

All oligonucleotide primers were purchased from IDT. PCR reactions, with a final volume of 50µl, were carried out in sterile 0.2ml PCR tubes in a G-Storm thermal cycler. Phusion polymerase and buffers were used (NEB), a typical reaction mixture is shown. Reactions were made up following manufacturer's guidelines.

Component	Volume for 50µl reaction	Final concentration
Sterile deionised H_2O	To 50 µl	-
HF buffer	10 µl	1x
10 mM dNTPs	1 µl	200 µM each
Primer A (10 µM)	1 µl	0.5 µM
Primer B (10 µM)	1 µl	0.5 µM
DMSO	1 µl	2% v/v
Phusion polymerase	0.25 µl	0.02 U/µl
Template DNA	0.5-2 µl	~20 ng final

Table 2.4. A typical PCR mixture

Touchdown PCR was performed to negate the need for optimisation of annealing temperature. The reaction was typically carried out with an initial denaturation for 1 minute at 98 °C followed by 35 cycles of denaturation at 98 °C for 10 seconds, annealing at a temperature reducing from 65 °C to 55 °C for 20 seconds and elongation at 72 °C for 20 seconds per kilobase. A final extension at 72 °C for 5 minutes completed the reaction.

2.3.2 Quantitative Real-Time PCR

Plasmid copy number per chromosome estimates were determined by quantitative real time PCR as described by Lee *et al.* 2006 (100) and Skulj et al 2008 (101). Primers were designed to amplify short amplicons from the plasmid and *G. thermoglucosidans* genomic DNA. Primers KanR F and R amplified a 182 bp aplicon from the kanamycin resistance marker on the plasmid and primers SigA F and R amplified a 168 bp

amplicon from the SigA gene on the *G. thermoglucosidans* genome (sequences in Table 2.5 below). The *G. thermoglucosidans* C56-SY93 genome (98) was used for sequence information.

Name	Sequence
KanR forward	ggtgtttatggctctcttgg
KanR reverse	tctgattccacctgagatgc
SigA forward	ttgaagaccaagaagcgacg
Sig A reverse	cttttccgacttcttcgagc

Table 2.5 Primers used for quantitative PCR

For the reactions, Kapa SYBR FAST qPCR mix (Kapa biosystems) was used according to manufacturer's instructions. An Eppendorf Mastercycler Realplex qPCR machine was used with the following PCR program: 95 °C for 15 s then 40 cycles of 95 °C for 2 s, 60 °C for 20 s, 72 °C for 30 s. Cycle threshold (Ct) values were calculated automatically by the Realplex software. Amplification efficiency of the primers was calculated from five 10-fold serial dilutions of plasmid and genomic DNA preparations. A relative standard curve was constructed placing the log value of the relative amount of of DNA (determined according to dilution) on the x axis and threshold cycles on the y axis. The slope of the line fitted to this graph is then used to calculate efficiency (E) according to:

$$E = 10^{(-1/slope)}$$

Samples for plasmid copy number estimation were prepared as described by Skulj et al 2008 (101). Culture media was boiled at 100 °C for 15 minutes then frozen at -20 °C and thawed before use. These samples were then diluted 1000x in distilled water and used as template for the PCR reactions. This preparation avoids any bias in chromosomal vs. plasmid DNA (101). From the calculated Ct values, technical triplicates were averaged and plasmid copy number for each culture sample was estimated based on the equation:

$$\text{Plasmid Copy number} = (Ec^{Ctc})/(Ep^{Ctp})$$

Where Ec and Ctc are the amplification efficiency and cycle threshold for the amplification from the chromosome and Ep and Ctp are the amplification efficiency and cycle threshold for the amplification from the plasmid.

2.3.2 Mutagenic PCR

The reaction comprised 10x standard Taq Mg- free reaction buffer (NEB), 50 mM $MgCl_2$, 200 µM of dPTP and 8-oxo-dGTP, 1 mg/ml gelatine and 5 U/µl Taq polymerase. 8-Oxo-dGTP can mispair with adenine, leading to A-to-C and G-to-T transversion mutations. dPTP in combination with 8-Oxo-dGTP can cause both transition mutations (A-to-G and G-to-A) and transversion mutations (A-to-C and G-to-T). PCR settings were: 98 °C for 2 minutes, then cycles of 98 °C for 1 minute (denaturation), 55 °C for 1.5 minutes (annealing), 72 °C for 5 minutes (elongation), for 20 cycles giving a mutation rate of approximately 10% and finally 5 minutes of extension at 72 °C. The mixes were then treated with 0.5 µl DpnI (NEB) at 37 °C for 1 hour to digest template DNA and used as template for a further PCR amplication with phusion polymerase to increase concentration and add overlap sequences for Gibson assembly into the backbone vector.

2.3.3 Site directed mutagenesis

To mutate a single basepair in CatE, phosphorylated primers were ordered (IDT) including the mutated base on the end of one primer. The whole template plasmid backbone was amplified was amplified via PCR with the phosphorylated primers then template DNA was digested with DpnI restriction enzyme (NEB UK). The PCR product was self-ligated using T4 ligase (NEB UK) according to manufacturers' instructions and transformed into *E. coli* with appropriate antibiotic selection.

2.3.4 Gibson Assembly

Gibson Assembly is a DNA ligation technique developed at the JCVI by Dan Gibson et al. in 2009. It uses three enzymes to ligate two or more sequences of DNA that have overlapping end sequences at the joining point. These overlapping regions can be added to the ends of DNA fragments by PCR.

Typically 40 bp oligos (20 bp annealing and 20 bp to create overlap) were used to amplify fragments by PCR before a Gibson assembly reaction. For each reaction 5µl total of the parts to be joined was added to 15 µl of 1.33x master mix (preparation below). The mix was then incubated at 50 °C for one hour, cooled and then 1 µl was mixed with competent cells for transformation.

1.33x Gibson assembly mix	Volume/µl
Taq ligase (40 u/µl)	50
5x isothermal buffer	100
T5 exonuclease (1 u/µl)	2
Phusion polymerase (2 u/µl)	6.25
Nuclease-free water	216.75
Total:	375

Table 2.6. Preparation of Gibson assembly master mix

2.3.5 Restriction/Ligation Cloning

Restriction enzymes were purchased from New England Biolabs (NEB UK). Digests were carried out according to manufacturer's instructions. DNA was
purified through 1 % agarose gel electrophoresis and gel-extracted using a Qiagen Gel Purification kit (Qiagen). The ligation reactions were carried out with T4 ligase (NEB) according to manufacturer's instructions. Before transformation the ligase was denatured at 65 ºC for 15 minutes.

2.3.6 Agarose Gel Electophoresis

Agarose gel electrophoresis was used to visualize and purify DNA restriction fragments and PCR products. 1% w/v agarose in 1x TAE buffer gels were used. Agarose was dissolved in 1xTAE buffer (50x TAE Buffer; 242.0 g/l Tris Base, 57.1 ml/l glacial acetic acid, 100.0 ml/l 0.5 M EDTA, pH 8.0) by microwaving. After cooling to ~50 °C and pouring, SYBR safe gel stain was added (1 µl/ml) was added.

5x gel loading buffer (30% v/v glycerol, 0.25% bromophenol blue) was added to samples before loading onto the gel along with 2 µl of molecular weight marker (2-log DNA ladder, NEB). Gels were run in BioRad Gel Electrophoresis tanks with 110 V

power supply from a BioRad powerpack. DNA fragments were visualised on a short wave UV transilluminator and photographed on a UV transilluminator BioDoc-it system with an attached analogue thermal printer (UVP). When gel purifying required, a blue light transilluminator (Invitrogen) was used instead and bands were excised with a scalpel. DNA was purified from the gel slice gel using a gel extraction kit (Qiagen) according to manufacturer's instructions.

2.3.7 DNA Sequencing

Sequencing of DNA was carried out externally by the Source Bioscience Sanger sequencing service.

2.4 Synthetic Biology Methods

2.4.1 Promoter and Ribosome Binding Site Characterisation

Parts were cloned into plasmid pUCG16 with the sfGFP reporter. For *G. thermoglucosidans* cultures were grown from single colonies in 5 ml of 2SPYNG media in 50 ml tubes at 55 °C overnight with shaking at 200 rpm. Cultures were then diluted 100x into fresh media and grown to stationary phase (maximum OD600). For each tube, a 200 µl aliquot of culture was added to a clear, flat bottom 96-well plate (Corning Life Sciences) and GFP fluorescence plus OD600 measurements were made with a BioTek Synergy HT plate reader (BioTek). For *E. coli*, LB media at 37 °C was used and cells were grown in 200 µl of media in 96-well plates. After subtracting for media autofluorescence GFP readings were divided by OD600 readings to give an estimate of GFP fluorescence per cell. Alternatively, fluorescence was measured by flow cytometry.

2.4.2 Flow Cytometry

1 µl of bacterial culture was diluted into 1 ml of water and analysed on a modified Becton-Dickinson FACScan flow cytometer. Samples were run on high flow rate until

30000 events were observed or 30 seconds had elapsed. Flow cytometry settings were FSC sensor E01, SSC voltage 350, SSC threshold 52 and FL1/FL5 voltage 700.
For sfGFP, excitation was with a yellow/green laser (488 nm) and detection via filter Fl-1, (530 nm). For mCherry, excitation was with a yellow/green laser (561 nm) and detection via filter Fl-5, (610 nm). Data was analyzed using FlowJo with a forward scatter/side scatter gate applied to select for a homogeneous population size.

2.4.3 Fluorescence Microscopy

Cells were diluted harvested and resuspended in water with a 5 µl sample spread onto a microscopy slide. A Nikon Eclipse Ti inverted microscope with the 60x CPI60 objective was used to image the cells. The the GFP Green channel was used for sfGFP with wavelengths 480 nm excitation and 535 nm emission and the Cy3 (Red) channel was used for mCherry with 532 nm excitation and 590 nm emission. The images were viewed using the software NIS-Elements Microscope Imaging Software (Nikon).

2.4.4 Hyaluronic Acid Extraction

Hyaluronic acid (HA) was extracted and purified from microbial culture with a protocol adapted from Yu *et al.* 2008 (103). Culture media samples were diluted with an equal volume of 0.1% w/v sodium-dodecyl-sulfate (SDS) and incubated at room temperature for 10 minutes to solubilise the HA and separate it from other exopolysaccarides. Samples were then filtered through a 0.2 µm syringe filter to remove the cells. 2 volumes of chilled ethanol was then added and samples were incubated at 4 °C for 1 hour to precipitate HA. Samples were centrifuged to pellet the HA and the pellet was washed washed twice by resuspending and in 2 volumes of 70% ethanol centrifuge concentrating. Pellets were then redissolved in acetate buffer (described below) for quantification.

2.4.5 Hyaluronic Acid Quantification

HA concentration was quantified using the turbidity assay as described in Song *et al.* 2009 (104). Acetate buffer and CTAB solution were prepared as described below.

Acetate buffer: 0.2 M sodium acetate-acetic acid, 0.15 M NaCl, pH 6.0.
0.2 M Na acetate solution (~pH 9) should first be made and 0.2 M acetic acid added until the pH reaches 6.0. Solid NaCl to 0.15 M can then be added.
CTAB solution: 2.5% CTAB, 2% NaOH
2% NaOH should first be made and cetyltrimethylammonium bromide reagent (CTAB) (Sigma-Aldrich) added to this. CTAB solution was stored at 37 °C.

HA samples and standards in acetate buffer were incubated at 37 °C for 10 minutes with shaking at 200 rpm to fully dissolve the HA. A clear, flat-bottomed 96-well microplate (Corning Life Sciences) and the CTAB solution were also prewarmed to 37 °C. 100 µl of each sample was added to the plate then 100 µl of the CTAB solution to precipitate the HA. OD600 readings were taken in a microplate reader (BioTek Synergy HT) at 37 °C after shaking at 200 rpm for 30 seconds to mix the samples. OD600 values of samples extracted from cultures were compared to values from diluted streptococcal HA standards to determine HA concentration.

Chapter 3: Working with Geobacilli, Protocols and Reporter Genes

Summary

Many different protocols exist for growing and transforming *Geobacillus* species in the laboratory. Possible transformation methods for *G. thermoglucosidans* were reviewed and electroporation and mineral nanofibre transformation were tested. Electroporation was found to be reasonably efficient as previously reported whereas the mineral nanofibre method was not found to be effective. Efficient conjugation has very recently been reported for *G. thermoglucosidans* (105) and so this is a promising method for future work. For synthetic biology, reporter genes are vital to give biological parts and systems an output for characterisation. Superfolder Green Fluorescent Protein was found to be stable in *G. thermoglucosidans* grown under aerobic conditions; fluorescence could not be recovered from anaerobic cultures however. An alternative red fluorescent protein, mCherry was found to be less thermostable but functional at 50 °C. Three possible anaerobic fluorescent protein variants were tested but not found to be functionally expressed in *G. thermoglucosidans*.

Chapter Aims

- To test growth of *G. thermoglucosidans* in common rich and minimal media preparations
- Review current bacterial transformation methods and test appropriate protocols with *G. thermoglucididans.*
- Test thermostability of fluorescent proteins *in vitro* and *in vivo*
- Search for and test possible anaerobic fluorescent reporters in *G. thermoglucosidans* to report protein expression under low oxygen conditions.

3.1 Introduction

3.1.1 Strain and Media Choice

A range of *Geobacillus* species and strains have been researched and tested for several applications. Of these, *G. thermoglucosidans* and *G. kaustophilus* have the best-characterised genetic parts available. *G. thermoglucosidans* species have been shown to be most suitable for metabolic engineering applications (60,66,80) and are more versatile as facultative anaerobes whereas *G. kaustophilus* species are strictly aerobic. Strains based on *G. thermoglucosidans* DL33 and DL44 (Materials and Methods 2.1.1) have been used in several previous studies that this work builds upon (60,92,106) and so were chosen for this work. Strain DL33 is a wild-type isolate from Britsh soil and strain DL44 is a lactate dehydrogenase knock-out strain of DL33. Under anaerobic conditions DL44 does not produce lactate and so metabolic flux can be redirected to other products making this a particularly useful strain for metabolic engineering (60,107). All experiments with *G. thermoglucosidans* in this study used strain DL44 unless stated otherwise.

Simple yeast extract plus tryptone/peptone media has been most commonly used for for *Geobacillus* species growth in the laboratory. Several media preparations were tested (Materials and Methods 2.2.2) and the simplest, 2SPYNG media (alternatively called 2-SPY) was chosen for all *Geobacillus* species growth in this study unless stated otherwise. Similarly, LB media was used as standard for *Escherichia coli.*

3.1.2 Transformation

The first successful transformation of a *Geobacillus* species was protoplast transformation of *G. stearothermophilus* (then *Bacillus stearothermophilus*) in 1982 by Imanaka et al. (81). The process is theoretically very efficient (up to 10^5 transformants per µg plasmid DNA) but even with later refinements (108) is very laborious, requiring protoplast generation with lysozyme before a polyethylene glycol (PEG) induced transformation step and recovery in liquid media, then two further recovery steps (at different temperatures) on solid media.

Conjugation into this species from an *E. coli* donor was then demonstrated in 1991 (109) with a modified conjugative transposon. Conjugation is a promising method for transferring plasmids to thermophiles as the protocol is comparatively simple once a suitable mesophilic donor strain is known. A rise in temperature simultaneously recovers the recipient thermophile whilst removing the mesophilic donor. This particular method did not take off for genetic modification however, as the transposon is large (18 kb), complex, and does not integrate in a particularly site specific or stable manner (110). Significant modifications would be required, such a making the integrase expression inducible, for this be a viable tool. Conjugation was later revived however in *G. kaustophilus* with conjugative plasmids and is a valuable technique here but only with specific, modified donor *E. coli* strains (94).

Electroporation is a standard transformation method for many non-model bacterial chassis including other thermophilic gram positives such as *Clostridia* and *Thermoanaerobacter* species (111,112). It is also the most popular protocol in published *Geobacillus* species studies so far. The method has room for improvement however, key drawbacks include the need for careful preparation of cells, buffers and DNA to keep them ion-free (for low conductivity) and specialised equipment - electroporators and cuvettes are comparatively expensive. Electroporation also relies on DNA entering the cells through temporary pores produced by the electrical shock and so there is an upper size limit to the plasmids which can pass through (113).

Strain	Method	Plasmid	Efficiency in cfu/µg plasmid DNA, unless otherwise stated.	Reference
G. stearothermophilus				
ATCC 12980	Protoplast transformation	pUB110	5.9×10^5	Imanaka et al. (1982) (81)
		pTB90	1.3×10^7	
		pTHT15	2.5×10^{-4} transformants/regenerant 6.1×10^{-2} transformants/regenerant	Hoshino et al. (1985) (114)
NRLL 1174	Protoplast transformation	pBST22	3×10^5	Liao and Kanikula (1990) (82)
BR219 (DSMZ 6285)	Conjugative transfer of a transposon	pAM120	2.6×10^{-7} /recipient	Natarajan and Oriel (1991) (109)
NUB36	Protoplast transformation	pTHT15	4×10^8	Wu and Welker (1989) (108)
		pLW05	2×10^7	

Strain	Method	Plasmid	Efficiency in cfu/µg plasmid DNA, unless otherwise stated.	Reference
		pRP9	6×10^5	De Rossi et al. (1994) (96)
	Electroporation	pUCG18	1.4×10^2	Kananavičiūtė et al. (2014) (115)
G. thermodenitrificans				
K1041	Electroporation	pUB110	5.8×10^5	Narumi et al. (1992) (96)
		pIH41	7.2×10^4	
		pSTE12	5.1×10^4	Nakayama et al. (1992) (116)
		pSTE33	2.8×10^6	Narumi et al. (1993) (83)
G. kaustophilus				
HTA426	Conjugative transfer	pUCG18 carrying ***oriT***	10^{-3} /recipient	Suzuki and Yoshida (2012) (94)
		pSTE33 carrying ***oriT***	10^{-6} /recipient	
G. thermoglucosidans				
DL33	Electroporation	pBST22	3.9×10^3	Taylor et al. (2008) (90)
		pUCG18	9.8×10^3	
NCIMB 11955	Electroporation	pUCG3.8	2.8×10^5	Bartosiak-Jentys et al. (2013) (86)

Table 3.1. Transformation procedures in *Geobacillus* species so far, adapted from Kananavičiūtė *et al.* 2014 (115)**.**

Whilst the three methods listed in Table 3.1 (protoplast transformation, conjugation and electroporation) are the only reported successes in transforming Geobacilli, a great variety of alternative procedures have been successful in other bacteria. All possible procedures were reviewed here for suitability with *Geobacillus* species. Criteria considered included the simplicity of the protocol, speed of the protocol and current or possible efficiencies with *Geobacillus* species.

Chemical Transformation

Once optimised, methods for chemotransformation in Gram-negative bacteria can be highly efficient with over 10^9 cfu/µg not uncommon for plasmids in *E. coli*. The protocols are particularly difficult to optimise however with many parameters to consider. Particular combinations of cations and other ingredients can have

unpredictable effects on efficiency and optimum combinations may be highly species or strain specific depending on the particular features of the plasma membrane and cell wall (117,118)

Gram positives are often less amenable to chemotransformation, however *B. subtilis*, arguably the model Gram-positive organism, happens to have very high natural competency which can be additionally improved by certain chemicals or media conditions (119). This is sadly not the case for Geobacillus species and all current protocols for chemotransformation in thermophiliic gram positives require protoplast generation (Alex Pudney, TMO Renewables Ltd. personal communication). Developing a simplified protocol would demand a huge amount of optimisation and may then only be applicable to one particular strain.

Modified "super-competent" strains have been produced of other chassis organisms by upregulating competence genes (120–122). In *B. subtilis* the regulatory protein *comS* and transcription factor *comK* promote competence and competence is increased when they are upregulated (122). *G. thermoglucosidans* has homologues of these genes that would be primary targets for overexpressed to produce a super-competent laboratory strain. Such a strain might gain natural competence or have increased chemical transformation or electroporation efficiency. However, *comK* is a high-level transcription factor with many targets: constitutive or leaky inducible expression could cause undesirable effects on the health and metabolism of the cells. Due to the lack of apparent natural competence in *Geobacillus* species and the complexity of inducing it or optimising chemical transformation otherwise, this method was not pursued in this chapter.

Sonoporation

Bacterial transformation mediated by ultrasound was reported in 2007 by Song *et al*. (123) and is a hugely promising technique. In aqueous solutions low power ultrasound forms cavitation bubbles that create transient pores in the cell wall enabling translocation of molecules into the cell interior. The protocol is fast, simple and high efficiency (~10^7 cfu/μg) for the mesophilic, Gram-negative bacteria (*E. coli, P. putida,*

and *P. fluorescens*) tested. Thermophilic, Gram-positive bacteria are significantly more recalcitrant to transformation (115) however. Lin *et al*. 2010 reported successful sonoporation in such a Gram-positive thermophile, *Thermoanaerobacter* sp. X514 but despite significant optimisation of the protocol varying multiple parameters, efficiency remained comparatively low, around 6 x 10^2 cfu/µg DNA. The main advantages for *Thermoanaerobacter* and other anaerobes is the simple protocol requiring minimal steps risking aerobic exposure. Sonoporation can take place with the bacteria in a sealed tube of growth media rather than requiring transfer to an electroporation cuvette or chemical transformation buffer. For *Geobacillus* species this is not a concern and with efficiency apparently limited this technique was therefore not pursued here.

Mineral Nanofibre Transformation

Mineral nanofibre mediated transformation of *E. coli* was first shown by Yoshida *et al.* in 2001 (124) but at very low efficiency (<10^3 cfu/µg pUC18 plasmid DNA) and utilising chrysotile asbestos – known to be carcinogenic in humans. In 2010 however, Tan *et al.* (99) optimised the method and showed successful transformation with the less dangerous mineral sepiolite. The suggested mechanism is shown in Figure 3.1.

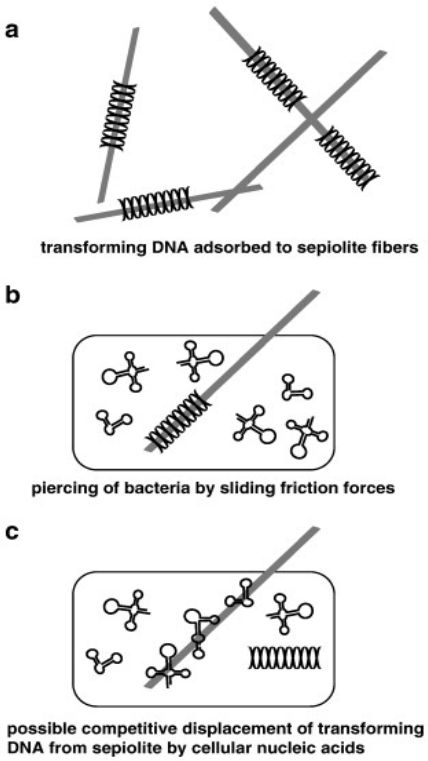


Figure 3.1. The suggested mechanism by which sepiolite mediated transformation of bacteria occurs, adapted from Yoshida et al. 2001 (124).

Sepiolite is an inexpensive and widely available, naturally occurring magnesium silicate mineral. The transformation protocol is simple involving vortex mixing of exponential phase cells in media with sepiolite powder and plasmid DNA. Reasonable efficiencies have been achieved with *E. coli* (~10^5 cfu/µg pET15b plasmid DNA) and the process is fast and cost effective compared to electroporation. It also may be able to transform with plasmids too large to enter by electroporation (99). Given its simple protocol and potential for high efficiencies this method was investigated here with *G. thermoglucosidans.*

Electroporation

Electrical fields at around 5–10 kV/cm for 5–10 μs induce the formation of temporary pores in the plasma membranes of cells. This technique was first used to transform DNA into mammalian cells (125) later optimised for bacteria. High efficiencies can be achieved with minimal need for optimisation; 10^5 cfu/μg plasmid DNA was achieved in *E. coli* the first time the technique was reported and it was quickly found to be applicable to a huge range of Gram-positive and Gram-negative bacteria (126,127). Subsequently electroporation has been shown to potentially transform virtually any cells, from eubacteria and archaea to protozoans, plants, animals, fungi. Due to its versatility electroporation is the go-to technique for transforming new host organisms.

Electroporation of a *Geobacilllus* species was shown by Narumi *et al*. in 1992 for *G. thermodenitrificans* (then classified as *B. stearothermophilus* K1041) (96). High efficiencies were achieved 5.8×10^5 cfu/μg plasmid however strain K1041 was selected for its unusually high competence from a total of 67 isolated *Geobacillus* strains. Strains isolated for other properties are not likely to be as competent.

G. thermoglucosidans DL33 is a more industrially relevant strain selected for being particularly hardy and resistant to solvent stress. With a revised protocol based on the high osmolarity method developed for bacillus species (128) Taylor *et al*. achieved electroporation efficiencies of around 10^4 cfu/μg with plasmid pUCG18 (Taylor et al. 2008). Electroporation of cells in a 1 molar sugar solution improves efficiency correlated with increased survival rate. This suggests the solution protects or improves recovery of cells rather than increasing their competence. This then allows stronger electric fields to be applied (128). As this method is the most used protocol in previous studies with *G. thermoglucosidans* it was tested in this study.

Conjugation

Despite apparent recalcitrance towards transformation, the genomes of *Geobacillus* species show evidence of significant horizontal gene transfer (129). It could in part be

that Geobacilli acquire *B. subtilis*-like natural competence under some as yet undiscovered conditions (and indeed they do possess competence gene homologues) however most of this gene transfer is likely to have occurred through conjugation and transfection. The Tn916 conjugative transposon, found to integrate DNA into the *G. stearothermophilus* genome (109), is not a viable biotechnological tool however the conjugative plasmids developed by Suzuki *et al*. for transformation of *G. kaustophilus* (94) are very promising. In order to improve conjugation efficiency, two *G. kaustophilus* methylases were heterologously expressed in the *E. coli* donor strain. Conjugation from this modified donor strain was very recently shown to be efficient for a range of *Geobacillus* species by Tominaga *et al.* 2016 (105). The conjugation protocol is a slightly slower than transformation methods requiring growth of the donor and recipient together, however no specialised equipment is required thus reducing costs. In addition, large constructs (>10 kb) can enter more easily than with electroporation for which efficiency is highly size dependent (113). Due to time constraints conjugation was not tested in this study however this could be a valuable method for future use and optimisation.

Alternative Novel Methods

A wide variety of alternative transformation methods exist and have been tested in other bacteria but not in *Geobacillus* species. They were reviewed but not found to be promising competitors for the currently established methods (electroporation and conjugation). Freeze thaw transformation is a very simple protocol involving flash freezing cells then thawing at 42 °C. It has been successful in an eclectic variety of organisms (118) but seems to have a limited, low efficiency (~10^3 cfu/µg) despite optimisation attempts (130). Liposome mediated transformation (or lipofection) has been successful at moderate efficiencies in bacteria (131) but is better suited to mammalian cells as bacterial cell walls interfere with the liposome fusion – a greater problem in gram positives. Also generation of the liposomes adds unnecessary complexity. Bombarding cells with DNA coated biolistics or magnetic nanoparticles is another niche method. Reasonable efficiencies can be again be achieved in bacteria (132) but with a complex and costly process, better suited to larger plant or mammalian

cells. The microprojectiles are around 0.1-0.3 μm in diameter with a bacterial cell only around 1.0 μm diameter.

3.1.3 Reporter Genes

Reporter genes code for proteins not normally present in the test chassis organism and crucially they enable detection and measurement of gene expression. Typically they may report expression through colourimetric assays, luminescence or fluorescence.

Enzymatic Reporter Genes

Enzyme reporters such as the commonly used *E. coli* β-galactosidase gene give colourimetric outputs with the addition of a particular substrate. For β-galactosidase the colourless synthetic compound o-nitrophenyl-β-D-galactoside (ONPG) is cleaved to give galactose and o-nitrophenol, which has a yellow colour. Production of o-nitrophenol can be determined with absorbance measurements. When ONPG is in excess over the enzyme, the production of coloured o-nitrophenol per unit time is proportional to the concentration of β-galactosidase so the rate of yellow colour production can be used to determine the concentration of the reporter enzyme. Thermostable β-galactosidase genes exist and a variant from *G. stearothermophilus* was used to test gene expression in *G. kaustophilus* (93). The assay was informative however a unwanted slight growth defect was noted. Alternative thermostable reporters include the *G. stearothermophilus* α-amylase gene amyE which reports expression qualitatively via decolouring iodine stained starch or more quantitatively via cleaving a fluorescently labelled starch substrate (93). The latter is more complex and costly and many *Geobacillus* species display native amylase activity so the reporter would not be broadly useful. The most promising enzymatic reporter already characterised in *G. thermoglucosidans* is the *Geobacillus stearothermophilus* catechol 2,3-dioxygenase gene, *pheB*. This produces a yellow colour when cleaving colourless catechol substrate and its activity is not seen in most other *Geobacillus* species (92). Activity can be quantified simply by spectroscopy and the reporter is functionally produced under all oxygen conditions.

Alternatives to colourimetric output are luminescent luciferase reporters. The advantages of a luciferase assay are high sensitivity, low background and wide dynamic range. The most versatile and common reporter gene is the luciferase of the North American firefly *Photinus pyralis*. The protein requires no posttranslational modification for enzyme activity, is non-toxic even at high concentration and can be used in a variety of organisms (133). Firefly luciferase catalyzes the bioluminescent oxidation of the luciferin in the presence of ATP, Magnesium and Oxygen, a reaction that produces bioluminescence. The main drawback of luciferase reporters is the requirement for addition of comparatively expensive luciferin substrate. Most luciferase enzymes used in biotechnology are from mesophilic organisms. Significant attempts have been made to engineer improved thermostability in firefly luciferases (134,135) but the most stable mutants are only functional up to 45 °C. The most naturally thermostable luciferase characterised so far is from the lantern fish *Benthosema pterotum.* It retains up to 50% activity over 50 °C but requires high pH and magnesium concentrations (136). Luciferase has been used previously as a reporter for a thermophilic bacterium. The luciferase enzyme of the bacterium *Photorhabdus luminescens* was shown to be functional in the thermophilic cyanobacteria *Thermosynechococcus elongatus* but only up to 43 °C. For temperatures above this, cultures were incubated at 30 °C for 1 hour before readings were taken which seemed to recover luminescence (137). For *Geobacillus* species, reporters that are simple to measure and stable over 60 °C would be preferred.

Fluorescent Reporters

The most versatile and popular reporter genes in modern biotechnology are fluorescent proteins. The discovery and characterisation of Green Fluorescent Protein (GFP) from the jellyfish *Aequorea victoria* and its application in reporting expression in heterologous hosts (138) revolutionised biological imaging and won the Nobel prize for chemistry in 2008. A huge variety of GFP variants have since been engineered with specialised properties including improved stability. Reporters mutated for improved stability include eGFP (139), folding reporter, frGFP (140) and superfolder, sGFP or sfGFP (141). These potentially have higher thermostability and could be suitable for reporting in thermophiles.

eGFP is a widely available GFP variant selected for its more intense fluorescence compared with wild-type GFP, but was shown not to be thermostable in *Thermus thermophilus* at high temperature (142). frGFP was further improved to generate superfolder GFP (141). It folds efficiently when fused to poorly folded polypeptides and exhibits an improved resistance to chemical attack and improved folding kinetics. It was shown to be functional in *Thermus thermophilus* up to 70 °C (142) and so as the most promising fluorescent protein from the literature, it was selected for use in *Geobacillus* species in this study. A major drawback of GFP and all related fluorescent proteins is a strict requirement for molecular oxygen as a cofactor for the synthesis of the chromophore (143). This limits GFP to reporting under aerobic conditions. For an industrial strain engineered for bioreactor fermentations gene expression under anaerobic conditions would also be of interest and so alternative reporters were also considered.

Flavin based fluorescent proteins (FbFPs) are a different class of fluorescent reporter gene functional under aerobic and anaerobic conditions. The reporters are based on the LOV protein domain (Light, Oxygen or Voltage sensing) first characterized in plant phototropins, a class of blue-light receptors. The domain is also present in bacterial proteins and binds flavin mononucleotide (FMN) chromophores noncovalently. When irradiated with blue light (~450 nm), these proteins undergo a photocycle involving the reversible formation of an FMN-cysteine linkage (resulting in a thiol adduct) and exhibit a weak autofluorescence with a maximal emission wavelength of 495 nm (a green/cyan colour) (144). When studying the biochemical function of the phototropin Phot1 from *Avena sativa* (the common oat), Swartz *et al*. mutated the crucial cysteine residue to an alanine in order to test its role in the photocycle (145). They noted that this mutant LOV domain did not undergo the normal photocycle as it could not form the thiol adduct and because of this (as well as reduced quenching of the FMN chromophore) it displayed significantly increased fluorescent emission.

It was 6 years later that the potential of LOV domains, beyond appealing to photobiophysicists (146) was realised. Drepper *et al*. showed that the cysteine to alanine mutation of the *Avena sativa* LOV domain had the same effect of increasing fluorescent output on LOV domains of bacterial proteins and these new, compact

domains could be used to report gene expression (147). These flavin-based fluoresecent proteins (FbFPs) have many advantages in addition to fluorescence when folded in the absence of oxygen. The functional LOV domain alone is significantly smaller than GFP and other related beta-barrel fluorescent proteins (~13 kDa to GFPs 27 kDa). This makes FbFPs useful for fusion proteins as they cause less steric interference with the folding and movement of proteins attached to them (148,149). Denaturation and refolding experiments suggest FbFPs have a faster maturation time than even rapid folding GFP variants. As small and stable domains, FbFPs can refold in 2-3 minutes whilst sfGFP requires around 10 minutes to recover fluorescence (148). This fast maturation makes FbFP reporters better suited to studies that require precise temporal reporting such as monitoring short-lived proteins and early detection of promoter activation. Additionally, whilst the intrinsic pH-sensitivity of the GFP chromophore can be exploited in designing pH-responsive probes (150), the loss of fluorescence at low pH complicates measurements of biological processes in acidic environments. The broad pH tolerance of FbFPs (spanning pH 4–11) again increases their versatility.

Finally, even as a reporter of promoter activity in normal aerobic growth conditions – a current standard application for GFP – FbFPs have advantages. Drepper *et al.* showed a codon optimised FbFP based on the LOV domain of the *B. subtilis* protein YtvA could outperform YFP (a yellow GFP variant) as a fluorescent reporter for real time monitoring of gene expression in *E. coli* grown in rich media. Fluctuations in oxygen levels occurred as the *E. coli* cells transitioned from actively respiring exponential phase growth to stationary phase, resulting in inaccurate assessment of promoter activity by the oxygen-sensitive fluorescence of YFP. By contrast, better agreement was shown between FbBP fluorescence levels and mRNA measurements by quantitative reverse transcription PCR (151). Early FbFPs were not as bright as GFP and particularly dim compared to eGFP and sfGFP variants but their stability and brightness is being improved (152). An FbFB variant functional in *G. thermoglucosidans* would be hugely valuable for molecular and synthetic biology in this organism and so several variants were investigated.

3.2 Results

3.2.1 Growth Media

A range of rich media previously used for *Geobacillus* species were tested as growth media (Materials and Methods 2.2.2). In all rich media except LB and LB +2% glucose *G. thermoglucosidans* grew to stationary phase from inoculation with a single colony at 55 °C in under 12 hours. Formulating LB or LB glucose with mineral water rather than distilled water allowed considerably faster growth on liquid and solid media however, suggesting that trace elements are limiting (Figure 3.2).

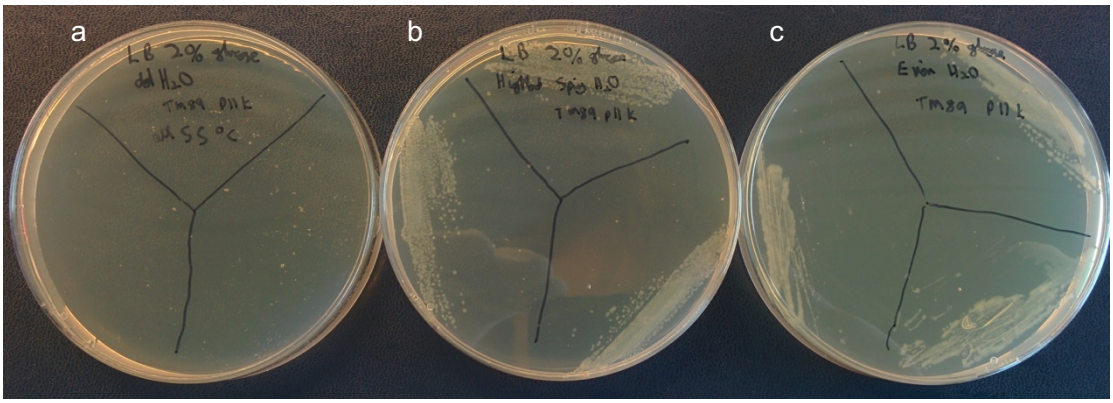


Figure 3.2. LB + 2% glucose agar plates (a) with distilled water (b) with Highland Spring™ mineral water (c) with Evian® mineral water all streaked with single colonies of *G. thermoglucosidans* and incubated at 55 °C for 12 hours.

Of the rich media that gave rapid growth, 2TY and 2SPYNG (also known as 2SPY) are the simplest to prepare requiring only three ingredients, which can all be combined before autoclaving. Both media have been widely used for previous *Geobacillus* studies and so there is little to recommend one above the other. Soy peptone is generally cheaper than casein tryptone and so 2SPYNG was chosen as the rich undefined media for *Geobacillus* species growth in this study.

For growth in defined media, ammonium salts media, ASM is a very complex minimal media but has been used previously for growth of *Geobacillus* species under aerobic and anaerobic conditions (106,153) and hence this was used in this study. Spizizens minimal medium has been used previously for growth and characterisation of *Bacillus subtilis* (154). It is a more simplified media, which is far easier to prepare than ASM.

However, this did not give growth of *G. thermoglucosidans* possibly due to a lack of essential trace elements (data not shown).

3.2.2 Transformation

Electroporation

Preparation of competent cells and electroporation of *G. thermoglucosidans* as reported by Taylor *et al* 2009 was (90) was tested and similar efficiencies (~10^4 cfu/µg plasmid DNA) were achieved as described previously.

Plasmid	Size/kbp	Reference	Transformation efficiency CFU/µg DNA
pUCG18	6.3	Taylor et al. 2008 (90)	4.9×10^3 ± 15%
pUCG3.8	3.8	Bartosiak-Jentys et al. 2013 (86)	5.2×10^3 ± 21%
pG1K	3.7	This study	5.3×10^4 ± 17%

Table 3.2. Electroporation efficiencies. The protocol is described in Materials and Methods 2.2.8. All plasmids were selected for on 2SPYNG agar plates with 12 µg/ml kanamycin. Percentage standard deviations from three biological repeats are given.

Due to the need for specialised cuvettes, electroporation is a comparatively expensive transformation method for large-scale use. Recycling of electroporation cuvettes was investigated with used cuvettes either washed or autoclaved, dried and reused. The cuvettes (BioRad UK, 0.1 cm gap width) deformed slightly on autoclaving (121 ºC/103 MPa 15 minutes) so were not reusable by this method. Washing with ethanol, bleach and distilled water has reportedly allowed cuvettes to be reused up to ten times for transformations with *E. coli* (155). The high osmolarity method used for *Geobacillus* species transformations requires higher field strength and recycled cuvettes were found to cause arcing (a rapid, narrow electrical discharge between the electrodes) so could not be used. This could be due to salt impurities remaining after the wash steps which lower conductivity or effects on the electrical contacts – aluminium oxide forms on the cuvettes electrodes, thus changing the electrical parameters.

Mineral Nanofiber Transformation

A method adapted from Tan *et al.* (2010) (99) for mineral nanofiber mediated transformation with sepiolite was tested with varying sepiolite concentration but was not found to be effective (data not shown). Although colonies arose on recovery plates following transformation, these were not seen to be kanamycin resistant after reculturing on solid or liquid media and plasmid could not be recovered from them.

3.2.3 Reporter Genes

Anaerobic fluorescent proteins based on the LOV domain have huge potential and as such characterising a variant for use in *Geobacillus* species was a priority of this study. Difficulties were encountered expressing an FbFP in *G. thermoglucosidans* and so four different sequence variants were tested.

hotLOV, a Thermosynechococcus LOV domain

The well-characterised Drepper *et al*. proteins BsFbFP and PpFbFP were derived from LOV domain containing genes from mesophilic bacteria – the blue light sensing YtvA protein of *B. subtilis* and the sensory box 2 protein of *P. putida,* which likely has a similar light detecting function. These proteins are known to thermally denature between 50 and 60 °C (148) so are not promising reporters for thermophiles. LOV domain-containing proteins exist in many bacterial species, however, and a suitable domain was found in a light sensing protein of the thermophilic cyanobacteria *Thermosynechococcus elongatus.*

A plasmid containing this thermophile LOV domain, with the cysteine to alanine mutation was kindly donated by Prof. John Christie (University of Glasgow) and was reported to produce a functional and fluorescent LOV protein in *E. coli* under aerobic and anaerobic conditions (John Christie, personal communication). The sequence was obtained in an *E. coli* plasmid and was in a form designed for fusion proteins. Therefore upon receipt, the hotLOV domain was amplified via PCR with primers to add an ATG and TAA start and stop codons, and this was then cloned into the shuttle vector pUCG16 with the moderate strength promoter pUP1 (see Chapter 4) immediately upstream (Sequences in Appendix 10.2).

E. coli cells transformed with this construct in the shuttle vector were noticeably fluorescent when exposed to blue light under an orange filter whereas *G. thermoglucosidans,* grown at 45 to 55 °C was not fluorescent when containing this construct. The peak excitation and emission wavelengths of LOV domain proteins (excitation 450 nm and emission 495 nm) are similar enough to the wavelengths for GFP (excitation 485 nm and emission 530 nm) that filters and setting for GFP detection by plate reader and flow cytometry also allow detection of LOV proteins. (147). No fluorescence above background levels was detected by plate reader or flow cytometry for hotLOV expressed in *G. thermoglucosidans,* however The promoter was shown to be effective with alternative reporters, but the hotLOV protein may not be translated at high levels due to problems with the RBS sequence or with codon usage. Alternatively, if it is translated then it may not be folding correctly, may be degraded quickly or may not be fluorescent due to no cofactor binding or no cofactor availability. It may also be not as thermostable as predicted and thus misfolding at thermophilic temperatures. To rule out this last consideration thermostability was tested in comparison to standard fluorescent proteins.

A joint E. *coli/B. subtilis* codon optimised version of sfGFP (156) listed in the iGEM Registry of Standard Biological Parts (157), and monomeric red fluorescent protein (mRFP) (158) were similarly cloned into the pUCG16 shuttle vector with the pUP1 promoter and transformed into *E. coli* DH10B. Stationary phase cultures expressing the fluorescent proteins were suspended in a lysis buffer and lysed by sonication. The insoluble fraction was removed by centrifugation and the fluorescent proteins in the supernatant were tested for thermostability. Lysates were incubated for 30 minutes at temperatures between 35 and 95 °C (Figure 3.3).

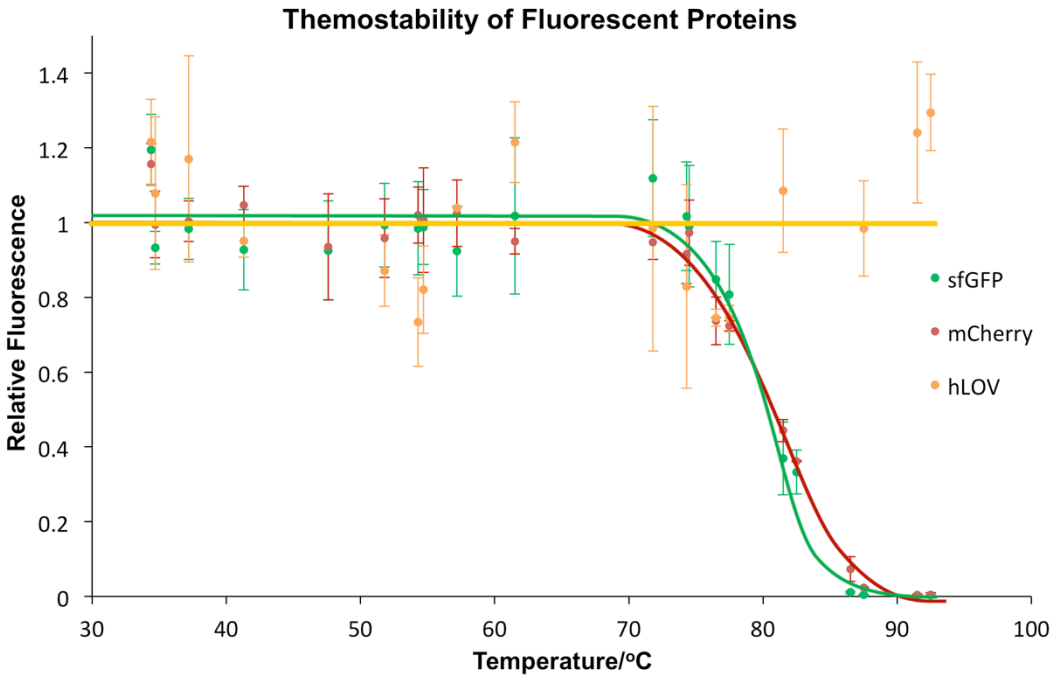


Figure 3.3. *In vitro* thermostability of fluorescent proteins in *E. coli* cell lysate. Proteins were exposed to the given temperature for half an hour with fluorescence measured by plate reader.

In this thermostability assay from *E. coli* cell lysates, hotLOV shows even greater thermal stability than sfGFP, known to be functionally expressed in thermophilic bacteria up to 70 °C (142) and so thermostability is unlikely to be the factor preventing functional hotLOV expression in *G. thermoglucosidans*. These data also confirm the known, high stability of sfGFP and also suggest mCherry could be a viable alternative reporter for thermophiles. Both of these were later investigated further.

cohLOV, a Codon Optimised LOV Domain

The initial hotLOV sequence that was used had been codon optimised for *E. coli*. As a Gram negative very distantly related to Gram positive bacilli, codon usage between *E. coli* and *Geobacillus* species is highly divergent. The sequence was therefore codon optimised using the Entelachon software tool (Fischer n.d.) with *Geobacillus* codon data taken from the codon usage database (Nakamura et al. 2000), more details on theory and parameters are given in Chapter 8. Nucleotide changes made in this codon optimisation are shown in Figure 3.4.

The codon optimised hotLOV protein (cohLOV) was synthesized (IDT, UK) and restriction cloned into pUCG16 with either the strong pLdh promoter or the moderate strength pUP1 promoter driving expression.

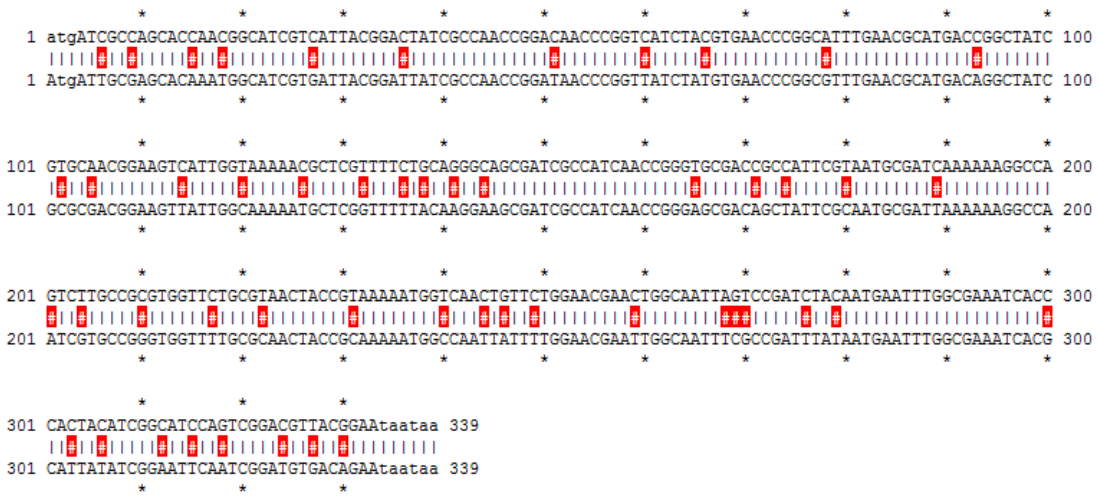


Figure 3.4. Nucleotide sequence of *E. coli* optimised hotLOV, top; compared with the *G. thermoglucosidans* optimised cohLOV, bottom. Alignment generated with the NCBI Blast tool (159).

Once cloned into the vectors in *E. coli*, fluorescence could again be visualised from colonies by eye with blue light excitation, however, when used to transform *G. thermoglucosidans* no fluorescence with any construct could be detected (Figure 3.5). Settings for GFP were used on the flow cytometer (Materials and Methods 2.4.2) and possible expression of both hLOV sequence variants was compared to sfGFP expression from the very weak promoter pUP7 (described in Chapter 4) as a positive control. No fluorescence was observed therefore even with improved codon optimisation a stronger promoter, hLOV is not functionally expressed *G. thermoglucosidans*.

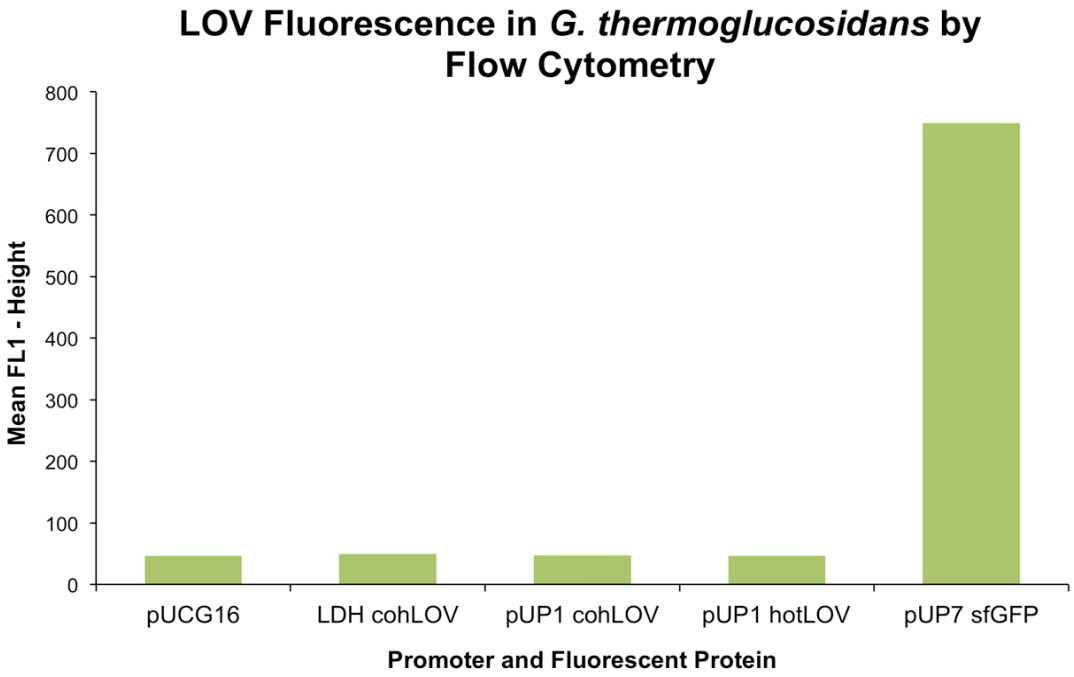


Figure 3.5. Flow cytometry readings for LOV proteins expressed from shuttle vectors in *G. thermoglucosidans*. Even with high gain settings fluorescence could not be detected above the background, pUCG16 empty vector levels. The positive control is sfGFP expressed from a synthetic weak pUP promoter called pUP7, which is weaker than both the pLDH and pUP1 promoters (promoters are detailed further in Chapter 4).

Both of the promoter/RBS combinations used gave visible expression with the sfGFP reporter and so the lack of fluorescence is unlikely to be due a complete lack of transcription or translation. The FbFP is likely therefore to be present and expressed but is either not fluorescent or is being targeted for degradation, possibly due to misfolding, aggregation or due to some other factor.

Full Length, Thermostable BsFbFP

In a final attempt to find a usable anaerobic fluorescent protein, a thermostable variant of the full length *B. subtilis YtvA* protein was codon optimised and tested for expression in *G. thermoglucosidans*. FbFPs are usually the LOV domain alone, though a full-length protein may be more stable. So far hotLOV has only been seen to fluoresce here in *E. coli,* but the original *B. subtilis* LOV reporter (BsFbFP) has been shown to be effective in a range of hosts including Gram-positive *Clostridium* species (160). The main limitation of BsFbFP for this study is its low thermostability, however in 2013 Song *et al*. engineered a more thermostable YtvA fluorescent protein using a combined computational and experimental method (161). Protein structure prediction combined

with a free energy model was used to identify 18 single amino acid changes that could increase protein stability. These mutants were tested and the most effective substitutions combined to give a final mutant with three amino acid changes (N107Y-N124Y-M111F) and an improved thermostability from a Tm of 50 °C to 75 °C. Therefore, this full length YtvA mutant was codon optimised for *G. thermoglucosidans* (and named bhLOV) and produced by gene synthesis (IDT UK) and cloned into shuttle vectors as before before being transformed into *E. coli* and then *G. thermoglucosidans*.

This final variant, as with the previous iterations, showed visible fluorescence in *E. coli* but not in *G. thermoglucosidans* even at the lower than usual temperatures of 45-55 °C (preliminary experiments, data not shown). Several explanations may explain this and significant further work would be required to determine the cause. This was not pursued here and alternative reporters were investigated.

Superfolder Green Fluorescent Protein

As mentioned above, the GFP variant sfGFP (141) is highly stable, fast maturing and has previously been used to report gene expression the thermophile *Thermus thermophilus* (142) and in *G. stearothermophilus* (72). The joint *E. coli/B. subtilis* codon optimised sfGFP generated as part of the 2008 Cambridge iGEM project (156) was known to function in *G. thermoglucosidans* before this study began (106) and the pUCG16 plasmid containing this reporter expressed from the *G. stearothermophilus* Ldh promoter was kindly shared (Elena Martinez-Klimova, Department of Life Sciences, Imperial College London). This construct was used for promoter characterisation in this study (Chapter 4), and sfGFP was observed to be brightly fluorescent (Figure 3.6) and well expressed at 45 to 65 °C (Figure 3.8) and also caused no noticeable growth defects in *G. thermoglucosidans*. Its only limitation as the reporter of choice for *Geobacillus* species is the need for aerobic conditions for its fluorescence.

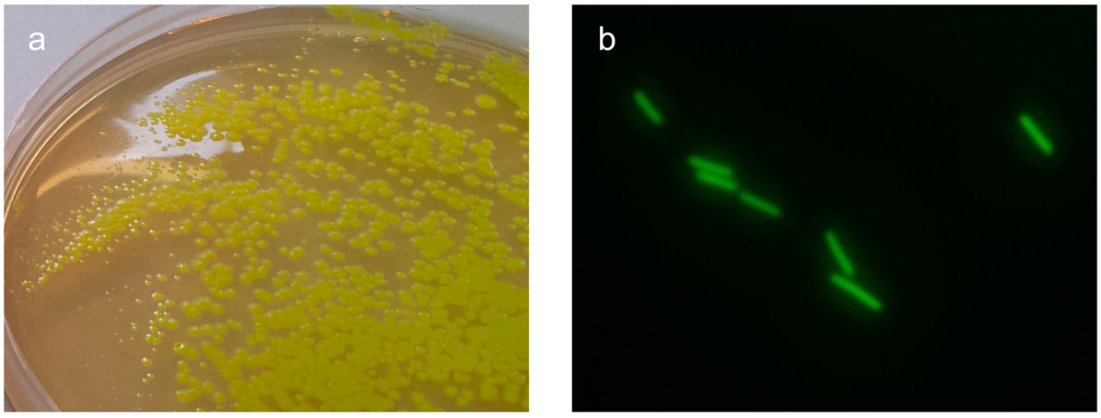


Figure 3.6. superfolderGFP expression in *G. thermoglucosidans* a) colonies on solid media and b) cells viewed by fluorescence microscopy (Materials and Methods 2.3.3).

Anaerobic Fluorescence Recovery of sfGFP

Zhang et al. 2005 reported a technique referred to as anaerobic fluorescence recovery (AFR) whereby GFP produced in cells grown anaerobically could have limited fluorescence recovered by subsequent exposure to oxygen (162). They applied the method to estimate relative cell density of a particular species in a coculture but here this method was adapted to see if it may be applicable to characterise promoter activity under low oxygen conditions for *G. thermoglucosidans*. pUCG16 plasmids with sfGFP expressed from three different promoters (pLdh, pUP2 and pUP5) were transformed into *G. thermoglucosidans* and aerobic pre-cultures were grown in tubes. These were each diluted 1000x into fresh media in triplicates and grown in media-filled, sealed Hungate tubes (Belco Glass, USA) at 55 °C to create microaerobic conditions. After 12 hours growth (late exponential phase) cells were harvested by centrifugation, washed and resuspended in PBS, shaken aerobically at room temperature for 1 hour as described in Zhang *et al.* 2005, then fluorescence measurements were taken by plate reader (Material Methods 2.4.1). In preliminary experiments only very low levels of fluorescence were ever observed, however, and the variability between triplicate repeats was high (Figure 3.7).

While this was a disappointing start, it was discovered from previous work that GFP that is initially folded anaerobically can eventually recover fluorescence with subsequent oxidation (163) and sfGFP has been shown to do this after being denatured (164). Therefore denaturation followed by subsequent aerobic refolding was

considered. Anaerobic samples were prepared as above and then cells were boiled at 100 °C for 20 minutes in media (or harvested and boiled in PBS). The protein was then recovered by shaking under aerobic conditions for 1 hour. Once again in this preliminary experiment, only very low-level fluorescence was recovered (Figure 3.7), around 1 suggesting this is not a viable method for characterising promoter strength under anaerobic conditions.

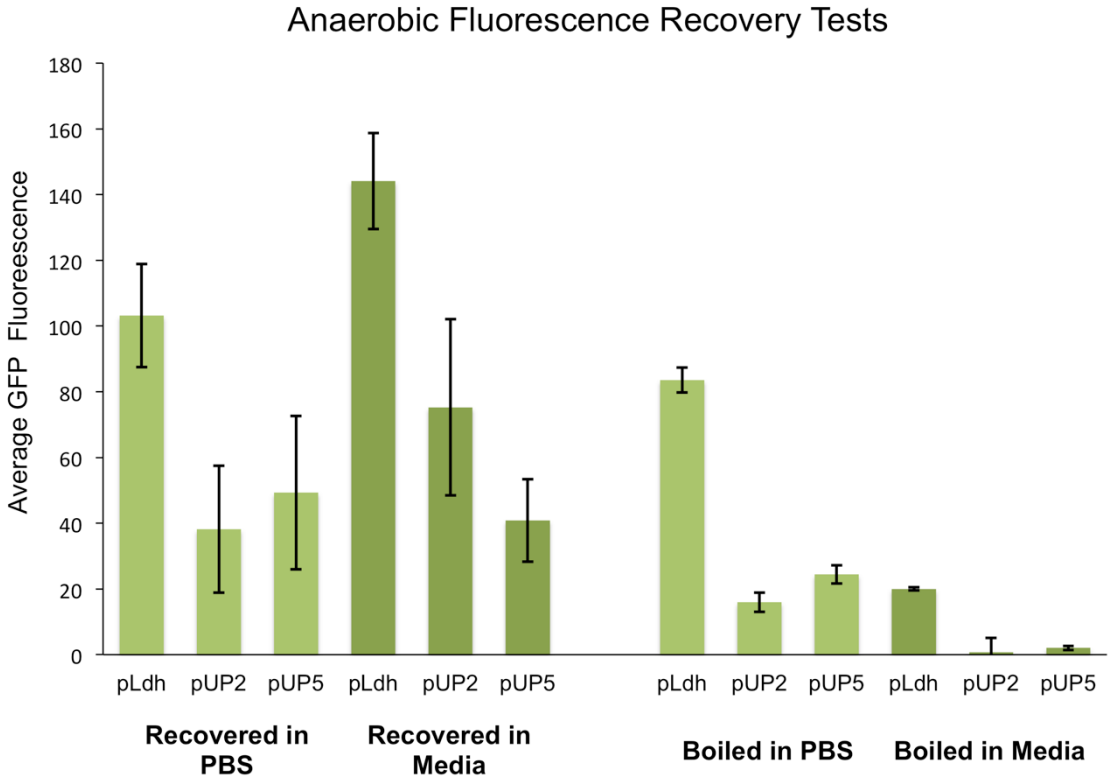


Figure 3.7. Fluorescence data from attempted anaerobic recovery of GFP fluorescence. *G. thermoglucosidans* transformed with plasmid pUCG16 expressing sfGFP from the three promoters shown was grown without oxygen. Growth was halted when cultures reached stationary phase and cells were harvested and oxygenated with shaking in media or PBS with or without boiling to denature proteins. Error bars show standard deviations from three biological replicates. All fluorescence readings were very low hence error margins are high.

Red Fluorescent Protein, mCherry

While not being able to solve the issue of a GFP or LOV reporter protein that works for anaerobic conditions was a set-back, having multiple different aerobic reporter proteins for expression in *G. thermoglucosidans* was still a goal. Several, separately quantifiable reporter proteins expressed from the same host can allow more sophisticated characterisation of biological systems. As fluorescent proteins have emerged the most popular reporter choice – particularly in synthetic biology – a range

of colours have been developed to allow simultaneous measurement of different reporters. Fluorescent markers with distinct excitation and emission wavelengths can be independently assayed within the same cell with minimal cross talk. Red fluorescent proteins have minimal spectral overlap with GFP reporters and the monomeric, cherry red fluorescent protein mCherry (158) was observed earlier in this chapter to display comparable *in vitro* thermostability to sfGFP. Thus, mCherry was cloned into pUCG16 expressed from the strong RplS promoter (details in Chapter 4) and transformed in to both *E. coli* and *G. thermoglucosidans.* Transformed cells containing this construct showed that mCherry was well expressed in *E. coli* at 37 °C (data not shown), however despite promising *in vitro* stability as shown before, when mCherry was expressed in *G. thermoglucosidans*, its fluorescence declined with increasing growth temperatures used *in vivo* (Figure 3.8). Very little red fluorescence is detected for cells grown above 50 °C, in contrast to green fluorescence from sfGFP where the protein is functionally expressed up to 65 °C.

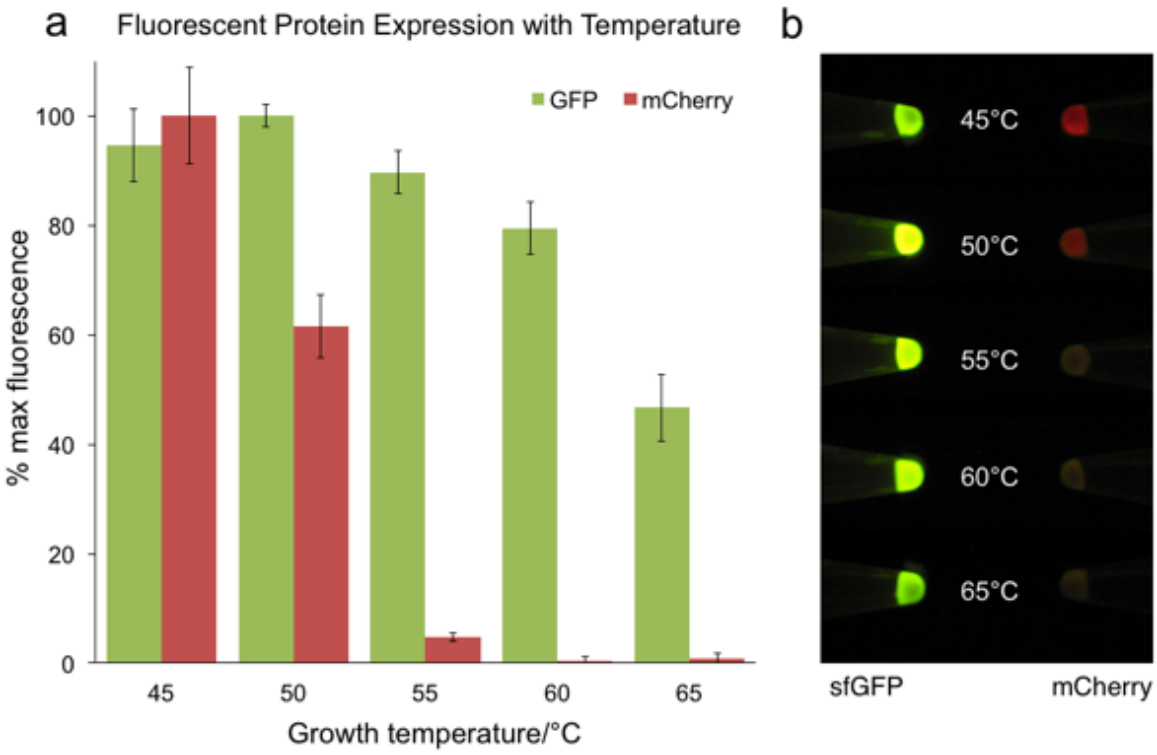


Figure 3.8. a) Fluorescent plate reader data for sfGFP and mCherry expressed in *G. thermoglucosidans.* Cells with fluorescent proteins expressed from the strong RplsWT promoter on the pUCG16 plasmid were grown to stationary phase at various temperatures and fluorescence readings taken. Error bars show standard deviations from three biological rpeats **b) Centrifuge pelleted *G. thermoglucosidans* cells grown for 3.9a.** Cells are illuminated by blue light and images taken through an orange filter.

To further explore this, flow cytometry was used (Materials and Methods 2.3.2). *G. thermoglucosidans* expressing mCherry with the strong pRplS promoter from plasmid pUCG16 was grown in liquid culture at 50 and 55 °C, the temperatures between which mCherry fluorescence seems to be lost (Figure 3.8). Average fluorescence readings of the whole population show this significant decrease (Figure 3.9a). Histograms of cell populations show fluorescence does not decrease evenly in all of the cells however (Figure 3.9b). At 55 °C two sub populations emerge with suggesting that whilst some cells were functionally expressing mCherry, in another the denatured/misfolded protein was likely aggregating or the stress from misfolded protein was causing the plasmid to be lost or mutated. mCherry could only be viable for dual-colour fluorescent reporting in thermophiles at 50 °C or below.

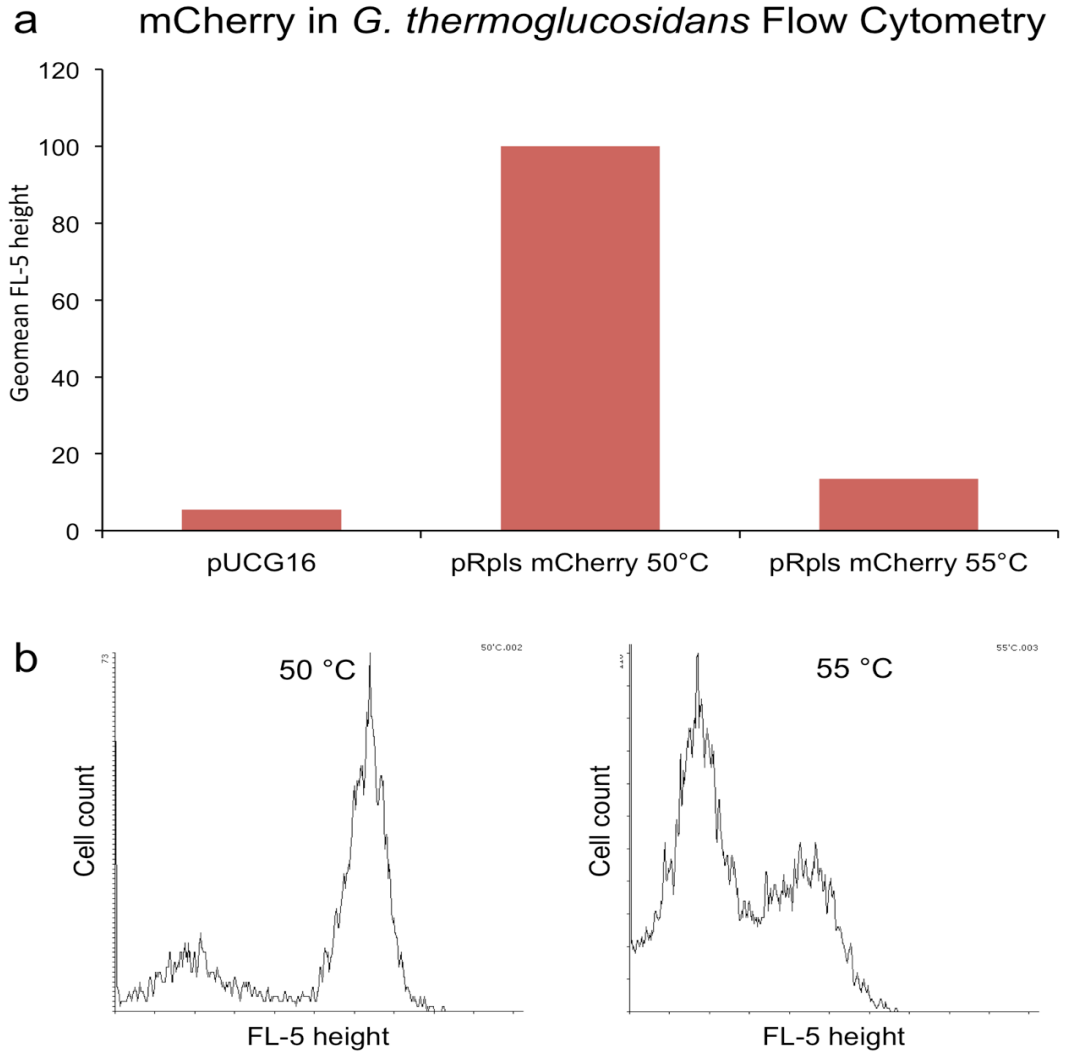


Figure 3.9. Flow cytometry data for mCherry expression in *G. thermoglucosidans* grown at different temperatures a) Geometric mean fluorescence output, the average fluorescence at 610nm of the whole population of cells. **b) Histograms of cell count against fluorescence level on a logarithmic scale**, the distribution of fluorescence levels for cells in the population can be seen. Excitation was with a yellow/green laser (561 nm) and detection via filter Fl-5, (610 nm).

3.3 Discussions and Future Work

In this chapter growth media was briefly tested and then transformation methods were comprehensively reviewed. Electroporation was found to be the most suitable method for *G. thermoglucosidans* in this study however conjugation is likely to be valuable in future. Anaerobic FbFps could not be functionally expressed in *G. thermoglucosidans* and this seems to be due to lack of protein rather than lack of function. Superfolder GFP was found to be the most suitable current reporter for *G. thermoglucosidans* though only in aerobic conditions. The red fluorescent protein mCherry could provide an alternative but only at temperatures of 50 °C or below.

3.3.1 Growth Media

G. thermoglucosidans seems to be more dependent on trace elements for growth than standard chassis organisms such as *B. subtilis* and *E. coli*. Industrial, lignocellulosic feedstocks are likely to have abundant trace elements whereas in laboratory media these may be limiting. This dependence may be important to consider when optimising media or feedstocks for future *Geobacillus* applications. Developing a media preparation that is simple to produce in the lab but more representative of typical feedstocks generated industrially from lignocellulosic biomass pretreatment could be valuable in future.

3.3.2 Transformation Methods

A range of transformation methods have been developed for bacterial species however currently electroporation and conjugation are most suitable for *G. thermoglucosidans*. Current electroporation efficiency is workable but could be improved. Optimisation of this protocol is a complex multifactorial problem however with many interdependent variables – cell preparation media, OD600 of harvested cells, electroporation buffer ingredients, pulse conditions and recovery conditions (165). Whilst many established, standardised protocols tend to be workable in a broad range of species, further specific optimisations tend to only give host specific improvements (118). Improving transformation in *Geobacillus* species to the degree that they could serve as a primary cloning host – negating the need for plasmid construction in *E. coli* shuttle vectors – would require around 1,000-fold increase in efficiency, which may not be possible.

Instead, in this study attempts were made to improve cost and reproducibility and to develop improved vectors. Electroporation efficiency is known to drop off steeply for larger plasmids. This is thought to be due to plasmids getting trapped in pores across the membrane killing the cells or being physically less able to diffuse into the cytoplasm (166). Efforts to develop compact plasmids for electroporation are described in Chapter 7.

Even with a standardised protocol electroporation efficiency is notoriously variable. This is perhaps due to a significant dependence on factors that are difficult to keep constant. The precise stage in the growth phase at which cells are harvested significantly affects competence (118) the actual electrical field strength received by the cells is dependent on the temperature of the cuvette (warmer solutions have higher conductivity) and salt contamination to the cells or DNA. Finally recovery time, if cells from a certain batch recover particularly quickly they may be able to undergo rounds of division before being plated, artificially raising the colony count and apparent efficiency. When novel electroporation protocols are reported for non-standard organisms they often include long recovery times of 2 hours or more in growth media before antibiotic selection. This is more than long enough, accounting for the lag in initiating growth under the new conditions, for many bacteria (*Geobacillus* species included) to undergo rounds of cell division. Electroporation efficiencies in many studies could be over estimated because of this though perhaps only by 2 or 4 times. Due to these inherent reproducibility issues and the possibility for over estimation, efficiencies in all studies cited here and indeed in this thesis should not be taken at face value and only provide an order of magnitude indication.

In this study efficiencies anywhere between 0 and $>10^5$ cfu/μg were achieved with the same strain and plasmid. Anecdotally a significant factor seemed to be temperature differences - room temperature plates gave no transformants whereas pre-warmed plates and plating quickly, close to the incubator improved efficiency. Slight differences in handling time when plating recovered cells may account for efficiency variation and so efforts were made to standardise this. Recovery time was then cut to 1.5 hours to reduce the chance of cell divisions in this step. The revised protocol is described in Materials and Methods 2.2.9.

Efficient conjugation to *G. thermoglucosidans* from an *E. coli* donor strain was recently demonstrated by Tominaga *et al*. 2016 (105). Due to the limited efficiency and difficulties optimising electroporation, conjugation may become the preferred method for transformation of *Geobacillus* species in future.

3.3.3 Reporter Proteins

LOV Protein Reporters

Further work is required to determine the reason for lack of LOV protein expression in G. thermoglucosidans. These proteins may be causing stress to the cells so become silenced or mutated upon transformation. However no growth defect or drop in transformation efficiency was observed so this is unlikely. The protein may instead be misfolding due to elevated temperatures in combination with unfavourable conditions in the cytoplasm, or due to the absence of necessary chaperone proteins in this host. Alternatively, it may well be interacting with native regulatory proteins and as such gets targeted for degradation. The issue could also be related to depletion of the FMN cofactor. To further assess LOV protein expression, fusion proteins with other reporters that are well expressed such as mCherry (at 50 °C) or PheB could be made. This may help to stabilise the LOV protein and assaying for the other reporter would help to identify the problem.

An anaerobic fluorescent reporter would be hugely valuable and is ultimately necessary for synthetic biology in organisms to be used for anaerobic fermentations. By understanding the issue with expression a stable LOV variant could be hopefully be found or engineered.

mCherry

Having several fluorescent reporters that can be independently measured is very valuable for synthetic biology and so improving mCherry thermostability could be considered in future. A computational design method similar to the process used by Song *et al*. to improve bsLOV thermostability (161) could also stabilise mCherry.

Alternatively, a library of mutant mCherry proteins could be expressed in *G. thermoglucosidans* and stable variants selected by fluorescence-activated cell sorting (FACS).

Superfolder GFP

Superfolder GFP should be the reporter of choice for thermophile synthetic biology though is limited only by oxygen dependence. The anaerobic fluorescence recovery technique reported by Zhang et al. 2005, allows GFP proteins that have been produced but not yet fully folded to oxidise and become fluorescent. This provides an estimate of the GFP production rate. Folding times are quite different between GFP variants however. The particular GFP used by Zhang et al. has a maturation rate of around 90 minutes (though interestingly maximal fluorescence was achieved significantly faster than this when exposed to oxygen) (162). Superfolder GFP, the only variant known to be stable in thermophiles, has a maturation time under 10 minutes (141) meaning far less unfolded protein would is present for oxidation. Highly sensitive GFP measurements would be required for this to be viable with sfGFP and optimisation of the protocol would be necessary to eliminate background fluorescence. In this initial test the majority of the fluorescence was likely from GFP produced during the aerobic preculture. Several rounds of anaerobic culturing would be required to dilute this out during which the plasmid could be lost or mutated. Also very strict anaerobic conditions would have to be maintained with repeated sparging of the media as GFP is very effective at scavenging any available oxygen. Should a slower folding GFP variant be shown to function in Geobacillus species then this technique may be viable however currently alternative reporters or methods must be used for anaerobic conditions.

In future, chemical denaturation and refolding in a buffer to reduce aggregation could be considered. Guanidine hydrochloride or guanidine thiocyanate are effective denaturing agents (164) however high concentrations (~5 M) are required and then must be diluted out (also diluting fluorescence) to allow refolding (167). The process would require considerable optimisation and is complex compared to other reporter options however this may be necessary should a suitable anaerobic fluorescent protein not be discovered.

Alternative Anaerobic Reporters

As a suitable fluorescent reporter could not be found for characterising anaerobic gene expression in *G. thermoglucosidans,* alternative, enzymatic reporters must be considered instead. Beta-galactosidase and alpha amylase reporters have been used in *Geobacillus* species however beta-gal caused a growth defect in *G. kaustophilus* (93). Amylase does not give a very quantitative output when decolouring iodine stained starch and synthetic fluorescently labelled substrates are expensive. The *G. stearothermophilus* catechol 2,3-dioxygenase gene, *pheB* is the best thermophile enzymatic reporter gene currently available (92) and a plasmid with this reporter was kindly shared (Elena Martinez-Klimnova, Department of Life Sciences, Imperial College London) and it was included in the shuttle vector toolkit (Chapter 6).

Chapter 4: Promoters and Promoter Libraries

Summary

Tuning of promoter strength is the primary control point for controlling gene expression in engineered biological systems. Characterising the strength of promoters allows rational design of future genetic circuits using those parts. Two methods for determining promoter strength, endpoint vs. synthesis rate calculations were reviewed and compared. Few promoters have been characterised in *G. thermoglucosidans* to date and no promoter libraries exist to fine tune expression. Constitutive promoter candidates were reviewed and two novel promoter libraries generated by different methods: degenerate oligonucleotides and mutagenic PCR.

Aims

- To find constitutive promoters, with strong expression in *G. thermoglucosidans.*
- To make libraries of promoters with a wide range of strengths to allow fine tuning of expression.
- To compare current methods for characterising promoter strength with fluorescent reporter genes: endpoint fluorescence and synthesis rate

4.1 Introduction

4.1.1 Promoter Selection

Promoter strength is a key control point for determining production levels of proteins and functional RNAs. Previous engineering in *G. thermoglucosidans* has mostly been limited to overexpression from strong natural promoters (60,80) though some lower strength and partially inducible promoters have been described (79,86). For more complex synthetic biology applications, precise tuning of expression over a wide dynamic range is required. This can be achieved through controlled induction of characterised inducible promoters or by selection of candidate promoters for a characterised promoter library. Inducible promoters allow construction of dynamic genetic circuits and can be useful for rapid testing or expression of toxic products. Fine-tuning gene expression with inducible promoters can be challenging however due to possible inducer hypersensitivity and population heterogeneity in expression. Also, inducer levels and hence expression levels may not remain constant as the inducer is consumed, degrades or is diluted out. Constitutive promoter libraries allow precise tuning without the need for addition of comparatively expensive inducers and can have steady expression rates in homogeneous populations with little variability in transcript levels between cells (168). They are useful for engineering stable, complex genetic circuits and for metabolic engineering applications and so were prioritised in this study. Many metabolic pathways are highly sensitive to levels of gene expression and small changes can cause a complete loss of activity (169). Fine graded control of gene expression using characterised promoter libraries is the best technique to avoid this.

Mutation of natural promoter sequences usually reduces strength and so to generate a promoter library with a wide expression range, the starting promoter must be very strong. The strongest and most commonly used promoter in *G. thermoglucosidans* previously is the *G. stearothermophilus* lactate dehydrogenase promoter pLdh (60,90), however the promoter could not be considered constitutive as expression levels are highly influenced by redox conditions (76,92). An alternative strong constitutive promoter was sought.

Many strong, constitutive promoters for *B. subtilis* are already available in the literature and the registry of standard biological parts. These could be a useful resource for *Geobacillus* species. *B. subtilis* is quite distantly related to the Geobacilli however and promoter function is dependent on DNA melting and polymerase binding which are highly temperature dependent (170). Promoters reported to be strong and constitutive in *B. subtilis* may have neither of those properties in a thermophile and so natural *Geobacillus* species promoters were considered instead. The decision to use novel promoters was also influenced in part by commercial considerations. This work was initially sponsored by TMO Renewables Ltd. (now in administration). Developing new strains with novel promoters makes intellectual property issues simpler than including promoters owned or produced elsewhere.

Some of the strongest characterised promoters in other hosts come from viruses (171) and so phages of *Geobacillus* species were reviewed. A handful of these viruses have been sequenced (172,173) but their genomes contain large sections of unknown sequence and hence are poorly annotated. Finding strong promoter candidates would therefore be hugely challenging.

The genomes of *Geobacillus* species themselves are the most promising source for promoters that definitely function in Geobacilli. Many genomes sequences are available but annotation of promoters is limited, as no transcriptomics data has been published. Annotations are from automated genome annotation programs and based entirely on primary sequence features. As such, all candidates were cross-referenced with known promoter sequences in well-characterised *Bacillus* species.

4.1.2 Promoter Library Generation

Two methods have previously been reported for generating promoter libraries, degenerate oligonucleotides and mutagenic PCR. The degenerate oligonucleotide method is based on the observation that varying bases between and around the -10 and -35 sequences modulates expression (174). Resynthesizing the promoter construct in a PCR amplification with the core promoter encoded in a primer with degenerate sequence was shown to be a simple and effective method for library generation (175).

Originally the method was used to investigate natural metabolic pathways or to over express proteins that may be toxic to cells; when the library is transformed in, only the clones with viable expression levels of the target gene will grow (175). It was later adopted by synthetic biology to generate libraries of parts for characterisation (176). The alternative method is a mutagenic PCR step using nucleoside analogues to vary the promoter sequence more randomly. The method was first developed to produce protein libraries (177) and adapted to produce promoter libraries for synthetic biology applications (178). The oligonucleotide method is a little more reliable in producing "useful" promoters, sequences similar to the starting promoter but variable in expression strength because base pair changes by the mutagenic PCR method, being more random, have a higher chance of producing promoters with either no change in expression from the WT or, by mutating critical bases in the -10 or -35, no expression at all (179). The mutagenic PCR method therefore demands screening of a larger number of colonies to produce the desired fine-graded library. Additionally, the degenerate oligonucleotide method tends to produce sequences with less homology between library members and so for building complex pathways, many promoters from the same library could be used with low risk of recombination.

4.2 Results

4.2.1 pUP Library Generation

The uracil phosphoribosyltransferase enzyme is vital for nucleotide metabolism in bacteria and is constitutively expressed. The promoter from this gene in *B. cereus* has been used as a useful part for synthetic biology giving strong and constitutive expression in *B. subtilis* and *E. coli* (180,181). The gene and its promoter are highly conserved across the Bacillaceae family and so the promoter from *G. thermoglucosidans* was selected from the genome sequence (98). The pUP promoter architecture is very typical of *Bacillus* species SigA promoters with -10 and -35 sequences close to the consensus and an T-rich upstream promoter element (UP element) . A relatively short sequence (86 bp) was selected as the template for a promoter library (Figure 4.1).

Figure 4.1. pUP promoter natural sequence and oligo with degenerate nucleotides for synthesis of the promoter library. The -10 and -35 sequences are held constant to keep constitutive expression but the surrounding nucleotides are varied modulating the expression strength. The natural pUP RBS was also kept constant.

The degenerate oligonucleotide method was chosen. As the promoter is small (86 bp), purifying the fragment after mutagenic PCR would be difficult whereas the whole promoter could be cheaply ordered as a single degenerate oligonucleotide. The oligonucleotide was ordered with 86 bp of the degenerate promoter sequence followed by 29 bp of GFP coding sequence to prime the template. A NotI site was added to the 5' end of the primer. The reverse primer was designed to bind upstream of the promoter on the template and also included a NotI site. The pUCG16+pLdh+sfGFP construct was used as the template and the whole construct was amplified, exchanging the Ldh promoter for the pUP library. The linear product was digested with NotI and self-ligated to reproduce a circular plasmid. The ligation was transformed into *E. coli* DH10B and approximately 1000 colonies scraped from plates for plasmid purification. This amplified library was then transformed into *G. thermoglucosidans.*

48 colonies with a range of apparent fluorescent outputs were selected for further characterisation with a bias for selecting strongly fluorescent colonies. Geobacillus cultures were grown, part saved as glycerol stocks for later characterisation and part used for plasmid preparation. Library member plasmids were then retransformed into *E. coli* and colonies grown up in liquid cultures with part saved for glycerol stocks and part used for plasmid preparation. These preparations were then sent for sequencing. Library members that failed to transform in or give plasmid preparations at any stage were discounted. After sequencing, duplicate sequences were removed, as were sequences with insertions or deletions. After characterisation, members with no detectable expression or expression equal to pUPWT in either species were also removed. This left only a small remaining library (sequences in Table 4.1). Expression from the promoters was characterised by reading sfGFP fluorescence using (and assessing) different characterisation methods.

Name	Sequence (after the NotI site up to the start codon)
pUPWT	GTGTTTTTTTGTTTGCGG**TTGAAT**ATATGCGTATTTTTCGG**TAGAAT**TTATGGAAGTGTAACCCGTAACAAAGAGGAGAGTGCAAAATG
pUP1	GTGTTTTTTTTGCTTGAC**TTGAAT**GTCTCTAGATGACATAC**TAGAAT**CGAAtCCCGGGATTTCCGTAACAAAGAGGAGAGTGcAAAATG
pUP2	CGGCCGCATAGCTCATAG**TTGAAT**TTCTGTTATAATGTTAG**TAGAAT**TATTTTGAGTGGACCCCGTAACAAAGAGGAGAGTGCAAAATG
pUP3	GTGTTTTTTTGTCAATGA**TTGAAT**GATACCGATGTTTGTAA**TAGAAT**GGTGTTTAGGAAAAGCCGTAACAAAGAGGAGAGTGCAAAATG
pUP4	GTGTTTTTTTAGCACAGT**TTGAAT**TTACATCTCCATTGTAA**TAGAAT**AAAATTATCCTACGACCGTAACAAAGAGGAGAGTGCAAAATG
pUP5	GTGTTTTTTTAGTCACAT**TTGAAT**ATTAGTCGGTGAGCTGT**TAGAAT**ATCAGACGAAGACATCCGTAACAAAGAGGAGAGTGCAAAATG
pUP6	GTGTTTTTTTACTATTTA**TTGAAT**CTTCATGTGACAATGGG**TAGAAT**AAATGGATAGCAGAACCGTAACAAAGAGGAGAGTGCAAAATG
pUP7	GTGTTTTTTTGTTGGGAG**TTGAAT**AGCGAAGTGAAATGCAG**TAGAAT**GTTAGTGCAGGGGGGCCGTAACAAAGAGGAGAGTGCAAAATG
pUP8	GTGTTTTTTTTTGCTTTG**TTGAAT**CGCTTGACCGTGGACAT**TAGAAT**GAGAACGGGGGAGAACCGTAACAAAGAGGAGAGTGCAAAATG

Table 4.1. Sequences of pUP promoter library promoters the -10 and -35 box sequences (in bold) were kept constant whilst bases around them were randomised to vary expression from the promoters

4.2.2 Methods for Characterising Promoter Strength

Promoter strength data is vital for rational design of genetic circuits. Protocols for the measurement of promoter strength are varied, however, with no agreed standard. The objective unit for promoter strength, polymerases per second (PoPS) is hugely challenging to estimate and so promoter strength is usually reported in arbitrary units or strength relative to a standard promoter (182).

Estimating promoter strength using a reporter protein requires a complex model but by considering only relative strength and choosing the appropriate reporter this can be hugely simplified. sfGFP is the reporter of choice because it has a very fast maturation rate (under 10 minutes) and is highly stable. Assuming no degradation of the reporter, that it is only diluted, allows promoter strength to be simply estimated from cell density and fluorescence measurements (models in (178,183) vs. (182)).

There are still two different methods for determining promoter strength, simple endpoint measurements or synthesis rate calculations from time course measurements. Endpoint measurements are simple and consider GFP per cell once the culture reaches stationary phase. Synthesis rate measurements consider the maximum rate of change of GFP per cell over a particular time interval, usually 1-hour (182).

$$\text{Synthesis Rate} = \frac{\text{GFP}(x)_{t1} - \text{GFP}(x)_{t2}}{\text{OD600}(x)_{average}}$$

The maximum synthesis rate occurs in mid exponential phase for most promoters and so time course measurements with GFP and OD600 readings taken at short intervals across the whole growth curve must be obtained to calculate the maximum rate. The methods and relative advantages are illustrated with an example promoter in Figure 4.2. Results for sfGFP expression from the pUP1 promoter on the pUGC16 plasmid in three *E. coli* cultures are shown and analysed by either the endpoint or synthesis rate method.

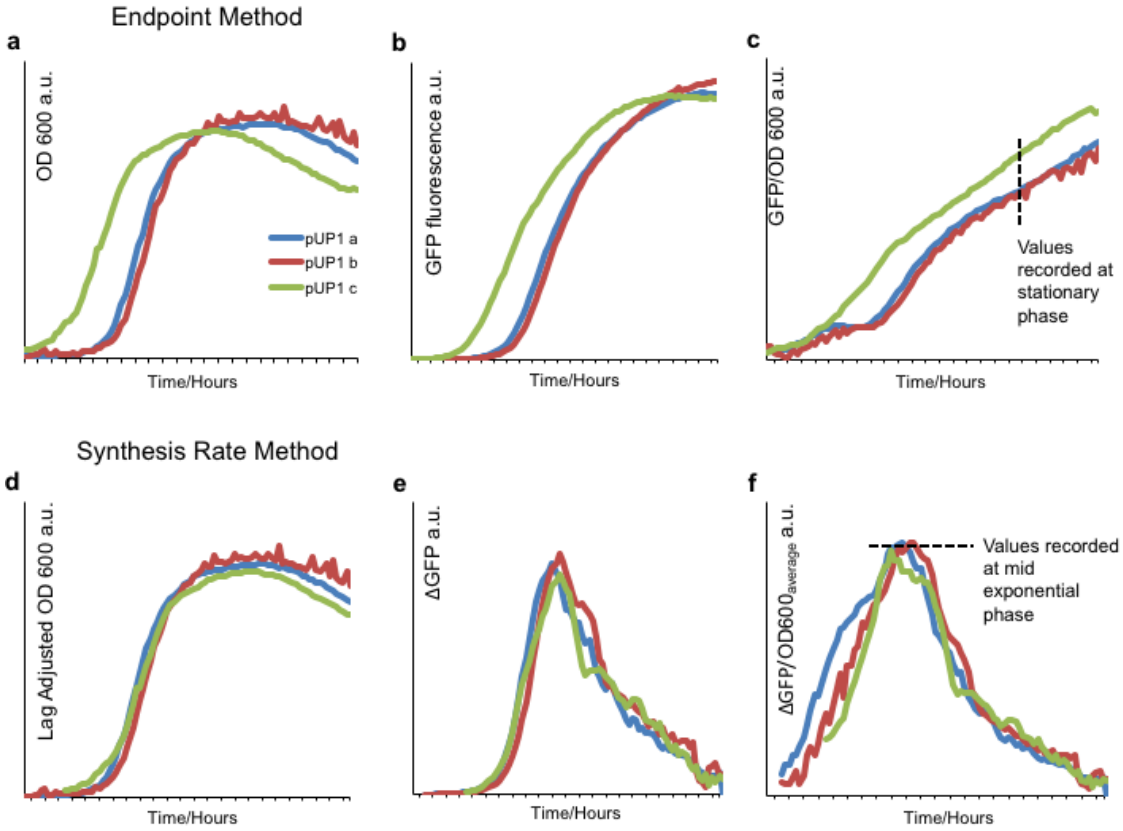


Figure 4.2. A comparison of the two methods for estimating promoter strength from GFP fluorescence and OD600 data, the graphs all show plate reader data from the same three biological replicates of the pUP1 promoter in *E. coli*, all x-axes are time with y-axes labelled. a-c) Data for estimating promoter strength by the endpoint method. GFP/OD600 readings are taken once the cultures reach stationary phase. d-f) Data for estimating promoter strength by the synthesis rate method. By taking OD600 readings over time the growth curves can be shifted to account for lag. Change in GFP fluorescence for each 1-hour period is calculated, ΔGFP = GFP(t) – GFP(t-1). Synthesis rate per cell is then ΔGFP divided by the average OD600.

Significant inaccuracies in the endpoint method can be caused by lag in the growth curves as shown in Figure 4.2a. All three replicates were a 1:100 dilution from the same outgrowth but the lag before exponential phase still differs. Replicate pUP1 c starts growth earlier and so promoter strength is overestimated when taking the endpoint reading (Figure 4.2c). Differences in growth curves between promoters due to burden

also cause inaccuracies. Lag can be accounted for by taking time course measurements and shifting growth curves during data analysis or it can be reduced experimentally by having a shorter (~5 hour) outgrowth step then re-diluting the cultures down to the same OD values before measurements are taken. This requires time course measurements or an additional complex liquid handling step that may not always be possible with the particular organism being studied. Despite these limitations the endpoint method is widely used due to its simplicity (184,185). Calculating synthesis rate is particularly difficult for non-model organisms that may be difficult to grow in multi-well plates and therefore monitor throughout their growth using a plate reader. Flow cytometry gives more sensitive data than plate reader measurements and also gives information on the population distribution. However, determining synthesis rates by flow cytometry is particularly challenging and so endpoint measurements are preferred for this method of data collection (186).

On the whole, synthesis rate calculations are far more accurate and reproducible than endpoint measurements which only provide an estimation of strength over the whole growth curve. Synthesis rates can be correlated to PoPS or GFP molecules produced per cell per second. Rate can also account for lag and burden and is now becoming the best practice for characterisation in synthetic biology (182). Beyond the major limitation of requiring time-course measurements, the shorter time window used when calculating rates makes the results more sensitive to instrument error and fluctuations (Figure 4.2e, f).

In this work, attempts were first made to achieve time course measurements for *G. thermoglucosidans.* However, the high temperature growth requirement proved a significant obstacle. A high temperature plate reader was not available and culturing Geobacilli in large well microplates was found to be problematic due to condensation and reduction in oxygen availability (and hence GFP fluorescence). Growing *G. thermoglucosidans* in flasks and manually sampling every 15 minutes was also unsuitable, as it caused fluctuations in growth temperature due to regular sampling. Results for promoter strength estimation performed via the endpoint method were found to be similar to synthesis rates whenever these were successfully taken (Figure 4.3) and so endpoint measurements were used for all further characterisation.

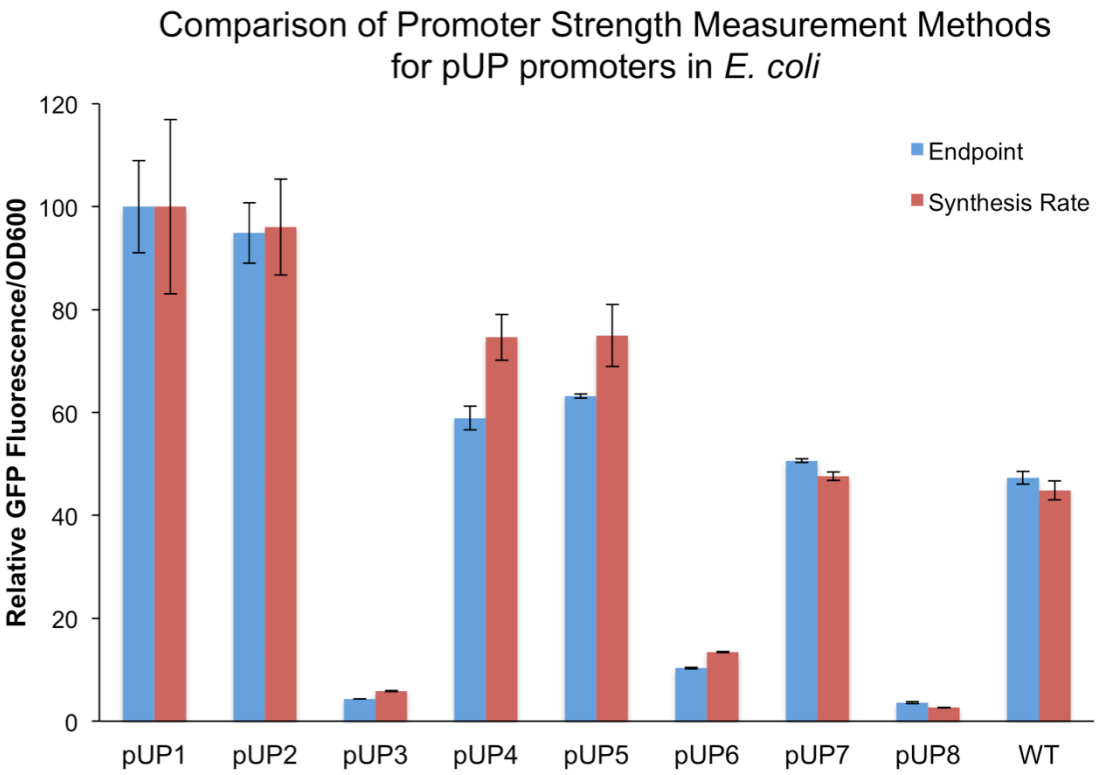


Figure 4.3. Comparison of values for promoter strength in *E. coli* determined by lag adjusted promoter synthesis rate over 1 hour in mid exponential phase or from early stationary phase endpoint fluorescence readings. Data are normalised to pUP1 at 100%. Promoters were named in rank order based on strength in *G. thermoglucosidans* (Figure 4.4) and this rank does not correlate well with the *E. coli* data shown here.

For the small pUP library expressed in *E. coli* the characterisation results from the two methods, endpoint and synthesis rate, appeared to be broadly similar. Differences were significant but the rank order of promoter strengths was preserved. Therefore, for this study endpoint measurements via plate reader will be used to characterise promoter strength. Samples will also be analysed by flow cytometry to verify that they are single populations. This will help identify promoters that have stochastic activity with large natural variability between cells or cases where mutation and/or burden has caused a loss or deletion of the construct. For optimising systems in synthetic biology, low variance, single populations are desired and flow cytometry helps to identify these characteristics.

4.2.3 pUP Promoter Characterisation

The pUP library was characterised in *G. thermoglucosidans* at 55 °C in 2SPYNG media. Results from plate reader endpoint fluorescence measurements are shown

(Figure 4.4a), normalised to pUP1 at 100%. Cultures were also analysed by flow cytometry to assess the cell population (Figure 4.4b).

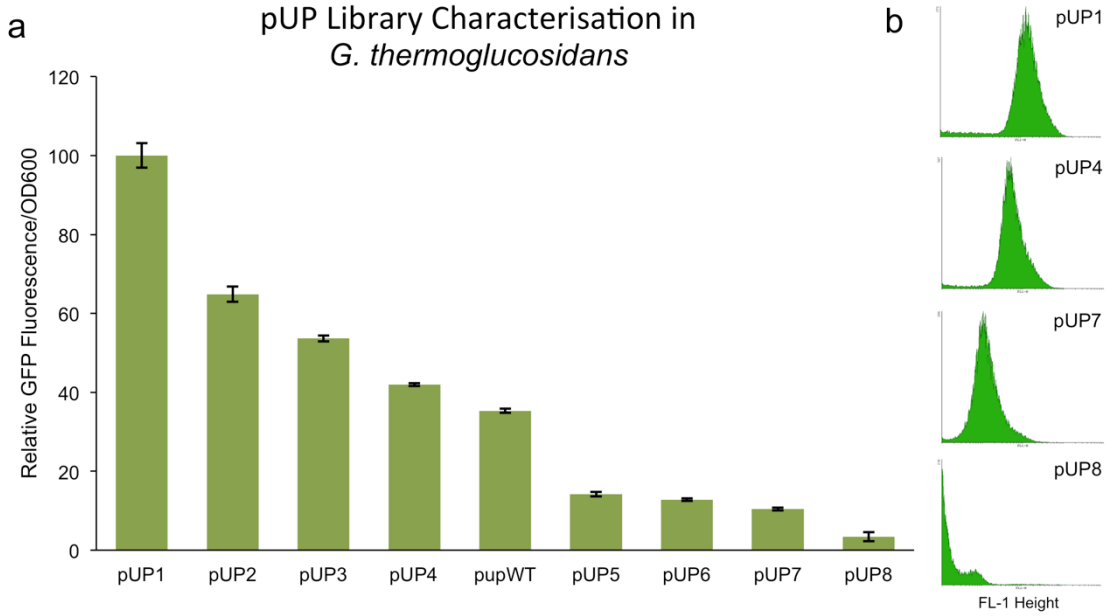


Figure 4.4. pUP library characterisation in *G. thermoglucosidans.* a) Average population fluorescence levels. Endpoint cultures were measured, Results were normalised to pUP1 at 100%, error bars show standard deviations for 3 biological repeats. **b) Flow cytometry data, selected histograms for four promoters show population distributions.** x- axes are FL-Height, the detected emission strength at 530nm after excitation with a laser at 488nm. y - axis are cell count.

From the data, we can see that several promoter library variants were able to improve on the strength of the wild type promoter. These variants were specifically selected for when picking colonies but the strength improvement of over 2-fold when characterised is quite striking. This suggests pUP has not evolved to be high strength and can be mutated to give greater strengths. Indeed even the improved pUP1, which is double the strength of pUP is still under half the strength of pLdh, suggesting that there is room for finding stronger promoters. To this end, efforts were turned to finding a stronger starting promoter for further mutation (section 4.2.3).

As well as repeatable endpoint plate-reader data, flow cytometry data was also taken for the pUP library and showed single populations for promoters pUP1 - 7 (examples in Figure 4.4b). However the weakest member pUP8 showed greater variance than the rest of the library, possibly indicating some stochastic effects and/or a sub-population of cells not expressing GFP from the promoter. The range of expression between pUP1 and pUP7 proved to be only one order of magnitude and so this library would

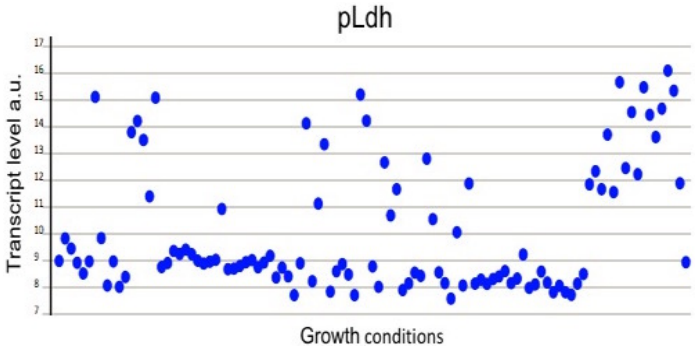
unfortunately not allow tuning of expression strength over a wide range. The library is also quite small in terms of numbers and so an improved library was next made (see below). While this is a small library, the pUP library promoters are still helpful parts for Geobacillus research and as such they have been already shared and used in other studies (86,106). The promoters are almost totally synthetic so do not risk recombination. They are also small so can easily be added to constructs by PCR with oligonucleotides encoding their sequences.

4.2.3 Stronger Constitutive Promoters

Alternative natural Geobacillus promoters were next reviewed as choices for a stronger promoter library. Candidates from this review are summarised in Table 4.2. As no large scale expression data has been published for *Geobacillus* species as of yet, data for *B. subtilis* was used, particularly the transcriptomic analysis by Nicolas *et al.* (76). *B. subtilis* transcription was profiled under a huge range of conditions including varied media, carbon sources, oxygen conditions, growth phases, sporulation/germination, heat, cold, salt, antibiotic, ethanol, and oxidative stress. The data set was made freely available and was mined to inform this study. Nicolas *et al.* note that graphs of expression data for all annotated promoters were generated for the '*subtiwiki'* database (187) and so these graphs were reproduced for the candidate promoters and are shown in Table 4.2 below.

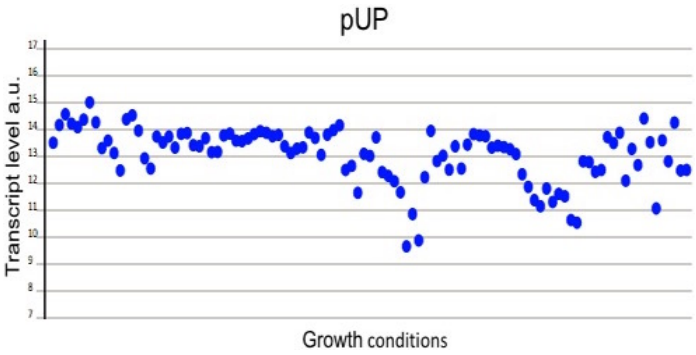
These data allow useful comparisons of relative strengths and constitutive vs. conditional expression. All expression values are from microarray datasets and the scale is non-linear with all values falling between approximately 7 and 17 arbitrary units. As all of the genes transcribed by these promoters are well conserved across the Bacillaceae family, similar expression profiles could be expected for the below genes in *Geobacillus* species.

Promoter Expression Level Graphs

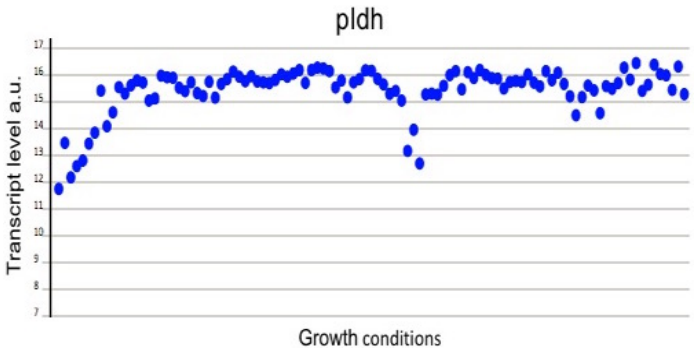


Notes

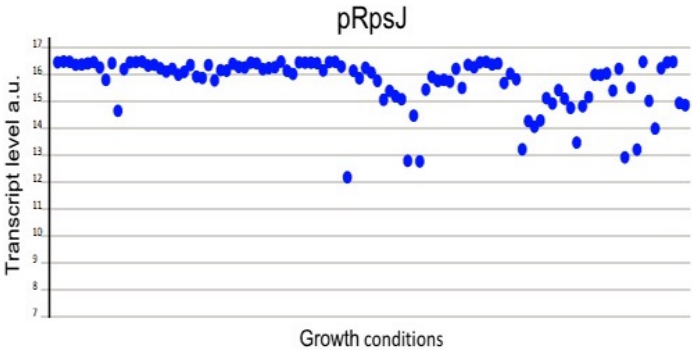
pLdh, Lactate dehydrogenase promoter.
Strong but not constitutive. Highest in LB, aerobic conditions. Low in alternative carbon sources, after glucose exhaustion or when stressed by heat, cold or ethanol.



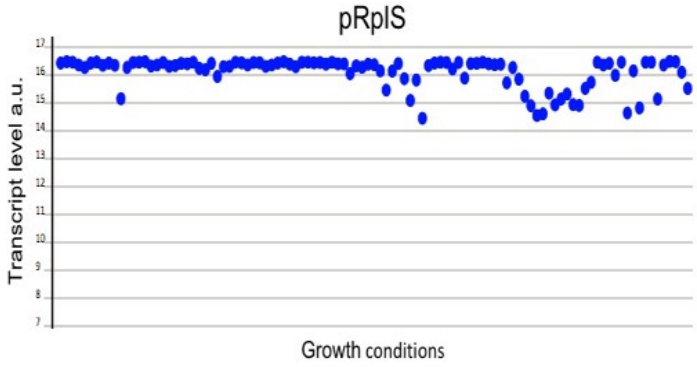
pUP, Uracil phosphoribosyltransferase promoter.
Quite strong and constitutive. Low during sporulation and stationary phase. Highest during exponential growth.



pIdh, Isocitrate dehydrogenase
Strong and quite constitutive. High even in stationary phase, lower under certain stresses (antibiotics, oxidative stress) not significantly affected by heat, cold and ethanol.



pRpsJ, promoter of the largest ribosomal protein operon.
Very strong, very constitutive. Highest in exponential growth, lower in stationary phase and sporulation. M17 type.



pRplS, Ribosomal protein promoter.
The most constitutively expressed gene in the study and the highest average expression. Only slightly lowered in stationary phase and sporulation.

Table 4.2. Comparison of promoter candidates using *B. subtilis* transcriptomics data from Nicolas et al 2012 (76). Graphs show estimated transcript levels (arbitrary units) on the y-axis under different growth condition along the x-axis. All the above genes are highly conserved between *B. subtilis* and *Geobacillus* species and so their transcription is also likely to be conserved.

The highly conditional nature of pLdh expression can be clearly seen from the Nicholas *et al.* data. This makes it a poor choice for industrial use as conditions in bioreactors can fluctuate. It also makes it unsuitable for synthetic biology applications as it would be difficult to accurately characterise. The *Bacillus* version of pUP is seen to be far more constitutive but not as strong as the other promoters shown in the table. pIdh has a similar profile to pUP but is stronger. However, the strongest promoter averaged across all conditions in this data set is the RplS ribosomal protein promoter followed by other ribosomal protein promoters with pRpsJ shown in the above table. Other strong candidates not shown here included elongation factors and other housekeeping genes such as adenylate kinase. These had similar expression profiles to pIdh. The isocitrate dehydrogenase promoter (pIdh) was reported to be strong and constitutive in *G. thermoglucosidans* (Dr. Alex Pudney, TMO Renewables Ltd. personal communication) and so this was selected for further testing. Of the ribosomal promoters, pRplS and pRpsJ were highly conserved between *B. subtilis* and *G. thermoglucosidans* and clearly annotated in the *G. thermoglucosidans* genome sequence (98). For these reasons these two were also chosen for further testing.

Primers were designed to amplify the three selected promoters as the 200 bp of DNA directly upstream of their gene's start codon on the *G. thermoglucosidans* DL44 genome. Promoters were cloned into pUCG18 expressing the sfGFP reporter, plasmid DNA was prepared from *E. coli* then transformed into *G. thermoglucosidans* for characterisation. Three promoters were then chosen for further characterisation. The *G. stearothermophilus* Ldh promoter is the strongest previously reported promoter used in

G. thermoglucosidans. Two variants were tested, the sequence with the promoters natural RBS as used by Cripps *et al*. 2009 (60) and the variant with an alternative RBS sequence from Bartosiak-Jentys et al. 2012 (92). Here the natural Ldh promoter's RBS was replaced with the RBS from the G. stearothermophilus PheB gene and this combination was shown to give very strong protein expression. When tested with the sfGFP reporter this alternative RBS was seen to increase expression around 2.5 fold and so was also used to replace the natural RBS of the other promising candidate promoters pIdh and pRlpS (Figure 4.5).

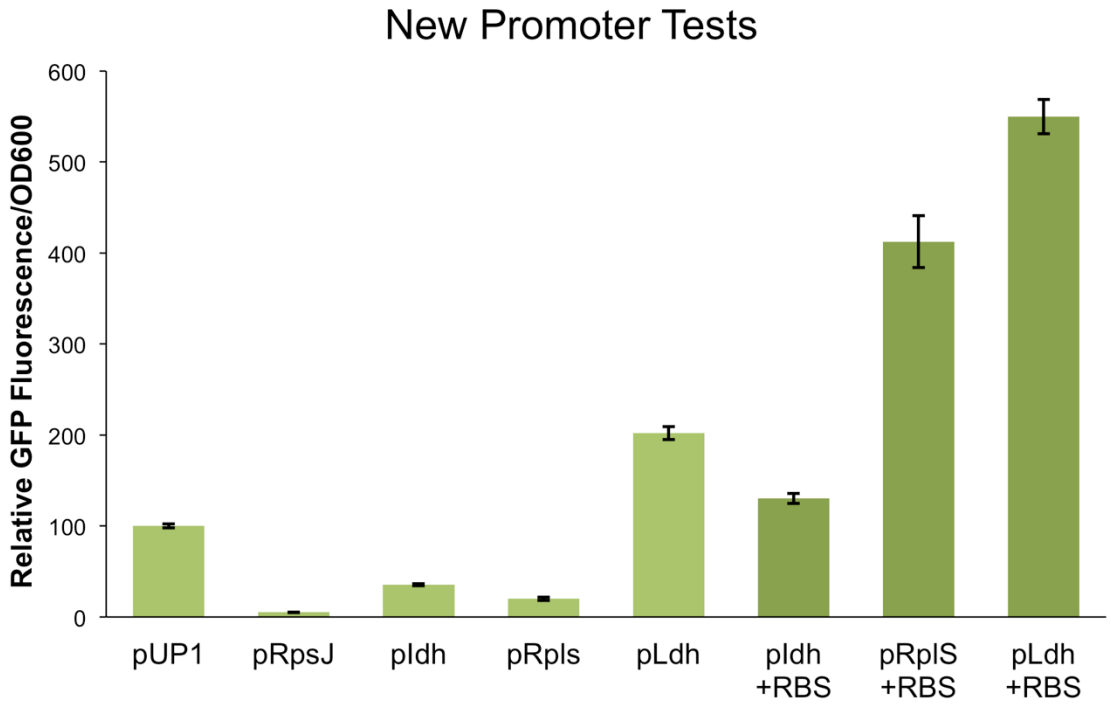


Figure 4.5. Characterisation of alternative promoters in *G. thermoglucosidans* bu plate reader fluorescence measurements.

From the endpoint characterisation data (Figure 4.5), the RpsJ promoter was found to be unexpectedly very weak; fluorescence detected was barely above background. This suggests the promoter region cloned for this experiment may have been incorrect, and instead been DNA that was miss-annotated in the genome. RpsJ is predicted to be the first protein from a large, complex operon of 21 ribosomal proteins but may actually be in the middle of an operon transcribed by the promoter of an upstream gene. It was dropped from further investigation.

The RplS promoter was also expected to be very strong, however initially it seemed to be weak. Analysis of the sequence using the RBS calculator software to predict translation initiation rate suggested only a very weak RBS. Addition of the strong RBS from the *G. stearothermophilus* PheB gene (92) greatly increased expression (all sequences available are in Appendix section 1). As seen in Figure 4.5 the new RBS greatly increases expression strength in all promoters, particularly pRplS. The reason for this may be because the genome sequence around *rplS* could also be miss-annotated. An alternative in-frame start codon exists downstream of the predicted promoter and translation may occur from this instead of the first start codon. This would have caused this natural RBS not to be included in the 200 bp sequence selected and amplified. Once a strong RBS is added, pRplS+RBS gives both strong and constitutive expression. These are the properties desired for generating a further promoter library. pLdh+RBS is also significantly stronger under these conditions: (aerobic growth in rich media at optimum growth temperature) but as the *B. subtilis* data suggest (Table 4.2) it may only be strong only under some conditions like those of this experiment and would likely have much weaker expression in other conditions.

4.2.4 An RplS Promoter Library

The pRplS sequence selected is longer than pUP and lacks an obvious consensus -10 and -35 box. Because of these two features, it was determined that the mutagenic PCR method would be preferable for library generation. The wild-type genomic RplS promoter (pRplSWT) sequence was amplified by PCR with nucleotide analogues 8-oxo-dGTP and dPTP included in the reaction. 8-Oxo-dGTP can mispair with A, leading to A-to-C and G-to-T transversion mutations whilst dPTP, in combination with 8-Oxo-dGTP can cause both transition mutations (A-to-G and G-to-A) and transversion mutations (A-to-C and G-to-T). The number of PCR cycles was calibrated to give approximately a 10% mutation rate, meaning that around 20 mutations would be expected over the 200 bp promoter region. The library fragments were then further amplified by standard PCR with primers to add overlap with the backbone and a pUCG16+sfGFP plasmid backbone fragment with complementary overlaps was also amplified. The two fragments were then joined by Gibson Assembly. The assembled plasmid library was transformed into *E. coli*, plate-scraped, prepped and transformed

in to *G. thermoglucosidans*. 96 colonies were picked and were tested or excluded following the same conditions as for the pUP library. 20 unique promoter sequences came through the screening and sequencing rounds and these were characterised in both species by endpoint fluorescent measurement as before (Figure 4.6).

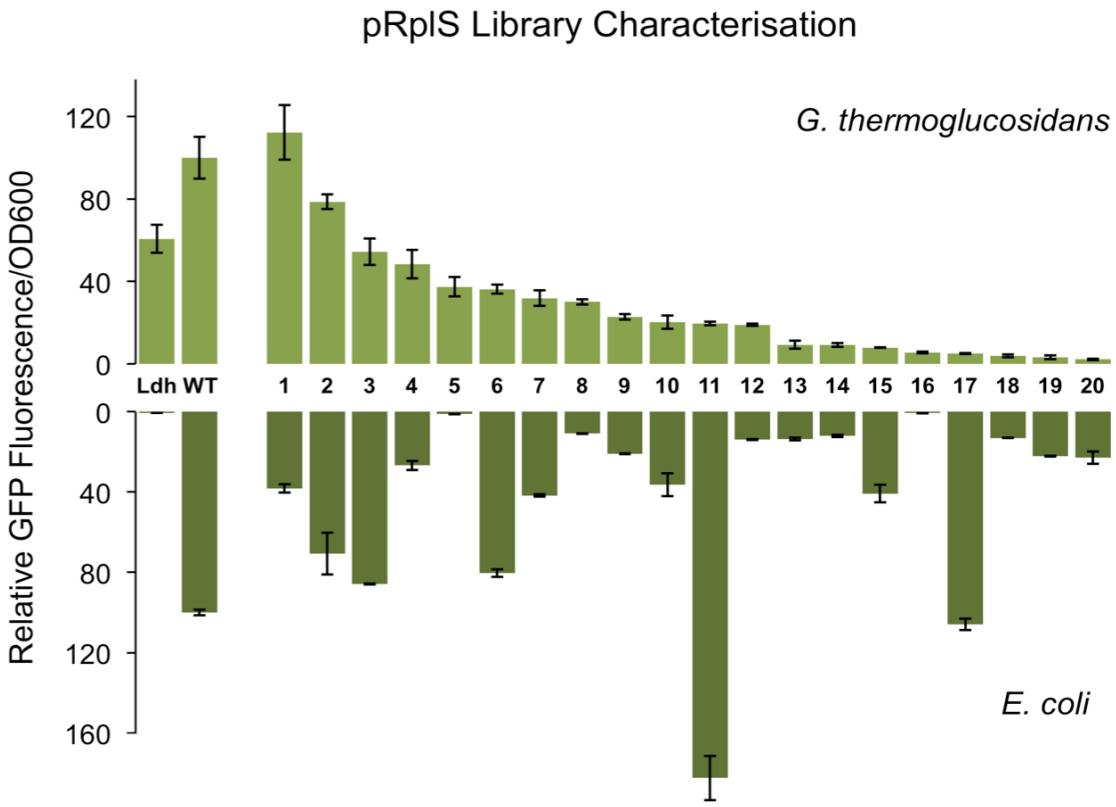


Figure 4.6. Characterisation of the pRplS library in *G. thermoglucosidans* above the axis and *E. coli* below. All outputs are normalised to pRplSWT at 100%. Error bars show standard deviations from three biological repeats.

This library proved to be a great improvement on the pUP library, spanning a 100-fold range in expression strengths. The library also includes promoter pairs with similar strengths but significantly different promoter sequences (for example RplS1 & RplWT or RplS5 & RplS6). These are valuable as these could be used within the same genetic circuit to perform similar expression levels, but without risk of recombination between each other's sequences (promoter sequences are in Appendix section 1). The pRlpS+RBS library adds a large number of new promoters for *G. thermoglucosidans* and will theoretically allow far more precise tuning of gene expression than was previously possible from the handful of published promoters.

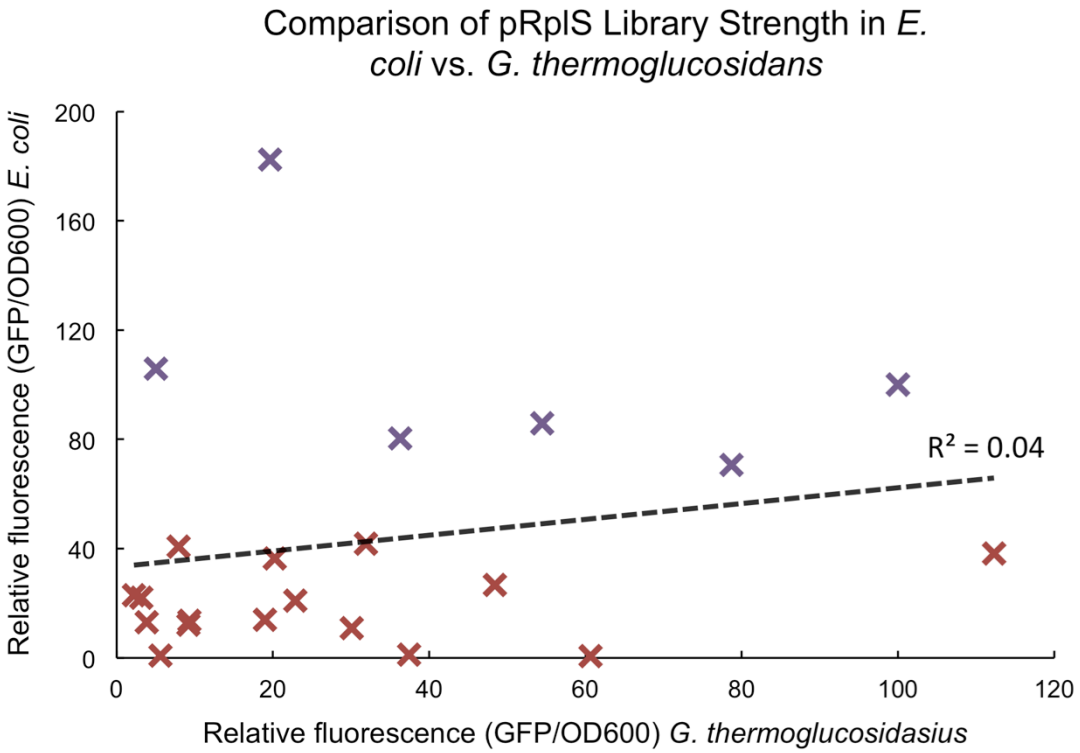


Figure 4.7. Correlation between promoter outputs in the two species. A very weak positive correlation of $R^2 = 0.042$ is seen. This may be useful for example, when cloning a construct that might cause stress to the cells, promoters with low expression in E. coli (coloured red) could be used to maintain high cloning efficiency.

Interestingly, a comparison of characterised expression levels recorded from experiments the two species (*E. coli* and *G. thermoglucosidans*) for the pRplS library promoters shows a very low correlation ($R^2 = 0.042$ Figure 4.7). Some library members that are weakly expressing in *E. coli* are strongly expressed in *G. thermoglucosidans* and vice versa. These differences between expression levels are likely to do with differences in the basic transcription machinery of the cells (e.g. sigma factors) and they can actually be of benefit. For example, the set of promoters coloured red in Figure 4.7 have low strength in *E. coli* but cover the full range of expression in *G. thermoglucosidans*. This subset could be useful for cloning certain genes, as low *E. coli* expression will reduces stresses caused to this host during cloning and plasmid propagation. This should therefore improve the efficiency of cloning of these plasmids, especially when they are being constructed to contain large or potentially toxic proteins for expression in Geobacilli.

4.3 Discussion and Future Work

The previous promoters used for metabolic engineering with Geobacillus species, variants of the lactate dehydrogenase promoter pLdh, are very strong and expression may fluctuate under different conditions. With the first promoter libraries generated for *Geobacillus* species presented here, more constitutive expression levels can now be predictably tuned. This will allow development of more complex metabolic engineering and other applications that demand sensitive tuning of expression.

4.3.1 Promoter Characterisation Methods

Since being popularised by Kelly *et al.* 2009 (182) the synthesis rate method has gradually become the best practice for parts characterisation in synthetic biology and may eventually be considered the standard. It is certainly the most comprehensive method and provides data from which the ideal unit of promoter activity, PoPS can be calculated. Currently however, the vast majority of promoter characterisation data available in the literature is from endpoint methods. The analysis performed here suggests that endpoint measurements are comparable enough to be similarly informative for future work with promoters such as those measured here, and are likely to only give incorrect values when there are problems with lag times or if growth media is exhausted too early so that a real endpoint is not reached. Biological parts are highly context dependent and so promoter strengths will be considerably different from the values obtained when expressing GFP from a plasmid when they are instead used for metabolic engineering or for building complex genetic circuits. Characterisation data is therefore currently only a guide for designing and predicting future constructs. In this regard, both characterisation methods (synthesis rate and endpoint) are equally informative at this stage in that they give us useful information but cannot capture every eventuality.

As our ability to understand and predict context effects improves (see Chapter 3 for discussion on RBS prediction), gaining the most precise possible promoter strength information may become more valuable. Higher-throughput time course data with *G. thermoglucosidans* could potentially be generated in the future with new high temperature plate readers (such as the PTI Inc. FluoDia™ T70) and with baffled

multiwell plates with breathable films that maintain oxygenation. Alternatively, the development of anaerobic fluorescent reporters (see Chapter 3) would reduce the need for oxygenation and could provide a different method to obtaining large amounts of expression measurement data.

4.3.2 pUP and pRplS Promoter Libraries

Both previously reported methods for producing promoter libraries were found to be effective in this chapter. The preference for one over the other was dependent on promoter length and promoter architecture. The pUPWT promoter was not found to be particularly strong and so is not the best candidate for a library, considering that most mutations decrease expression. It also did not yield an appropriately large library. Ideally far more colonies should have been screened to give a larger library. However, despite being a small library, with a narrow range of expression outputs, the generated pUP library promoters have still been useful to others however. Bartosiak-Jentys *et al.* used this library in *G. thermoglucosidans* with difficult to express cellulases in a 2013 paper in *Microbiology* (86). They found that the weaker promoter, pUP7 (then provisionally named pUP2n38) gave improved activity of the endoglucanase Cel5A. The pUP promoters have the advantages of being small and synthetic and also function in both *E. coli* and *G. thermoglucosidans* unlike pLdh. This is useful for checking expression from constructs in the *E. coli* cloning host before further plasmid preparation and transformation. Being functional in both these organisms - a Gram-positive and Gram-negative only very distantly related - suggests these promoters could likely function in a variety of chassis and may be useful when building broad host range constructs.

The Bacillus characterisation data from past experiments that is shown above for other promoters suggests that the ribosomal protein promoters are excellent starting candidates when searching for strong promoters for a novel chassis. If viral promoters are not available and constitutive expression is important it appears that these would be the primary option – they are highly expressed and maintain this expression in most

conditions. However, it appears that care should be taken over their annotation in the genome.

The pRplS library is the most useful parts set produced in this study and could be very valuable for future synthetic biology in *G. thermoglucosidans*. The ability to independently vary expression strength in the two hosts is advantageous and, as with the pUP library, the expression in both species suggests these promoters could be useful for broad host range applications. To help make these parts widely available to the community this library is included with a toolkit of plasmids (Chapter 6) that have been submitted for publication. Sequences are available from the NCBI database with plasmids available from Addgene.

4.3.3 Future Work

Further characterisation of promoter expression across a broader range of conditions would be valuable. Characterising redox effects on promoter strength is particularly important for promoters used in strains for bioreactor fermentations and so comparing expression with the promoters here to pLdh and other redox sensitive promoters would be interesting. This would require using the PheB enzymatic reporter should a suitable anaerobic fluorescent reporter not be developed. Characterising expression under other stresses, different growth temperatures and with different media – particularly cellobiose media or a pretreated lignocellulosic feedstock based media – would also be valuable. Improving upon the existing inducible promoters is an area worth considerable future study. Existing, natural inducible promoters taken from the genomes of Geobacillus species are not strongly inducible with a 12-fold expression change at best upon induction (79). If an alternative strongly inducible system cannot be adapted from a thermophile then existing, mesophilic systems could be adapted by evolving thermostability in the transcription factor. Improved inducible promoters would valuable for industrial overproduction of a protein of interest and could allow more complex synthetic biology devices and systems to be built such as feedback loops, timer switches and genetic logic gates (23). Such genetic systems can help the production of complex or composite products and are necessary to develop future applications in biosensing and bioremediation (188).

Chapter 5: Ribosome Binding Sites

Summary

Tuning RBS strength is vital for accurate control of gene expression. However, RBS sequences are highly context dependant so cannot be characterised independently. Instead when rationally designing novel genetic constructs, RBS sequence strength can be computationally predicted. Many tools already exist for RBS prediction and all have their limitations, but together they have so far proved to be invaluable for synthetic biology. Current tools do not account for translation in thermophiles or Gram-positive organisms but may still be useful and will be improved over time to give better predictions in a wider range of chassis.

Aims

- To investigate translation rate calculator tools and their application for synthetic biology.

- To test the utility of current translation rate calculator tools with *G. thermoglucosidans*.

- To determine limitations of current calculators and suggest improvements to better predict translation rate in Gram-positive thermophiles.

5.1 Introduction

Protein expression levels are affected by both the transcription and translation rates. Early genetic engineering approaches usually focused solely on transcription (Lipniacki, 2006) due to its heavy dependence on promoter strength and due to the relative ease of estimating the binding affinity of RNA polymerase (Alper *et al*., 2005). For synthetic biologists to have accurate, efficient and predictable control over protein production in any chassis, translation rates must also be considered.

Translation initiation is usually rate limiting in the translation process and plays a large role in determining the overall translation rate (Kudla *et al*., 2009). While other factors such as the elongation rate, termination rate and ribosome turnover also significantly affect translation (Lithwick & Margalit, 2003, Mehra & Hatzimanikatis 2006), the initiation rate is of particular interest for synthetic biology as it provides an opportunity to tune protein production over many orders of magnitude by only varying the relatively short sequence at the start of the mRNA. Modelling this step is therefore hugely valuable for designing biological systems. Software tools can potentially design synthetic RBS sequences far stronger or more predictably than was previously possible by manual design or by copying natural sequences (183,184). These tools have so far only been applied to mesophiles and so potential for use in thermophilic Geobacilli was reviewed and tested in this chapter.

5.1.1 Principles of Translation Initiation

Modelling translation initiation requires an accurate understanding of ribosome interactions with the mRNA 5' untranslated region (5'-UTR) ahead of protein synthesis. When a ribosome docks with an mRNA the 30S subunit of the ribosome binds the 5'-UTR. The 16S ribosomal RNA (rRNA) within this subunit binds to a sequence in the 5'-UTR known as the RBS (Ribosome Binding Site), while the initiator transfer RNA (fMET-tRNA) binds to the start codon (AUG) of the protein-coding sequence. The spacing between these sites on the 5'-UTR is important, with a distance of 5-8 nucleotides between the RBS and AUG being optimal (185). Within the RBS, the 3' end of the 16S rRNA is complementary to a short sequence named the Shine-Dalgarno (SD) sequence.

Factors that influence the rate of translation initiation can be grouped into three categories (Figure 5.1). Firstly, the global folding and unfolding of transcribed mRNAs; these secondary structures can hinder the access of the ribosome (De Smuit & Van Duin, 1990). Secondly, regional folding and unfolding of nucleotides in the RBS region: the ribosome docking site, a sequence roughly thirty nucleotides around the start codon, must be unfolded and exposed for the ribosome recognition sequence to bind. Lastly, there is the efficiency of ribosome binding itself, which is determined by the binding affinities between the SD sequence and the complementary 16S rRNA anti-SD sequence (Na, Lee, and Lee, 2010).

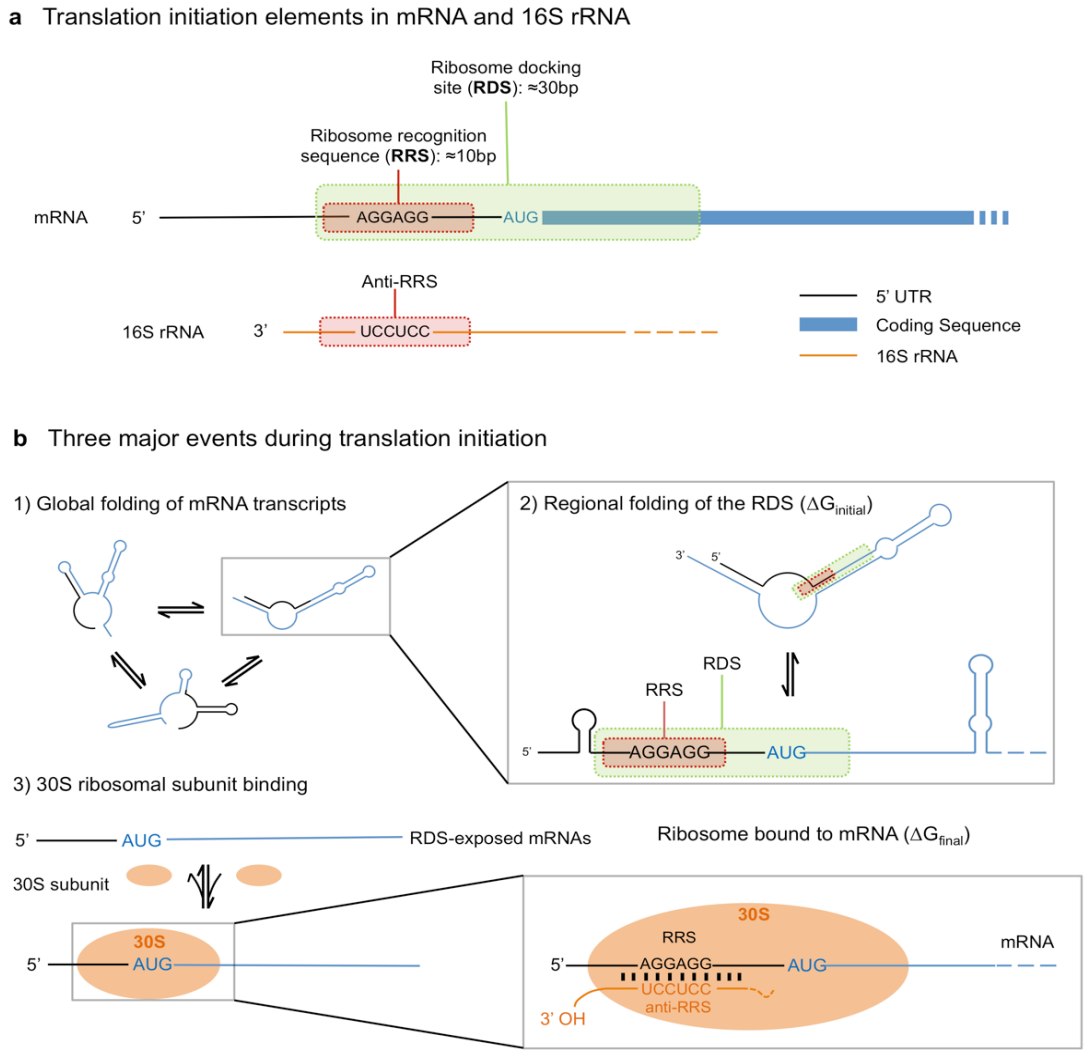


Figure 5.1. An illustration of factors affecting translation initiation, adapted from Reeve et al 2014 (184) a) The 5' untranslated region (5'-UTR) of an mRNA b) and the three major events that affect prokaryotic translation initiation. All three calculators estimate translation initiation by considering the difference in free energies between the initial state (unbound mRNA folded into secondary structures) and final (mRNA bound to a ribosome) state.

5.1.2 Translation Rate Calculator Tools

In the past decade, three different translation rate calculators have been developed to model the interactions between RNA and the ribosome at translation initiation. These were reviewed to consider their general functionality, value for synthetic biology applications and possible applicability to prediction in *Geobacillus* species. The first prediction tool, released in 2009 is the RBS Calculator (186). Next is the RBS Designer (187) and in 2013, Seo *et al.* developed the UTR Designer (188). The RBS Calculator and UTR Designer both use a statistical thermodynamic model considering free energies for key molecular interactions in translation initiation to give an estimation of translation rate. The RBS Designer makes similar free energy calculations but has a different method for calculating the translation rate. To find free energy values for mRNA secondary structures and for interactions between mRNA and rRNA, all three use one of the nucleic acid secondary secondary software suites, NUPACK (189), ViennaRNA (190) or UNAFold (191). Subsequently, all of the translation rate calculators use a proportional scale for their estimated translation initiation rate rather than any definitive units.

All three calculator models were initially designed for "reverse-engineering" – i.e. predicting the translation initiation rates and estimating protein expression from a given mRNA sequence. Each calculator also incorporates a "forward-engineering" feature, where a 5'-UTR sequence (if required) and coding sequence are inputted with a desired translation initiation rate. An algorithm is then used to generate a suitable RBS sequence to go between the 5'-UTR sequence and coding sequence to give the desired rate (or to maximise the rate) of translation initiation.

The Salis RBS Calculator and Postech UTR Designer employ similar equilibrium statistical thermodynamic models. These use free energies of the key molecular interactions involved in translation initiation. The models describe two states, an initial state in which a free 30S complex and folded mRNA strand exist, and a final state in which the assembled 30S initiation complex is attached to the mRNA. These states are separated by a reversible transition. The two states exhibit a change in the Gibbs free energy, labelled as ΔG_{total}. This is comprised of several different ΔG components, each governed by a particular aspect of the transition between the two states. The models

differ in how they calculate these ΔG components but have the same exponential relationship between the translation rate and the ΔG value for the ribosome binding transition.

The Salis lab RBS Calculator has twice been updated in, 2010 (183) and 2014 (192) to revise the method of calculating these ΔG components and to add extra functionalities including RBS library design (where degenerate RBS sequences can be analysed or designed to give a range of possible RBS strengths). Another advantage of this calculator over other software tools is the estimation of confidence that is given for calculations, and various codes are given to indicate potential inaccuracies. For example, inaccuracy may arise when there are multiple closely-spaced or overlapping start codons that could cause unpredictable ribosome-ribosome interactions. The calculator will automatically annotate results when it detects that this may be the case. In reverse-engineering these are identified to alert the user, whereas in forward-engineering these are avoided to offer the most accurate predictions. This tool is now the most popular amongst the synthetic biology community and the most highly cited.

The Postech UTR-designer is very similar to the RBS calculator and also features a UTR Library Designer that designs degenerate sequences to give translation rates across a specified range. Unlike other calculators however, the UTR Designer can also alter the codons of the coding sequence in order to reduce secondary structures and improve the translation rate when variation of the sequence of the 5′-UTR cannot satisfy the desired expression levels.

The Dokyun lab RBS Designer model is rather different, using a probability-based translation efficiency model. The probability that a given mRNA is bound to a free ribosome is calculated based on the stability of possible mRNA secondary structures vs. the ribosome binding to the exposed ribosome-docking site; this is assumed to be proportional to the protein production level. Whereas the Salis and Postech calculators only consider the most stable unbound mRNA secondary structures this model considers a range of structures and includes very long range interactions within the mRNA requiring >300 nucleotides of mRNA sequence to be entered. This offers potentially greater accuracy but is far more computationally intensive, hence the tool must be downloaded and run locally unlike the other browser-based calculators.

When creating synthetic ribosome binding sites by forward-engineering all of the currently available calculators show similarly accurate predictions compared with the experimental data in their respective publications (R^2 values of 0.8 to 0.9). With these high levels of accuracy all these calculators can have great value for synthetic biologists who need to predictably design new, synthetic sequences. For forward-engineering the calculators can deliberately avoid motifs known to be less predictable - such as overlapping stop codons or strong hairpins - and hence display high accuracy. When used in reverse-engineering this is not the case and the calculator's predictive accuracy is far lower (183,184).

Feature	**RBS Calculator Salis 2011; Salis *et al* 2009**	**UTR Designer Seo *et al.* 2013**	**RBS Designer Na and Lee 2010**
Location	Online https://salislab.net/software/forward	Online http://sbi.postech.ac.kr/utr_designer	Locally run, available from http://ssbio.cau.ac.kr/web/?page_id=195
Software used for RNA free energy calculations	NUPACK (v1.0) ViennaRNA (v1.1 & v2.0)	NUPACK	UNAfold
Original publication citations (as of 01/03/2016)	582	36	39
Advantages	Frequently updated, gives indications of confidence, support for library design and operons	Can edit codon usage to limit unwanted secondary structures	Considers very long range interactions within the mRNA, requires 300+ bp of mRNA sequence to be entered
Disadvantages	Online servers often busy	No support for operon design	No library or operon design

Table 5.1. Comparison of available translation rate calculators

On balance the Salis Lab RBS calculator was deemed to be the most useful and so this will be tested for applicability with *G. thermoglucosidans*. First, the underlying model was considered in detail to understand its temperature dependence.

5.1.3 The Salis Lab RBS Calculator Model

The RBS Calculator software available at the Salis lab website is continually revised to best account for the hugely complex array of factors affecting translation initiation.

However, at its heart remains a simple thermodynamic model (183,186,192). The model describes two states: an initial state in which a free 30S complex and folded mRNA strand exist and a final state in which the assembled 30S initiation complex is attached to the mRNA. These states are separated by a reversible transition and the change in the Gibbs free energy for the transition to the final state labelled as ΔG_{total} (Figure 5.2). This value determines how favourable the binding of a ribosome to the mRNA is and so can be used to predict translation initiation rate and thus give a good approximation of the protein expression rate.

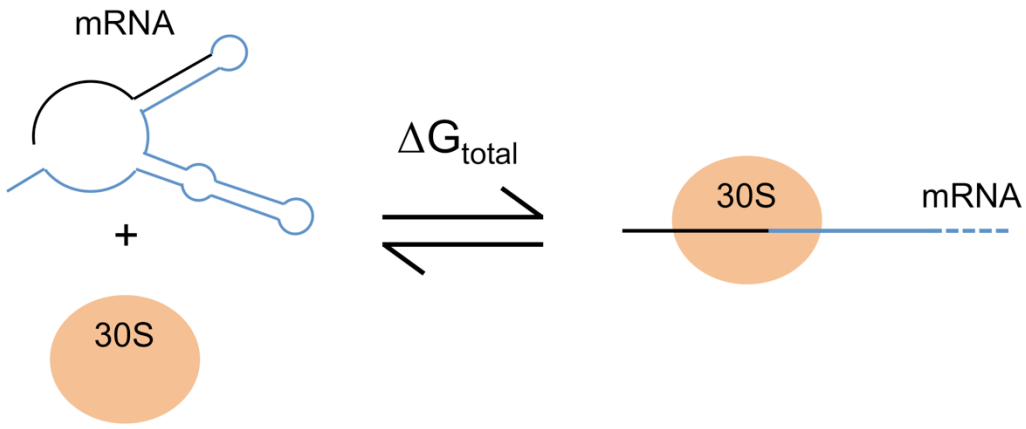


Figure 5.2. Simplified model used by the RBS Calculator to estimate translation initiation rate. Translation rate is considered to be a function of the Gibbs free energy change associated with unfolding of the mRNA and binding of the 30S ribosomal subunit.

ΔG_{total} is comprised of five different ΔG components, each governed by a particular part of the translation initiation process. The five components are:

- $\Delta G_{mRNA:rRNA}$, the energy released when the SD sequence hybridizes to the 16S rRNA anti-SD

- ΔG_{start}, the energy released when the start codon of the coding sequence hybridizes to the initiator tRNA

- $\Delta G_{spacing}$, the energetic penalty for compressing or stretching the ribosome when binding to the Shine-Dalgarno sequence when the start codon is not optimally spaced

- $\Delta G_{standby}$, the work required to unfold secondary structures that sequester a ribosome standby site (usually located four nucleotides upstream of the RBS)

- ΔG_{mRNA}, the work required to unfold the local mRNA sequence around the ribosome docking site when it folds to its most stable secondary structure

ΔG_{total} is related to these ΔG terms by the relationship:

$$\begin{aligned}\Delta G_{total} &= \Delta G_{final} - \Delta G_{initial} \\ &= (\Delta G_{mRNA:rRNA} + \Delta G_{start} + \Delta G_{spacing} - \Delta G_{standby}) - \Delta G_{mRNA}\end{aligned}$$

The translation initiation rate is then exponentially related to ΔG_{total} according to the simple function:

$$r \propto e^{-\beta \Delta G_{total}}$$

Where *r* is the translation initiation rate and ß is the Boltzmann factor for the system. Similarly, the total protein expression is then proportional to the translation initiation rate *r* by a constant *k* that accounts for ribosome, mRNA and metabolic interactions that are independent of the 5'-UTR sequence and parameters unaffected by translation (Salis. 2011).

5.2 Results

5.2.1 RBS Library Design for *G. thermoglucosidans*

The Salis Lab RBS Calculator does not currently account for different temperatures but does allow 16S rRNA sequences for any bacterial species to be entered. To test the calculator's predictions for *G. thermoglucosidans* a small, 4-member RBS library was designed and used for expression of the sfGFP reporter from the RplS promoter. The calculator was allowed 40 bp of sequence between the promoter and start codon to design. Sequence that would be part of the mRNA, 30 bp of 5'-UTR from the promoter

sequence (after the transcriptional start site but before the RBS), and 75 bp of the coding sequence (after the start codon) was also entered into the calculator. This allows the calculator to make secondary structure predictions. The RBS calculator v2.0 was used with the 'maximise' translation rate feature to design RBS A. RBS D was designed to have minimal detectable expression and RBS B and C were designed to have strengths between these extremes, approximately 30% and 5% of the maximum (Table 5.2).

The synthetic sequences were ordered as phosphorylated oligonucleotides (IDT) and used as one primer in PCR amplification of a pUCG16+pRplS+(G.st RBS)+sfGFP template. The reverse primer was designed to bind directly upstream in the promoter and together these primers amplified divergently, to copy the entire plasmid template. The original template was then removed by DpnI digestion and the product self-ligated with T4 DNA ligase to produce plasmid constructs containing the novel RBS sequences.

Name	Sequence	Predicted Strength, RBS calculator v2.0
G. st	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGT GATTCGAATG	6,110
A	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCAT AGAATTTATCAAGAAGGAGGTACAATAATG	24,500
B	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCAT AGAATTTATCTGAAAGGAGGTCCCACAATG	6,700
C	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCAT AGAATTTATCGAGGGGGTTTCCGGGATATG	1,680
D	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCAT AGAATTTATCAACCGCGGAGATCCCGAATG	247

Table 5.2 Natural *G. stearothermophilus pheB* RBS and synthetic library RBS sequences and predicted strength with the RBS Calculator v2.0. Sequence from the transcription start site to the start codon is shown, with RBS sequences in green.

All novel constructs as well as the previous plasmid containing the G. st RBS (as a positive control) were transformed into *E. coli,* prepped and then transformed into *G. thermoglucosidans.* Expression of GFP from these constructs in both *E. coli* and *G. thermoglucosidans* was measured as in the previous chapter and used for characterisation of the RBS strength. It is shown in Figure 5.3.

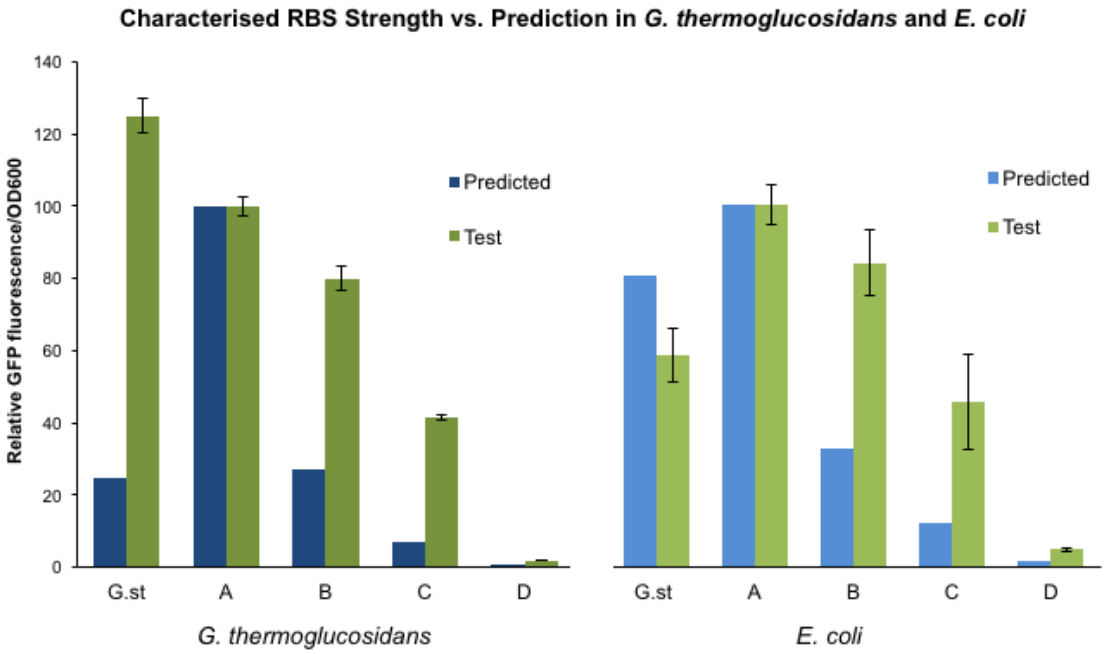


Figure 5.3. Graph of the *in vivo* strength of the natural G.st RBS and designed RBS sequences (A-D) in blue, compared to predictions of their relative strengths from the RBS Calculator software. Output was characterised as described for promoters in materials and methods (2.4.1) with endpoint fluorescence and OD measurements by plate reader. All results are standardised to A at 100% and error bars are standard deviations from three biological repeats.

The RBS calculator gives estimates of strength in arbitrary units, and *in vivo* characterisation gives relative fluorescence and so all data are standardised to RBS A, which is set at a value of 100. Overall the calculator's predictions are useful with limitations. Firstly, prediction of strength for the natural G.st RBS is largely inaccurate in *G. thermoglucosidans*. The Salis lab acknowledge that reverse-engineered predictions are less accurate than those obtained by forward-engineering, however, in this case the strength is around 5x stronger than predicted (compared to RBS A) and the predicted rank order is incorrect with RBS G.st predicted to be weaker than RBS B (when *in vivo* it is actually stronger than A and B). One possible explanation is that the spacing between the start codon and Shine-Dalgarno (SD) sequence can have a large impact and spacing requirements being different between Gram-positives and Gram-negatives (note that the RBS Calculator was designed for Gram-negatives). However, in this case the spacing, 6 bp, is the same for the natural and synthetic RBSs (see Table 5.3) and so this is not the cause of error.

Name	Sequence
RBS G.st	5' UTR - TAGATA AGGAGTG ATTCGA ATG – Coding sequence
RBS A	5' UTR - TCAAGA AGGAGGT ACAATA ATG – Coding sequence
RBS B	5' UTR - TCTGAA AGGAGGT CCCACA ATG – Coding sequence
Maximum 16S binding	5' UTR - NNNNAA AGGAGGT NNNNNN ATG – Coding sequence

Table 5.3. Comparison of core Shine-Dalgarno sequences and spacing for the three strongest RBS sequences. The spacing is equal for all sequences. RBS G.st is further for maximum strength 16S RNA binding and contains a possible alternative GTG start codon.

The Geobacillus 16S rRNA sequence used by the calculator (ACCUCCUUU) is correct according to the *G. thermoglucosidans* genome sequence (92) and so the much higher strength of RBS G.st must be due to some feature of translation not fully predicted by the RBS Calculator. Perhaps there is significant association of the ribosome to an upstream standby site in the 5'-UTR included with the RplS promoter sequence. Designed RBS sequences from the calculator are large, being 35 bp in length, and this might push this possible standby site too far upstream from the start codon to exert its effect. Alternatively, there is some evidence to suggest that if the binding between the RBS and 16S rRNA is too strong, this actually reduces translation rate as elongation is delayed (187). This effect depends on the machinery of translation initiation and elongation and so is likely to be different between organisms and thus difficult to predict. This effect is not factored into the Salis RBS Calculator model. If this were to be the case then the designed RBSs A and B would be too close to the consensus sequence (Table 5.3), whereas RBS G.st, optimised by evolution rather than by algorithms, would have the optimum affinity between the mRNA and rRNA for maximal protein production. Another factor that may account for the discrepancy is the coupling or interference between nearby start codons. The G.st RBS includes an alternative, GTG start codon 6 bp upstream from the intended ATG start codon. Translation from this GTG is not predicted to be strong but would produce in-frame, functional sfGFP. Perhaps this start codon helps to recruit ribosomes to an upstream standby site (192) and then these slide into position over the primary ATG.

In *E. coli* the natural RBS G.st proved to be weaker than predicted. It could be that the above possible standby sites are in some way *Geobacillus*-specific or it could be to do with known limitations of the calculator such as not accounting for long-range mRNA secondary structures or interactions with other RNAs in the cell.

For the synthetic mini-library of four sequences, the prediction proved to be reasonably good; the rank order is correct and the "maximised" RBS A is indeed very strong (although disappointingly it could not beat the strong natural G.st sequence). Inaccuracies in quantitative prediction of relative strength could largely be due to biological factors not directly related to translation initiation and so very difficult for the model to predict. RBS sequences A and B are both very strong and translate a codon-optimised reporter protein transcribed from a very strong promoter on a high copy plasmid (high copy in *E. coli*, medium copy in *G. thermoglucosidans*, see Chapter 6 for plasmid details). This very high level of expression could prove to be a considerable burden to the cells, although a significant growth defect was not observed. Certainly for RBS A and also for RBS B, protein expression levels may likely be approaching the maximum possible rates. Expression from these very strong RBS sequences may be limited by the cell's expression capacity and the limits of shared resource pools (193). When expressed from a weaker promoter and/or lower copy plasmid, the relative strength predictions would likely become more accurate as the strength of A (and B) could reach much higher levels relative to C and D without being held back by these burden constraints. The strong promoter was chosen because maximising expression is a common goal for metabolic engineering applications, and so testing strong promoter/RBS combinations is useful for informing this work (see Chapter 8).

Despite the above limitations, the RBS calculator was able to rationally design a ribosome binding site sequence library with varied expression and a predictable rank order in *G. thermoglucosidans* and so, even in its current form, it can be recommended as a tool for synthetic biology in this organism.

5.2.2 Temperature Effect on mRNA Secondary Structure

Temperature affects all free energy calculations and so will have a significant effect on translation rate. In the Salis RBS calculator model, translation strength is particularly affected by the competition between, two terms: ΔG_{mRNA} (the free energy from the mRNA folding into secondary structures) and $\Delta G_{mRNA:rRNA}$ (the free energy from the

ribosome binding to the mRNA), to investigate how higher temperature affects changes in these terms, the RBS mini-library members, A, B, C and D were resynthesised to introduce a hairpin in the 5'-UTR. This hairpin makes ΔG_{mRNA} more negative and would be expected to significantly reduce translation rate. Sequence upstream of SD sequence was edited to contain the reverse complement of the core SD sequence and form a stable hairpin. The core SD sequence was kept constant to keep the $\Delta G_{mRNA:rRNA}$ term constant. These hairpins, shown in Figure 5.4 with full sequences in Table 5.3, were predicted by the RBS Calculator to reduce translation efficiency by approximately a factor of 10.

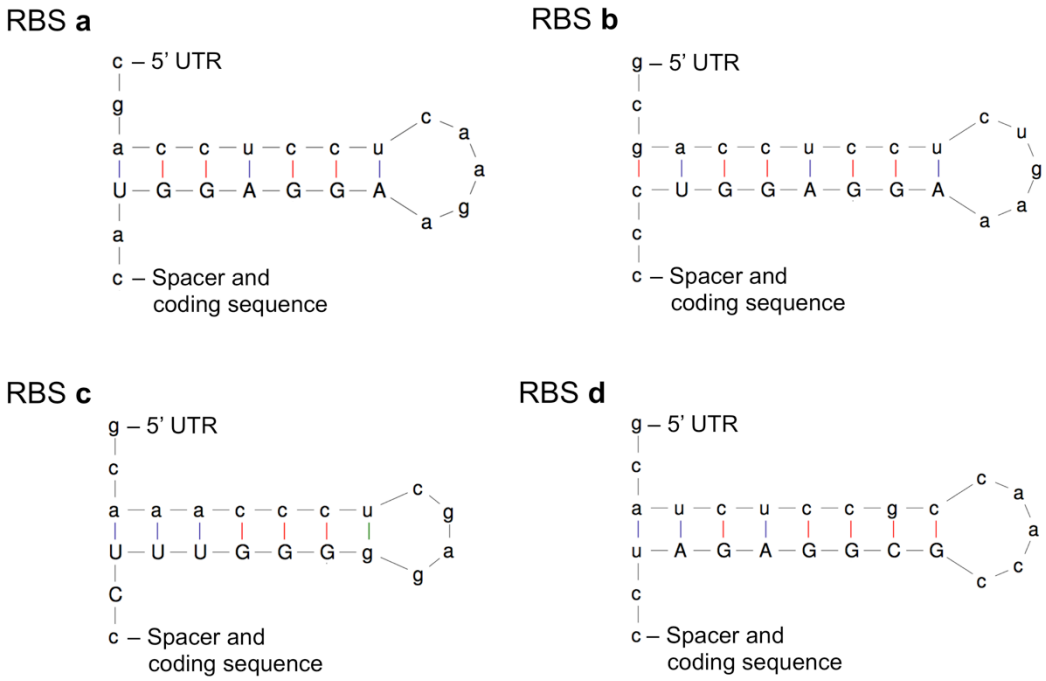


Figure 5.4. Secondary structure predicted by UNAfold (191) for the new synthetic library RBSs a, b, c and d around the ribosome recognition sequence. The Shine Dalgarno sequence (capitalised) is blocked by a hairpin formed with sequence upstream in the 5'-UTR.

Name	Sequence	Predicted Strength, RBS Calculator v2.0
a	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCATCGACCTCC TCAAGA**AGGAGGT**ACAATAATG	2,100
b	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCAGCGACCTCC TCTGAA**AGGAGGT**CCCACAATG	727
c	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCCCGCAAACCC TCGAGG**GGGTTTC**CGGGATATG	8
d	TGTTCCGCTGCAATGATGAAAAAGCATTGTCTCATAGAGTCAGCATCTCCG CCAACC**GCGGAGA**TCCCGAATG	15

Table 5.3 Synthetic library RBS sequences with reverse complement Shine-Dalgarno sequence added to create a hairpin. Sequence from the transcription start site to the start codon is shown, with the designed RBS sequence in green. Core SD sequence is in bold. Translation rate values predicted with the RBS Calculator v2.0 are shown. Predicted strength predicted below 100 units is effectively no expression.

Sequences were ordered as phosphorylated oligonucleotides and constructs were generated by PCR as with the first library. These were then transformed into *E. coli* and *G. thermoglucosidans* for characterised as before. Relative RBS strength was calculated as before from the fluorescence data obtained in order to determine the effects of these introduced secondary structures at different temperatures (Figure 5.5).

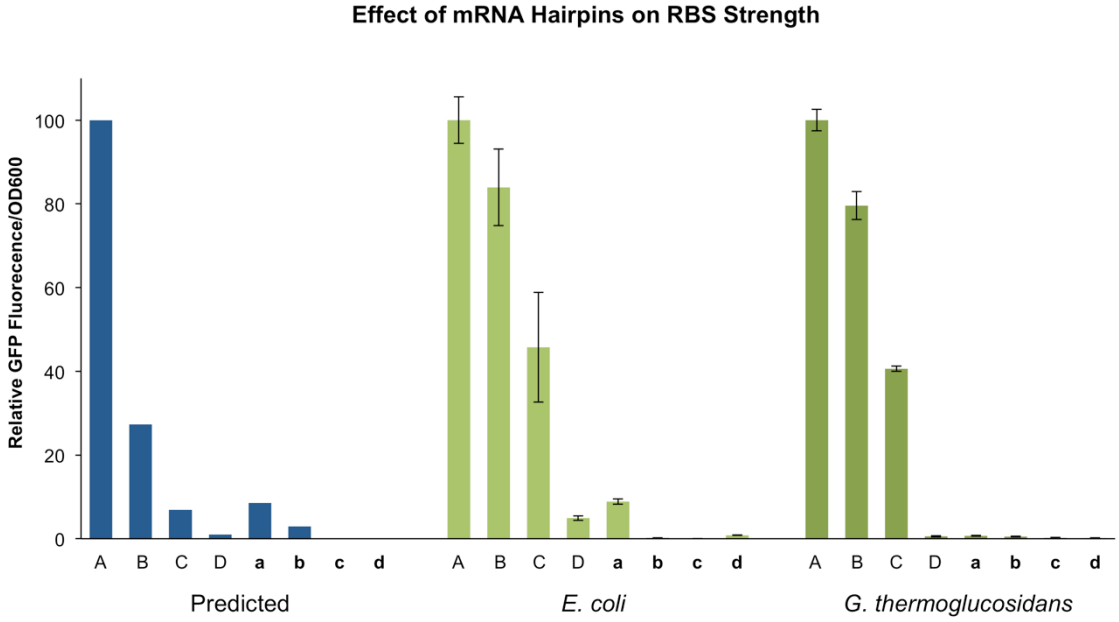


Figure 5.5. Graph of relative RBS strength for library sequences (A-D) and library sequences with added hairpins (a-d). *In vivo* data from both species is shown, predictions for these sequences were very similar when imputing either *G. thermoglucosidans* or *E. coli* 16S rRNA sequence and so only G. thermoglucosidans 16S rRNA predictions are shown for simplicity. Output was characterised as described previously. All results are standardised to RBS A at 100 and error bars are standard deviations from three biological repeats.

The negative effects on translations strength from introducing hairpins are generally underestimated by the RBS Calculator for both species. For *E. coli,* the RBS Calculator predicts **a** will be stronger than C and **b** stronger than D, which is not the case. Only **a** allows significant expression from the sequences containing the hairpins. In *G. thermoglucosidans* the secondary structures abolish sfGFP expression completely. Strong, local secondary structures affect ribosome access to the SD sequence and also alter the global folding of the mRNA in less predictable ways. The more accurately predicted initial RBS library was designed by randomly varying nucleotides in and around the SD sequence; this impacts the more predictable mRNA/rRNA binding strength more significantly and hence predictions were better. The predictions were far more accurate for *E. coli* and whilst this could in part be due to differences in translation

machinery, the free energy model used suggests temperature is a significant factor. Future modifications to the RBS Calculator to account for temperature would therefore be hugely valuable and improve predictive accuracy in thermophiles.

Interestingly, one might expect that the introduced hairpins would have less of an effect on translation rate in *G. thermoglucosidans* as they would melt open more easily at higher temperatures, in fact the opposite was observed. Depending on the calculation of ΔS, the model, as discussed below, could have predicted this.

When considering Gibbs free energy, a reaction or interaction is favoured if it has a negative Gibbs free energy change (negative ΔG). For a system at constant temperature and pressure separate from, but thermally connected to, its surroundings (both reasonable assumptions in this case), temperature affects ΔG according to the Gibbs energy equation:

$$\Delta G = \Delta H - T\Delta S_{internal}$$

Where ΔH is the enthalpy change of the interaction and $\Delta S_{internal}$ the entropy change of the internal subsystem (not including the surroundings). Temperature is in Kelvin and so the difference between 37 and 55 °C (310 to 328 K) is relatively small (~6%).
The relationship between translation initiation rate and ΔG is, however, exponential: $r \propto e^{-\beta \Delta G_{total}}$ and so temperature can have a large effect on translation rate.

In the RBS Calculator model, ΔG_{total} for translation initiation is the difference between two states (shown earlier in Figure 5.2). The initial state with a folded mRNA and free ribosome ($\Delta G_{initial}$) and the final state with the ribosome bound to unfolded ribosome recognition sequence (ΔG_{final}).

$$\Delta G_{total} = \Delta G_{final} - \Delta G_{initial}$$

If we expand these ΔG terms in relation to the Gibbs energy equation and simplify we get:

$$\Delta G_{total} = (\Delta H_{final} - T\Delta S_{final}) - (\Delta H_{initial} - T\Delta S_{initial})$$

$$\Delta G_{total} = \Delta H_{final} - \Delta H_{initial} - T(\Delta S_{final} - \Delta S_{initial})$$

$$\Delta G_{total} = \Delta H_{final} - \Delta H_{initial} - T(\Delta S_{total})$$

The translation rate is exponentially related to ΔG_{total} and the more negative ΔG_{total} is, the stronger the translation rate. Base pairing of RNA sequence is a favourable exothermic reaction and has a negative enthalpy change, negative ΔH. For strong ribosome binding sites, the mRNA to 16S rRNA binding is more favourable than mRNA secondary structure folding and so ΔH_{final} is more negative than $\Delta H_{initial}$, making the total enthalpy change negative and giving a negative ΔG_{total}. ΔS_{total} is the entropy change for the binding of the ribosome to the mRNA and as the two particles coming together reduces "disorder", ΔS_{total} will generally be a relatively small negative value for this reaction. The $-T(\Delta S_{total})$ term is then positive and makes ΔG_{total} less negative. When T is larger, this entropy term is greater and so (all else being equal) temperature increase decreases translation rate.

When adding the hairpin we make the $\Delta H_{initial}$ term significantly more negative. For RBSs **c** and **d** with low ΔH_{final} his makes $\Delta H_{initial} \leq \Delta H_{final}$ and so ΔG_{total} is not negative and expression is abolished. For RBS **a**, ΔH_{final} (rRNA to mRNA binding) is still sufficiently negative to give a negative ΔG_{total} in *E. coli* and so we observe significant expression with RBS **a**. In *G. thermoglucosidans* however, the $-T(\Delta S_{total})$ term is larger - large enough to overcome the ΔH_{total} and so ΔG_{total} is not negative and we do not get expression. Accounting for temperature in the RBS Calculator may thus have predicted the lack of expression seen in our experimental results.

To determine how temperature could be considered in future work, I next investigated possible options. The RNA structure prediction software used by the different versions of the RBS Calculator (NUPACK and ViennaRNA) both allow alternative temperature inputs, so updated $\Delta G_{mRNA:rRNA}$, ΔG_{mRNA}, and $\Delta G_{standby}$, terms could be calculated from the sequence information, and these could be added to a modified version of the

RBS Calculator. Despite the Python source code from the RBS Calculator being available, rebuilding an edited calculator for prediction for *Geobacillus* species was not straightforward. As even setting-up the RBS Calculator to run on a computer proved to be a significant challenge, we determined that this idea would be a lengthy computational project and beyond the scope of this research. However, as an alternative, the owner of the RBS Calculator website, Prof Howard Salis kindly allowed us to run some sequences on an early build of his RBS Calculator 2.0 software and for this he set the temperature for ΔG value predictions to 60 °C rather than the standard 37 °C. Predicted values from this are shown in Figure 5.6.

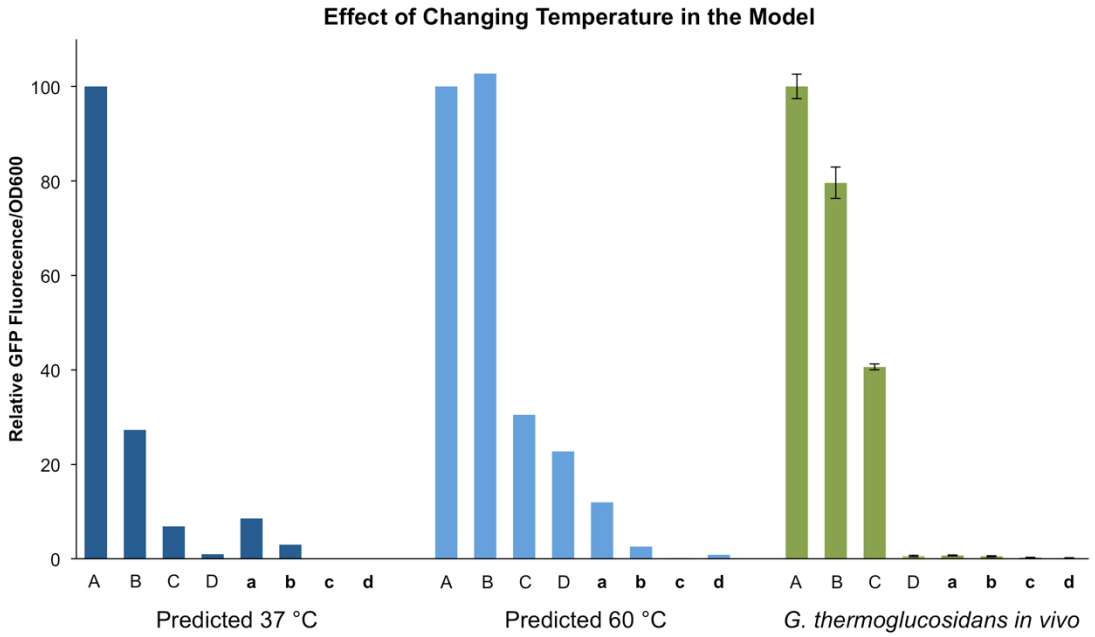


Figure 5.6. Predicted RBS strengths from the RBS calculator v2.0 software with temperature for RNA folding and ΔG calculations at 37 °C and adjusted to 60 °C. Predictions are compared to *in vivo* data from *G. thermoglucosidans.*

Unfortunately, simply changing the temperature does not necessarily improve predictive accuracy. Predictions for the relative expression of stronger sequences, A, B and C seems more accurate but for the sequences with hairpins predictions are considerable overestimates.

Given these limitations, we concluded that significant revisions to the model and relative weighting of different terms would be required to improve accuracy in *G. thermoglucosidans.* Parameters in the current model were calculated by fitting predictions to experimental data in *E. coli* at 37 °C, and this may need to be repeated

for *G. thermoglucosidans.* The model at 60 °C overestimates the output of lower strength RBSs. There could be two main reasons for this error:

Firstly, in the competition between ΔH_{final} and $\Delta H_{initial}$, the binding of the ribosome (ΔH_{final}) is aided by the non-sequence-specific affinity of the ribosome for the mRNA. This is not factored into the ΔG calculations and so is not altered by the temperature change but is accounted for by their relative weighting when calculating translation rate. At higher temperatures this could cause the effective relative favourability of the final state to be overestimated, increasing the relative rate.

Secondly and perhaps most significantly, slight errors in ΔS prediction have a greater effect at higher temperatures and indeed the RNA folding prediction programs can less accurately predict entropy values at higher temperatures (Prof Howard Salis, personal communication). For mRNAs with stable secondary structures, the binding of the ribosome could be entropically favourable if it displaces these structures. That would give a positive ΔS and so the $-T(\Delta S_{total})$ term would contribute to making total ΔG more negative. Slight overestimation of this ΔS value could account for the error seen in *G. thermoglucosidans* predictions, particularly at the higher temperatures.

5.3 Discussion

5.3.1 Future Improvements for Gram-Positive Thermophiles

The Salis Lab RBS Calculator model is a simplification of an incredibly complex process (183,184). Despite this, accuracy for forward engineering design of 5'-UTR sequences for use in *E. coli* is reasonably good, hence the RBS Calculator is widely used and highly cited. Predictions are helpful but less accurate for *G. thermoglucosidans* and so some revisions to the model would likely be necessary. The differences in translation between *E. coli* and *G. thermoglucosidans* that could be considered include:

16S rRNA Sequence

The RBS Calculator helpfully already allows entry of different 16S rRNA sequences. For organisms with highly divergent sequences this would make a considerable difference to compared translation rates. Incidentally the sequence for *E. coli* (ACCUCCUUA) and *G. thermoglucosidans* (ACCUCCUUU) are almost identical and hence predicted translation rates differ only slightly between these species.

Translation Machinery

The ribosome is aided by a collection of associated factors with various roles and the presence or structure of these differs considerably between bacterial species. Ribosomal S1 protein for example binds mRNA, stabilising the rRNA/mRNA and the interaction is also supported by various translation initiation factors. For simplicity these effects are not accounted for in the RBS Calculator. Their effects are subtle and so difficult to incorporate into the model. Hopefully future revisions will include some of the larger contributions and consider differences between Gram-negatives and positives but this is not the largest source of error and so not a priority.

Spacing Parameters

The difference in ribosomes and associated translation machinery between Gram-positive and Gram-negative bacteria is particularly apparent in the spacing requirements between the start codon and Shine-Dalgarno sequence. Gram-positives are less tolerant to deviations from optimum spacing (185). This will be accounted for in future versions of the calculator with a modified $\Delta G_{spacing}$ term (Prof Howard Salis, personal communication). Currently for strong, forward-designed synthetic ribosome binding sites, spacing is kept constant so will not affect relative strengths. For the libraries generated in this study spacing is not primarily responsible for the inaccuracies in prediction.

Temperature

At higher temperatures the effect of the ΔS term is more significant and so accurate calculation of the entropy of RNA is more important. Improvements in understanding the underlying biophysics and refitting parameters to experimental data would likely be required.

5.3.2 General Summary and Future Prospects

The current RBS Calculator is valuable for designing novel RBS sequences to predictably vary expression in *E. coli* and *G. thermoglucosidans*. Relative strength predictions are typically qualitatively correct and as the RBS Calculator is updated, more quantitative predictions will be possible. Significant modifications to the model will be required to improve accuracy for Gram-positive thermophiles but broadening the range of organisms is planned for future updates (Prof Howard Salis, personal communication).

The Salis RBS Calculator is by far the most useful and feature-packed calculator available, but a new calculator (194) EMOPEC, has recently been released. This calculator is also designed for *E. coli* at 37 °C and lacks some of the Salis RBS Calculator functionalities but claims to be more accurate (though only within 2x for most sequences).

RBS strength is highly context dependent and so design of novel genetic system for any organism should involve *in silico* RBS strength prediction if possible to inform design or flag up potential problems with expression. Alternatively, if many designs can be screened, degenerate RBS sequences can be designed to vary translation strength within a predictable margin. The data here suggest predictions with the current Salis Lab RBS calculator are valuable for these purposes in *G. thermoglucosidans*. Relative strength predictions are qualitatively correct and as the calculator is updated more quantitative predictions will be possible. Significant modifications to the model will be required to improve accuracy for Gram-positive thermophiles but broadening the range of organisms is planned for future updates (Prof Howard Salis, personal communication). For *E. coli* the accuracy of the RBS calculator's reverse engineering feature is such that translation strength for natural proteins can be predicted from the genome sequence or transcriptomics data to give interesting insights (183). This is certainly not possible for *Geobacillus* species at this stage however. Reverse engineering is generally less accurate and so significant improvements would be required.

For rough testing in *G. thermoglucosidans* the RplS promoter plus G.st RBS combination is a recommended starting point for strong constitutive expression. This combination also showed strong expression of mCherry and PheB reporters and so provides a reliable initial choice of parts likely to express most proteins strongly. Transcription can then be tuned down with alternative sequences from the pRplS library with translation tuned down by design of new RBS sequences or degenerate sequences by with the RBS calculator.

Chapter 6: Plasmid Vectors

Summary

Plasmids previously used to transform *Geobacillus* species are a mix of large natural themophile plasmids and broad host vectors from mesophiles. Many novel plasmids have also been created as fusions between these previous vectors or between parts of them. However, these vectors are generally large with limited characterisation data available. Inspired by synthetic biology principles and previous modular vectors designed for non-standard chassis organisms, a modular plasmid set was designed for *Geobacillus* species. Minimal modular parts were designed and constructed into shuttle vectors and characterised. A collection of these vectors, named "The Geobacillus plasmid set" was made available to the research community. This work was a continuation of research begun by Dr Martinez-Klimova, Imperial College (99) who provided the initial modular parts. Generation of multiple cloning site design, the improved chloramphenicol resistance gene, and assembly plus characterisation of the plasmids took place as part of the work of this thesis.

Aims

- Build and test a set of modular shuttle vectors with interchangeable origins of replication and antibiotic resistance markers for *G. thermoglucosidans* and other *Geobacillus* species

- Test copy number, temperature stability and the compatibility of the included origins of replication

- Test and improve the thermostability of the antibiotic resistance markers

- Test the electroporation efficiency of these new vectors compared to those of existing plasmids

6.1 Introduction

6.1.1 Plasmid Architectures

For efficient, rational future engineering with *Geobacillus* species a more complete genetic toolkit is necessary. In other non-model chassis such as *Clostridium* and *Pseudomonas* species, sets of modular vectors have been used to help standardise characterisation and exchange of genetic parts (62,195). A similar toolkit for *Geobacillus* species, in combination with the reporters and promoters developed in the previous chapters would promote *G. thermoglucosidans* as a chassis for synthetic biology and help accelerate the development of novel parts and applications with this organism.

Vector sets such as the pMTL Clostridia vectors (62), the broad host pSEVA vectors (195) and the more recent pHsal archaeal vectors (196) have similar architectures where distinct modules for minimal replicons, antibiotic selection and additional cargos are flanked by rare cutter restriction sites to allow exchange of modules by conventional restriction/ligation cloning (Figure 6.1). This system allows many different plasmid variants to be created combinatorially from standard parts, allowing maximum flexibility in how the parts are used. Newly-characterised parts added to the toolkit can also be easily combined with any previously characterised modules. The pSEVA architecture has been particularly successful; this flexible modular design has allowed the vector collection to grow in functionality beyond the initial *E. coli* and *P. putida* chassis. SEVA standard vectors have been used in an astoundingly diverse range of Gram-negative organisms, including the nitrogen fixing *Azoarcus communis* (197), the insect enteric bacterium *Shimwellia blattae* (198), marine bacterium *Alcanivorax borkumensis* (199) and the bacterial cellulose producing *Gluconacetobacter hansenii* (200). Parts and characterisation data that then comes from work in these organisms feeds-back to improve the collection even further.

For *Geobacillus* species, shuttle vectors capable of replication in both *E. coli* and *Geobacillus* species were required. No minimal replicons that are functional in both of

these species are currently known and so, as with the previous pMTL vectors (Figure 6.1), two minimal replicon module sites will be required; one for *E. coli* and one for *Geobacillus*. To expand functionality, two interchangeable *Geobacillus* minimal replicon modules and two selectable markers for this species were planned. The vectors can then be used to host the reporter genes and promoters characterised in previous chapters as 'cargo' modules.

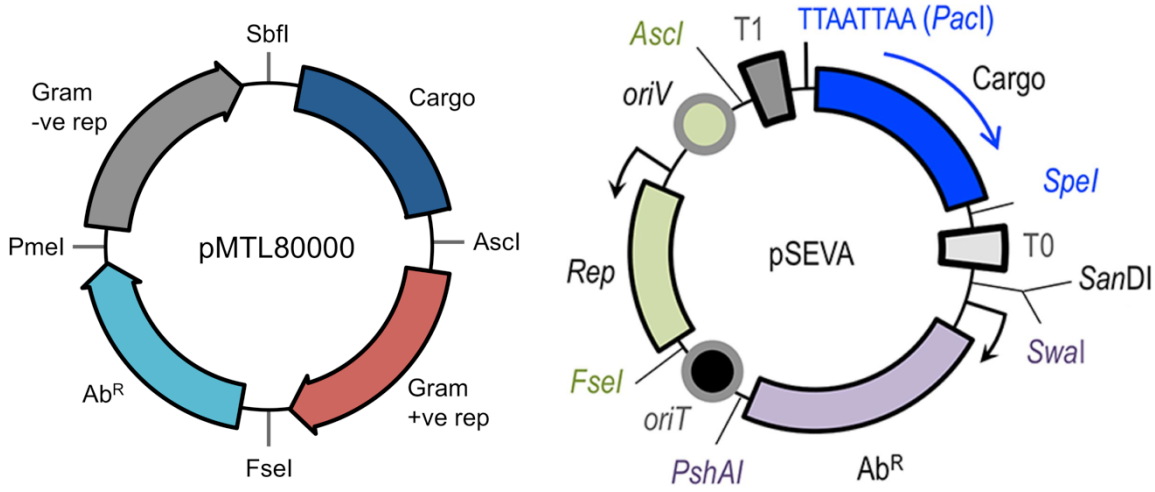


Figure 6.1. The plasmid architectures of *Clostridim* species pMTL plasmids (left) and the broad host Gram-negative pSEVA plasmids (right). Adapted from (62,201). Modules are separated by rare cutter restriction sites.

6.1.2 Plasmid Replication

An essential sequence for plasmid propagation is the replicon module which encodes the origin of replication (or 'Ori') where plasmid replication is initiated. Protein coding or functional RNA genes responsible for replication and maintenance of copy number then usually surround the Ori. The smallest section of sequence containing these elements that can support stable propagation of a plasmid is known as the minimal replicon. These parts can be used to generate new plasmids with similar replication properties and equivalent copy number (plasmid copies per cell) to the minimal replicon's original plasmid.

When designing the Geobacillus plasmid set, minimal replicons were selected from plasmids previously shown to replicate stably in *Geobacillus* species, listed in Table 6.1 below.

Replicon name	Original plasmid name and host	Replicon notes	Plasmids containing this replicon	Original plasmid reference
repBSTI	pBST1 from *G. stearothermophilus*	Stable at 68 ºC	pBST22, pUCG18, pUCG3.8	Liao *et al.* 1986 (202)
repSTK1	pSTK1 from *G. stearothermophilus*	Stable at 67 ºC	pSTE3	Narumi *et al.* 1993 (76)
repBC1	pBC1 from *B. coaglulans*	Stable at 60 ºC	pRP9, pNW33N, pTMO19	De Rossi *et al.* 1992 (203)
repB	pUB110 from *Staphylococcus aureus*	Temperature sensitive, no replication above 68 ºC	pUB190, pTMO31, pLW05	Gryczan *et al.* 1978 (204)
repTHT15	pTHT15 from an unknown thermophilic *Bacillus* isolate	Stable at 60 ºC, copy number ~45 per chromosome	pIH14, pSTE12	Hoshino *et al.* 1985 (107)
repTB19	pTB19 from an unknown thermophilic *Bacillus* isolate	Stable up to 65 ºC, copy number	pTB90, pTB913	Imanaka *et al.* 1981 (205)

Table 6.1. Replicons of previous *Geobacillus* species vectors. For the vectors developed in this study, repBSTI and RepB were considered for testing and characterisation.

6.1.3 Selectable Markers

Many types of selectable marker are available for genetic engineering in microorganisms, however, resistance genes for common antibiotics are the most popular choice in bacteria. Antibiotics can be cheap, broad spectrum and selection does not require modifications to the host genome. This allows appropriate genes to give selectable resistance across a broad range of chassis organisms. For mesophilic organisms, many well-characterised resistance markers are already available. The choice of selectable markers in thermophiles, however, is limited both by the thermostability of resistance proteins produced and thermostability of the antibiotic compound required to place a selective pressure. In *Geobacillus* species, kanamycin, chloramphenicol and tetracycline resistance genes have all been previously reported for selection of transformation and maintenance of plasmids (101,107). Resistance to many other common antibiotics including ampicillin and streptomycin has been found in *Geobacillus* species isolates but the resistance genes have not been characterised (206).

Markers known to allow selection in both *G. thermoglucosidans* and *E. coli* were most desirable for creating the modular plasmid set for this study. The thermostable kanamycin resistance TK101 marker gene has been used in several previous *Geobacillus* species vectors and was shown to function in *E. coli* on the pUCG3.8 shuttle vector (79). The sequence was amplified from pBST22 (75) and is a variant of the mesophilic gene found on pUB110 (207). Previous vectors containing this marker lacked a transcriptional terminator after the gene, however this was corrected for the plasmids produced here. A chloramphenicol resistance gene was chosen as the second marker. CatE is originally from the *Staphylococcus aureus* plasmid pC194 (208) and is included on the pNW33N vector, which has previously been used as a shuttle vector between *E. coli* and *G. stearothermophilus*.

6.2 Results

6.2.1 The Geobacillus Plasmid Set Architecture

A collection of nine parts were amplified from original plasmids by PCR using primers that add flanking rare cutter restriction sites. The parts were then assembled in either a 4-part or 5-part architecture. The plasmid set includes origins of replication for *E. coli* and *G. thermoglucosidans*, selectable markers and a choice of three reporter genes expressed from the novel RplS promoters developed in Chapter 4 (Figure 6.2).

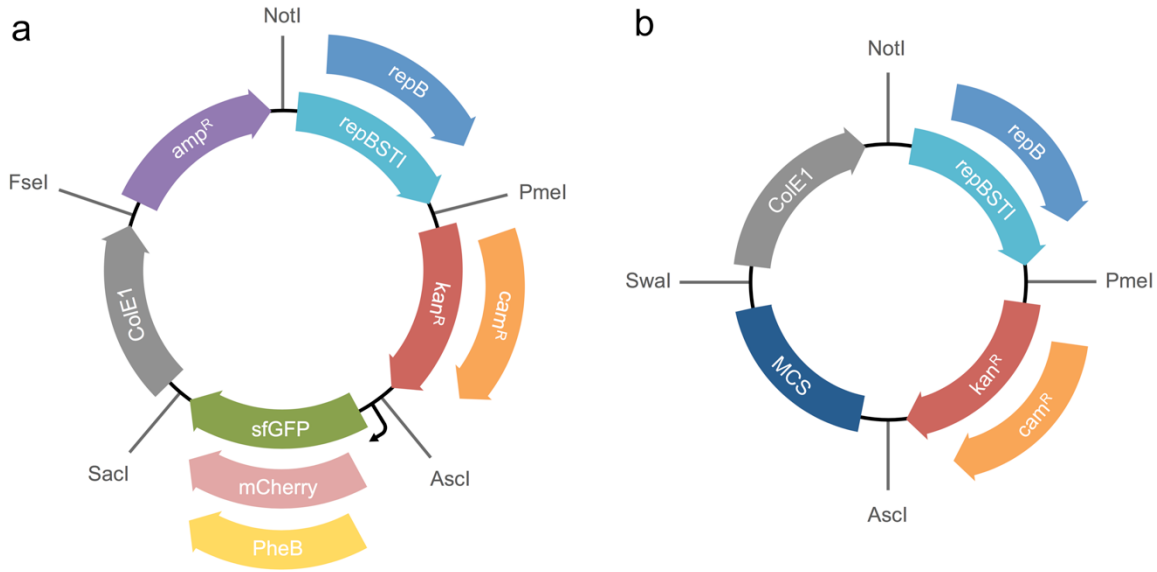


Figure 6.2. The *Geobacillus* plasmid set architecture. (a) Diagram of the 5-part plasmid including two antibiotic resistance markers, the pRplS promoter plus a selection of reporter genes (b) Diagram of the 4-part plasmid including synthetic Multiple Cloning Site (MCS).

For propagation in *E. coli* during plasmid cloning and construction, a high copy ColE1 origin of replication module is included in all plasmids (209), this ensures that DNA constructed by cloning in *E. coli* can be extracted from this organism in high yields. For propagation in *G. thermoglucosidans* a choice of two replication origins is provided; repBST1 and repB. For selection in both *G. thermoglucosidans* and *E. coli* the options are TK101: a thermostable kanamycin resistance gene (kan^R) or CatE: the chloramphenicol resistance gene (cam^R). The three reporter proteins shown to be useful in *G. thermoglucosidans*: sfGFP, mCherry and PheB (Chapter 3) expressed from the pRplS promoter (Chapter 4) are also included as modular parts. The 5-part plasmids also include the *bla* ampicillin resistance gene to enable higher efficiency cloning in *E. coli* (209), whereas the 4-part plasmids omit this to reduce plasmid size and to increase transformation efficiency in *G. thermoglucosidans*. As electroporation efficiency is negatively correlated with plasmid size (159), compact vector backbones increase efficiency and theoretically allow larger cargoes such as multi-gene operons to be carried whilst maintaining workable transformation efficiencies.

To enable cloning of cargo DNA into the 4- and 5-part plasmids, a novel Multiple Cloning Site (MCS) was designed containing many commonly used restriction sites (Figure 6.3) and insulated at either end with transcriptional terminators, rho1 and rho2. Both have previously been used in constructs for *Geobacillus* species. Terminator rho1 is taken from pUCG18 (83) and rho2 is taken from plasmids based on pUCG3.8 that are used for secretion from *G. thermoglucosidans* (79). They resemble typical rho-independent terminators consisting of a stable hairpin structure followed by polyT sequence (210).

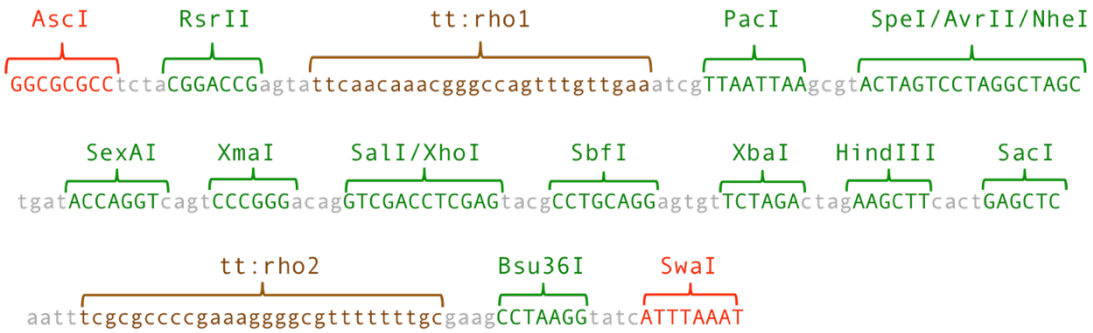


Figure 6.3. Sequence of the novel multiple cloning site included in the Geobacillus plasmid set.

The naming convention for these plasmids in the "Geobacillus plasmid set" is given in Table 6.2. All plasmids follow the naming convention of *pGxxx-cargo* where the first variable is the replicon, repBSTI = 1, repB = 2, followed by letters indicating the selection marker(s) present and then finally the cargo. The plasmids in Table 6.2 have been deposited with AddGene to be available for other researchers to use and their DNA sequences have been submitted to the NCBI database (accession numbers given in Appendix Section 2). Alternative module combinations can be created through simple restriction cloning from the main seven combinations given in Table 6.2

Plasmid Name	***Geobacilli* replicon**	**Selection marker(s)**	**Cargo**
pG1K	repBSTI	kan^R	MCS
pG2K	repB	kan^R	MCS
pG1C	repBSTI	cam^R	MCS
pG1AK	repBSTI	kan^R, amp^R	MCS
pG1AK-sfGFP	repBSTI	kan^R, amp^R	sfGFP
pG1AK-mCherry	repBSTI	kan^R, amp^R	mCherry
pG1AK-PheB	repBSTI	kan^R, amp^R	PheB

Table 6.2. The Geobacillus plasmid set naming convention. MCS = multiple cloning site

6.2.2 Plasmid Replicon Testing

repBST1 originates from pBST22 (75) which was derived from *Bacillus stearothermophilus* cryptic plasmid pBST1. This replicon is known to be stable up to 68 °C and has previously been used reliably in the *G. thermoglucosidans* plasmids pUCG18 (83) and the more compact variant pUCG3.8 (79). repB was selected because this replicon is not related to repBSTI and it initiates replication via rolling circle rather than theta-replication (65). repB is temperature sensitive and inactive over 65 °C, which is useful for the creation of knock-out strains (60) or potentially for varying copy number by changing growth temperature. The sequence for repB was obtained from pUB110 (207), originally a cryptic *Staphylococcus aureus* plasmid. For these two parts, minimal replicon sequences flanked by rare cutter restriction sites were generated as previously (99). These were cloned for characterisation in the new plasmid backbones.

6.2.3 Segregational Stability

To be useful for synthetic biology applications the plasmids must be stably maintained in the cell population over many generations. To test this *G. thermoglucosidans* was transformed with plasmids pG1AK-sfGFP and pG2AK-sfGFP. Colonies were picked and inoculated in triplicates (a, b and c) in 10 ml of media with or without kanamycin (at 12 µg/ml) in a 50 ml tube then incubated at 55 °C. Every 12 hours an aliquot was taken and diluted 200x into fresh media. At each time point aliquots of cells were also plated onto 2SPYNG agar plates without antibiotic selection and incubated at 55 °C until colonies were visible. sfGFP expressing colonies vs. non-sfGFP colonies were counted and the percentage of sfGFP expressing colonies, retaining the plasmid, was recorded (Figure 6.4).

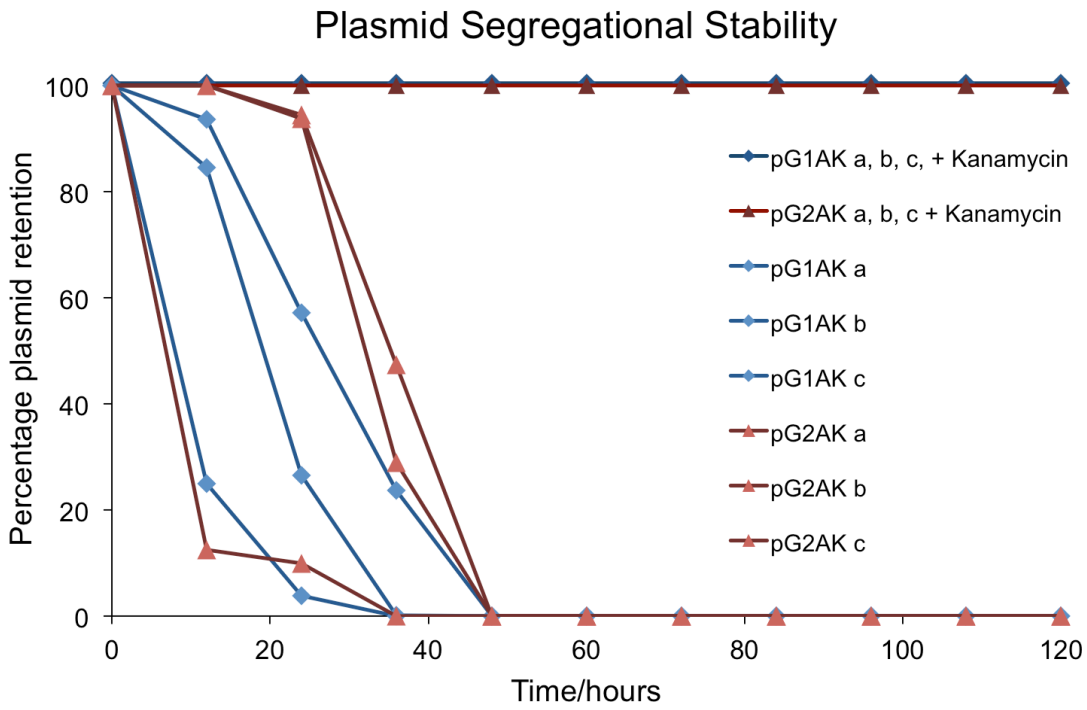


Figure 6.4. Plasmid segregational stability of pG1AK (repBSTI) and pG2AK (repB) both expressing sfGFP from the strong RplsWT promoter at 55 °C. Plasmids loss without antibiotic selection appears to occur within 48 hours. With antibiotic selection applied, both plasmids are stably maintained beyond five days and ten round of passaging, despite the likely burden of strong expression of sfGFP from the plasmids.

Both replicons allow plasmids to be stably maintained with antibiotic selection for over 120 hours of growth, long enough for most uses in microbiology, synthetic biology research and batch fed industrial processes. Plasmids are stably maintained despite

strong expression of GFP from the plasmid placing a burden on cellular resources. Without antibiotic selection however the plasmids are fairly rapidly lost, in under 48 hours, likely due to this burden.

6.2.4 Copy Number

To further characterise the minimal replicon parts of these plasmids, plasmid copy number (PCN) was estimated. Many methods exist for PCD estimation including southern blots, agarose gel based methods and PCR based methods. For rapid and accurate assays in bacteria, quantitative real time PCR as described by Lee *et al.* 2006 (100) and Skulj *et al.* 2008 (101) is the preferred method. Total DNA was prepared from *G. thermoglucosidans* cultures transformed with the plasmids and grown with antibiotic selection to early stationary phase, this was used as a template for quantitative real-time PCR. Short aplicons from the plasmid and genomic DNA were amplified in the reaction and their amplification efficiency and threshold cycle was used to estimate PCN.

From the calculated Ct values, technical triplicates were averaged and plasmid copy number for each culture sample was estimated based on the equation:

$$PCN = (Ec^{Ctc})/(Ep^{Ctp})$$

Where Ec and Ctc are the amplification efficiency (calculated from 10-fold dilutions of template) and cycle threshold for the amplification from the chromosome and Ep and Ctp are the amplification efficiency and cycle threshold for the amplification from the plasmid (details in Methods 2.3.2). Data displayed in figure 6.5 are averages of the three biological triplicates with error bars showing the standard deviation.

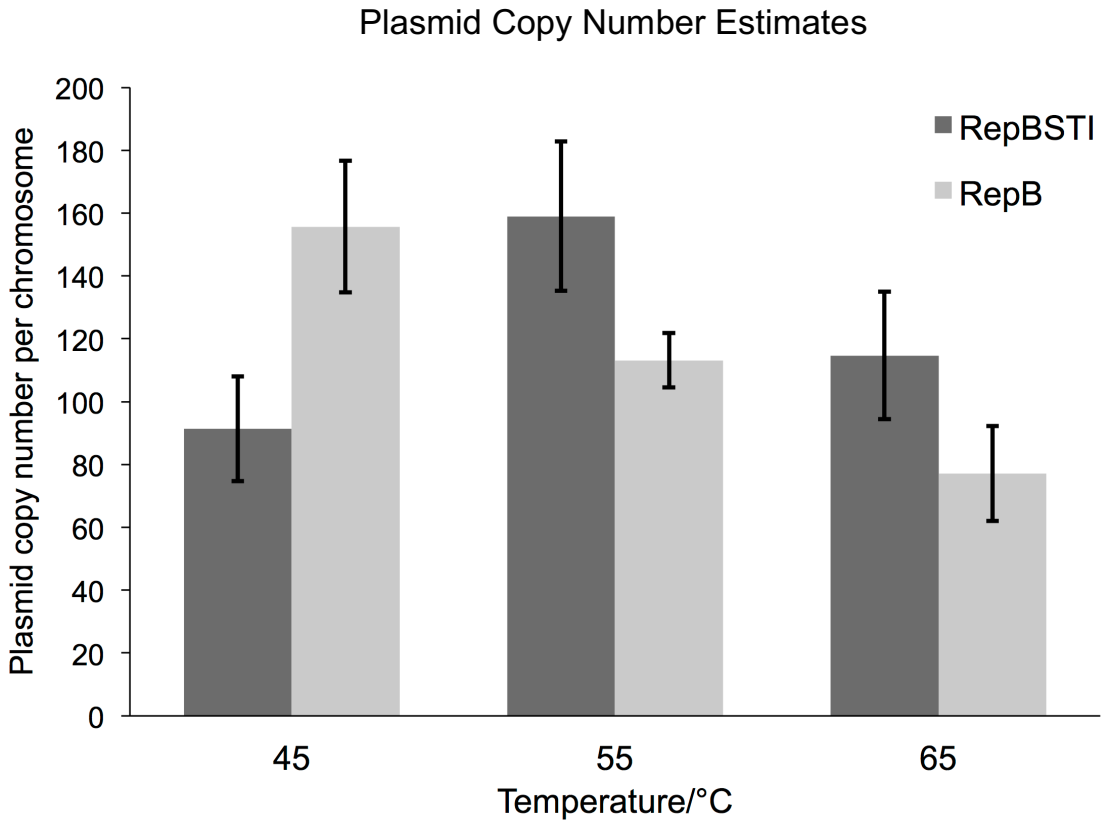


Figure 6.5. Plasmid copy number per chromosome estimated by qPCR for plasmids pG1AK and pG2AK with the two different Geobacilli replicons, repBSTI and repB respectively at different growth temperatures. Error bars show standard deviation of three biological repeats.

Both replicons give medium to high plasmid copy numbers of around 80 to 160 copys per chromosome. RepBSTI is higher copy number at 55 and 65 °C whereas repB, from a mesophilic plasmid (pUB110), is higher at 45 °C with copy number reducing as temperature rises.

With two different replicons, being able to stably propagate two different plasmids within the same cell becomes a possibility. This would be useful for *Geobacillus* species research and for applications requiring adding lots of DNA to the cells, and so replicon compatibility was tested once the two replicons had been established within the collection. Unfortunately, the two replicons were found not to be compatible; pG2K-sfGFP (repB, kanamycin resistance) and pG1C-mCherry (repBSTI, chloramphenicol resistance) could not be transformed into the same cells in either order (data not shown). Competent cells containing pG2K were made and then pG1C was electroporated into these and similar work was done *vice versa.* The inability to create strains containing both plasmids was an unfortunate and unexpected failure. The

replicons were thought to be good candiates for being compatible as they are very different: repBSTI initiates theta-replication with RNA priming at the ori, whereas repB uses the alternative rolling circle replication mechanism. These two different mechanisms are highly divergent and unlikely to interfere with one another. Perhaps cross talk occurs between the copy number maintenance systems of these replicons and this causes the incompatibility. The replicons may actually be compatible together in other scenarios, but in our hands they are not compatible with these particular plasmid backbones. When tested here, both plasmids were expressing fluorescent proteins from the strong RplSWT promoter and so perhaps the burden of gene expression was too great in this case for two plasmids to be maintained within every cell. Alternatively, the challenge of both chloramphenicol and kanamycin in combination in the growth media may have adversely affected the cells despite both expressing the necessary resistance genes.

6.2.5 Antibiotic Resistance Markers

Kanamycin and chloramphenicol have good thermostability compared to other commonly used antibiotics in bacterial growth media at thermophilic temperatures (212). Resistance markers previously shown to allow selection in both *G. thermoglucosidans* and *E. coli* were chosen. TK101, a thermostable kanamycin resistance gene (kan^R), was obtained from pBST22 (75) and is a variant of the mesophilic gene found on pUB110 (207). Previous vectors containing this marker lacked a transcriptional terminator after the gene, so this was corrected for the plasmids produced here. A chloramphenicol resistance gene was chosen as the second marker gene for the kit. CatE is originally from the *Staphylococcus aureus* plasmid pC194 (208) and the marker is present on the vector pNW33N, which has been used previously for *Geobacillus* species (88). The marker was cloned onto the pUCG16 backbone in place of the TK101 kan^R marker for testing. The minimum inhibitory concentration of chloramphenicol for *G. thermoglucosidans* for overnight growth in 2SPYNG at 50 °C was found to be 6 μg/ml. Cultures transformed with the plasmid could grow without a reduction in growth rate at 12 μg/ml chloramphenicol and so this concentration was used for selection. This chloramphenicol marker did not seem to function at more the optimal *Geobacillus* growth temperatures of 55 or 60 °C (even with only 6 μg/ml

chloramphenicol provided). This is likely due to misfolding of the mesophilic CatE protein, and so attempts were next made to improve its thermostability.

The pC194 CatE gene was amplified by mutagenic PCR to generate a library of mutant variants, using the same method as used with pRplS previously (Chapter 3). Four different reactions were run with the number of cycles varied to give lower mutational loads than for the promoter library (approximately 1%, 2.5%, 5% and 10% mutation rates). The results of these four mutagenic PCR amplifications were then mixed in equimolar amounts to give a single, large variability library that was cloned back into the pUCG16 backbone by Gibson assembly and then transformed into *E. coli.* Resulting colonies that grew on LB-agar plus chloramphenicol (approximately 20,000 colonies) were scraped from the plates and purified plasmids were prepared from these cells. The *E. coli* step served to amplify the library before the comparatively inefficient transformation into *G. thermoglucosidans* and to select out all mutants in which the cam^R gene became non-functional. The library, prepared from *E. coli,* was then transformed into *G. thermoglucosidans* cells. These were recovered and plated at 53 °C with chloramphenicol selection at 7 μg/ml. Surviving colonies were then re-streaked and grown at higher temperatures (55 to 59 °C) with their growth compared to that of *G. thermoglucosidans* transformed with plasmids containing the wild-type cam^R sequence. No mutants with improved growth were detected (data not shown).

The lack of more stable mutants could be due to the library simply not being large enough, or the mutational load not being optimal. Further testing and optimisation was not pursued, however, as a different group was able to achieve the desired outcome by a different method, published in 2015. Kobayashi *et al*. (81) expressed the pC194 cam^R in a previously engineered error-prone *G. kaustophilus* strain and selected for improved thermostability. The *G. kaustophilus* strain MK480 has four DNA repair genes deleted giving a mutation rate that is over 100-fold higher than that in the wild-type strain (80). By sub-culturing this strain, and expressing cam^R from a plasmid at increasing temperature and chloramphenicol concentrations, a mutant version of the gene allowing growth with chloramphenicol at up to 65 °C was isolated. This was found to have a single base substitution (G to A) at base 412. This changed amino acid 138 from an alanine to a threonine (A138T). The change probably improves hydrogen-bonding interactions in the protein backbone, and was shown to increase stability without

compromising catalytic activity (81). The exact coding change reported in this 2015 paper was recreated here by site directed mutagenesis of the *cam*^{R} gene on the plasmid with phosphorylated primers designed to incorporate the mutation and inverse-amplify the whole plasmid backbone prior to self-ligation (Materials and Methods 2.5.1). As reported with *G. kaustophilus*, *G. thermoglucosidans* was now able to grow with chloramphenicol selection at 65 °C when transformed with this mutated *cam*^{R} gene expressed in this case from plasmid pG1C. Resistance in *E. coli* could also be selected for at 37 °C as normal.

For additional flexibility when selecting in *E. coli*, the commonly used *bla* gene for ampicillin resistance (*amp*^{R}) (209) is also included in the vectors. This marker was found to give the highest transformation efficiencies in *E. coli* (table 6.2) and so is included in the 5-part format. In including this optional part, the 5-part vector format is more flexible as this non-essential *bla* gene, currently flanked by FseI and NotI restriction sites, could be replaced by alternative modules for example, an origin of transfer (OriT) for conjugation, without needing to expand the backbone and introduce new restriction sites.

Transformation Efficiencies

Having prepared all of the DNA modules, determined that they were functional and then constructed the Geobacillus plasmid set, the transformation efficiencies of the plasmids from this set were determined for both *E. coli* and *G. thermoglucosidans* and compared with previous vectors (Table 2). The new vectors all show good transformation efficiencies, with pG1K giving over an order of magnitude improvement in colonies produced compared to the best previously published

			Transformation efficiency CFU/µg DNA	
Plasmid	Size/kbp	Antibiotic selection	*G. thermoglucosidans*	*E. coli*
pUCG18	6.3	Kanamycin (Ampicillin)	4.9×10^3	1.6×10^6 (4.4×10^6)
pUCG3.8	3.8	Kanamycin	5.2×10^3	1.9×10^6
pG1K	3.7	Kanamycin	5.3×10^4	3.4×10^6
pG2K	3.8	Kanamycin	1.1×10^4	3.5×10^6
pG1C	3.9	Chloramphenicol	3.9×10^3	9.6×10^4
pG1AK	4.7	Kanamycin (Ampicillin)	5.8×10^3	3.0×10^6 (7.4×10^6)

Table 6.2 Transformation efficiencies of Geobacillus plasmid set plasmids compared to previous vectors. Efficiencies given are the average of three biological repeats.

In addition to the more compact size of the new plasmids (especially the 4-part plasmids), the improvement in transformation efficiency could also be due to the improved TK101 *kan*^{R} expression, as a transcriptional terminator was added to this that was previously absent, likely improving its expression. pG1K was also found to transform other *Geobacillus* species *G. stearothermophilus,* and *G. thermodenitrificans* though with efficiency around two orders of magnitude lower (~10^2 CFU/µg DNA).

6.3 Discussion and Future Work

The Geobacillus plasmid set developed here provides more compact and versatile vectors with higher electroporation efficiency than previously available plasmids, as demonstrated in Table 6.2. Existing *E. coli* to *Geobacillus* species shuttle vectors such as pNW33N and pUCG3.8 have been shown both previously and here to be functional but use of these vector for new applications would likely require bespoke re-design, requiring significant cloning. Parts are in these previous shuttle vectors were not well insulated in the sense that in some cases they even lacked transcriptional terminators. This means that gene expression from these plasmids could be affected by the neighbouring sequences of the other plasmid parts (e.g. the selectable markers). The refined vectors developed here have already been shared with several other groups and have been made available through Addgene and via depositing their sequences with NCBI (accession numbers are given in Appendix 9.2). If adopted by many groups, this

plasmid set could help facilitate exchange and standardised characterisation of parts within the Geobacilli research community and enable others who are interested in extending synthetic biology and metabolic engineering into thermophiles.

While the work here demonstrates the construction and initial verification of the Geobacillus plasmid set, the characterisation data taken here only represents the measurement of the basic main properties of the plasmids (i.e. their transformation rate, and ability to propagate at different temperatures). Accurate measurement of the copy number of plasmids containing each replicon is a further priority, and ideally these measurements would be per cell or per chromosome. Refining the method of DNA extraction from protoplasts would be required for this, as current methods for plasmid preparation from *Geobacilli* are low-yield and high-noise. The stabilities of the plasmids over long term culturing and further subculturing at different temperatures (with or without antibiotic selection) would also be important future information to be added. Testing the host range and transformation efficiencies of these plasmids in other chassis cells for example, testing their function in a wider panel of *Geobacillus* species would be a useful next step. This panel could also be designed to include some more distantly related but industrially useful stains such as *Bacillus smithii*, an acid tolerant bacterium useful for production of organic acids (213) or *Anoxybacillus* species that have useful bioremediation capabilities (214).

The plasmid set developed in this chapter also has potential for many new parts to be added to it. In terms of potential new parts to be added, an origin of transfer (OriT) module that directs conjugation would be a priority. An existing *E. coli* OriT has been shown to work well for transfer to *G. kaustophilus* (87). If amplified flanked by FseI and NotI sites this part could be cloned into the 5-part architecture in place of ampicillin. This could potentially then allow conjugation from *Geobacillus* to other *Geobacillus*, or versions could be made that allow conjugation from *E. coli* into *Geobacillus*. Conjugation may be especially desirable for transferring large genetic constructs into cells or for aiding integration into the genome.

For expanding the plasmid set, more options for minimal replicon and antibiotic resistance modules would also be valuable. Any of the other replicons from Table 6.1 could be useful, particularly if they are shown to have a different copy number, a

particularly broad host range or are compatible with repBSTI or repB to allow two or more plasmids to be present per cell. The tetracycline resistance gene of pTHT15 (101) would be the next best resistance marker to include in an expanded set and recently a gene conferring resistance to thiostrepton was shown to be an effective marker in *G. kaustophilus,* although only up to 55 °C (82). Alternative replicons to the *E. coli* ColEI module would also be useful for cloning steps in *E. coli*. Lower or inducible copy number replicons would aid cloning of constructs that could cause a burden on the host cell. Broader host range replicons could also expand the versatility of the set, allowing vectors to shuttle between many different hosts. Entirely new modules such as the many already available for the SEVA collection could also be added. A promising example would be to incorporate a toxin/antitoxin system to reduce horizontal gene transfer (215). This would enable these plasmids to be useful beyond metabolic engineering applications, as *Geobacillus* species could also be useful as a chassis for biosensing or bioremediation applications. Measures to reduce horizontal gene transfer may aid in ensuring controlled release into the environment by reducing the escape of synthetic genetic material and therefore possibly preventing unintended effects on the ecosystem. Adding a broad-spectrum toxin module onto the plasmid and adding a corresponding antitoxin expressed from the host cell chromosome would prevent the plasmid from surviving in other hosts (as they do not express the antitoxin).

The Geobacillus plasmid set developed in this chapter was initially inspired by the Clostron/pMTL system for Clostridial species, where a dedicated architecture and collection of specialised modular parts are made available specifically for this particular chassis. By contrast, the SEVA collection, whilst initially designed for *P. putida,* has now become useful for a very broad range of hosts. Its conserved architecture remains in most cases but the collection also includes many species-specific parts. This outcome is more true to the ideals of promoting standardisation in synthetic biology and encourages the development of broad host range modules. This can ultimately lead to host-independent design of genetic circuits followed by testing in a panel of hosts to select the host that is most optimal for the desired function. The current SEVA collection is focussed on Gram-negative bacteria, however, and the architecture currently only allows a single replicon module. The list of forbidden restriction sites is large and so the Geobacillus plasmid set modules are not currently fully cross-

compatible with the SEVA standard. However, with re-synthesis or site-directed mutagenesis the modules could be made compatible between the two plasmid sets. If the SEVA collection expands to become the *de facto* standard for plasmids in synthetic biology and accommodates multiple replicon plasmids (i.e. shuttle plasmids) then it would be beneficial to merge the Geobacillus plasmid set modules into the SEVA collection. Alternatively, the Geobacillus plasmid set presented here could expand or evolve to become the Gram-positive standard architecture. The replicon repB is already known to function in *Bacillus subtilis* and *Staphylococcus aureus* (216) and the antibiotic resistance markers used in the set are broad host range. Plasmid pG2K is likely to propagate in a wide range of hosts both thermophilic and mesophilic. This could be expanded further if the ColEI replicon was exchanged for a broader host range Gram-negative replicon. The emergence and adoption of standards is unpredictable, but the Geobacillus plasmid set provides many useful parts that are certain to find future applications. It therefore offers a useful architecture standard for *Geobacillus* species and represents the current state-of-the-art for these thermophiles.

Chapter 7: Metabolic Engineering

Summary

To demonstrate the application of the tools developed in this study and to test the feasibility for *G. thermoglucosidans* to be used as a chassis for more complex chemical production, metabolic engineering for the production of hyaluronic acid was attempted. Hyaluronic acid is a valuable biopolymer with a growing market in healthcare and consumer products. It was selected as an attractive target as successful recombinant microbial production of hyaluronic acid has been reported previously in standard laboratory organisms, but this could potentially be improved upon by using a thermophilic chassis. For production in *G. thermoglucosidans,* a potentially thermostable candidate hyaluronan synthase enzyme from *Streptococcus thermophilus* was cloned, refactored and tested. This was combined in an artificial operon with two native genes for upregulation of precursor sugar synthesis and together these were tested for hyaluronic acid production in *E. coli* and *G. thermoglucosidans.* Unfortunately, initial construct designs could not propagate in *G. thermoglucosidans* likely due to the burden imposed by the extra genes provided. However, promising yields (~78 mg/l) were achieved in initial tests in *E. coli*, which shows that the synthetic operon is functional. A future strategy for high throughput genetic optimisation of the hyaluronic acid production operon was also designed.

Aims

- Test the utility of parts and plasmids developed in this study for an industrially relevant application

- Assess the feasibility of *G. thermoglucosidans* as a chassis for production of a complex product - the biopolymer hyaluronic acid

- Test the functionality of the *S. thermophilus* hyaluronan synthase in a non-native host and assess the potential for thermophilic production of hyaluronic acid using this enzyme

7.1 Introduction

Previously *G. thermoglucosidans* has been used as a chassis for production of simple molecules used in biofuels, with ethanol and isobutanol the only two published products (60,73). The metabolic engineering in these production strains only required gene overexpression or knockouts. With the tools produced in this study, production of simple molecules could be optimised by fine-tuning of expression, but also more complex molecules requiring more sophisticated engineering become new possible targets. Products produced by multi-gene pathways require precise tuning of the relative expression levels of each enzyme to give balanced metabolic flux and not overburden the cells. Fine control of expression levels can then maximise the production of products by optimising the trade off between host cell growth and biosynthesis. To prove the utility of the new genetic parts developed here for *G. thermoglucosidans,* a novel metabolic engineering target was chosen and a pathway for its production was designed, refactored and tested in *E. coli* and *G. thermoglucosidans.*

7.1.1 Metabolic Engineering Targets

A literature review was conducted to determine appropriate candidates for a metabolic engineering project in *G. thermoglucosidans*. The findings from this review are very briefly summarised in Table 7.1.

Compound	Value	Engineering	Details
Aromatic amino-acids	Medium	Overexpression of native or heterologous genes	Yields tend to be very low without significant optimisation.
Other amino acids, valine, threonine	Low	Overexpression of native genes and/or knockouts	Already well optimised in other species. Productivity gains often made by editing media/bioreactor conditions rather than genetics
Nucleotides and derivatives – ionosine etc.	Low	Overexpression of native genes	Already very efficiently produced by yeast
Sugars, succinate, pyruvate, lactate	Low	Overexpression of native genes and knockouts	Potential, as use of low value feedstocks is important. The genetics comparatively simple, process optimisation would be the main challenge
Sugar alcohols, xylitol, mannitol	Low	Heterologous expression of key enzyme(s)	Possible problems with enzyme stability

Anti-microbial peptides	High	Overexpression of heterologous genes – many from *Geobacillus* species though.	Some may only be well expressed in thermophiles, could be harmful to the cells though
Poly lactic acid, PLA	Low	Complex for direct production or simpler for two step lactate production then polymerisation	Direct production probably too challenging currently. Just lactate production too simple though
Polyhydroxyalkanoates, PHAs	Low	Expression of at least three heterologous enzymes, thermophilic variants known but poorly characterised	Enzymes may not function in *Geobacilli*
Hyaluronaic acid	Medium	One heterologous enzyme required. Upregulation of native enzymes improves yield	Good production in *B. subtilis* has been achieved. Strategy could be replicated/improved upon if a thermostable synthase can be found

Table 7.1. Possible new targets for metabolic engineering in *G. thermoglucosidans.*

From the many exciting candidate products, hyaluronic acid (HA) was chosen. Hyaluronic acid is a complex biomolecule but one that has been successfully produced by metabolic engineering in mesophilic organisms. The ability to utilise cheaper feedstocks and the faster feedstock conversion could give *G. thermoglucosidans* an advantage over production with established mesophile chassis. Quantification of hyaluronic acid is comparatively simple and in previous studies hyaluronic acid could be produced at detectable yields (mg/l) even with minimal optimisation. However, higher yields demanded optimised expression of at least three enzymes and because of this, would only be possible in *G. thermoglucosidans* with parts and tools developed in the previous chapters.

7.1.2 Hyaluronic Acid

Hyaluronic acid (HA) is a naturally occurring linear polysaccharide of repeating N-acetylglucosamine and glucuronic acid units and is found in the connective tissue and epithelium of eukaryotic organisms. It plays a structural role, lubricates joints and has many functions in tissue repair, adherence, development, cell motility and angiogenesis. Many products based on or

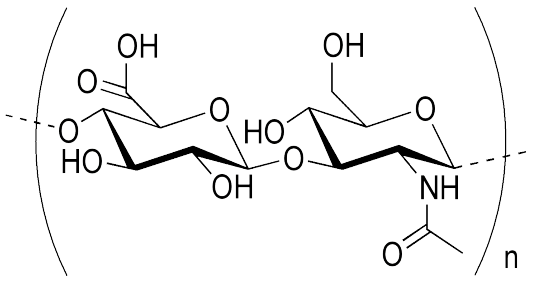


Figure 7.1. Chemical structure of hyaluronic acid, a polymer of N-acetylglucosamine and glucuronic acid.

including HA have been developed and are now widely used in orthopaedics, rheumatology and dermatology. The global market is growing rapidly, from an estimated $5.3 billion in 2012 to over $10 billion by 2020 (217).

Hyaluronic acid, chemically identical to that synthesised by eukaryotes, is also produced by *Streptococci* and a few other bacterial species to form part of the cell capsule of these organisms. Bulk production of HA for medical and cosmetic products is now largely achieved from bacterial fermentations, particularly with the high yielding *Streptococcus equisilimis* subsp. *Zooepidemicus* (218), the most popular strain for HA biosynthesis. Many genetically modified streptococcus strains exist with increased yields achieved by modifying endogenous genes to boost precursor production and by deleting hyaluronidases (218–220). Streptococcal fermentation however, has many drawbacks. These species are more difficult to culture than standard laboratory workhorse microbes, requiring comparatively expensive supplemented media. As potential pathogens, they also have many toxins and immunogenic molecules that must be carefully removed from the final product. Additionally, with the HA forming part of the cell's capsule, it must be separated from other capsular components which complicates downstream processing adding substantially to cost (221). Safer modified strains have been engineered with reduced production of contaminating toxins and with pathogenicity factors knocked out, however these still suffer from dependence on expensive media supplements and issues with purification of the product from capsules. Because of this there has been interest in hyaluronic acid production in common laboratory and industrial bacterial species such as *Escherichia coli* (96,222,223) and *Bacillus subtilis* (221,224). Most bacteria naturally produce the HA precursors, N-acetylglucosamine and glucuronic acid, as these are also precursors for cell wall polymers such as peptidoglycan, teichoic acids or other exopolysaccharides. In recombinant production strains, streptococcal hyaluronic acid synthase enzymes (HAS) are expressed and the native genes for N-acetylglucosamine and glucuronic acid production are upregulated. These recombinant *E. coli* and *B. subtilis* strains can produce reasonable yields of HA and importantly, as the chassis cells do not naturally produce cell capsules, the product is excreted into the media, greatly simplifying downstream purification. These organisms are also comparatively cheap to culture and generally recognised as safe.

Industrial HA production with these recombinant strains is growing and the most successful process with recombinant *B. subtilis* (221) has been used by Novozymes for large scale HA production since 2012 who claim it gives greater purity and more consistent molecular weight than previous processes (225). There are still limitations to this approach, however. Feedstock conversion and yields are lower than with streptococcal fermentation and molecular weight of the products is lower which makes this less commercially valuable. HA secreted into the media significantly increases viscosity, causing problems with mixing, aeration and downstream processing. The existing best strain, a modified *B. subtilis*, could potentially be improved upon by similar engineering in a *G. thermoglucosidans* chassis. The thermophile has reduced risk of contamination due to growth at a higher temperature. It also has faster feedstock conversion giving potentially greater yields per hour, and can grow on cheaper feedstocks improving long-term economic viability. The viscosity of HA in aqueous solution drops sharply with increasing temperature (226) and so a thermophilic process would also avoid this key complication that is seen in the mesophilic biosynthesis process.

6.1.3 Thermophilic Production of Hyaluronic Acid

In naturally HA-producing *Streptococcus* species up to five genes for HA production, *HasA* to *HasE* are encoded on an operon and are co-expressed. Genes *HasB* to *HasE* upregulate precursor sugar production (Figure 7.2) and *HasA* is the hyaluronan synthase enzyme. Other Gram-positive bacteria such as *Bacilllus* and *Geobacillus* species have close homologues of *HasB* to *HasE* and produce significant quantities of N-acetylglucosamine and glucuronic acid to build cell wall polysaccharides. This makes them attractive alternative production strains. These endogenous homologous genes are also attractive targets for upregulation to boost HA production (221,224). In both previous strategies for production in *B. subtilis*, the *Streptococcus equi* subsp. *zooepidemicus HasA* was used. Streptococcal *HasA* genes codes for a type I hyaluronan synthase, a large transmembrane enzyme that binds the precursor sugars, synthesizes the HA chain and secretes the polymer out into the capsule. This enzyme is surprisingly highly conserved across the kingdoms of life with the bacterial version displaying

significant homology to mammalian synthases apart from the C-terminal domains involved in mammalian cell signalling (227).

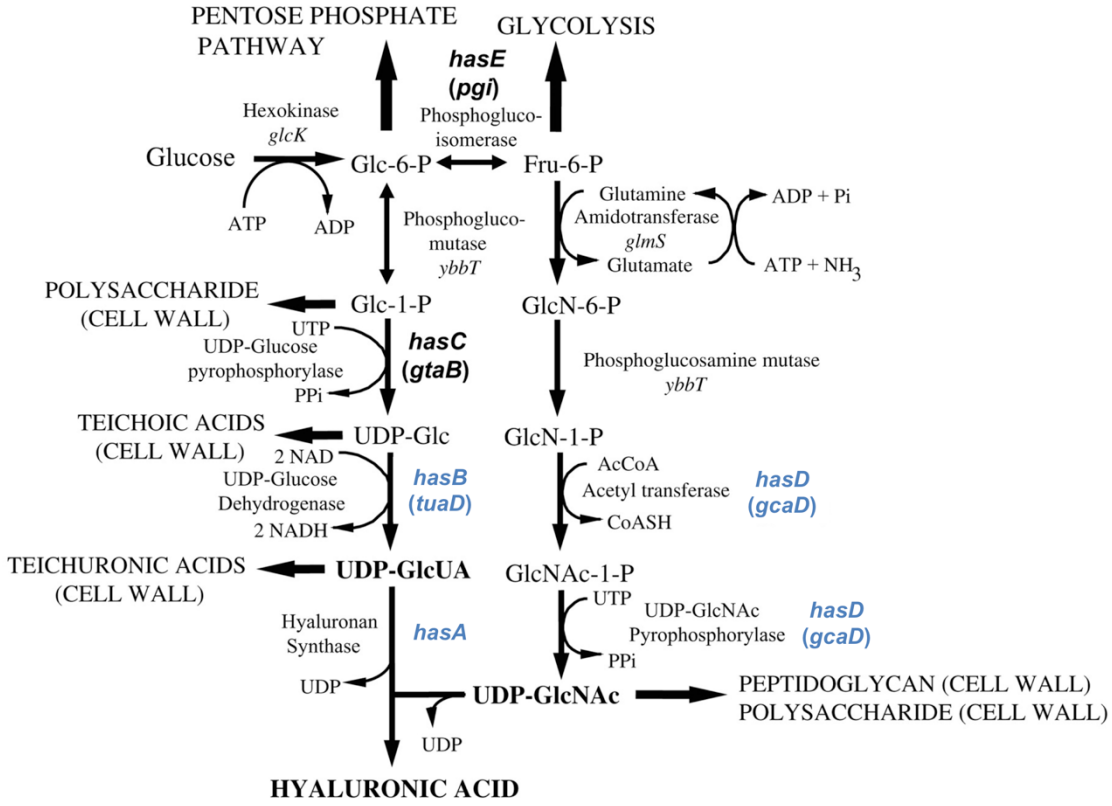


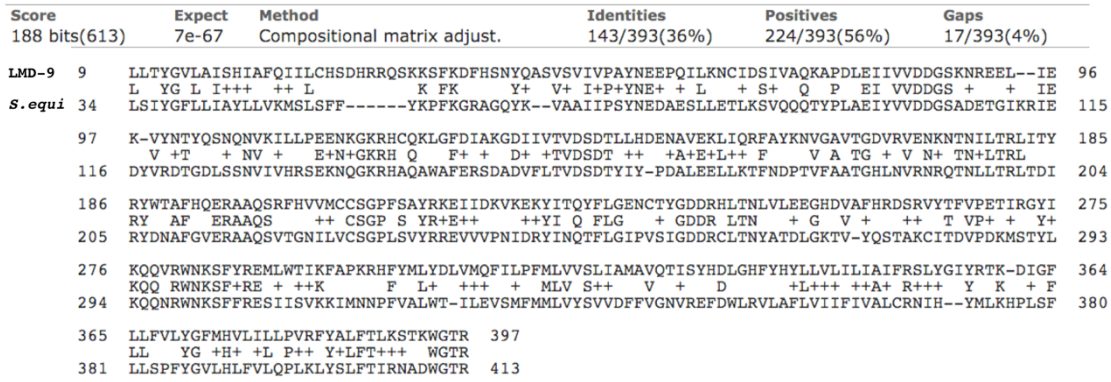
Fig 7.2. Hyaluronic acid biosynthesis. Genes from the *S. equ subsp. zooepidemicus* operon *hasA* to *E* are labelled with *G. thermoglucosidans* homologues underneath in brackets if present. The genes in blue, *tuaD* and *gcaD* will be upregulated in *G. thermoglucosidans* and *hasA* will be heterologously expressed. Pathway adapted from Widner *et al* 2005 (221).

Production of HA in recombinant *E. coli* has also involved upregulation of native *has* gene homologues and heterologous expression of a hyaluronan synthase, either a type I streptococcal enzyme (96,222) or the only know Gram negative synthase, a type II enzyme from *Pasteurella multocida* (223). The *P. multocida* enzyme is cytoplasmic rather than transmembrane and has been used for *in vitro* enzymatic HA synthesis methods, however yields for this are very low. These processes are not likely to be a serious competitor for microbial synthesis in the near future (228).

Whilst this project aims to replicate and improve upon the previous metabolic engineering in these well-characterised hosts, engineering in any new chassis is challenging. In this case, the thermostability of the heterologous enzyme used is an

added issue that needs to be considered. The previous *B. subtilis* and *E. coli* strategies used hyaluronan synthases from mesophilic organisms with optimum growth temperatures of 37 °C and so these proven enzymes are unlikely to be suitable for expression in *G. thermoglucosidans*. As a Gram-positive chassis, a type I enzyme from a fellow Gram-positive *Streptococcus* species is most likely to be functionally expressed. The *Streptococcus* species with the highest potential growth temperature is the moderate "thermophile" *S. thermophilus*, capable of growth up to 50 °C (229) though optimum growth temperature is around 40 °C. Certain *S. thermophilus* strains do produce hyaluronic acid and the species is generally well studied due to its importance in cheese and yoghurt manufacture. Around twenty strains have full genome sequences available and of these, three strains (LMD-9, ND03 and TH982) have predicted glycosyl transferase enzymes that may be hyaluronan synthases based on homology to known mesophilic *hasA* genes. HA production has not been specifically reported in these sequenced strains however it has been studied in other wild type isolates.

Izawa *et al*. isolated and characterised 46 new strains from dairy food products and found six to be HA producing with strain YIT2084 a particularly high producer (230). This strain was tested in fermentations to produce HA, and although no genome sequence is available, a predicted *hasA* gene was amplified and sequenced. The YT2084 gene showed 100% identitywith a predicted LMD-9 glycosyl-transferase and when overexpressed in YIT2084, it boosted HA production confirming the gene's predicted function (231). The recombinant strain was however not as high-yielding as the recombinant *S. equisimilis* strains. This may be due to less optimised media and bioreactor conditions during the experiment. *S. thermophilus* is however, non-pathogenic so requires less stringent containment and will not contaminate the product with exotoxins. Interestingly the LMD-9 (and YIT2084) *hasA* is comparatively divergent from the *hasA* of *Streptococcus equi* subsp. *zooepidemicus*, the most popular fermentation strain and the source of *hasA* for previous metabolic engineering attempts. Despite having the same function in species of the same genus only 36% sequence homology is seen in the protein sequence (Figure 7.3). Some of these differences may be due to increased thermostability.



Score	Expect	Method	Identities	Positives	Gaps
188 bits(613)	7e-67	Compositional matrix adjust.	143/393(36%)	224/393(56%)	17/393(4%)

```
LMD-9   9    LLTYGVLAISHIAFQIILCHSDHRRQSKKSFKDFHSNYQASVSVIVPAYNEEPQILKNCIDSIVAQKAPDLEIIVVDDGSKNREEL--IE  96
             L  YG L I+++  ++ L          K FK     Y+  V+ I+P+YNE+ + L   + S+  Q  P  EI VVDDGS +   +  IE
S.equi  34   LSIYGFLLIAYLLVKMSLSFF------YKPFKGRAGQYK--VAAIIPSYNEDAESLLETLKSVQQQTYPLAEIYVVDDGSADETGIKRIE  115

        97   K-VYNTYQSNQNVKILLPEENKGKRHCQKLGFDIAKGDIIVTVDSDTLLHDENAVEKLIQRFAYKNVGAVTGDVRVENKNTNILTRLITY  185
               V +T   + NV +   E+N+GKRH Q   F+ +  D+ +TVDSDT ++  +A+E+L++ F    V A TG + V N+ TN+LTRL
        116  DYVRDTGDLSSNVIVHRSEKNQGKRHAQAWAFERSDADVFLTVDSDTYIY-PDALEELLKTFNDPTVFAATGHLNVRNRQTNLLTRLTDI  204

        186  RYWTAFHQERAAQSRFHVVMCCSGPFSAYRKEIIDKVKEKYITQYFLGENCTYGDDRHLTNLVLEEGHDVAFHRDSRVYTFVPETIRGYI  275
             RY  AF  ERAAQS    ++ CSGP S YR+E++    ++YI Q FLG   + GDDR LTN   + G  V +   ++  T VP+ +  Y+
        205  RYDNAFGVERAAQSVTGNILVCSGPLSVYRREVVVPNIDRYINQTFLGIPVSIGDDRCLTNYATDLGKTV-YQSTAKCITDVPDKMSTYL  293

        276  KQQVRWNKSFYREMLWTIKFAPKRHFYMLYDLVMQFILPFMLVVSLIAMAVQTISYHDLGHFYHYLLVLILIAIFRSLYGIYRTK-DIGF  364
             KQQ RWNKSF+RE + ++K      F  L+  +++  +  MLV S++   V  +   D      +L+++ ++A+ R+++  Y  K  + F
        294  KQQNRWNKSFFRESIISVKKIMNNPFVALWT-ILEVSMFMMLVYSVVDFFVGNVREFDWLRVLAFLVIIFIVALCRNIH--YMLKHPLSF  380

        365  LLFVLYGFMHVLILLPVRFYALFTLKSTKWGTR  397
             LL   YG +H+ +L P++ Y+LFT+++  WGTR
        381  LLSPFYGVLHLFVLQPLKLYSLFTIRNADWGTR  413
```

Figure 7.3. Protein sequence alignment of *S. thermophilus* LMD-9 *hasA* and *S. equi* subsp. *zooepidemicus hasA*. The sequences are quite divergent despite having the same function in closely related species with only 36% homology seen. Alignment generated with the NCBI Blast tool (152).

The *S. thermophilus* LMD-9 *hasA* gene is promising for HA synthesis in *G. thermoglucosidans.* For upregulation of precursor sugar production, the entire LMD-9 *has* gene operon could be introduced, however native *G. thermoglucosidans* genes are more likely to be well-expressed and will be of course thermostable. Two *G. thermoglucosidans* genes (*tuaD* and *gcaD*) were chosen to build an artificial operon in a strategy inspired by the previously successful HA production in *B. subtilis* by Widner *et al.* (221). Here the *B. subtilis* homologues of *hasB* (*tuaD*), *hasC* (*gtaB*) and *hasD* (*gcaD*) were expressed in various combinations in an operon preceded by the *S. equisimilis hasA* gene. In this previous study, the artificial *hasABD* operon gave the highest yields though inclusion of the hasB homologue (*tuaD*) gave by far the greatest increase, suggesting that UDP-glucuronic acid availability was limiting. The *hasA*, *tuaD*, *gcaD* operon was expressed from the genome with a strong *B. subtilis* promoter and strong RBS for *hasA,* with *tuaD* and *gcaD* translated from their natural RBS sequences. This gene order was replicated for *G. thermoglucosidans* in this study but with the operon expressed from a plasmid and the RplS promoter library used to tune expression. The genes were refactored, cloned, tested in *G. thermoglucosidans* and *E. coli,* and then a strategy for optimising production was planned.

6.1.4 Optimising Production

For engineering a new biological function into a host, the chosen genes must first be refactored into suitable sequences for cloning and expression in the new chassis, then

high-throughput testing and optimization techniques can be used to maximise product production (162). Operon design tools and codon optimisation were used to refactor HA production for *G. thermoglucosidans*.

Genetic Refactoring: Codon Optimisation

When expressing a synthetic circuit in a chassis, it is depletion of the ribosome pool during translation which places the most significant burden on the host (193). Maximising translation elongation efficiency is therefore vital for synthetic construct design in order to maximise functionality with minimum stress to the host.

Due to different 16S ribosomal RNA sequences and codon usages, elongation efficiencies for a particular sequence will vary greatly between different chassis. The natural genes of a particular host are likely to have evolved to be efficiently translated in that host (to minimise burden), however when expressed in a different host that sequence is likely to be less efficient. Through changing synonymous codons, a gene sequence may be re-optimised for a new host. In this case the *S. thermophilus hasA* required optimising for expression in *G. thermoglucosidans*.

The exact contributions of codon usage and sequence motifs in the mRNA to translation efficiency are not fully understood but useful design rules have emerged. Many strategies exist for codon optimisation and these give different "optimal" sequences based on their approach. A variety of computational tools have been produced to aid sequence optimisation. The most simple strategy, used in the GeneDesign software (232), includes the most frequently found codon in the genome of the new chassis for all instances of an amino acid in the sequence to be optimised. Codons frequently found in the genome are likely to have complementary tRNAs at higher proportions in the tRNA pool so will be able to bind a tRNA more quickly. Other software such as Gene Designer (233) and EuGene (234) instead adjust codon usage so that it is proportional to the natural distribution of the original host organism. They also match the positions of proportionally slower codons (codon harmonization) to try and maintain regions of slower translation thought to be important for protein folding. More recent approaches decide slow vs. fast codons based on better data for cytoplasmic tRNA concentrations rather than genomic frequency, or avoid codon pairs known to translate slowly (235).

Altering mRNA sequence can also affect translation efficiency through effects on mRNA secondary structure; stable hairpins potentially block access to the ribosome binding site or impede the ribosomes progress along the transcript (236). Additionally codon choice affects mRNA stability by altering the competition between protein elongation and mRNA degradation (237). A final design constraint of particular importance is the affect of sequences within the mRNA that have affinity for the anti-Shine-Dalgarno sequence of the ribosomal RNA. Such sequences are thought to cause pausing of the ribosome leading to reduced elongation rates and decreasing the free ribosome pool, causing burden and fitness defects (193,238). Many computational sequence design tools are available that attempt to balance some or all of these factors, with varying degrees of manual input. Due to the considerable industrial interest in sequence optimisation, particularly in overproduction of recombinant proteins, the most advanced algorithms and software are proprietary and commercial. When optimising sequence for *G. thermoglucosidans* in this study a free, open strategy and software was chosen. Many tools exist but most only optimise for expression only in common chassis organisms. The Entelachon software tool (239), was chosen as all parameters could be customised and any data from the codon usage database (240) could be used to determine codon frequencies.

The most important features suggested by the literature for efficient translation were used to define the design rules to guide optimisation with the Entelachon tool:

Avoiding very rare codons – genomic codon frequency does not correlate well with a codon's effect on translation efficiency (235,237) as there are many factors at play and more research is required to fully understand this relationship. It is known however, that very rare codons definitely reduce translation rate (237,241). As such, these rarely used codons were removed or avoided.

Removing Shine-Dalgarno like sequences – sequence likely to bind the anti-Shine Dalgarno sequence of the ribosome stall translation (238). mRNA sequence similar to the Shine-Dalgarno sequence, AAGGAGGU for *G. thermoglucosidans* and *E. coli*, were removed.

Avoiding strong secondary structures – of the many secondary structures that mRNA forms, strong, stable hairpins are known to particularly reduce translation efficiency. Structures affecting the 5'-UTR and first ~16 codons have a particularly significant effect (186,237) due to reduction of translation initiation so should be avoided. Very stable hairpins elsewhere in the structure should also be reduced.

Genetic Refactoring: A New Operon

For initial testing, an operon with three genes, *hasA*, *tuaD*, *gcaD* was designed and expressed from the Ldh promoter and a panel of RplS library promoters. As *tuaD* and *gcaD* are natural *G. thermoglucosidans* genes their natural RBS and coding sequences were not changed. The operon was designed to take advantage of translational coupling to increase translation efficiency and this was checked *in silico* with the Salis Lab Operon Calculator (242) a tool based on the model of the RBS Calculator discussed previously (Chapter 5). Translational coupling is the increase in translation rate of genes in an operon caused by translation of the upstream genes. This is due to the translation machinery unfolding the secondary structure of the mRNA which increases ribosome access to the RBS of other coding sequences (243). Additionally translation re-initiation can occur where, after terminating translation of one gene in an operon, the ribosome does not dissociate completely and instead scans along the mRNA to initiate translation of the following gene (244). The Salis Lab Operon Calculator (242) predictably accounts for these effects on translation initiation. The tool was used to inform spacer design, check RBS strength for the initial construct and develop a strategy for future optimisation of relative expression.

Arranging the genes in an operon rather than with separate promoters also reduces transcriptional noise (245), is more simple to construct, keeps plasmid size smaller for better electroporation efficiency and reduces any risk of recombination between similar promoter sequences.

7.2 Results

7.2.1 Genetic Refactoring

The *S. thermophilus* LMD-9 *hasA* gene sequence from the genome (246) was codon optimised for expression in *G. thermoglucosidans.* The Entelachon software tools was used to help remove rare codons, unwanted restriction sites, RBS like sequences and any secondary RNA structures (239) (Figure 7.4).

For codon usage optimisation, *G. thermoglucosidans* genomic codon frequency data from the codon usage database was used (247). Parameters on Entelachon were set to change codons below 50% expected frequency to the most frequent codon. This avoids only rare codons, for example: alanine has 4 codons, these settings will define rare codons as those that occur <12.5% of the time for alanine in the *G. thermoglucosidans* genome. The software tool allows forbidden sequences to be entered and so restriction sites likely to be used in future cloning and those used for modular vector construction (Chapter 7) were disallowed. RBS-like sequences can reduce translation efficiency by stalling the ribosome and so any sequence within two mismatches of AAAGGAGGT, the consensus RBS sequence complementary to the 16S rRNA sequence were also removed. Suggested sequences were then checked for strong secondary structures in early mRNA sequence with UNAFold (191). In the final sequence no strong hairpins (more than 5 complementary base pairs) including the first 16 codons were predicted in the mRNA.

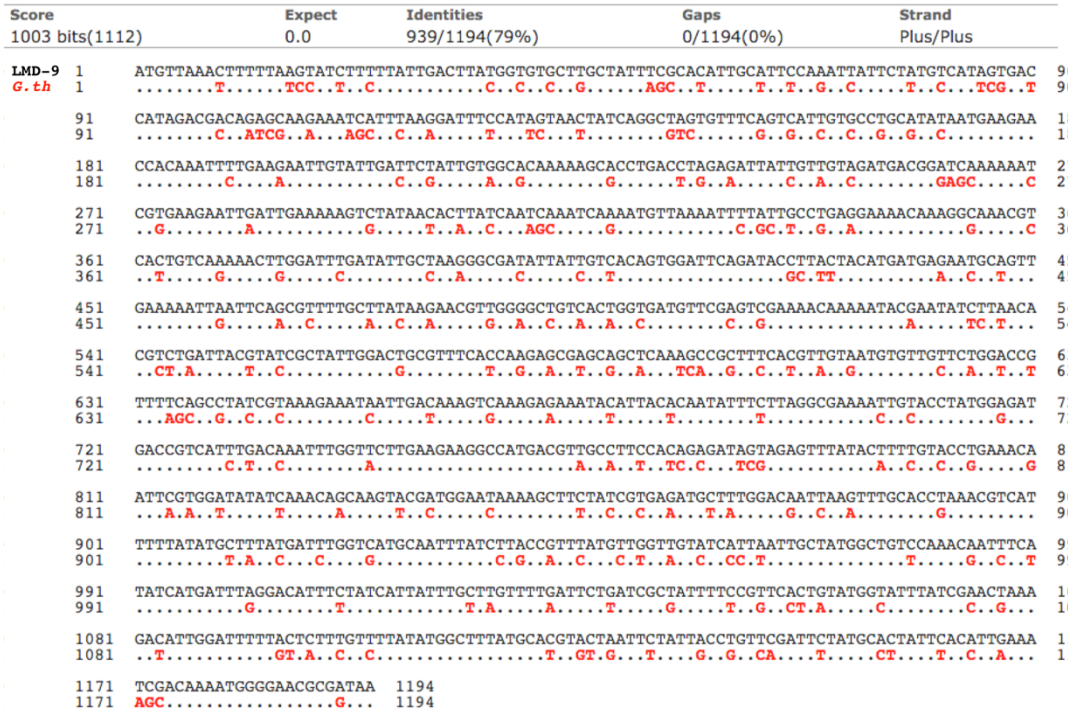


Figure 7.4. Natural *S. thermophilus* LMD-9 *hasA* sequence (top row) aligned to the same sequence codon optimised for expression *G. thermoglucosidans* (bottom row, changes shown in red). Considerable changes were needed to the sequence (~20% bases) in order to avoid rare codons and forbidden sites. Alignment generated with the NCBI Blast tool (152).

The optimised sequence was ordered as two DNA fragments (GeneArt "Strings") and cloned by Gibson Assembly expressed from the G.st RBS sequence with either the Ldh promoter or a moderate or low strength RplS library promoter (pRplS 5 or 16). These promoters all have very low strength in *E. coli* to limit burden when cloning in this host.

The *G. thermoglucosidans* genome sequence (92) was then searched for *hasB* and *hasD* homologues. *tuaD*, a UDP glucose 6-dehydrogenase producing UDP-glucuronic acid and *gcaD*, a bifunctional N-acetylglucosamine-1-phosphate uridyltransferase and glucosamine-1-phosphate acetyltransferase which increases UDP-glucuronic acid levels were identified. Two possible *hasB*/*tuaD* homologues were identified and so the gene with closest homology to the *B. subtilis tuaD* expressed successfully by Widner *et al*. (221) was chosen. Primers were designed to amplify these gene sequences from the genomic DNA and include their natural RBS sequences, considered to be the 30 base pairs upstream of the start codon (186). To space the genes further apart and improve cloning efficiency, 20 base pair neutral linker sequences were designed using the R2oDNA Designer software (248). This tool generates biologically neutral

sequences free of restriction sites and without homology to other known sequences (or other generated linkers) to reduce recombination. Linkers have a balance of the different nucleotides to reduce sequencing, synthesis or assembly errors and can have defined GC content. The 20 base pair linkers were designed with 40% GC content as this has been shown to give optimal assembly efficiency (249). The linker sequences were added to the primers used to amplify *tuaD* and *gcaD* to generate overlap sequences for construction of the operon by Gibson Assembly. The 30 base pair RBS sequence plus 20 base pair linkers generated an operon with 50 base pair spacing between each of the three genes (Figure 7.5). This was large enough to optimise sequence for assembly and leave space to substitute in different RBS designs but is small enough to take advantage of translational coupling (250).

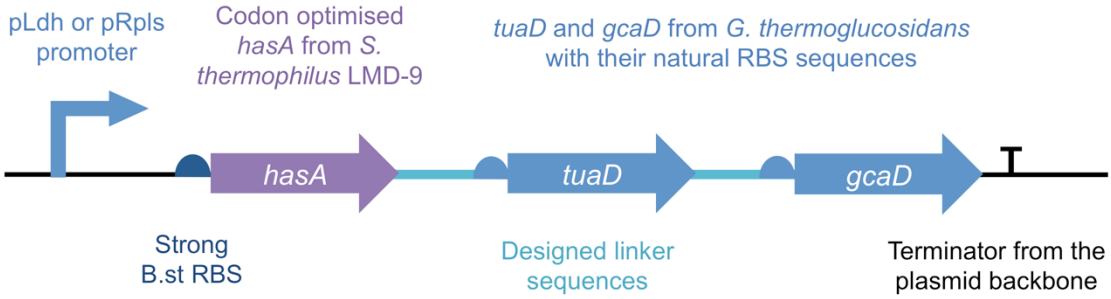


Figure 7.5. Diagram of the synthetic HA synthesis operon designed and constructed in this study. This operon was cloned in the pG1AK plasmid backbone.

7.2.2 Operon Construction and Cloning

Genomic DNA extraction from *G. thermoglucosidans* using commercial preparation kits has previously given relatively low quality, low yield DNA (Dr Martinez-Klimova Imperial College, personal communication) and so an alternative procedure was used. Chelex™ (Bio-Rad Inc.) is a chelating agent capable of lysing cells and inhibiting DNAse enzymes by chelating metal ion cofactors. Procedures to isolate DNA with Chelex were originally developed for forensic samples (251) but later adapted for preparations from bacteria (252). An adapted Chelex genomic DNA preparation method was used to prepare genomic DNA from *G. thermoglucosidans* (Materials and Methods 3.2). This prepared DNA was successfully used as template DNA for PCR amplification of the *tuaD* and *gcaD* genes with the added linkers.

Primers were also designed to amplify the vector backbones of the previously cloned constructs containing codon-optimised *hasA* expressed from the three different promoters pLdh and pRpls5 and 16 (Table 7.2). The primers added complementary overlap sequences (30 base pairs) with the fragments amplified from the genome. The complete three gene operon constructs were then assembled in 3-part Gibson Assembly reactions. Successful clones for all three constructs were confirmed by test digestion of plasmids obtained from *E. coli* colonies, and subsequent sequencing.

	Promoter strength relative to pRplSWT at 100%	
Promoter	***E.coli***	***G. thermoglucosidans***
pLdh	0	130
pRplS 5	5	37
pRplS 16	0.8	5

Table 7.2. Approximate relative promoter strengths in *E. coli* and *G. thermoglucosidans* of the promoters used for initial operon construction and cloning.

7.2.3 HA Detection and Testing Yields

Many methods for detection of HA production from microbial chassis have been reported. The simplest qualitative method is observing a change on agar plates of colony morphology from normal to a mucoid colony phenotype (221,224). Phenotypic changes were not significantly noticeable in the strains produced here, though that may be due to lower yields or the different media used in this study compared to in previous work.

Quantification of yields from liquid cultures required filtering out cells then precipitating any HA with isopropanol, followed by collection by centrifugation, washing and resuspension in a suitable buffer (Materials and Method 2.3.1). Quantification of precipitated HA is then traditionally achieved by the carbazole assay (253) where the polymer is hydrolysed by sulphuric acid then the glucuronic acid content is quantified based on a chemical complex with carbazole reagent to form a violet chromophore that can be assayed by spectrophotometry (254). Other sugars such as glucose and sucrose also react in this assay, however, giving the assay poor specificity. From microbial samples, background levels can be high due to sugar

content in the media or other microbial polysaccharides present. Thus quantification is dependent on extracting very pure HA. More recently a simpler method using safer chemicals based on the precipitation of the polyanionic HA with the ammonium cations of cetyltrimethylammonium bromide (CTAB) was developed (97,255). The precipitated HA causes turbidity that can be measured by spectroscopy and is proportional to its concentration. This method is similarly sensitive to the carbazole assay but far more specific. Therefore this method was used for quantification in this study (Materials and Methods 2.3.2). The protocol was tested and standard curves were generated with streptococcal HA (Sigma-Aldrich) in order to give quantitative values.

7.2.5 Testing in *E. coli*

To test the function of the designed operon with *S. thermophilus* hyaluronan synthase in a novel chassis, HA synthesis in recombinant *E. coli* DH10B cells was tested. No obvious colony phenotypes were observed with any of the three *has*-operon constructs on solid LB media. Strains with the *has*-operon expressed from pRplS5, the strongest promoter of the four in *E. coli* with a relative strength of 5% wildtype pRplS were used to test HA isolation and quantification from liquid cultures. 50 ml tubes containing 5 ml LB + 1% glucose media were inoculated in triplicate with both the test strain and the empty vector control strain. The cultures were grown at 37 °C with shaking (200 rpm) for 16 hours. 2.5 ml aliquots of each culture were then taken for HA extraction (Materials and Methods 2.3.1). The potential extracted HA was redissolved in 2 ml of acetate buffer and HA was quantified with the turbidity assay in a plate-reader (Materials and Methods 2.3.2). 100 µl of HA samples were added to 100 µl of CTAB solution in 96-well microplate wells and OD600 readings were taken. A Streptococcal HA standard (Sigma-Aldrich) was serially diluted and used to determine equivalent HA concentration from the OD600 readings (Figure 7.6).

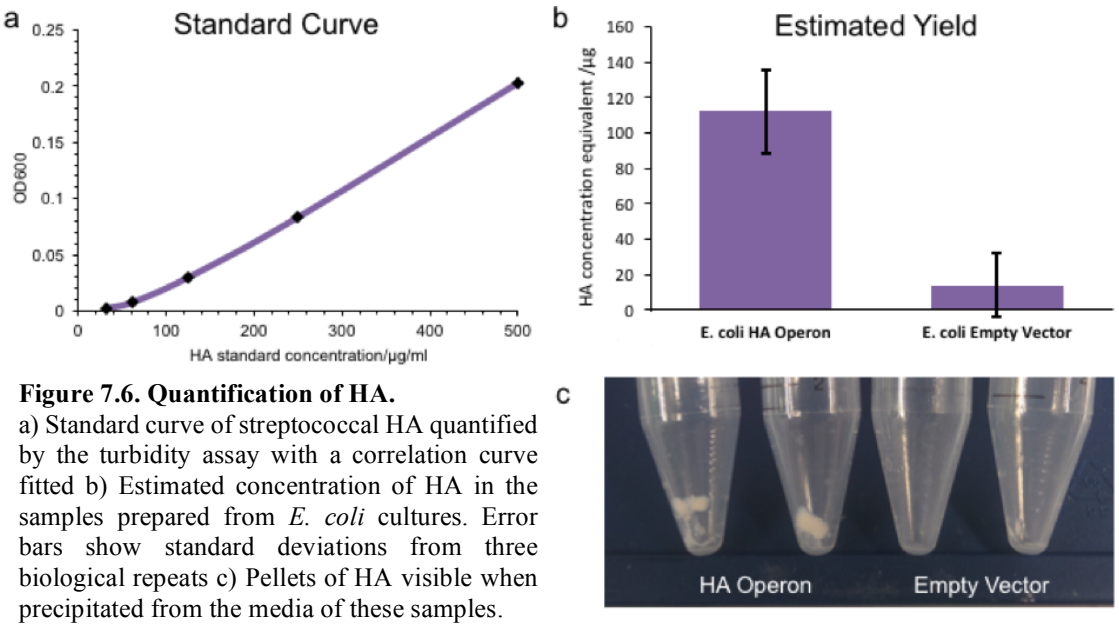


Figure 7.6. Quantification of HA.
a) Standard curve of streptococcal HA quantified by the turbidity assay with a correlation curve fitted b) Estimated concentration of HA in the samples prepared from *E. coli* cultures. Error bars show standard deviations from three biological repeats c) Pellets of HA visible when precipitated from the media of these samples.

Significant quantities of produced HA could be observed when pelleted during the extraction process (Figure 7.6c) and these were detected with the turbidity assay (Figure 7.6b). By comparing OD600 readings for the extracted samples to the standards (Figure 7.6a) the equivalent HA concentration can be determined (Figure 7.4b). The test samples have an average equivalent HA content of 114 µg/ml with the empty vector controls giving a background turbidity equivalent to 12 µg/ml of HA, likely due to contaminants in the purification process. The test samples minus the empty vector background give an estimated yield of HA at 98 µg/ml. The total HA solid extracted was redissolved in 2 ml buffer for testing so approximately 196 µg total solid HA was extracted. 2.5 ml of overnight culture was used for each extraction and so the total culture generated an estimated yield of 78 mg/l.

7.2.5 Testing in *G. thermoglucosidans*

The three operon constructs with *hasA+tuaD+gcaD* expressed from the three different promoters (shown in Table 7.3) were constructed, cloned in the pG1AK plasmid backbone and plasmid preparations made via *E. coli*. Unfortunately none of these were able to successfully transform *G. thermoglucosidans*. Strains DL33 and DL44 were both tested with recovery on plates consisting of 2SPYNG or 2SPYNG + 2% glucose media at 45 and 55 °C. However, in all conditions no colonies arose. As the

overexpression of the native genes *tuaD* and *gcaD* genes may be causing too much metabolic burden on the cells, constructs with only the optimised *hasA* gene expressed from the same three promoters were also tested (Table 7.3). pLdh *hasA* was similarly unable to transform *G. thermoglucosidans* however *hasA* expressed from both RplS library promoters did give colonies with both strains under all conditions.

Construct	Colonies with *G. thermoglucosidans*?
pG1AK pLdh + *hasA* + *tuaD* + *gcaD*	✗
pG1AK pRplS5 + *hasA* + *tuaD* + *gcaD*	✗
pG1AK pRplS16 + *hasA* + *tuaD* + *gcaD*	✗
pG1AK pLdh + *hasA*	✗
pG1AK pRplS5 + *hasA*	✓
pG1AK pRplS16 + *hasA*	✓

Table 7.3. Possible HA production constructs transformed into *G. thermoglucosidans* DL44 and DL33. Electroporated cells were recovered on and plated on 2SPYNG or 2SPYNG + 2% glucose media at 45 and 55 °C

No significant colony phenotypes were observed on 2SPYNG or 2SPYNG + 2% glucose. Purification and quantification of possible HA produced by these transformed strains was not possible due to time constraints but is a priority for future work.

7. Discussion and Future Work

7.2.1 Hyaluronic Acid Production in *E. coli*

The yield achieved for HA production by *E. coli* with the construct designed in this study (78 mg/l) compares quite favourably to previous production of HA with *E. coli*. Yields of 21 mg/l (222) 48, 160 and 190 mg/ml (96) were achieved by different strategies with some optimisation of constructs and conditions. The highest reported yield in *E. coli* to date is 561 mg/l, although this required considerable strain engineering and process optimisation with very rich media. Optimised batch processes with recombinant *B. subtilis* can achieve yields around 2 g/l (221,224) whilst the best streptococcal strains, under highly optimised conditions, can achieve yields up to 7 g/l (256). Considering the construct, strain and growth conditions have not been optimised

the initial yield is very promising and suggests the *S. thermophilus* hyaluronan synthase is a good candidate for recombinant HA production, even in *E. coli*.

Further testing to check the quality of the product is required. Contaminants such as peptidoglycan may also give turbidity in the assay – as seen with the empty vector control samples – and so the presence of HA and its purity needs to be confirmed. This could be achieved using a commercial HA detection kit (Corgenix, Inc. product #: 029-001), infrared spectroscopy or mass spectrometry. The molecular weight of the product is also important higher molecular weight product (>1kDa) is more valuable with a greater range of applications.

7.2.2 Hyaluronic Acid Production with *G. thermoglucosidans*

The lack of detected HA production with *G. thermoglucosidans* is disappointing as this would have confirmed the potential of this chassis as a production organism for higher value products. The two pG1AK pRplS + *hasA* constructs were able to give transformed colonies with *G. thermoglucosidans* but are unlikely to produce significant HA yields on glucose media as production of the precursor sugars is not upregulated. Testing the hyaluronan synthase in *G. thermoglucosidans* by growing these strains on media supplemented with N-acetyl-glucosamine and glucuronic acid would be the first priority. The full operon constructs (pG1AK *hasA, tuaD, gcaD*) are quite large, around 9 kb and so may simply be too big to efficiently transform *G. thermoglucosidans* by electroporation. Cloning the operon into the more compact modular plasmid pG1K may help. Alternatively, addition of an origin of transfer to the plasmid would allow it to be transferred by conjugation, which is not as size dependent. More likely however, the operon is causing too much metabolic burden and the cells are not viable. HA synthesis has been shown to cause burden and significantly lower growth rate in *B. subtilis* (221). pG1AK has the repBSTI replicon and so is quite high copy in *G. thermoglucosidans*; even with relatively weak pRplS library promoters expression may be too high. Replacing the current promoters with the weakest pRplS promoters and tuning down translation strength (informed by the RBS calculator) could give viable strains. Equally, lowering the copy number by changing the plasmid replicon module and/or raising the temperature would also reduce expression though temperature changes could affect the stability of the *hasA* protein.

7.3 Future Work

If good quality HA were detectably produced by a *G. thermoglucosidans* strain then the genetics and growth conditions could be optimised and the process scaled up to test if yields are comparable to previous production strains.

A high throughput screening assay for HA production with *E. coli* was shown by Mao et al. (223) and this could be adapted for optimisation with *G. thermoglucosidans*. Libraries of constructs generated with different pRplS promoter variants and variable RBS sequences for each gene – designed with the RBS library calculator (183) – could be constructed by Golden Gate assembly (257) and transformed into *G. thermoglucosidans.* Each transformed colony could could be grown at small-scale (in 5 ml tubes) and approximate HA content of the media assayed using the dye alcian blue. This shows a slight colour change, detectable by spectroscopy, upon binding HA (223). Promising clones could then be grown at larger scales in shake flasks and HA production more accurately quantified with the turbidity assay. If yields comparable with the best alternative recombinant strains can be achieved (approximately 2 g/l) then further scale up and optimisation of growth conditions could take place.

Chapter 8: General Discussion and Future Work

8.1 General Discussion

G. thermoglucosidans is a thermophilic bacterium of industrial importance with significant potential as a strain for production of biobased chemicals from cheap, renewable lignocellulosic feedstock. Previous genetic tools for engineering *G. thermoglucosidans* were limited and this restricted the potential products that could be produced with this chassis. This study was successful in its aims creating and testing tools for synthetic biology in *G. thermoglucosidans* and initiated more ambitious metabolic engineering in this host.

Key Results

- Thermostability of reporter proteins sfGFP and mCherry was characterised *in vivo* and *in vitro.*

- For *G. thermoglucosidans*, the use of LOV based fluorescent proteins and anaerobic fluorescence recovery with sfGFP was discounted.

- Two useful promoter libraries functional in both *E. coli* and *G. thermoglucosidans* were generated and characterised.

- The RBS Calculator tool was shown to be useful to predictably design 5'-UTR sequence for *G. thermoglucosidans*. The limitations of translation rate prediction tools were considered and a review paper on this topic was published (Reeve *et al*. 2014 Predicting translation initiation rates for designing synthetic biology. *Frontiers in bioengineering and biotechnology* 2, p.1.) (184)

- A set of modular shuttle vectors based on existing parts was constructed and characterised in *G. thermoglucosidans.*

- A toolkit of parts including the modular vectors, the RplS promoter library and several reporter proteins was assembled and has been shared with other researchers in academia and industry. These parts are currently available via Addgene and their sequences deposited in the NCBI database (accession numbers are given in Appendix 1.1). This toolkit and its associated characterisation data has been submitted for publication in the journal ACS Synthetic Biology.

- A new hyaluronan synthase enzyme, *S. thermophilus* LMD-9 HasA was shown to be functional in a heterologous host.

- An operon for potential hyaluronic acid production in *G. thermoglucosidans* was designed and constructed.

Impact

- The toolkit of parts produced could help to improve current biotechnological applications with *Geobacillus* species, such as production of ethanol, production of isobutanol or production of thermophilic proteins.

- The tools and methods demonstrated will accelerate the development of future applications including production of more complex molecules like hyaluronic acid or engineered strains for other activities such as bioremediation.

- The tools can be applied to the study of fundamental *Geobacillus* species biology to help understand the interesting biochemistry, evolution and ecology of this genus.

- The tools and methods could potentially also be applied to accelerate research and development of applications with many other related and industrially useful *Bacillus* species.

- *G. thermoglucosidans* has many advantages as a chassis organism and this study helps to establish this species as a new chassis for more in the synthetic biology community to consider using.

- This study encourages synthetic biologists more generally to look beyond model-organisms and provides a blueprint for establishing non-standard organisms as tractable chassis strains.

- This study now places *G. thermoglucosidans* as the leading chassis for high-temperature synthetic biology research and applications.

8.1.1 Overview of Parts and Protocols

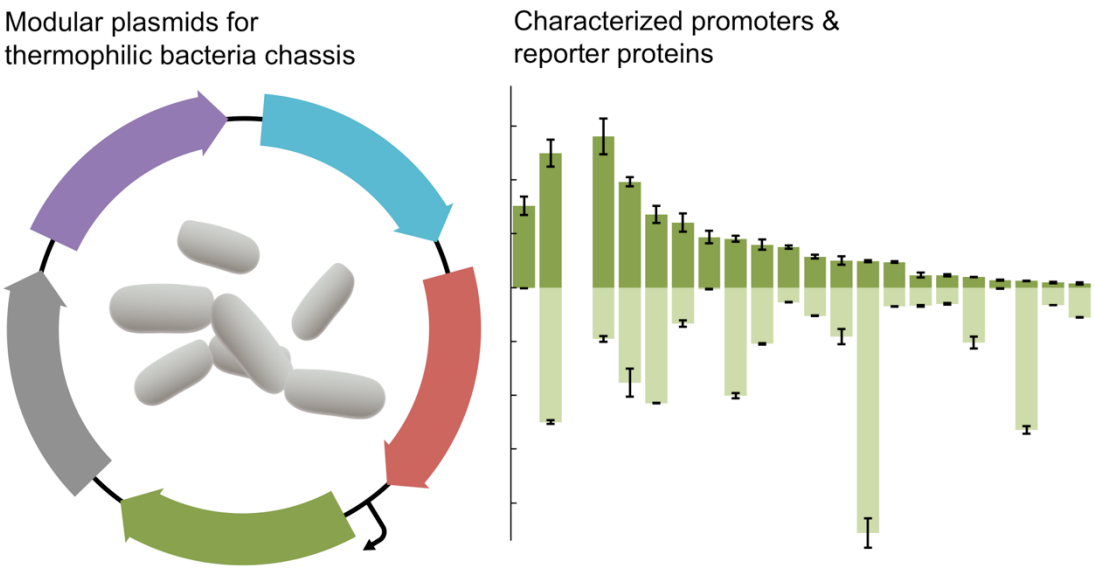


Figure 8.1 An overview of the genetic parts generated in this study

A significant legacy of this study is the toolkit of characterised parts and modules that has been generated and now has been made available to the research community. The pUP promoters and RplS promoters have already been used in a journal publication by Bartosiak-Jentys et al. (2013) to express hydrolases for secretion from *G. thermoglucosidans* (79) and these were also used to express reporter genes in PhD thesis of Dr Elena Martinez-Klimova (99). The modular plasmids have been shared with and used by several other academic groups and also with researchers in industry

at TMO Renewables Ltd. (now ReBio Technologies Ltd.) and at Corbion N.V. The toolkit of modular plasmids, reporter proteins and the RplS promoter library is currently available via Addgene with DNA sequences deposited in the NCBI database (accession numbers are given in Appendix 1.1). With these novel parts, the potential for more complex metabolic engineering has been shown here, with several promising targets identified. Unfortunately due to time constraints, the production of hyaluronic acid (HA) was not shown in *G. thermoglucosidans*, however the designed operon was shown to function and produce HA in *E. coli*. In preparing this operon, the *S. thermophilus* LMD-9 *hasA* hyaluronan synthase was also shown to be a promising new gene for heterologous HA production. The parts developed here will thus hopefully enable and inspire more ambitious engineering in *G. thermoglucosidans* in future.

Alongside the DNA parts developed in this study, protocols for research with *G. thermoglucosidans* have also been tested and revised, with updated methods for both electroporation and genomic DNA preparation demonstrated and detailed (Materials and Methods 2.2.9 and 2.2.11). In addition, different synthetic biology methods commonly used in other organisms have been tested and compared here. Both previously reported methods for promoter library generation were used and compared, as were multiple methods for promoter characterisation with a fluorescent reporter. The use of computational tools for predictable control of translation rate was investigated and a general review of these methods was published (184). Useful application of all of these methods with *G. thermoglucosidans* was shown and the Salis Lab RBS calculator in particular is recommended for informing sequence design for future metabolic engineering or other applications in *G. thermoglucosidans* (or indeed with other related organisms). Developing a modified. thermophile specific RBS calculator was unfortunately beyond the scope of this study, however, the work done here does promote thermophiles as production chassis and so this may encourage future developments in translation rate prediction for these alternative hosts.

Interest in thermophilic bacteria has grown considerably in the past decade and novel strains with unique advantages are constantly being sequenced and characterised. The parts and tools developed here are likely to be directly useful in many other species. All of the tools are likely to be directly applicable in closely-related species such as *Bacillus smithii,* an acid tolerant thermophile useful for production of organic acids

(213) and also in *Anoxybacillus* species, which are particularly suited to certain bioremediation applications (214). Other industrially useful Gram-positive thermophiles could benefit, particularly from the promoters and modular vectors developed here. *Thermoanaerobacter* species for example can grow on a broad range of feedstocks including syngas (33) and *Caldicellulosiruptor* species can utilise untreated lignocellulosic biomass (258). Certain parts, particularly the shuttle vectors or the resistance and reporter modules could find applications in more diverse thermophilic bacteria such as *Thermus* or *Thermosynecoccus* species. However, these parts would likely require some recharacterisation and/or the addition of new chassis specific modules (e.g. promoters and RBS sequences) to enable functioning in these hosts.

8.1.2 The Wider Impact of this Study

A Thermophilic Chassis for Synthetic Biology

This study has helped to promote *G. thermoglucosidans* at the thermophile chassis of choice for synthetic biology. Work presented here is arguably the first study with a foundational, parts-based synthetic biology approach to work with a thermophile chassis. This brings the chassis (*G. thermoglucosidans*) and its toolkit to the attention of the synthetic biology community and encourages synthetic biologists working with other chassis to consider applications in a thermophilic host.

The "model thermophile" in general microbiology is *Thermus thermophilus* due to decades of study as a source of thermostable enzymes and comparatively simple protocols for transformation and manipulation in the laboratory. However, this study and other recent works (65,79,99) argue for *G. thermoglucosidans* to be the thermophile of choice for industrial biotechnology and synthetic biology. In many ways *T. thermophilus* could be considered the thermophilic *E. coli* – well characterised and simple to work with in the laboratory, whereas *G. thermoglucosidans* is becoming the thermophilic *B. subtilis* – the preferred industrial production strain.

Alternative thermophiles such as the photosynthetic cyanobacterium *Thermosynechococcus elongatus* and hyperthermophilic archaeon *Pyrococcus furiosus*

have also been better-studied than *Geobacillus* species for understanding thermophile biology and as a source of thermophilic enzymes. They have unique advantages and can be transformed and genetically manipulated (259,260). However, they do not make attractive chassis organisms for applications beyond production of thermophilic proteins (261). Laboratory testing and genetic manipulations are generally more difficult than with *Geobacillus* species and, beyond being thermophilic, they lack many of the advantages of *Geobacilli* such as wide tolerance to stresses and feedstock flexibility. Strong rivals for thermophilic industrial applications include thermophilic *Clostridium* or *Thermoanaerobacter* species. These chassis bacteria may be more applicable for exploiting specific feedstocks or for producing particular products related to these. However, the *Geobacillus* genus' affinity for lignocellulosic biomass, their inhibitor and product tolerance and their broad potential applications in chemical production and bioremediation (65,70) arguably make them a more suitable, and flexible set of thermophilic bacteria for synthetic biology.

This work in no way intends to discourage work in these other species however, quite the opposite. This study aims to generally promote the use of alternative chassis organisms beyond the "pantheon of established production strains" (7). Diversifying the choice of available chassis is essential for the success of existing challenging applications and broadens the possibilities of synthetic biology into new areas. This work argues and demonstrates that parts can be developed for a new chassis based on existing synthetic biology techniques and principles, and that existing tools designed for model organisms can also be applied to these alternative strains. This process of testing existing parts and generating and characterising novel chassis specific modular parts to be used and shared can be replicated for any other novel organism of interest.

Due in part to this study and also due to other recent advances with *Geobacillus* species there are indications the synthetic biology community is indeed adopting Geobacilli as a potential new chassis. The SynBioMine tool (262), that is currently under development allows synthetic biology focussed mining and analysis of biological data and DNA sequence and is based on the successful InterMine system (263). It currently includes genomic data from *E. coli*, *Bacillus* and *Geobacillus* species. Additionally the RBS Calculator tool (186) that is discussed in Chapter 4, will soon be updated to

account for Gram-positive organisms and for thermophiles (Prof. Salis, Penn State, personal communication).

The choice of *G. thermoglucosidans* as a novel production chassis for hyaluronic acid in this study follows a wider trend in diversifying the chassis choice for microbial biological products. Initially, industrial production of a valuable product usually takes place in a natural strain found to produce it, *Streptococcus zooepidemicus* in the case of HA (264). The natural host is then improved genetically to boost production (219). Later enzymes from this strain or elsewhere are heterologously expressed in a model organism, usually *E. coli*, *B. subtilis* or *Saccharomyces cerevisiae* as these can be grown cheaply and are accompanied with many existing tools for high throughput strain optimisation and for increasing yield. As understanding of the underlying process then improves, production may later be attempted in a more ambitious, non-natural and/or non-model strain with specific advantages for the particular application. *G. thermoglucosidans* represents this third-stage host for HA production, with the potential for rapid (renewable) feedstock conversion at low viscosity. Other products that have followed a similar pattern of progress include polyhydroxybutyrate, (PHB) a promising biodegradable plastic. Industrial production was originally achieved in one of the many species of naturally PHB-producing bacteria such as *Azotobacter* (265) which were then evolved and improved. The process was then developed in model organisms such as *E. coli* (266). More ambitious non-native hosts have since been investigated. PHB has now been produced in a variety of chassis including oilseed rape *Brassica napus,* which allows direct production from sunlight and CO_2 (267) and the non-model yeast *Yarrowia lipolytica,* which is able to produce different polymer variants depending on the feedstock (268). HA is one of many products currently made by model organisms that could be explored in a wider range of chassis; these may allow more efficient production of HA or biosynthesis of interesting HA variants or composites of HA with other biopolymers. This study will hopefully encourage more non-model, non-native production chassis to be considered.

One of the commonly voiced concerns in the synthetic biology community is the need for the field to justify its “hype” and deliver “real-world applications” (9,19). As the discipline matures, perceived success of the engineering approach to biology hinders on impact outside of the laboratory. This work takes a firmly foundational synthetic

biology approach but is highly application focussed. The potential for thermophilic production of HA is promising and the parts generated have already proved to be useful to researchers in industry: TMO Renewables Ltd. (now ReBio Ltd.) for producing bioethanol from Geobacilli, and Corbion N.V. for producing organic acids from other thermophilic Bacilli. A related concern is that as the field of synthetic biology grows, broadens and becomes more applied, its founding principles – an engineering approach with rational design, modularity and abstraction, becomes lost or diluted (269)(270). This project demonstrates and argues the advantages of a fundamental synthetic biology approach and aims to encourage these principles in the study of other non-standard chassis organisms.

Synthetic Biology Approaches to Non-Model Organisms

Many non-model organisms, particularly extremophiles such as *G. thermoglucosidans* have small, tight knit, research communities built up around them. In demonstrating the benefit of a synthetic biology approach to modifying non-model chassis and for sharing parts within the community, this study hopes to encourage uptake of a similar approach within communities around other non-model chassis. Standardisation of protocols and characterisation methods improves collaboration and reproducibility within the community. Modularisation and abstraction allows easier sharing of parts and improves the ability to build on previous work. Finally, rational model-guided design allows more precise predictable engineering but is a rarely applied in non-standard organisms. The use of tools such as the Salis Lab RBS Calculator and Operon Designer demonstrated here shows these tools have utility beyond the model strains for which they were designed.

The synthetic biology approach can also help to break down the barriers between separate research communities, promoting the sharing of parts and comparisons of data. This study took parts (and inspiration) from other non-model organisms with sfGFP selected due to its previous characterisation in *T. thermophilus* (135) and the modular vector architecture based on that of the pMTL Clostridial plasmids (62). The parts and data generated in this study have in turn been shared with researchers outside the Geobacilli community (Elleke Bosma and colleagues, Wageningen University). With a synthetic biology approach to sharing parts and tools between different hosts as well

as modular designs that can be more easily refactored for alternative chassis, researchers could become less wedded to their favourite organisms and be able to instead consider a broader range of chassis, and to select one with particular application-specific advantages. As researchers continue to explore the microbial diversity of our planet, new extremophiles like thermophiles are constantly being sequenced and characterised. The synthetic biology community will hopefully begin to take more interest in this growing group of organisms. Similarly, for researchers characterising new strains, studies such as this one encourage the inclusion of synthetic biology relevant information. Traditional reporting of a new strain typically has included phenotypic data, e.g. colony morphology, preferred growth media, antibiotic resistances, etc. Synthetic biology relevant data such as the functionality of existing plasmids and reporter genes in the new organism could now become increasingly important essential information.

Expanding Research with *Geobacillus* Species

The parts and tools developed in this study can also aid wider biotechnology and microbiology research with the *Geobacillus* genus. A huge range of valuable, highly stable enzymes are derived from *Geobacillus* species including restriction enzymes, hydrolases, enzymes for bioconversion of valuable commodity chemicals, for bioremediation and many others area (122). However, due to a lack of genetic parts for expression in Geobacillus species, genes are usually cloned into heterologous expression host such as *E. coli* (271). Many enzymes however, due to codon usage, temperature requirements on folding or presence of specific cofactors or chaperones, are better expressed in their native *Geobacillus* species host (272,273). The promoters and shuttle vectors presented here, functional in both *Geobacillus* species and *E. coli,* will allow researchers to easily test overexpression in either chassis. This will speed up discovery and development of thermophilic enzymes, an area of huge industrial importance.

Beyond direct industrial value, *Geobacillus* species are biochemically, ecologically and evolutionarily fascinating. Geobacilli and their spores are amongst the most durable, long-lived and widely distributed organisms on the planet (68). They have an amazing ability to thrive in challenging environments, with enzymes and biochemistry that are

extraordinarily tolerant to stresses. We still have much to learn which will no doubt benefit future biotechnology applications but also improve our fundamental understanding of microbial ecology and evolution. This work can further our understanding of more general extremophile biology. The compact broad host plasmids and promoters generated here could be used to study gene transfer or to upregulate or knock down native gene expression to study its effect on phenotype.

8.2 Future work

Immediate future work from this study would be testing for HA production in *G. thermoglucosidans* with the constructed operon expressed from a weaker promoter or with weaker RBS sequences. If successful, expression from the operon would then be optimised and production in a bioreactor tested and assessed with different feedstocks. Other work would focus on improving the characterisation of existing parts and the development of new parts that expand the functionalities of the toolkit. More foundational biology to better understand the chassis strain chosen for this study (*G. thermoglucosidans*) would be valuable alongside this. Finally, other applications including production of alternative products or bioremediation could be explored.

8.2.1: Improving Protocols and Reporter Genes

Protocols for research with Geobacillus species are generally well established. Improving transformation efficiency is a priority however. Electroporation efficiency was low with *G. thermoglucosidans* (10^2 to 10^4 CFU/µg plasmid DNA) such that library generation required transformation into *E. coli* first and efficiency was even lower with other *Geobacillus* species, *G. stearothermophilus* and *G. thermodenitrificans* (10^1 to 10^2 CFU/µg plasmid DNA). Large or burdensome plasmid constructs may decrease efficiency even further. Using DNA prepared from an *E. coli* strain expressing *G. kaustophilus* methylase enzymes has been shown to improve conjugation efficiency (98) and would likely allow a similar improvement for electroporation. Enabling conjugative transfer of the modular vectors would also be a priority. Work towards this goal has already been published with an origin of transfer

(OriT) in pUCG18, a plasmid containing the same replicon (repBSTI) and selectable maker (TK101 Kan^R) as pG1K (87). Thus, an origin of transfer (OriT) would be the next module to include in a future updated plasmid toolkit.

Along with improved plasmids and transformation methods, a modified, high transformation efficiency laboratory strain could also be developed. Most strains used in *E. coli* synthetic biology, for example DH5 Alpha or BL21-DE3, have extensive genome modifications that aid transformation rates and stable maintenance of plasmids. Upregulation of *Geobacillus* competence genes and knock out of any nucleases could improve transformation as it has for other bacteria. The strain could also have enzymes involved in recombination knocked out to improve construct stability and could be evolved to be fast growing on laboratory media. Constructs could be more rapidly tested and modified in an improved laboratory workhorse strain before final testing in a panel of more hardy, wild-type-like strains specialised for industrial or environmental applications.

For reporter genes, developing a functional anaerobic fluorescent reporter remains a priority. The absence of characterisation under different oxygen conditions represents a significant gap in the promoter characterisation data presented here, despite repeated attempts to solve this challenge. For high throughput characterisation and for measurements by flow cytometry, a fluorescent reporter is effectively essential. Eventually a LOV protein may be tested that works as expected; more of these are becoming available as their utility in bioscience research becomes more widely known. Indeed, a new anaerobic fluorescent LOV protein, CreiLOV, has recently been described and shown to be thermostable up to 60 °C (274) thus providing a further opportunity for testing an anaerobic reporter for *Geobacillus*. If this were also to fail, then as the biochemistry and molecular genetics of *Geobacillus* species becomes better understood the reason for the lack of LOV protein expression may become elucidated. This could then aid in the design of a functional LOV-based anaerobic reporter that circumvents the reason for the failures of the previous variants. Alternatively a totally novel class of anaerobic fluorescent reporter might emerge or improved thermostability could be evolved (81) or designed (154). Other useful characterisation tools currently used in mesophiles such as the fluorescent RNA aptamer “Spinach” could be adapted

for thermophiles. Spinach RNA allows fluorescence characterisation of transcription and of mRNA stability (275), and a thermostable Spinach RNA could be similarly evolved or designed for higher temperatures.

8.2.2 Further Promoters

For the existing promoter parts described in this thesis, improving characterisation across a broader range of conditions is a future priority. Characterising redox effects on promoter strength is particularly important for promoters used in strains for bioreactor fermentations. The previously used Ldh promoter is known to fluctuate significantly in its strength with changing redox conditions and comparing this to the UP and RplS promoters would be valuable. Currently the suggested constitutive nature of pUP and pRplS is based only on comparisons to homologues in *B. subtilis* and so requires confirmation in *G. thermoglucosidans.* Characterising expression under other stresses, different growth temperatures and with different media – particularly cellobiose media or a pretreated lignocellulosic feedstock based media – would also be valuable.

Expanding beyond constitutive promoters to parts with expression that are inducible based on temperature, redox state or addition of an inducer would be the next step. If a strongly inducible system cannot be adapted from a thermophile then existing, successful mesophilic systems such as the LacI/pLac and AraC/pBAD could be adapted by evolving thermostability in the transcription factor. If the Lac and araBAD promoters do not function in *G. thermoglucosidans,* then libraries could be made of these and improved expression selected. Alternatively, a transcription factor binding site could be added to a pUP or pRplS sequence to add regulation (likely repression) to these constitutive promoters. Simple, temperature inducible promoters could also be produced based on mesophilic repressor proteins where raising the growth temperature denatures the mesophilic repressor, thus allowing transcription from the promoter.

In particular, the development of inducible promoters and characterised transcription factors known to specifically activate or repress certain promoter sequences would then allow more complex synthetic biology devices and systems to be built such as feedback loops, timer switches and genetic logic gates (23). Such genetic systems can help the

production of complex or composite products and are necessary to develop future applications in biosensing and bioremediation (182).

Inducible promoters are also useful for industrial overproduction of a protein of interest. Allowing a culture to reach exponential growth before inducing protein production can give higher yields than constitutive expression, as significant overexpression reduces a cell's growth rate and thus lowers yields from batch production. For expression of certain native proteins, *Geobacillus* species may be the most suitable chassis, as the thermophilic protein may not be well expressed in a model mesophilic expression strain. The promoters developed in this study could improve protein production in *Geobacillus* species but an inducible promoter would be even more advantageous. Combining the rapid feedstock conversion and cheap feedstock requirements of *G. thermoglucosidans* with a suitable inducible expression system would allow this chassis to be considered for a range of protein production applications.

High yield protein production in mesophilic chassis often uses the T7 expression system (276) where production of the viral T7 polymerase is induced and then specifically transcribes the protein of interest highly specifically from its cognate T7 promoter at very high levels. A thermostable T7 polymerase variant which functions at 50 °C has not been published but has been patented (277) and testing this or a similarly evolved thermostable variant in *G. thermoglucosidans* could be very valuable.

For characterisation of current and future promoters in *G. thermoglucosidans* and other thermophiles, more high throughput protocols would greatly improve the quality and quantity of characterisation data. Developing protocols for growth in multiwell plates, in a high temperature plate reader would allow florescence measurements to be taken in real-time giving better, more reproducible characterisation data. It would allow larger libraries to be screened and also give dynamic data, important for characterising inducible systems or for more complex genetic circuits.

8.2.3 Improving RBS Sequence Design

The utility of translation rate prediction for *G. thermoglucosidans* was demonstrated in this study but also shown to be far from ideal. For future work with metabolic engineering or for other applications described below, the use of the Salis Lab RBS Calculator to inform design is recommended but could be improved. In future, generating data to help refine the calculator for Gram-positive bacteria and thermophiles in particular would be valuable. Parameters in the current model were determined based on data for the *in vivo* translation rates of over 100 RBS sequences in *E. coli* (186). Repeating this data collection with the sequences in *G. thermoglucosidans* expressing the sfGFP reporter would allow the model to be adapted for this chassis. Measurements at temperature intervals between 45 and 65 °C would also allow temperature dependence to be better understood and potentially incorporated into the model. Advances in our fundamental understanding of thermophile biochemistry will also aid in improving the model, especially in terms of how complexes such as the translation initiation complex are constructed and stabilised at high temperatures. These insights will allow better predictions in the future.

8.2.4 Improved Plasmids and Modules

A huge range of alternative modules could be added to the vector set to expand its functionality. As mentioned above, an origin of transfer (OriT) would be a first priority. Conjugation from *E. coli* to *G. thermoglucosidans* with an OriT added to the vector pUCG18 (83), that has the same replicons and resistance markers as pG1K, has already been demonstrated (98) and so this OriT sequence would be the first choice. Conjugation was shown to give reasonable efficiency with a range of thermophilic *Bacillus* and *Geobacillus* species (98) and so would provide a solution to the problem of lower electroporation efficiencies with *Geobacillus* species other than *G. thermoglucosidans* with these vectors. Alternatives to current modules could also be provided such as a thiostrepton resistance marker (82) or repBC1, an alternative minimal replicon from the pBC1 plasmid (203) which is known to function in Geobacilli (88) and may be compatible with the current replicons repB or repBSTI.

To better standardise parts and plasmids across with other collection, resynthesizing the modules to be fully-compatible with the increasingly popular SEVA Standard (195,201) would also be valuable. This could encourage the *Geobacillus* plasmid set vectors to be used more-widely and allow the creation of novel vectors with modules taken from both collections. It is also worth noting that the highest electroporation efficiency in this study was seen with plasmid pG1K and that this is free of BsaI restriction sites. This plasmid can therefore be amplified and adapted to be used as an entry vector and first destination vector for simple Golden Gate cloning (257) (Figure 8.2 a). To accelerate the development of applications that require more complex genetic constructs in *Geobacillus* species, a modular cloning (MoClo) kit of parts would ideally be generated that can be combinatorially combined by hierarchical assembly following the Golden Gate DNA assembly method (278). With such a MoClo kit, libraries of parts are generated and stored in entry vectors with flanking standardised, 4 base-pair sequences for assembly followed by type IIS restriction sites (BsaI for example). When cut out, these parts can then be assembled in a standard configuration (promoter-RBS-coding sequence) into a suitable level-1 destination vector that again flanks the construct with sites for assembly and alternative type IIS restriction sites (BpiI for example). These compound parts may then be further cut out and assembled into a level-2 destination vector to created large multi-gene constructs (278) (Figure 8.2 b).

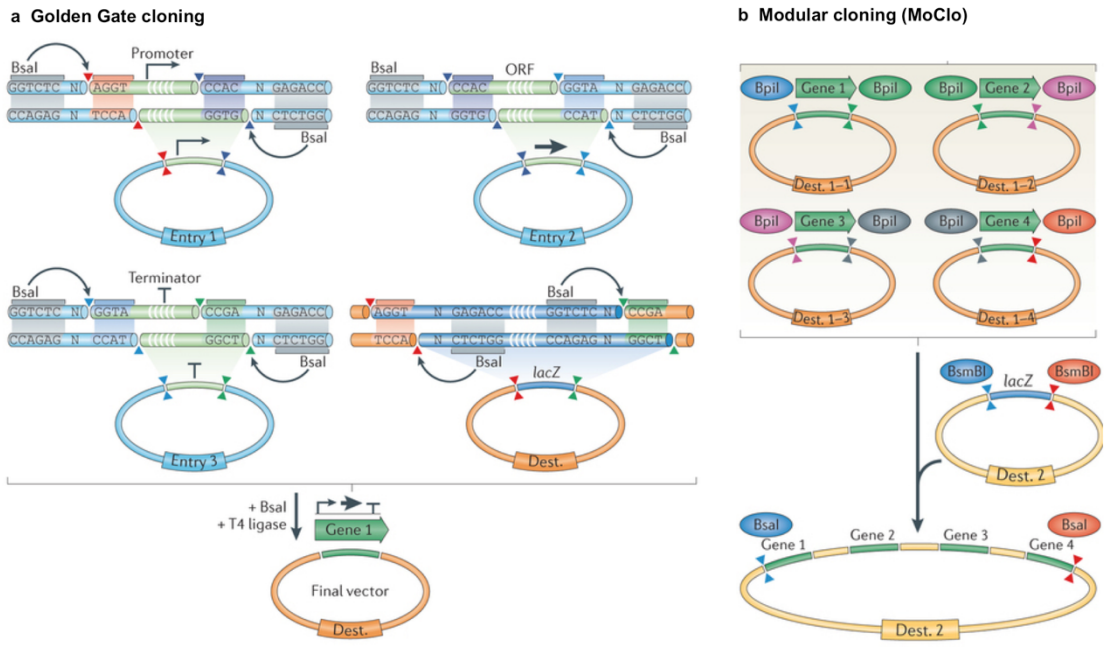


Figure 8.2. a) Golden Gate assembly. Modular parts stored in entry vectors are cut out using BsaI a type IIS restriction enzyme that cuts outside its recognition sequence allowing overhang sequences that can be chosen. Complementary overlap sequences are designed such that fragments are ligated together, into the similarly cleaved destination vector (labelled 'Dest.'), in a defined order. **b) Modular cloning.** Destination vectors can be designed to flank the assembled construct from a first round Golden Gate reaction with alternative type IIS restriction sites. A second reaction then assembles multigene constructs in a similar manner. With these methods, libraries of parts stored in entry vectors can be combined combinatorially into large complex constructs. Figure adapted from Casini *et al.* 2015 (279).

In future, mutation of a single BsaI site and BpiI site in pG2K would allow this plasmid to then be used as a level 1 and level 2 destination vector for modular cloning. Reformatting the RplS library promoters and RBS library 5'-UTR sequences as modules in pG1K entry vectors for modular cloning by this method would be valuable.

For targeted genome modifications in Geobacillus species, adding modules encoding a functional CRISPR/Cas9 system could hugely accelerate strain engineering (280). Natural CRISPR systems are abundant in thermophilic bacteria (281) however they have not been used for genome editing in thermophilic species so far. The CRISPR system of *S. thermophilus* has been well characterised however (282), and the StCas9 protein has been used for genome modification in many hosts including, *E. coli*, *S. cerevisiae* and human cell lines (283,284). StCas9 is likely to also function in *G. thermoglucosidans* and would be a valuable addition to the modular toolkit. *S. thermophilus* is only a moderate thermophile but thermostability in this case is not critical. The Cas9 would only need to be expressed around 45 °C (a temperature at which both *S. thermophilus* and *G. thermoglucosidans* can grow well, though not optimally) for the permanent genome modifications to be made.

Should the StCas9 be stable at higher temperatures (or alternatively, a Cas9 from a true thermophile shown to function heterologously) a "dead"-Cas9 (dCas9) for targeted gene knockdowns (repression) and upregulation (activation) could also be generated (285). Here a mutation is made to the Cas9 nuclease catalytic domain that "kills" nuclease activity but retains targeted DNA binding activity. Designing a guide RNA to targeting dCas9 to a promoter region or the start of a coding sequence, downregulates

gene expression by stearic repression, whereas targeting a dCas9 fused to the ω subunit of RNA polymerase to a sequence upstream of a promoter region can upregulate transcription from that promoter (285). These advanced tools in *Geobacillus* would enable the study of gene function, improved strain engineering and the construction of more complex synthetic gene networks.

8.2.5 Advancing Metabolic Engineering

Given the progress made in Chapter 7, testing for hyaluronic acid (HA) expression in *G. thermoglucosidans* with the *S. thermophilus hasA* or with a more weakly expressed synthetic *has* operon would be the first priority for future work in this area. There are significant possible advantages to a thermophilic process for HA biosynthesis and the *hasA* enzyme is expected to be stable to at least 50 °C. Reducing expression with a weaker promoter or ribosome binding site sequence should reduce burden enough for *G. thermoglucosidans* to be transformed. Possible production of HA from both the single synthase and 3-gene operon would be tested at 45+ °C on a variety of media including with N-acetyl-glucosamine and glucuronic acid supplementation to reduce the metabolic burden of their production. Should HA production not be detected at all, no expression, or non-functional expression of *hasA* would be the most likely explanation. This would argue against the aims of this thesis to promote *G. thermoglucosidans* as a production chassis for a range of biological products beyond proteins or alcohols. However, with *S. thermophilus* being moderately thermophilic and its *hasA* shown to be functionally expressed in a heterologous host (*E. coli*), functional expression in *G. thermoglucosidans* seems likely. Temperature and media would then be optimised at 50 ml tube or small flask scale. Ideally a cheap, non-supplemented media would be found, e.g. LB with autoclaved tap water plus glucose or sucrose. The operon would then be optimised by generating a library of constructs with variable promoter and RBS sequences and screening for high HA production (as described in Chapter 7.4). This would then be a decision point, if yields are comparable to those with recombinant *B. subtilis* (221,224) then scale up and further optimisation could take place. Small scale bioreactor tests would be particularly interesting to test the impact of viscosity on stirring and aeration and see how this changes with temperature. Thermophiles may present an interesting solution for the otherwise challenging

production of viscous products. Alternatively, the lowered oxygen solubility at higher temperatures might reduce yields, and aeration has generally been found to improve HA production elsewhere (256), although that may not be the case with *G. thermoglucosidans*.

Product quality would also need to be assessed. Consistent, pure, high molecular weight (>2 MDa) HA is the desired product. In heterologous hosts molecular weight tends to be lower but more consistent. Molecular weight can be increased by improving precursor synthesis relative to HA synthase expression (256) and so this may be another factor which must be considered during optimisation. If the product looks promising then further process optimisation, feedstock optimisation and strain improvement will be required before testing at larger scales could be attempted.

Even if *G. thermoglucosidans* does not seem to be a promising HA production chassis, the *S. thermophilus* LMD-9 *hasA* enzyme could be a better candidate than the previously used *S. equisimilis* hyaluronan synthases for recombinant HA production in *E. coli* or other chassis bacteria. Biochemical analysis of this *hasA* enzyme compared to those of other production strains would be useful, although this work is typically challenging for transmembrane enzymes. Comparisons of thermostability, temperature optima, catalytic rates and substrate affinities would be valuable data.

Beyond HA there are many promising targets for thermophilic production in *G. thermoglucosidans* that could be explored in future. Polyhydroxyalkanoates (PHAs) are group of microbially produced polymers with many applications as biodegradable plastics. An operon for their production in *Thermus thermophilus* HB8 has been studied (286). Refactoring this operon for production in *G. thermoglucosidans* offers all the benefits of this host (fast feedstock conversion, reduced contamination etc.) as well as potentially increased yields as *G. thermoglucosidans* lacks the enzymes that naturally degrade these polymers. PHA production is a higher volume, lower value biotechnology, so production from cheap feedstocks such as lignocellulosic biomass becomes important for this case.

8.2.6 Improving Chassis Characterisation

This study focussed on the development and characterisation of modular DNA parts with *G. thermoglucosidans* chosen as the chassis. However, characterisation of the chassis itself (not just the parts) is equally important. A rigorous comparative phenotypic study of thermophilic *Bacillus* and *Geobacillus* species would be particularly helpful further work to better inform future choices for thermophilic production strains. Factors to test could include growth rate, optimal growth temperatures, carbon sources, the effects of pH, oxygen and solvent stresses and DNA competence. This group of organisms has general advantages but the best chassis strain will depend based on the particular feedstock and application being developed.

Improved our understanding and characterisation of *Geobacillus* species biology would also aid future applications with this strain. For *G. thermoglucosidans*, genome scale metabolic models have been developed (287) and metabolic fluxes during fermentation for ethanol production have been studied (67,288). Wider metabolomic profiling under a range of conditions would be helpful for broader future applications and aid in tying together the known genome sequence with the results of metabolic modelling. To date no transcriptomics data has been published and this would be hugely valuable to give a better understanding of gene expression and regulation. It could also improve genome annotation, promoter prediction and aid in an understanding of how to engineer the chassis strain to alter its metabolism.

8.2.7 Future Outlook

G. thermoglucosidans has previously been successfully engineered to produce simple, low value molecules (alcohols for biofuel) from lignocellulosic biomass (60,73). As the tools and chassis characterisation have improved, production of higher value products is now possible. The use of cheap lignocellulosic particularly benefits production of high volume products and so organic acids (lactic acid succinic acid etc.) and biopolymers (PLA, PHAs, HA) are the next most likely targets to be developed in the near future. *Geobacillus* species are also now promising expression hosts for the evolution and production of thermostable enzymes for biotechnological applications.

This is a growing area and as more thermophile genomes are sequences many more valuable enzymes will be discovered.

Further ahead, *Geobacillus* species could be engineered for production of intermediate and higher value products, particularly where the production or downstream processing steps could benefit from a thermophilic process. With their eclectic taste in feedstocks *Geobacillus* species could also be engineered to give valuable products from niche feedstocks such as industrial waste streams that may contain unusual contaminants. With the ability to degrade environmental pollutants such as long chain alkanes (289) and organophosphates (70), *Geobacillus* species also have potential for use in pollution control and bioremediation (71) (69). Engineered strains could have improved degrading capabilities and *Geobacillus* spores are highly stable and long lived so could be stored and transported easily. With these abilities to detect and uptake a variety of contaminants *Geobacillus* species could also make interesting chassis for whole cell biosensing.

8.3 Conclusion

The aim of this project was to improve the parts and tools available for synthetic biology in the thermophile *G. thermoglucosidans,* and then to test the potential of these tools and this microbial chassis for the production a complex higher-value product by metabolic engineering. The four foundational aims were achieved: reporter genes were tested and characterised, promoter libraries constructed and characterised, design software for RBS sequences was shown to be useful (with limitations) and minimal modular shuttle vectors were constructed, characterised and have been formatted into a shareable toolkit that is soon to be published. Considerable progress towards the applied goal of hyaluronic acid production was made with a new operon designed and constructed and a suitable hyaluronan synthase enzyme shown to function in a heterologous host. Due to this work, the *G. thermoglucosidans* bacteria is now far more accessible for synthetic biology applications and has a promising future as the thermophile chassis of choice for the production of renewable biobased chemicals and for other applications.

Chapter 9: Bibliography

1. Lacey P, Keeble J, McNamara R. Circular advantage: innovative business models and technologies to create value in a world without limits to growth. 2014.

2. Drzeniek-Hanouz M, Marti G, Galvan C. The Global Risks Report 2016. 2016.

3. Nieves LM, Panyon LA, Wang X. Engineering Sugar Utilization and Microbial Tolerance toward Lignocellulose Conversion. Front Bioeng Biotechnol. 2015 Jan;3:17.

4. Lopes MSG. Engineering biological systems toward a sustainable bioeconomy. J Ind Microbiol Biotechnol. 2015 Jun;42(6):813–38.

5. van der Ploeg F. Natural Resources: Curse or Blessing? J Econ Lit. 2011;49(2):366–420.

6. Langeveld H, Sanders J, Meeusen M. The Biobased Economy: Biofuels, Materials, and Chemicals in the Post-oil Era. Earthscan; 2012. 1-10 p.

7. National Research Council. Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals. Washington, D.C.: National Academies Press; 2015 Jun.

8. Wallace S, Balskus EP. Opportunities for merging chemical and biological synthesis. Curr Opin Biotechnol. 2014 Dec;30:1–8.

9. Church GM, Elowitz MB, Smolke CD, Voigt CA, Weiss R. Realizing the potential of synthetic biology. Nat Rev Mol Cell Biol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Apr;15(4):289–94.

10. Keasling JD. Synthetic biology and the development of tools for metabolic engineering. Metab Eng. 2012 May;14(3):189–95.

11. The European Bioeconomy in 2030 -Delivering Sustainable Growth by addressing the Grand Societal Challenges' [Internet]. 2011 [cited 2016 Feb 3]. Available from: http://www.epsoweb.org/file/560

12. Friedman DC, Ellington AD. Industrialization of Biology. American Chemical Society; 2015.

13. Clarke L. A synthetic biology roadmap for the UK. 2012.

14. Baldwin G, Bayer T, Dickinson R, Ellis T, Freemont PS, Kitney RI, et al. Synthetic Biology — A Primer: Revised Edition. World Scientific; 2015.

15. Endy D. Foundations for engineering biology. Nature. 2005 Nov 24;438(7067):449–53.

16. Pauwels E. Communication: Mind the metaphor. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013 Aug 29;500(7464):523–4.

17. de Lorenzo V. Beware of metaphors: chasses and orthogonality in synthetic biology. Bioeng Bugs. Taylor & Francis; 2011 Jan 1;2(1):3–7.

18. Shetty RP, Endy D, Knight TF. Engineering BioBrick vectors from BioBrick parts. J Biol Eng. BioMed Central; 2008 Jan 14;2(1):5.

19. Kwok R. Five hard truths for synthetic biology. Nature. Nature Publishing Group; 2010 Jan 21;463(7279):288–90.

20. Canton B, Labno A, Endy D. Refinement and standardization of synthetic biological parts and devices. Nat Biotechnol. Nature Publishing Group; 2008 Jul;26(7):787–93.

21. Federici F, Rudge TJ, Pollak B, Haseloff J, Gutiérrez RA. Synthetic Biology: opportunities for Chilean bioindustry and education. Biol Res. 2013 Jan;46(4):383–93.

22. Kelwick R, MacDonald JT, Webb AJ, Freemont P. Developments in the tools and methodologies of synthetic biology. Front Bioeng Biotechnol. 2014 Jan;2:60.

23. Ellis T, Wang X, Collins JJ. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat Biotechnol. Nature Publishing Group; 2009 May;27(5):465–71.

24. Nikel PI, Martínez-García E, de Lorenzo V. Biotechnological domestication of pseudomonads using synthetic biology. Nat Rev Microbiol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 May;12(5):368–79.

25. Goh E-B, Baidoo EEK, Keasling JD, Beller HR. Engineering of bacterial methyl ketone synthesis for biofuels. Appl Environ Microbiol. 2012 Jan 1;78(1):70–80.

26. Renninger N, McPhee D. Fuel compositions comprising farnesane and farnesane derivatives and method of making and using same. USA; US20080083158 A1, 2008.

27. Marliere P. Production of alkenes by enzymatic decarboxylation of 3-hydroxyalkanoic acids. US20110165644 A1, 2011.

28. Weyler W, Dodge TC. Microbial production of indigo. US6303354 B1, 2001.

29. Tawasha M. Eastman, Genencor to market new process [Internet]. ICS News. 1999 [cited 2016 Mar 29]. Available from: http://www.icis.com/resources/news/1999/08/12/91877/eastman-genencor-to-market-new-process/

30. Fahnestock SR, Irwin SL. Synthetic spider dragline silk proteins and their production in Escherichia coli. Appl Microbiol Biotechnol. 1997 Jan;47(1):23–32.

31. Xia X-X, Qian Z-G, Ki CS, Park YH, Kaplan DL, Lee SY. Native-sized recombinant spider silk protein produced in metabolically engineered Escherichia coli results in a strong fiber. Proc Natl Acad Sci U S A. 2010 Aug 10;107(32):14059–63.

32. Saha BC. Hemicellulose bioconversion. J Ind Microbiol Biotechnol. 2003 May;30(5):279–91.

33. Weghoff MC, Müller V. CO metabolism in the thermophilic acetogen Thermoanaerobacter kivui. Appl Environ Microbiol. 2016 Feb 5;

34. Alam F, Mobin S, Chowdhury H. Third Generation Biofuel from Algae. Procedia Eng. 2015;105:763–8.

35. Alvira P, Tomás-Pejó E, Ballesteros M, Negro MJ. Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolysis: A review. Bioresour Technol. 2010 Jul;101(13):4851–61.

36. Bommarius AS, Sohn M, Kang Y, Lee JH, Realff MJ. Protein engineering of cellulases. Curr Opin Biotechnol. 2014 Oct;29:139–45.

37. Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, et al. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science. American Association for the Advancement of Science; 2007 Feb 9;315(5813):804–7.

38. Chen R, Dou J. Biofuels and bio-based chemicals from lignocellulose: metabolic engineering strategies in strain development. Biotechnol Lett. 2015 Oct 14;

39. Ohgren K, Bengtsson O, Gorwa-Grauslund MF, Galbe M, Hahn-Hägerdal B, Zacchi G. Simultaneous saccharification and co-fermentation of glucose and xylose in steam-pretreated corn stover at high fiber content with Saccharomyces cerevisiae TMB3400. J Biotechnol. 2006 Dec 1;126(4):488–98.

40. Salehi Jouzani G, Taherzadeh MJ. Advances in consolidated bioprocessing systems for bioethanol and butanol production from biomass: a comprehensive review. Biofuel Res J. 2015 Mar 1;2(1):152–95.

41. Mills TY, Sandoval NR, Gill RT. Cellulosic hydrolysate toxicity and tolerance mechanisms in Escherichia coli. Biotechnol Biofuels. 2009 Jan;2:26.

42. Martinez A, Rodriguez ME, York SW, Preston JF, Ingram LO. Effects of Ca(OH)(2) treatments ("overliming") on the composition and toxicity of bagasse hemicellulose hydrolysates. Biotechnol Bioeng. 2000 Sep 5;69(5):526–36.

43. Wang X, Yomano LP, Lee JY, York SW, Zheng H, Mullinnix MT, et al. Engineering furfural tolerance in Escherichia coli improves the fermentation of lignocellulosic sugars into renewable chemicals. Proc Natl Acad Sci U S A. 2013 Mar 5;110(10):4021–6.

44. Zaldivar J, Nielsen J, Olsson L. Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. Appl Microbiol Biotechnol. 2001 Jul;56(1-2):17–34.

45. Chen GG-Q, Jewett MC. Editorial: Transforming biotechnology with synthetic biology. Biotechnol J. 2016 Feb;11(2):193–4.

46. Tesfaw A, Assefa F. Current trends in bioethanol production by Saccharomyces cerevisiae: substrate, inhibitor reduction, growth variables, coculture, and immobilization. Int Sch Res Not. 2014;

47. van Maris AJA, Abbott DA, Bellissimi E, van den Brink J, Kuyper M, Luttik MAH, et al. Alcoholic fermentation of carbon sources in biomass hydrolysates by Saccharomyces cerevisiae: current status. Antonie Van Leeuwenhoek. 2006 Nov;90(4):391–418.

48. Ito K, Yoshida K, Ishikawa T, Kobayashi S. Volatile compounds produced by the fungus Aspergillus oryzae in rice Koji and their changes during cultivation. J Ferment Bioeng. 1990 Jan;70(3):169–72.

49. Kang SW, Park YS, Lee JS, Hong SI, Kim SW. Production of cellulases and hemicellulases by Aspergillus niger KK2 from lignocellulosic biomass. Bioresour Technol. 2004 Jan;91(2):153–6.

50. Bokinsky G, Peralta-Yahya PP, George A, Holmes BM, Steen EJ, Dietrich J, et al. Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered Escherichia coli. Proc Natl Acad Sci U S A. 2011 Dec 13;108(50):19949–54.

51. Li XZ, Webb JS, Kjelleberg S, Rosche B. Enhanced benzaldehyde tolerance in Zymomonas mobilis biofilms and the potential of biofilm applications in fine-chemical production. Appl Environ Microbiol. 2006 Feb;72(2):1639–44.

52. Wojtusik M, Rodríguez A, Ripoll V, Santos VE, García JL, García-Ochoa F. 1,3-Propanediol production by Klebsiella oxytoca NRRL-B199 from glycerol. Medium composition and operational conditions. Biotechnol Reports. 2015 Jun;6:100–7.

53. Zhang X-Z, Zhang Y-HP. One-step production of biocommodities from lignocellulosic biomass by recombinant cellulolytic Bacillus subtilis: Opportunities and challenges. Eng Life Sci. 2010 Oct 2;10(5):398–406.

54. Bergey DH. Thermophilic Bacteria. J Bacteriol. 1919 Jul;4(4):301–6.

55. Vane LM, Alvarez FR. Membrane-assisted vapor stripping: energy efficient hybrid distillation-vapor permeation process for alcohol-water separation. J Chem Technol Biotechnol. 2008 Sep;83(9):1275–87.

56. Hild HM, Stuckey DC, Leak DJ. Effect of nutrient limitation on product formation during continuous fermentation of xylose with Thermoanaerobacter ethanolicus JW200 Fe(7). Appl Microbiol Biotechnol. 2003 Feb;60(6):679–86.

57. Takami H, Takaki Y, Chee G-J, Nishi S, Shimamura S, Suzuki H, et al. Thermoadaptation trait revealed by the genome sequence of thermophilic Geobacillus kaustophilus. Nucleic Acids Res. 2004 Jan;32(21):6292–303.

58. Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA. Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. Trends Biotechnol. 2009 Jul;27(7):398–405.

59. P S, T G, B A. Potential for using thermophilic anaerobic bacteria for bioethanol production from hemicellulose. Portland Press Ltd.; 2004 Apr 1;

60. Cripps RE, Eley K, Leak DJ, Rudd B, Taylor M, Todd M, et al. Metabolic engineering of Geobacillus thermoglucosidasius for high yield ethanol production. Metab Eng. 2009;11(6):398–408.

61. Jones DT, Woods DR. Acetone-butanol fermentation revisited. Microbiol Rev. 1986 Dec;50(4):484–524.

62. Heap JT, Pennington OJ, Cartman ST, Minton NP. A modular system for Clostridium shuttle plasmids. J Microbiol Methods. 2009 Jul;78(1):79–85.

63. Lin L, Xu J. Dissecting and engineering metabolic and regulatory networks of thermophilic bacteria for biofuel production. Biotechnol Adv. 2013 Nov 16;31(6):827–37.

64. Chen J, Zhang Z, Zhang C, Yu B. Genome Sequence of Geobacillus thermoglucosidasius DSM2542, a Platform Hosts for Biotechnological Applications with Industrial Potential. Journal of Biotechnology. 2015.

65. Kananavičiūtė R, Čitavičius D. Genetic engineering of Geobacillus spp. J Microbiol Methods. 2015 Apr;111:31–9.

66. Lynn Yarris. Driving for Biofuels: New Technique Speeds Search for Biofuel Microbes [Internet]. Berkeley, CA; 2009 [cited 2012 May 15]. Available from: http://newscenter.lbl.gov/2009/03/19/driving-for-biofuels-new-technique-speeds-search-for-biofuel-microbes/

67. Tang YJ, Sapra R, Joyner D, Hazen TC, Myers S, Reichmuth D, et al. Analysis of metabolic pathways and fluxes in a newly discovered thermophilic and ethanol-tolerant Geobacillus strain. Biotechnol Bioeng. 2009 Apr 1;102(5):1377–86.

68. Zeigler DR. The Geobacillus paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? Microbiology. 2014 Jan;160(Pt 1):1–11.

69. Zheng C, He J, Wang Y, Wang M, Huang Z. Hydrocarbon degradation and bioemulsifier production by thermophilic Geobacillus pallidus strains. Bioresour Technol. 2011 Oct;102(19):9155–61.

70. McMullan G, Christie J, Rahman T, Banat I, Ternan N, Marchant R. Habitat, applications and genomics of the aerobic, thermophilic genus Geobacillus. Biochem Soc Trans. 2004 Apr;32(Pt 2):241–7.

71. Tulasi S, Littlechild J, Kawarabayasi Y. Thermophilic Microbes in Environmental and Industrial Biotechnology: Biotechnology of Thermophiles. Springer Science & Business Media; 2013. 954 p.

72. Suzuki H, Yoshida K-I, Ohshima T. Polysaccharide-Degrading Thermophiles Generated by Heterologous Gene Expression in Geobacillus kaustophilus HTA426. Appl Environ Microbiol. 2013 Sep;79(17):5151–8.

73. Lin PP, Rabe KS, Takasumi JL, Kadisch M, Arnold FH, Liao JC. Isobutanol production at elevated temperatures in thermophilic Geobacillus thermoglucosidasius. Metab Eng. 2014 Jul;24:1–8.

74. Imanaka T, Fujii M, Aramori I, Aiba S. Transformation of Bacillus stearothermophilus with plasmid DNA and characterization of shuttle vector plasmids

between Bacillus stearothermophilus and Bacillus subtilis. J Bacteriol. 1982 Mar;149(3):824–30.

75. Liao HH, Kanikula AM. Increased efficiency of transformation ofBacillus stearothermophilus by a plasmid carrying a thermostable kanamycin resistance marker. Curr Microbiol. 1990 Nov;21(5):301–6.

76. Narumi I, Nakayama N, Nakamoto S, Kimura T, Yanagisawa T, Kihara H. Construction of a new shuttle vector pSTE33 and its stabilities in Bacillus stearothermophilus, bacillus subtilis, and Escherichia coli. Biotechnol Lett. 1993 Aug;15(8).

77. Atsumi S, Hanai T, Liao JC. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature. Nature Publishing Group; 2008 Jan 3;451(7174):86–9.

78. Turner P, Mamo G, Karlsson EN. Potential and utilization of thermophiles and thermostable enzymes in biorefining. Microb Cell Fact. 2007 Jan;6:9.

79. Bartosiak-Jentys J, Hussein AH, Lewis CJ, Leak DJ. Modular system for assessment of glycosyl hydrolase secretion in Geobacillus thermoglucosidasius. Microbiology. 2013 Jul;159(Pt 7):1267–75.

80. Suzuki H, Kobayashi J, Wada K, Furukawa M, Doi K. Thermoadaptation-directed enzyme evolution in an error-prone thermophile derived from Geobacillus kaustophilus HTA426. Appl Environ Microbiol. 2015 Jan 1;81(1):149–58.

81. Kobayashi J, Furukawa M, Ohshiro T, Suzuki H. Thermoadaptation-directed evolution of chloramphenicol acetyltransferase in an error-prone thermophile using improved procedures. Appl Microbiol Biotechnol. 2015 Jul;99(13):5563–72.

82. Wada K, Kobayashi J, Furukawa M, Doi K, Ohshiro T, Suzuki H. A thiostrepton resistance gene and its mutants serve as selectable markers in Geobacillus kaustophilus HTA426. Biosci Biotechnol Biochem. 2016 Feb;80(2):368–75.

83. Taylor MP, Esteban CD, Leak DJ. Development of a versatile shuttle vector for gene expression in Geobacillus spp. Plasmid. 2008/05/27 ed. 2008;60(1):45–52.

84. Thompson AH, Studholme DJ, Green EM, Leak DJ. Heterologous expression of pyruvate decarboxylase in Geobacillus thermoglucosidasius. Biotechnol Lett. 2008 Aug;30(8):1359–65.

85. Bartosiak-Jentys J, Eley K, Leak DJ. Application of pheB as a reporter gene for Geobacillus spp., enabling qualitative colony screening and quantitative analysis of promoter strength. Appl Environ Microbiol. 2012 Aug 15;78(16):5945–7.

86. Suzuki H, Murakami A, Yoshida K. Counterselection system for Geobacillus kaustophilus HTA426 through disruption of pyrF and pyrR. Appl Environ Microbiol. 2012 Oct;78(20):7376–83.

87. Suzuki H, Yoshida K. Genetic transformation of Geobacillus kaustophilus HTA426 by conjugative transfer of host-mimicking plasmids. J Microbiol Biotechnol. 2012 Sep;22(9):1279–87.

88. Blanchard K, Robic S, Matsumura I. Transformable facultative thermophile

Geobacillus stearothermophilus NUB3621 as a host strain for metabolic engineering. Appl Microbiol Biotechnol. 2014 Aug;98(15):6715–23.

89. Mee E., Welker NE. Cloning vector pNW33N, complete sequence [Internet]. NCBI Nucleatide Database. 2003 [cited 2016 Mar 30]. Available from: http://www.ncbi.nlm.nih.gov/nuccore/AY237122

90. Narumi I, Sawakami K, Nakamoto S, Nakayama N, Yanagisawa T, Takahashi N, et al. A newly isolatedBacillus stearotheromophilus K1041 and its transformation by electroporation. Biotechnol Tech. 1992 Jan 1;6(1):83–6.

91. Tai S-K, Lin H-PP, Kuo J, Liu J-K. Isolation and characterization of a cellulolytic Geobacillus thermoleovorans T4 strain from sugar refinery wastewater. Extremophiles. 2004 Oct;8(5):345–9.

92. Lucas S, Han J, Lapidus A, Cheng J-F, Goodwin L, Pitluck S, et al. Geobacillus thermoglucosidasius C56-YS93, complete genome - Nucleotide - NCBI [Internet]. NCBI nucletide database. 2011 [cited 2015 Sep 14]. Available from: http://www.ncbi.nlm.nih.gov/nuccore/CP002835

93. Brumm PJ, Land ML, Mead DA. Complete genome sequence of Geobacillus thermoglucosidasius C56-YS93, a novel biomass degrader isolated from obsidian hot spring in Yellowstone National Park. Stand Genomic Sci. 2015 Oct 5;10(1):73.

94. Tan H, Fu L, Seno M. Optimization of bacterial plasmid transformation using nanomaterials based on the Yoshida effect. Int J Mol Sci. Multidisciplinary Digital Publishing Institute (MDPI); 2010 Dec 3;11(12):4962–72.

95. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Meth. Nature Publishing Group; 2009;6(5):343–5.

96. Yu H, Stephanopoulos G. Metabolic engineering of Escherichia coli for biosynthesis of hyaluronic acid. Metab Eng. 2008 Jan;10(1):24–32.

97. Song J-M, Im J-H, Kang J-H, Kang D-J. A simple method for hyaluronic acid quantification in culture broth. Carbohydr Polym. 2009 Oct 15;78(3):633–4.

98. Tominaga Y, Ohshiro T, Suzuki H. Conjugative plasmid transfer from Escherichia coli is a versatile approach for genetic transformation of thermophilic Bacillus and Geobacillus species. Extremophiles. 2016 Mar 1;

99. Martinez--Klimova E. Synthetic biology approaches to the metabolic engineering of Geobacillus thermoglucosidans for isobutanol production. Imperial College London; 2015.

100. Taylor M. Metabolic Engineering Of Geobacillus Species For Enhanced Ethanol Production. Imperial College London; 2007.

101. Wu LJ, Welker NE. Protoplast transformation of Bacillus stearothermophilus NUB36 by plasmid DNA. J Gen Microbiol. 1989 May 1;135(5):1315–24.

102. Natarajan MR, Oriel P. Conjugal transfer of recombinant transposon Tn916 from Escherichia coli to Bacillus stearothermophilus. Plasmid. 1991 Jul;26(1):67–73.

103. Burrus V, Pavlovic G, Decaris B, Guédon G. Conjugative transposons: the tip of the iceberg. Mol Microbiol. 2002 Nov;46(3):601–10.

104. Olson DG, Lynd LR. Transformation of Clostridium thermocellum by electroporation. Methods Enzymol. 2012 Jan;510:317–30.

105. Peng H, Fu B, Mao Z, Shao W. Electrotransformation of Thermoanaerobacter ethanolicus JW200. Biotechnol Lett. 2006 Dec;28(23):1913–7.

106. Lin L, Song H, Ji Y, He Z, Pu Y, Zhou J, et al. Ultrasound-mediated DNA transformation in thermophilic gram-positive anaerobes. Xu S, editor. PLoS One. Public Library of Science; 2010 Jan 4;5(9):e12582.

107. Hoshino T, Ikeda T, Narushima H, Tomizuka N. Isolation and characterization of antibiotic-resistance plasmids in thermophilic bacilli. Can J Microbiol. 1985 Apr;31(4):339–45.

108. Kananavičiūtė R, Butaitė E, Citavičius D. Characterization of two novel plasmids from Geobacillus sp. 610 and 1121 strains. Plasmid. 2014 Jan;71:23–31.

109. Nakayama N, Narumi I, Nakamoto S, Kihara H. A new shuttle vector forBacillus stearothermophilus andEscherichia coli. Biotechnol Lett. 1992 Aug;14(8):649–52.

110. Hanahan D, Jessee J, Bloom FR. Bacterial Genetic Systems. Methods in Enzymology. Elsevier; 1991. 63-113 p.

111. Aune TEV, Aachmann FL. Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed. Appl Microbiol Biotechnol. 2010 Feb;85(5):1301–13.

112. Dubnau D. Genetic competence in Bacillus subtilis. Microbiol Mol Biol Rev. 1991 Sep 1;55(3):395–424.

113. Rahmer R, Morabbi Heravi K, Altenbuchner J. Construction of a Super-Competent Bacillus subtilis 168 Using the P mtlA -comKS Inducible Cassette. Front Microbiol. Frontiers; 2015 Jan 21;6:1431.

114. Mirończuk AM, Kovács ÁT, Kuipers OP. Induction of natural competence in Bacillus cereus ATCC14579. Microb Biotechnol. 2008 May;1(3):226–35.

115. Hahn J, Luttinger A, Dubnau D. Regulatory inputs for the synthesis of ComK, the competence transcription factor of Bacillus subtilis. Mol Microbiol. 1996 Aug;21(4):763–75.

116. Song Y, Hahn T, Thompson IP, Mason TJ, Preston GM, Li G, et al. Ultrasound-mediated DNA transfer for bacteria. Nucleic Acids Res. 2007 Jan;35(19):e129.

117. Yoshida N, Ikeda T, Yoshida T, Sengoku T, Ogawa K. Chrysotile asbestos fibers mediate transformation of Escherichia coli by exogenous plasmid DNA. FEMS Microbiol Lett. 2001 Feb 20;195(2):133–7.

118. Neumann E, Schaefer-Ridder M, Wang Y, Hofschneider PH. Gene transfer into mouse lyoma cells by electroporation in high electric fields. EMBO J. 1982 Jan;1(7):841–5.

119. Luchansky JB, Muriana PM, Klaenhammer TR. Application of electroporation for transfer of plasmid DNA to Lactobacillus, Lactococcus, Leuconostoc, Listeria, Pediococcus, Bacillus, Staphylococcus, Enterococcus and Propionibacterium. Mol Microbiol. 1988 Sep;2(5):637–46.

120. Fiedler S, Wirth R. Transformation of bacteria with plasmid DNA by electroporation. Anal Biochem. 1988 Apr;170(1):38–44.

121. Xue G-P, Johnson JS, Dalrymple BP. High osmolarity improves the electro-transformation efficiency of the gram-positive bacteria Bacillus subtilis and Bacillus licheniformis. J Microbiol Methods. 1999;34(3):183–91.

122. Studholme DJ. Some (bacilli) like it hot: genomics of Geobacillus species. Microb Biotechnol. 2015 Jan;8(1):40–8.

123. Stepanov AS, Puzanova OB, Dityatkin SYa, Loginova OG, Ilyashenko BN. Glycine-induced cryotransformation of plasmids into Bacillus anthracis. J Gen Microbiol. 1990 Jul;136(7):1217–21.

124. Kawata Y, Yano S, Kojima H. Escherichia coli can be transformed by a liposome-mediated lipofection method. Biosci Biotechnol Biochem. 2003 May;67(5):1179–81.

125. Elliott AR, Silvert PY, Xue GP, Simpson GD, Tekaia-Elhsissen K, Aylward JH. Transformation of Bacillus subtilis using the particle inflow gun and submicrometer particles obtained by the polyol process. Anal Biochem. 1999 May 1;269(2):418–20.

126. Reeve B, Sanderson T, Ellis T, Freemont P. How synthetic biology will reconsider natural bioluminescence and its applications. Adv Biochem Eng Biotechnol. 2014 Jan;145:3–30.

127. Tisi L., White P., Squirrell D., Murphy M., Lowe C., Murray JA. Development of a thermostable firefly luciferase. Anal Chim Acta. 2002 Apr;457(1):115–23.

128. Mortazavi M, Hosseinkhani S. Design of thermostable luciferases through arginine saturation in solvent-exposed loops. Protein Eng Des Sel. Oxford University Press; 2011 Dec 1;24(12):893–903.

129. Homaei AA, Mymandi AB, Sariri R, Kamrani E, Stevanato R, Etezad S-M, et al. Purification and characterization of a novel thermostable luciferase from Benthosema pterotum. J Photochem Photobiol B. 2013 Aug 5;125:131–6.

130. Onai K, Morishita M, Itoh S, Okamoto K, Ishiura M. Circadian rhythms in the thermophilic cyanobacterium Thermosynechococcus elongatus: compensation of period length over a wide temperature range. J Bacteriol. 2004 Aug;186(15):4972–7.

131. Chalfie M, Tu Y, Euskirchen G, Ward W, Prasher D. Green fluorescent protein as a marker for gene expression. Science (80-). American Association for the Advancement of Science; 1994 Feb 11;263(5148):802–5.

132. Cormack BP, Valdivia RH, Falkow S. FACS-optimized mutants of the green fluorescent protein (GFP). Gene. 1996 Jan;173(1 Spec No):33–8.

133. Waldo GS, Standish BM, Berendzen J, Terwilliger TC. Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol. Nature America Inc.; 1999 Jul;17(7):691–5.

134. Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS. Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol. Nature Publishing Group; 2006 Jan;24(1):79–88.

135. Cava F, de Pedro MA, Blas-Galindo E, Waldo GS, Westblade LF, Berenguer J. Expression and use of superfolder green fluorescent protein at high temperatures in vivo: a tool to study extreme thermophile biology. Environ Microbiol. 2008 Mar;10(3):605–13.

136. Tsien RY. The green fluorescent protein. Annu Rev Biochem. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA; 1998 Jan 28;67:509–44.

137. Losi A, Polverini E, Quest B, Gärtner W. First evidence for phototropin-related blue-light receptors in prokaryotes. Biophys J. 2002 May;82(5):2627–34.

138. Swartz TE, Corchnoy SB, Christie JM, Lewis JW, Szundi I, Briggs WR, et al. The photocycle of a flavin-binding domain of the blue light photoreceptor phototropin. J Biol Chem. 2001 Sep 28;276(39):36493–500.

139. Losi A. Flavin-based Blue-Light photosensors: a photobiophysics update. Photochem Photobiol. 2007 Jan;83(6):1283–300.

140. Drepper T, Eggert T, Circolone F, Heck A, Krauss U, Guterl J-K, et al. Reporter proteins for in vivo fluorescence without oxygen. Nat Biotechnol. Nature Publishing Group; 2007 Apr;25(4):443–5.

141. Mukherjee A, Walker J, Weyant KB, Schroeder CM. Characterization of flavin-based fluorescent proteins: an emerging class of fluorescent reporters. PLoS One. Public Library of Science; 2013 Jan 31;8(5):e64753.

142. Chapman S, Faulkner C, Kaiserli E, Garcia-Mata C, Savenkov EI, Roberts AG, et al. The photoreversible fluorescent protein iLOV outperforms GFP as a reporter of plant virus infection. Proc Natl Acad Sci U S A. 2008 Dec 16;105(50):20038–43.

143. Bizzarri R, Serresi M, Luin S, Beltram F. Green fluorescent protein based pH indicators for in vivo use: a review. Anal Bioanal Chem. 2009 Feb;393(4):1107–22.

144. Drepper T, Huber R, Heck A, Circolone F, Hillmer A-K, Büchs J, et al. Flavin mononucleotide-based fluorescent reporter proteins outperform green fluorescent protein-like proteins as quantitative in vivo real-time reporters. Appl Environ Microbiol. 2010 Sep;76(17):5990–4.

145. Mukherjee A, Weyant KB, Walker J, Schroeder CM. Directed evolution of bright mutants of an oxygen-independent flavin-binding fluorescent protein from Pseudomonas putida. J Biol Eng. 2012 Jan;6(1):20.

146. Bartosiak-Jentys J. Metabolic engineering and Metabolic Flux Analysis of thermophilic, ethanologenic Geobacillus spp. Imperial College London; 2010.

147. Anagnostopoulos C, Spizizen J. REQUIREMENTS FOR TRANSFORMATION IN BACILLUS SUBTILIS. J Bacteriol. 1961 May;81(5):741–6.

148. Matsen JB. Lidstrom:Electroporation - OpenWetWare [Internet]. 2015 [cited 2016 Feb 19]. Available from: http://openwetware.org/wiki/Lidstrom:Electroporation

149. Stefan Milde, James Brown, Hugo Schmidt, Linda Boettger, Marie Chapart, Kevin Cheng, et al. Improved GFP, Cambridge iGEM 2008 [Internet]. 2008 [cited 2012 May 15]. Available from: http://2008.igem.org/Team:Cambridge/Improved_GFP

150. Parts Catalog, Registry of Standard Biological Parts [Internet]. igem.org. [cited 2016 Feb 20]. Available from: http://parts.igem.org/Catalog

151. Shaner NC, Campbell RE, Steinbach PA, Giepmans BNG, Palmer AE, Tsien RY. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nat Biotechnol. 2004 Dec;22(12):1567–72.

152. Madden T. The BLAST Sequence Analysis Tool. In: McEntyre J, Ostell J, editors. The NCBI Handbook. 2nd ed. National Center for Biotechnology Information (US); 2003.

153. Buckley AM, Petersen J, Roe AJ, Douce GR, Christie JM. LOV-based reporters for fluorescence imaging. Curr Opin Chem Biol. 2015 Aug;27:39–45.

154. Song X, Wang Y, Shu Z, Hong J, Li T, Yao L. Engineering a more thermostable blue light photo receptor Bacillus subtilis YtvA LOV domain by a computer aided rational design method. PLoS Comput Biol. Public Library of Science; 2013 Jan 4;9(7):e1003129.

155. Zhang C, Xing X-H, Lou K. Rapid detection of a gfp-marked Enterobacter aerogenes under anaerobic conditions by aerobic fluorescence recovery. FEMS Microbiol Lett. 2005 Aug 15;249(2):211–8.

156. Sniegowski JA, Lappe JW, Patel HN, Huffman HA, Wachter RM. Base catalysis of chromophore formation in Arg96 and Glu222 variants of green fluorescent protein. J Biol Chem. 2005 Jul 15;280(28):26248–55.

157. Stepanenko O V, Stepanenko O V, Kuznetsova IM, Shcherbakova DM, Verkhusha V V, Turoverov KK. Distinct effects of guanidine thiocyanate on the structure of superfolder GFP. PLoS One. Public Library of Science; 2012 Jan 7;7(11):e48809.

158. Löfblom J, Kronqvist N, Uhlén M, Ståhl S, Wernérus H. Optimization of electroporation-mediated transformation: Staphylococcus carnosus as model organism. J Appl Microbiol. 2007 Mar;102(3):736–47.

159. Sheng Y, Mancino V, Birren B. Transformation of Escherichia coli with large DNA molecules by electroporation. Nucleic Acids Res. 1995 Jun 11;23(11):1990–6.

160. Andrews BT, Schoenfish AR, Roy M, Waldo G, Jennings PA. The rough energy landscape of superfolder GFP is linked to the chromophore. J Mol Biol. 2007 Oct 19;373(2):476–90.

161. Hammer K, Mijakovic I, Jensen PR. Synthetic promoter libraries--tuning of gene expression. Trends Biotechnol. 2006 Feb;24(2):53–5.

162. Smanski MJ, Zhou H, Claesen J, Shen B, Fischbach MA, Voigt CA. Synthetic biology to access and expand nature's chemical diversity. Nat Rev Microbiol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2016 Feb 15;14(3):135–49.

163. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-

dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012 Mar 2;335(6072):1103–6.

164. Repoila F, Gottesman S. Temperature Sensing by the dsrA Promoter. J Bacteriol. 2003 Oct 31;185(22):6609–14.

165. Meijer WJJ, Salas M. Relevance of UP elements for three strong Bacillus subtilis phage phi29 promoters. Nucleic Acids Res. 2004 Jan 18;32(3):1166–76.

166. Wang Y, Zhang X. Genome analysis of deep-sea thermophilic phage D6E. Appl Environ Microbiol. 2010 Dec 1;76(23):7861–6.

167. Liu B, Zhou F, Wu S, Xu Y, Zhang X. Genomic and proteomic characterization of a thermophilic Geobacillus bacteriophage GBSV1. Res Microbiol. 2009 Mar;160(2):166–71.

168. Jensen PR, Hammer K. The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. Appl Environ Microbiol. 1998 Jan;64(1):82–7.

169. Solem C, Jensen PR. Modulation of Gene Expression Made Easy. Appl Environ Microbiol. 2002 May 1;68(5):2397–403.

170. Blount BA, Weenink T, Vasylechko S, Ellis T. Rational Diversification of a Promoter Providing Fine-Tuned Expression and Orthogonal Regulation for Synthetic Biology. Tuite MF, editor. PLoS One. Public Library of Science; 2012 Mar 19;7(3):e33279.

171. Zaccolo M, Williams DM, Brown DM, Gherardi E. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. J Mol Biol. 1996 Feb 2;255(4):589–603.

172. Alper H, Fischer C, Nevoigt E, Stephanopoulos G. Tuning genetic control through promoter engineering. Proc Natl Acad Sci U S A. 2005 Sep 6;102(36):12678–83.

173. Hammer K, Mijakovic I, Jensen PR. Synthetic promoter libraries--tuning of gene expression. Trends Biotechnol. 2006/01/13 ed. 2006;24(2):53–5.

174. Dunn AK, Handelsman J. A vector for promoter trapping in Bacillus cereus. Gene. 1999 Jan 21;226(2):297–305.

175. Goodman D. Part: BBa K090401 - parts.igem.org [Internet]. Registry of Standard Biological Parts. 2008 [cited 2016 Mar 6]. Available from: http://parts.igem.org/Part:BBa_K090401

176. Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, et al. Measuring the activity of BioBrick promoters using an in vivo reference standard. J Biol Eng. 2009 Jan;3(1):4.

177. Leveau JHJ, Lindow SE. Predictive and Interpretive Simulation of Green Fluorescent Protein Expression in Reporter Bacteria. J Bacteriol. 2001 Dec 1;183(23):6752–62.

178. Mimee M, Tucker AC, Voigt CA, Lu TK. Programming a Human Commensal Bacterium, Bacteroides thetaiotaomicron, to Sense and Respond to Stimuli in the Murine Gut Microbiota. Cell Syst. Elsevier; 2015 Jul;1(1):62–71.

179. Markley AL, Begemann MB, Clarke RE, Gordon GC, Pfleger BF. Synthetic biology toolbox for controlling gene expression in the cyanobacterium Synechococcus sp. strain PCC 7002. ACS Synth Biol. American Chemical Society; 2015 May 15;4(5):595–603.

180. Weenink T, Ellis T. Creation and characterization of component libraries for synthetic biology. Methods Mol Biol. 2013 Jan;1073:51–60.

181. Michna RH, Commichau FM, Tödter D, Zschiedrich CP, Stülke J. SubtiWiki-a database for the model organism Bacillus subtilis that links pathway, interaction and expression information. Nucleic Acids Res. 2014 Jan;42(Database issue):D692–8.

182. Wang B, Kitney RI, Joly N, Buck M. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. Nat Commun. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011 Jan 18;2:508.

183. Salis HM. The ribosome binding site calculator. Methods Enzymol. 2011 Jan;498:19–42.

184. Reeve B, Hargest T, Gilbert C, Ellis T. Predicting translation initiation rates for designing synthetic biology. Front Bioeng Biotechnol. Frontiers; 2014 Jan 20;2:1.

185. Vellanoweth RL, Rabinowitz JC. The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. Mol Microbiol. 1992 May;6(9):1105–14.

186. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol. Nature Publishing Group; 2009 Oct;27(10):946–50.

187. Na D, Lee D. RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. Bioinformatics. 2010 Oct 15;26(20):2633–4.

188. Seo SW, Yang J-S, Kim I, Yang J, Min BE, Kim S, et al. Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. Metab Eng. 2013 Jan;15:67–74.

189. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, et al. NUPACK: Analysis and design of nucleic acid systems. J Comput Chem. 2011 Jan 15;32(1):170–3.

190. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W70–4.

191. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Keith JM, editor. Methods Mol Biol. Totowa, NJ: Humana Press; 2008 Jan;453:3–31.

192. Espah Borujeni A, Channarasappa AS, Salis HM. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Res. 2014 Feb;42(4):2646–59.

193. Ceroni F, Algar R, Stan G-B, Ellis T. Quantifying cellular capacity identifies gene

expression designs with reduced burden. Nat Methods. Nature Publishing Group; 2015 May 6;12(5):415–8.

194. Bonde MT, Pedersen M, Klausen MS, Jensen SI, Wulff T, Harrison S, et al. Predictable tuning of protein expression in bacteria. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2016 Jan 11;13(3):233–6.

195. Silva-Rocha R, Martinez-Garcia E, Calles B, Chavarria M, Arce-Rodriguez A, de las Heras A, et al. The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. Nucleic Acids Res. 2012 Nov 23;41(D1):D666–75.

196. Silva-Rocha R, Pontelli MC, Furtado GP, Zaramela LS, Koide T. Development of New Modular Genetic Tools for Engineering the Halophilic Archaeon Halobacterium salinarum. PLoS One. 2015 Jan;10(6):e0129215.

197. Fernández H, Prandoni N, Fernández-Pascual M, Fajardo S, Morcillo C, Díaz E, et al. Azoarcus sp. CIB, an anaerobic biodegrader of aromatic compounds shows an endophytic lifestyle. PLoS One. 2014 Jan;9(10):e110771.

198. Felpeto-Santero C, Rojas A, Tortajada M, Galán B, Ramón D, García JL. Engineering alternative isobutanol production platforms. AMB Express. 2015 Dec;5(1):119.

199. Sevilla E, Yuste L, Rojo F. Marine hydrocarbonoclastic bacteria as whole-cell biosensors for n-alkanes. Microb Biotechnol. 2015 Jul;8(4):693–706.

200. Florea M, Reeve B, Abbott J, Freemont PS, Ellis T. Genome sequence and plasmid transformation of the model high-yield bacterial cellulose producer Gluconacetobacter hansenii ATCC 53582. Sci Rep. Nature Publishing Group; 2016 Jan 24;6:23635.

201. Martínez-García E, Aparicio T, Goñi-Moreno A, Fraile S, de Lorenzo V. SEVA 2.0: an update of the Standard European Vector Architecture for de-/re-construction of bacterial functionalities. Nucleic Acids Res. 2015 Jan;43(Database issue):D1183–9.

202. Liao H, McKenzie T, Hageman R. Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. Proc Natl Acad Sci U S A. 1986 Feb;83(3):576–80.

203. De Rossi E, Milano A, Brigidi P, Bini F, Riccardi G. Structural organization of pBC1, a cryptic plasmid from Bacillus coagulans. J Bacteriol. 1992 Jan;174(2):638–42.

204. Gryczan TJ, Contente S, Dubnau D. Characterization of Staphylococcus aureus plasmids introduced by transformation into Bacillus subtilis. J Bacteriol. 1978 Apr;134(1):318–29.

205. Imanaka T, Fujii M, Aiba S. Isolation and characterization of antibiotic resistance plasmids from thermophilic bacilli and construction of deletion plasmids. J Bacteriol. 1981 Jun;146(3):1091–7.

206. Sharma A, Pandey A, Shouche YS, Kumar B, Kulkarni G. Characterization and identification of Geobacillus spp. isolated from Soldhar hot spring site of Garhwal Himalaya, India. J Basic Microbiol. 2009 Apr;49(2):187–94.

207. McKenzie T, Hoshino T, Tanaka T, Sueoka N. The nucleotide sequence of pUB110:

some salient features in relation to replication and its regulation. Plasmid. 1986 Mar;15(2):93–103.

208. Horinouchi S, Weisblum B. Nucleotide sequence and functional map of pC194, a plasmid that specifies inducible chloramphenicol resistance. J Bacteriol. 1982 May;150(2):815–25.

209. Yanisch-Perron C, Vieira J, Messing J. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene. 1985 Jan;33(1):103–19.

210. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic Acids Res. 2001 Sep 1;29(17):3583–94.

211. Pavlostathis SG, Marchant R, Banat IM, Ternan NG, McMullan G. High growth rate and substrate exhaustion results in rapid cell death and lysis in the thermophilic bacterium Geobacillus thermoleovorans. Biotechnol Bioeng. 2006 Sep 5;95(1):84–95.

212. Peteranderl R, Shotts EB, Wiegel J. Stability of antibiotics under growth conditions for thermophilic anaerobes. Appl Environ Microbiol. 1990 Jun;56(6):1981–3.

213. Bosma EF, van de Weijer AHP, Daas MJA, van der Oost J, de Vos WM, van Kranenburg R. Isolation and screening of thermophilic bacilli from compost for electrotransformation and fermentation: characterization of Bacillus smithii ET 138 as a new biocatalyst. Appl Environ Microbiol. 2015 Mar;81(5):1874–83.

214. Goh KM, Kahar UM, Chai YY, Chong CS, Chai KP, Ranjani V, et al. Recent discoveries and applications of Anoxybacillus. Appl Microbiol Biotechnol. 2013 Feb;97(4):1475–88.

215. Wright O, Delmans M, Stan G-B, Ellis T. GeneGuard: A modular plasmid system designed for biosafety. ACS Synth Biol. American Chemical Society; 2015 Mar 20;4(3):307–16.

216. Boe L, Gros MF, te Riele H, Ehrlich SD, Gruss A. Replication origins of single-stranded-DNA plasmid pUB110. J Bacteriol. 1989 Jun;171(6):3366–72.

217. GVR. Hyaluronic Acid Market Analysis By Product, Application To 2020. San Francisco; 2015.

218. Krahulec J, Krahulcová J. Increase in hyaluronic acid production by Streptococcus equi subsp. zooepidemicus strain deficient in beta-glucuronidase in laboratory conditions. Appl Microbiol Biotechnol. 2006 Jul;71(4):415–22.

219. Kim J-H, Yoo S-J, Oh D-K, Kweon Y-G, Park D-W, Lee C-H, et al. Selection of a Streptococcus equi mutant and optimization of culture conditions for the production of high molecular weight hyaluronic acid. Enzyme Microb Technol. 1996 Nov;19(6):440–5.

220. Chen WY, Marcellin E, Hung J, Nielsen LK. Hyaluronan molecular weight is controlled by UDP-N-acetylglucosamine concentration in Streptococcus zooepidemicus. J Biol Chem. 2009 Jul 3;284(27):18007–14.

221. Widner B, Behr R, Von Dollen S, Tang M, Heu T, Sloma A, et al. Hyaluronic acid

production in Bacillus subtilis. Appl Environ Microbiol. 2005 Jul;71(7):3747–52.

222. Jongsareejit B, Bhumiratana A, Morikawa M, Kanaya S. Cloning of hyaluronan synthase (sz-has) gene from Streptococcus zooepidemicus in Escherichia coli. Sci Asia. 2007;

223. Mao Z, Shin H-D, Chen R. A recombinant E. coli bioprocess for hyaluronan synthesis. Appl Microbiol Biotechnol. 2009 Aug;84(1):63–9.

224. Chien L-J, Lee C-K. Enhanced hyaluronic acid production in Bacillus subtilis by coexpressing bacterial hemoglobin. Biotechnol Prog. 2007 Jan;23(5):1017–22.

225. Videbaek T (Novozymes). Hyaluronic Acid; Technology, Market & Timing. 2011.

226. Maleki A, Kjøniksen A-L, Nyström B. Anomalous Viscosity Behavior in Aqueous Solutions of Hyaluronic Acid. Polym Bull. 2007 Mar 23;59(2):217–26.

227. Weigel PH, Hascall VC, Tammi M. Hyaluronan synthases. J Biol Chem. 1997 May 30;272(22):13997–4000.

228. Boeriu CG, Springer J, Kooy FK, L.A.M. Broek van den, Eggink G. Production methods for hyaluronan. Int J Carbohydr Chem. 2013;2013.

229. Knight GC, Nicol RS, McMeekin TA. Temperature step changes: a novel approach to control biofilms of Streptococcus thermophilus in a pilot plant-scale cheese-milk pasteurisation plant. Int J Food Microbiol. 2004 Jun 15;93(3):305–18.

230. Izawa N, Hanamizu T, Iizuka R, Sone T, Mizukoshi H, Kimura K, et al. Streptococcus thermophilus produces exopolysaccharides including hyaluronic acid. J Biosci Bioeng. 2009 Feb;107(2):119–23.

231. Izawa N, Serata M, Sone T, Omasa T, Ohtake H. Hyaluronic acid production by recombinant Streptococcus thermophilus. J Biosci Bioeng. 2011 Jun;111(6):665–70.

232. Richardson SM, Wheelan SJ, Yarrington RM, Boeke JD. GeneDesign: rapid, automated design of multikilobase synthetic genes. Genome Res. 2006 Apr;16(4):550–6.

233. Villalobos A, Ness JE, Gustafsson C, Minshull J, Govindarajan S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinformatics. 2006 Jan;7:285.

234. Gaspar P, Oliveira JL, Frommlet J, Santos MAS, Moura G. EuGene: maximizing synthetic gene design for heterologous expression. Bioinformatics. 2012 Oct 15;28(20):2683–4.

235. Mauro VP, Chappell SA. A critical analysis of codon optimization in human therapeutics. Trends Mol Med. 2014 Nov;20(11):604–13.

236. Gaspar P, Moura G, Santos MAS, Oliveira JL. mRNA secondary structure optimization using a correlated stem-loop prediction. Nucleic Acids Res. 2013 Apr 1;41(6):e73.

237. Boël G, Letso R, Neely H, Price WN, Wong K-H, Su M, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. Nature Publishing

Group; 2016 Jan 13;

238. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol. 2013 Mar 29;30(3):549–60.

239. Fischer M. Entelechon codon optimisation tool [Internet]. Entelechon GmbH, Regensburg, Germany. [cited 2014 Jan 3]. Available from: http://www.entelechon.com/bttool/bttool.html

240. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 2000 Jan 1;28(1):292.

241. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res. 2014 Aug 23;42(14):9171–81.

242. Tian T, Salis HM. A predictive biophysical model of translational coupling to coordinate and control protein expression in bacterial operons. Nucleic Acids Res. 2015 Aug 18;43(14):7137–51.

243. Rex G, Surin B, Besse G, Schneppe B, McCarthy J. The mechanism of translational coupling in Escherichia coli. Higher order structure in the atpHA mRNA acts as a conformational switch regulating the access of de novo initiating ribosomes. J Biol Chem. 1994 Jul 8;269(27):18118–27.

244. Adhin MR, van Duin J. Scanning model for translational reinitiation in eubacteria. J Mol Biol. 1990 Jun 20;213(4):811–8.

245. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. Computational identification of operons in microbial genomes. Genome Res. 2002 Aug;12(8):1221–30.

246. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, et al. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A. 2006 Oct 17;103(42):15611–6.

247. Yasukazu N. Codon Usage Database [Internet]. [cited 2012 May 16]. Available from: http://www.kazusa.or.jp/codon/

248. Casini A, Christodoulou G, Freemont PS, Baldwin GS, Ellis T, MacDonald JT. R2oDNA designer: computational design of biologically neutral synthetic DNA sequences. ACS Synth Biol. American Chemical Society; 2014 Aug 15;3(8):525–8.

249. Casini A, MacDonald JT, De Jonghe J, Christodoulou G, Freemont PS, Baldwin GS, et al. One-pot DNA construction for synthetic biology: the Modular Overlap-Directed Assembly with Linkers (MODAL) strategy. Nucleic Acids Res. 2014 Jan 1;42(1):e7.

250. Levin-Karp A, Barenholz U, Bareia T, Dayagi M, Zelcbuch L, Antonovsky N, et al. Quantifying translational coupling in E. coli synthetic operons using RBS modulation and fluorescent reporters. ACS Synth Biol. American Chemical Society; 2013 Jun 21;2(6):327–36.

251. Suenaga E, Nakamura H. Evaluation of three methods for effective extraction of DNA from human hair. J Chromatogr B Analyt Technol Biomed Life Sci. 2005 Jun

5;820(1):137–41.

252. Yang J-L, Wang M-S, Cheng A-C, Pan K-C, Li C-F, Deng S-X. A simple and rapid method for extracting bacterial DNA from intestinal microflora for ERIC-PCR detection. World J Gastroenterol. 2008 May 14;14(18):2872–6.

253. Bitter T, Muir HM. A modified uronic acid carbazole reaction. Anal Biochem. 1962 Oct;4(4):330–4.

254. Cesaretti M. A 96-well assay for uronic acid carbazole reaction. Carbohydr Polym. 2003 Oct 1;54(1):59–61.

255. Chen Y-H, Wang Q. Establishment of CTAB Turbidimetric method to determine hyaluronic acid content in fermentation broth. Carbohydr Polym. 2009 Aug;78(1):178–81.

256. Liu L, Liu Y, Li J, Du G, Chen J. Microbial production of hyaluronic acid: current state, challenges, and perspectives. Microb Cell Fact. 2011 Jan;10(1):99.

257. Engler C, Kandzia R, Marillonnet S. A one pot, one step, precision cloning method with high throughput capability. PLoS One. Public Library of Science; 2008 Jan 5;3(11):e3647.

258. Chung D, Cha M, Guss AM, Westpheling J. Direct conversion of plant biomass to ethanol by engineered Caldicellulosiruptor bescii. Proc Natl Acad Sci U S A. 2014 Jun 17;111(24):8931–6.

259. Waege I, Schmid G, Thumann S, Thomm M, Hausner W. Shuttle vector-based transformation system for Pyrococcus furiosus. Appl Environ Microbiol. 2010 May 15;76(10):3308–13.

260. Iwai M. Improved Genetic Transformation of the Thermophilic Cyanobacterium, Thermosynechococcus elongatus BP-1. Plant Cell Physiol. 2004 Feb 15;45(2):171–5.

261. Chandrayan SK, McTernan PM, Hopkins RC, Sun J, Jenney FE, Adams MWW. Engineering hyperthermophilic archaeon Pyrococcus furiosus to overproduce its cytoplasmic [NiFe]-hydrogenase. J Biol Chem. 2012 Jan 27;287(5):3257–64.

262. Clark-Casey J. SynBioMine v3 Data mining for Synthetic Biology [Internet]. [cited 2016 Mar 22]. Available from: http://www.synbiomine.org/synbiomine/begin.do

263. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. 2012 Dec 1;28(23):3163–5.

264. Stoolmiller AC, Dorfman A. The Biosynthesis of Hyaluronic Acid by Streptococcus. J Biol Chem. 1969 Jan 25;244(2):236–46.

265. Page WJ, Knosp O. Hyperproduction of Poly-{beta}-Hydroxybutyrate during Exponential Growth of Azotobacter vinelandii UWD. Appl Envir Microbiol. 1989 Jun 1;55(6):1334–9.

266. Fidler S, Dennis D. Polyhydroxyalkanoate production in recombinant Escherichia coli. FEMS Microbiol Rev. 1992 Dec;9(2-4):231–5.

267. Houmiel K, Slater S, Broyles D, Casagrande L, Colburn S, Gonzalez K, et al. Poly(beta-hydroxybutyrate) production in oilseed leukoplasts of brassica napus. Planta. 1999 Oct;209(4):547–50.

268. Haddouche R, Poirier Y, Delessert S, Sabirova J, Pagot Y, Neuvéglise C, et al. Engineering polyhydroxyalkanoate content and monomer composition in the oleaginous yeast Yarrowia lipolytica by modifying the ß-oxidation multifunctional protein. Appl Microbiol Biotechnol. 2011 Sep;91(5):1327–40.

269. Morange M. A critical perspective on synthetic biology. Int J Philos Chem. 2009;

270. Serrano L. Synthetic biology: promises and challenges. Mol Syst Biol. 2007 Jan;3:158.

271. Leow TC, Rahman RNZRA, Basri M, Salleh AB. A thermoalkaliphilic lipase of Geobacillus sp. T1. Extremophiles. 2007 May;11(3):527–35.

272. Sifour M, Zaghloul TI, Saeed HM, Berekaa MM, Abdel-Fattah YR. Enhanced production of lipase by the thermophilic Geobacillus stearothermophilus strain-5 using statistical experimental designs. N Biotechnol. 2010 Sep 30;27(4):330–6.

273. Jiang T, Huang M, He H, Lu J, Zhou X, Cai M, et al. Bioprocess exploration for thermostable α-amylase production of a deep-sea thermophile Geobacillus sp. in high temperature bioreactor. Prep Biochem Biotechnol. 2015 Dec 17;

274. Mukherjee A, Weyant KB, Agrawal U, Walker J, Cann IKO, Schroeder CM. Engineering and characterization of new LOV-based fluorescent proteins from Chlamydomonas reinhardtii and Vaucheria frigida. ACS Synth Biol. American Chemical Society; 2015 Apr 17;4(4):371–7.

275. Pothoulakis G, Ceroni F, Reeve B, Ellis T. The Spinach RNA Aptamer as a Characterization Tool for Synthetic Biology. ACS Synth Biol. 2014 Mar 21;3(3):182–7.

276. Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol. 1986 May;189(1):113–30.

277. Greif M, Mueller R, Christian R. Novel T7 RNA polymerase variants with enhanced thermostability. 2011.

278. Werner S, Engler C, Weber E, Gruetzner R, Marillonnet S. Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system. Bioeng Bugs. Public Library of Science; 2012 Jan 1;3(1):38–43.

279. Casini A, Storch M, Baldwin GS, Ellis T. Bricks and blueprints: methods and standards for DNA assembly. Nat Rev Mol Cell Biol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015 Jun 17;16(9):568–76.

280. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013 Mar;31(3):233–9.

281. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin E V. Viral diversity threshold for adaptive immunity in prokaryotes. MBio. 2012 Jan 31;3(6):e00456–12.

282. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol. 2008 Feb 15;190(4):1401–12.

283. Xu K, Ren C, Liu Z, Zhang T, Zhang T, Li D, et al. Efficient genome engineering in eukaryotes using Cas9 from Streptococcus thermophilus. Cell Mol Life Sci. 2015 Jan;72(2):383–99.

284. Müller M, Lee CM, Gasiunas G, Davis TH, Cradick TJ, Siksnys V, et al. Streptococcus thermophilus CRISPR-Cas9 Systems Enable Specific Editing of the Human Genome. Mol Ther. American Society of Gene & Cell Therapy; 2016 Mar;24(3):636–44.

285. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. 2013 Aug 1;41(15):7429–37.

286. Pantazaki AA, Papaneophytou CP, Lambropoulou DA. Simultaneous polyhydroxyalkanoates and rhamnolipids production by Thermus thermophilus HB8. AMB Express. 2011 Jan;1(1):17.

287. Liberal R, Lisowska BK, Leak DJ, Pinney JW. PathwayBooster: a tool to support the curation of metabolic pathways. BMC Bioinformatics. 2015 Jan;16:86.

288. Niu H, Leak D, Shah N, Kontoravdi C. Metabolic characterization and modeling of fermentation process of an engineered Geobacillus thermoglucosidasius strain for bioethanol production with gas stripping. Chem Eng Sci. 2015 Jan;122:138–49.

289. Wang L, Tang Y, Wang S, Liu R-L, Liu M-Z, Zhang Y, et al. Isolation and characterization of a novel thermophilic Bacillus strain degrading long-chain n-alkanes. Extremophiles. 2006 Aug;10(4):347–56.

10. Index and Appendices

10.1 Accession Numbers

Plasmid name	NCBI Accession number	Addgene plasmid number	Components
pG1K	KU169262	71741	*ColE1, repBST1, kanR, Multiple cloning site (MCS)*
pG2K	KU169263	71742	*ColE1, repB, kanR,* MCS
pG1C	KU169261	71740	*ColE1, repBST1, camR, MCS*
pG1AK	KU169257	71736	*ColE1, repBST1, ampR, kanR, MCS*
pG1AK-sfGFP	KU169260	71739	*ColE1, repBST1, ampR, kanR, RplsWT, sfGFP*
pG1AK-mCherry	KU169258	71737	*ColE1, repBST1, ampR, kanR, RplsWT, mCherry*
pG1AK-PheB	KU169259	71738	*ColE1, repB, ampR, kanR, RplsWT, PheB*

Plasmids available on request from Addgene with sequences in the NCBI database.

10.2 Parts Sequences

hotLOV

ATCGCCAGCACCAACGGCATCGTCATTACGGACTATCGCCAACCGGACAA
CCCGGTCATCTACGTGAACCCGGCATTTGAACGCATGACCGGCTATCGTG
CAACGGAAGTCATTGGTAAAAACGCTCGTTTTCTGCAGGGCAGCGATCGC
CATCAACCGGGTGCGACCGCCATTCGTAATGCGATCAAAAAAGGCCAGTC
TTGCCGCGTGGTTCTGCGTAACTACCGTAAAAATGGTCAACTGTTCTGGAA
CGAACTGGCAATTAGTCCGATCTACAATGAATTTGGCGAAATCACCCACT
ACATCGGCATCCAGTCGGACGTTACGGAA

Sequence of hotLOV from John Christie, University of Glasgow – received in a pUC cloning vector with ampicillin resistance

Rpls Library Promoters

Name	Sequence
Rpls WT	CTGCAGAACAATCGTTAAAGCGGACGTTTTTGCGCCGCCCGGATTTGCTTGAAAACTACCCGCTGAC AGAAAAGCAAAAACGATGGATCGAAGAGTGGAAAAAAGAAAAACAGTAGCTATTGCGCATGATAC AAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTGCTTAAGCGAGGAAAACGATGTTCCG CTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
1	ctacagAGCAATCGCTAAAGCGGACGCCTTCGCGCCGCCCGGATTTGCCTGAGGACTACCCGCTGGCA GAAAAGCAGAAACGACGGATCGAAGAGTGGAAAAAAGAGGAACAGTAGCTATTGCGCATGATGCG AGTTTATGCTACTATATTCCTTGTGCAACCTAGACGACTTGCTTAAGCGAGGAAAACGGTGCTCCGC TGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
2	CTGCAGAACAATCGTTAAAGCGGACGTTTCTGCGCCGCCCGGATTTGCTTGAAAACTGCCCGCTGA CAGAAAAGCGGAGACGATGGATCGGAGAGTGGAAAAAAGAAGAGCAGTAGCTATTGCGCATGATA CAAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTACCTAAGCGAGGAAGACGGTATTCC GCCGCAGTGATGGAAAACGCTCGTCTAGATAAGGAGTGATTCGAATG
3	CTGCAGAACAATCGTTAAAGCGGACGTTTTTGCGCCGCCCGGATTTGCTTGAAAACTACCCGCTGAC AGAAAAGCAAAAACGATGGATCGAAGAGTGGAAAAGAGAAAAACAGTAGCTATTGCGCATGATAC AAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTGCTTAAGCGAGGAAAACGATGTTCCG CTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
4	CTGCAGGACAACCGTTAAAGCGGACGTTCTTGCGCCGCCCGGGTTTGCTCGAAAACTACCCGCTGA CAGAAAGGCAAAAACGGTGGATCGAGGAGTGGAGAAGAGGAAAATAGTAGCTATTGCGCATGATA CAAGTTCATGCTACTATATTCCTTGTGCAACCTAAGCGATTTGCTTAAGCGAGGAGAACGATGCTCC GCTGCAGTGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
5	CTGCAGGACAACCGTTAAAGCGGACGTTTTCGCGCCGCCCGGGTTTGCTTGAAGACTACCCGCTGA CAGAGAAGCAAAGGCGATGGATCGAAGAGTGGAAGAAAGAGAAACAGTAGCTATTGCGCATGGTA CAAGTTTATGCTACTATATTCCTTGCGCAACTTAAGCGATTTGCCTAAGCGAGGAAAGCGGTGCTCC GCTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
6	CTGCAGAACAACCGTCAAAGCGGGCGTCTCTGCGCCGCCCGGACTTGCCTGAGAGCCACCTGCCGG CAGAAGAGCAAAAACGATGGACCGAAGAGTGGAGAAAAGAGAAACAGTAACTATTGCGCATGATA CAGGTTTATGCTACTATACTCCTTGTGCAACTCAAGCGATTTGCTCAAGCGAGGAAAACGATGTCCC GCTGTAGTGATGAAAAAGCGTTGTCTAGATAAGGAGTGATTCGAATG
7	CTGCAGAACAATCGTTAAAGCGGACGTTTTTGCGCCGCCCGGATTTGCTTGAAAACTACCCGCTGAC AGAAAAGCAAAAACGATGGATCGAAGAGTGGAAAAAAGAAAAACAGTAGCTATTGCGCATGATAC AAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATCTGCTTAAGCGAGGAAAACGATGTTCCG CTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
8	ctgcagGACAATCGTTAAAGCGGACGCTTTTGCGCTGCCCGGACTTGCTTGAAAACTTTCCGTTGACAG AAAAGCAAAAACGATGGATCGGGGAGTGGGAAAAAGAGAAACGGTAGCTATTGCGCACGGCACA AGCTTATGTTACTATATTCTTTGTGCAACTTAAGCGATTCGCTTAGACGGGGAGGATGGTGTTCCGC CGTAATGGTGAAGGAGCATTGTCTAGATAAGGAGTGATTCGAATG
9	CTGCAGAGCAGTCGTCAAAGCGGACGTTTCTGCGCCGCCCGGATTTGCTTGAAAACTACCCGCTGG CAGAAAAGCGAAAGCGATGGATCGAAGAGTGGAAAAAAGAAAAGCAGTAGCTATTGCGCATGATA CAAGTTTATGCTACTATATTCCTTGTGCAGCTCAAGCGACTCGCCTAAGCGAGGAAAGCGATGCTCC GCTGTAATGATGAGAAAGCGCCGTCTAGATAAGGAGTGATTCGAATG
10	CTGCAGAACAATCGTCAGAGCGGGCGTCTTCGCGCCGCCCGGGCTTGCTTGAAAACTACCCGCTGA CAGGAAAGCAGAGGCGGTGGGCCGAAGAGTGGGAGAAAGGAGAACAGTAGCTATGGCGCATGAT ACGAGTTTATGCTACTATATTCCTTGTGCAGCTTAAGCGATTTGCTTGAGCGAGGAAAACGATGCTC CGCTGCCACGGTGAAAAAACATTGTCTAGATAAGGAGTGATTCGAATG
11	CTGCAGAGCAGTCGTTGAAGCGGACGTTCTCGCGCCGCCCGGGTTTGCTGGAAGACTACCCGCTGG CAGAAAAGCAAAGACGATGGATCGGAGAGTGGAAAAAAGAAAAGCAGTAGCTATTGCGCATGATA CAAGCTTATGCTACTATGTTCCTCGTGCGACTTGAGCAATTTACTTAAGCGGGGCAAACGGTGTTCC GCTGCAATGATGAAAGAGCATTGTCTAGATAAGGAGTGATTCGAATG
12	CTGCAGGACAACCGCTAGGGCGGACACTCCTACGCCGCCCGGACTTGCTTGGAGACTGCCCGCTGG CCGGAGAGCAGAAGCGACGGCTCGAAGGGTGGAGAAAGGAAAAACAGCAGCGATTGCGCATGAT ACAAGTTTATGCTACTATATCCCTTGTGCAACTTAAGCGGCTTGCTTAAGTGAGGAAGACGGTGCCC CGCTGCAATGATGAAAGAGCATTGTCTAGATAAGGAGTGATTCGAATG
13	CTGCAGAACAATCGTTAAAGCGGACGTTTTTGCGCCGCCCGGATTTGCTTGAAAACTACCCGCTGAC AGAAAAGCAAAAACGATGGGTCGAAGAGTGGAAAAAAGAAAAACAGTAGCTACTGCGCATGATAC AGGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTGCTTAAGCGAGGAAAACGATGTTCCG CTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
14	CTGCAGAACAGTCATTCAAGCGGACGTTTTTGCGCCGCCCGGACTTGCTTGAGGACTACCCGCTGAC AGGAAAGCAAGGGCGCTGGGTCGAAGAGTGGAAAAAAAAAAAGACAGTAGCTACTGCGCGTGATAC AAGTTTATGCTACCATATTCCTTGTGCAACTTAAGCAATTTGCTTAAGCGAGGAAAGCGATGCCCCG CTGCAATGATGAGAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
15	CTGCAGAACAATCGTTAAAGCGGACGTTTTTGCGCCGCCCGGATTTGCTTGAAAACTACCCGCTGAC AGAAAAGCAAAAACGATGGATCGAAGAGTGGAAAAAAGAAAAACAGTAGCTATCGCGCATGATAC AAGTTTATGCTACTATATTCCTTGTGCAACTTAAGCGATTTGCTTAAGCGAGGAAAACGATGTTCCG CTGCAATGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG

16	CTGCAGAACAGTCGTTAAAACGGGCGTCTCTGCGTCGCCCGGGTTCGCTTGAAAACTACCCGCTGG CAGAAAAGCGAGAACAGTGGATCGGAGGGTGGAAAAAAGAAAAGCAGTCACTATTGCGCATGATA CAAGTTTATGCTGCTATATCCCTTGTGCAACCTAAGCGACTTGCTTAAGCGGGGAGAGCGGTATCCC GCTGCAATGATGGAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
17	CTGCAGGGCAACCGTCAAAGCGGACGCTCTTGCGCCGCCCGGACTTGCTTGAAAGCTACCCGCTGA CAGGAGAGCGAGGACGATGGATCGAAGGGTGGGGAAGAGAGGAACAGGAGCTATTGCGCGTGGT ACAAGTTCATGCTACTATATCCCCTGTGCAACTTAAGCGGTCTACTTAAGCGAGGAGAACGGTGCC CCGCTGCAACGATGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
18	CTGCAGAACAATCGTTAGAGCGGGTGTTTTCGCGCCGCCCGGACTTGCTCGAAAGCTACCCGCTGA CAGAGAAGCAGAAACGACGGACCGAGGAGTGGAAAAAGGAAAAACAGTAGCTACTGCGCATGAT ACAAATCTATGCTACTGTGTTCCTTGTGCAACTTAAGCGGTTTGCTTAAGCGGGGAAAGCGATGTTC CGCTGCAATGATGGAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG
19	CTGCAGAACAATCGTTAAAGCGGGCGTTTTTGCGCCGCCCGGGTCTACTTGAAAGCTACCCGCTGA CAGAAAGGCAAGAACGATGGCTCGAAGAGTGGAAAAAAGAAGAATAGCAGCTATTGCGCATGATA CAAGCTTGTGCCACTATATTCCTTGCGCAGCTTAAGCGATTTGCCTAAGCGCGGAAAACGATGGTCC ACTGCAATGATGAGGGAGCATTGTCTAGATAAGGAGTGATTCGAATG
20	CTGCAGAACAGTCGTTGAAGCGGACGCCCTTGCGCCGCCCGGACTTGCCTGGAAACTACCCGCTGA CAGAAAAGCGAAGACGATGGATCGAAGGGTGGAAGGAAGAAAGACAGCAGCTATTGCGCATGATA CAAGCTTACGCTACTGTATTCCTCGTGCAACTTAAGCAACTTGCTTGGGCGAGGAGAACGATGTTCC GCTGCAATGACGAAAAAGCATTGTCTAGATAAGGAGTGATTCGAATG

Rpls Library sequences table. Library members are named in descending order according to output strength in G. thermoglucosidans. pRpls1 has the strongest output with pRpls20 the weakest.

The Ldh Promoter

GGCGGGACGGGAGCTGAGTGCTCCCGTTGTTTGCCGCGGCGTCTGTCATG
AAATGGACAAACAATAGTCAAACAATCGCCACAATCGCGCATGCATTGCG
GTGCGCCTTTCGCGTAAAATATTTATATGAAAGTGTTCGCATTATATTGAG
GGAGGATGAATCATATG

Sequence of the *G. stearothermophilus* NCA1503 Ldh promoter as used in Taylor *et al*. 2008 and Cripps *et al*. 2009 (60,83). This sequence was used to compare strength with promoters generated in this study.

The Ldh Promoter and PheB RBS

GCGGGACGGGAGCTGAGTGCTCCC
GTTGTTTGCCGCGGCGTCTGTCATGAAATGGACAAACAATAGTCAAACAA
TCGCCACAATCGCGCATGCATTGCGGTGCGCCTTTCGCGTAAAATATTTAT
ATGAAAGTGTTCGCATTATATTGAGGGAGGATTCTAGATAAGGAGTGATT
CGAATG

Sequence of the *G. stearothermophilus* NCA1503 Ldh promoter plus *G. stearothermophilus* DSM6285 PheB gene RBS (in red) separated by an XbaI restriction site (in blue) as used in Bartosiak-Jentys *et al*. 2012 (85). The use of this RBS sequence was shown in this study to increase protein production and this RBS was also used with the pRplS promoter.

The Idh Promoter

CGATTTTTGCCGTAAGCCGCATGTCTGGATGGCTTGCACATATTTTGGAAC
AATATGATAACAATCGCCTCATCCGTCCGCGTGCAGAATATACAGGTCCG
GAGAAGCGGACGTATGTTCCGATTGAACAACGAGGCTAAATTAGTTTATA
AAAGGTGAGAAGATAGTTCTATTCTCACCTTTCACAACAAAAATATATTG
GAGGTTGTTATG

Sequence of the *G. thermoglucosidans* Idh promoter, this sequence was used to compare strength with promoters generated in this study.

The Idh Promoter and PheB RBS

CGATTTTTGCCGTAAGCCGCATGTCTGGATGGCTTGCACATATTTTGGAAC
AATATGATAACAATCGCCTCATCCGTCCGCGTGCAGAATATACAGGTCCG
GAGAAGCGGACGTATGTTCCGATTGAACAACGAGGCTAAATTAGTTTATA
AAAGGTGAGAAGATAGTTCTATTCTCACCTTTCACAACAAAAATATATTG
GAGGTTGTTTCTAGATAAGGAGTGATTCGAATG

Sequence of the *G. thermoglucosidans* Idh promoter, plus *G. stearothermophilus* DSM6285 PheB gene RBS (in red) separated by an XbaI restriction site (in blue).