HEURISTIC DISCOVERY AND DESIGN OF PROMOTERS FOR THE FINE-CONTROL OF METABOLISM IN INDUSTRIALLY RELEVANT MICROBES

Submitted by James Gilman to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Biological Sciences
in April 2018

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

Abstract

Predictable, robust genetic parts including constitutive promoters are one of the defining attributes of synthetic biology. Ideally, candidate promoters should cover a broad range of expression strengths and yield homogeneous output, whilst also being orthogonal to endogenous regulatory pathways. However, such libraries are not always readily available in non-model organisms, such as the industrially relevant genus *Geobacillus*.

A multitude of different approaches are available for the identification and *de novo* design of prokaryotic promoters, although it may be unclear which methodology is most practical in an industrial context. Endogenous promoters may be individually isolated from upstream of well-understood genes, or bioinformatically identified *en masse*. Alternatively, pre-existing promoters may be mutagenised, or mathematical abstraction can be used to model promoter strength and design *de novo* synthetic regulatory sequences.

In this investigation, bioinformatic, mathematic and mutagenic approaches to promoter discovery were directly compared. Hundreds of previously uncharacterised putative promoters were bioinformatically identified from the core genome of four *Geobacillus* species, and a rational sampling method was used to select sequences for *in vivo* characterisation. A library of 95 promoters covered a 2-log range of expression strengths when characterised *in vivo* using fluorescent reporter proteins. Data derived from this experimental characterisation were used to train Artificial Neural Network, Partial Least Squares and Random Forest statistical models, which quantifiably inferred the relationship between DNA sequence and function. The resulting models showed limited predictive- but good descriptive-power. In particular, the models highlighted the importance of sequences upstream of the canonical -35 and -10 motifs for determining promoter function in *Geobacillus.*

Additionally, two commonly used mutagenic techniques for promoter production, Saturation Mutagenesis of Flanking Regions and error-prone PCR, were applied. The resulting sequence libraries showed limited promoter activity, underlining the difficulty of deriving synthetic promoters in species where understanding of transcription regulation is limited. As such, bioinformatic identification and deep-characterisation of endogenous promoter elements was posited as the most practical approach for the derivation of promoter libraries in non-model organisms of industrial interest.

Extended Abstract

Predictable output is a defining aspiration of synthetic biology, and collections of thoroughly characterised genetic parts are a fundamental requirement if this predictability is to be achieved. Libraries of robust promoter sequences, for example, can be used to precisely trigger the expression of a transgene or synthetic pathway. Ideally, candidate promoters for synthetic biology applications should yield consistent output in a range of genetic and environmental contexts, whilst also being orthogonal to endogenous regulatory pathways. Although collections of promoters with the required attributes have been reported in model organisms such as *Escherichia coli*, promoter activity is often poorly conserved between species, hindering the application of previously characterised promoters in alternative hosts. The development of species-specific promoters is therefore necessary if the synthetic biology approach is to be applied in non-model organisms, such as the industrially relevant genus *Geobacillus*.

Different approaches are available for the identification and *de novo* design of species-specific promoters. Endogenous promoters may be individually isolated from upstream of well-understood genes, or bioinformatically identified *en masse*. Alternatively, previously identified promoter sequences may be mutagenised, potentially resulting in novel activity. The data that result from the characterisation of these promoter libraries can also potentially be used to derive mathematical models of the relationship between DNA sequence and promoter function. Such models could accelerate promoter discovery by making *pre hoc* predictions of promoter activity and could also potentially enhance our fundamental knowledge of genetic regulation in complex systems.

To expand the *Geobacillus* promoter toolkit, 636 putative promoters were bioinformatically identified from the core genome of four *Geobacillus* species. Low transformation efficiencies precluded the *in vivo* characterisation of all 636 sequences. To maximise the portion of the promoter design space that was empirically explored, a phylogeny of the putative promoters was used to rationally select sequences for *in vivo* characterisation in *G. thermodenitrificans* and *G. thermoglucosidans*. Two fluorescent reporter proteins, GFP and the RFP derivative mOrange, were used to quantify promoter activity.

In total, 105 *promoter::GFP* fusions and 82 *promoter::mOrange* fusions were characterised in *G. thermoglucosidans*. Although a 2-log range of expression levels was observed for both reporter proteins, promoter activity was generally poorly conserved between the reporters. However, seven promoters, covering a four-fold range of expression levels, were shown to function

consistently regardless of the downstream coding sequence. Five of these seven sequences were also shown to function independently of culture aeration.

Data derived from the *in vivo* characterisation of the bioinformatically-identified promoters were used to train Artificial Neural Network (ANN), Partial Least Squares (PLS) and Random Forest partition models that quantifiably linked promoter DNA sequence to function. Although ANN and PLS models were obtained that returned accurate fits of training, validation and primary test data sets, predictive accuracy was low when the sequence-function models were applied to predicting activity levels of secondary test sets of bioinformatically identified putative promoters or *de novo* designed synthetic promoter sequences. The lack of predictive power displayed by the models was hypothesised to be the result of a lack of significant sequence homology in the training data and the relatively small size of the training data set as compared to the dimensionality of the promoter design space.

Although the obtained ANN and PLS models displayed limited predictive power, Random Forest partitioning produced useful descriptive models. By identifying sequence positions that were key in determining promoter output, the partition models served to increase understanding of *Geobacillus* promoter structure. In particular, sequence positions upstream of the canonical -35 and -10 motifs were shown to strongly influence promoter activity. This result suggested that UP-elements, which had previously been identified in *Bacillus subtilis* and *E. coli* promoters, play a role in regulating transcription in *Geobacillus.*

Additionally, two commonly used mutagenesis-based techniques for promoter production, error-prone PCR (epPCR) and Saturation Mutagenesis of Flanking Regions (SMFR), were applied. In both instances, the *G. thermodenitrificans ldhA* promoter, which had previously been applied for metabolic engineering in various *Geobacillus* species, was used as the template sequence. However, only 5% of the characterised epPCR-derived sequences and 10% of the characterised SMFR sequences showed statistically significant promoter activity. Additionally, both epPCR and SMFR showed a proclivity to reduce promoter strength as compared to the wild-type template. This tendency was corroborated by a review of 21 published mutagenesis-derived promoter libraries.

Acknowledgements

There are a great many people without whose help this thesis would not have been possible. My sincere thanks to them all.

I would like to extend particular gratitude to my supervisor, John Love, for giving me the opportunity to undertake this PhD. Your endless belief, encouragement and support have been invaluable, and your aquarium has provided a welcome thesis-avoidance tool. Thanks also to Shell Research Ltd., in particular Rob Lee, Dave Parker and Jeremy Shears, for their backing and financial sponsorship.

I am grateful to Tom Howard, for his role as supervisor and idea sounding board in the crucial first year of this project, and to Thomas Lux and Christine Sambles for their help with bioinformatics. I also owe a big thank you to Chloe Singleton for training me in the lab and putting up with endless stupid questions, and to Paul James and Richard Tennant for all of their help, humour and a seemingly limitless supply of nicknames. Thanks also to Karen Moore and Audrey Farbos for their assistance with qPCR.

Thank you to colleagues past and present from the Exeter Microbial Biofuels Group, Mezzanine and BioEconomy centre laboratories, particularly Steve Brown, Anja Nennninger and Stefan Sassmann for all of their help, in and out of the lab. I also owe a particular thanks to Peggy Dousseaud, for providing some much needed balance.

Last, but by no means least, I would like to thank my family, in particular my parents, for their never-ending love and support. I am where I am because of them.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN **Artificial Neural Network**
A family of machine learning algorithms that relate input variables to one or more response variables via a series of interconnected nodes.

BAM **Binary alignment map**
A text-based bioinformatic file format. The compressed, binary representation of a Sequence alignment map.

BDBH **Bidirectional best BLAST hit**
A clustering algorithm used to identify homologous gene families from the genomes of multiple species of interest.

bp **Base pairs**
Two complementary nitrogenous molecules that are connected by hydrogen bonds to form the building blocks of the DNA double helix.

CDS **Coding sequence**
The region of gene's DNA or RNA sequence that encodes a protein.

COG **Clusters of Orthologous Groups**
A clustering algorithm used to identify homologous gene families from the genomes of multiple species of interest.

Ct **Cycle threshold**
The number of PCR cycles required to detect a fluorescence signal that is greater than background levels in quantitative PCR.

CV **Cross validation**
A series of techniques used to evaluate and compare the predictive accuracy of statistical analyses. The original data set is partitioned; a training set is used to train the model(s), and a validation set is withheld from the training process and subsequently used to evaluate the ability of the model to generalise to an independent data set.

Cvar **Coefficient of variance**
A measure of the relative variability of individual data points in a data series around the mean value.

ddH2o **Double distilled water**

dGTP **Deoxyguanosine triphosphate**
An oxidised derivative of the nucleoside deoxyguanosine. Used to induce transversion mutations (*i.e.* purine ⇔ pyrimidine) in error-prone PCR.

DNA **Deoxyribonucleic acid**

dNTP **Deoxynucleotide**

DoE **Design of Experiments**
The statistical, typically multivariate, approach to planning, conducting and analysing the results of tests to determine the relationships between factors that affect a process and the output of that process.

dPTP **2'-Deoxy-P-nucleoside-5'-Triphosphate**
A triphosphate derivative of the mutagenic nucleoside dP. Used to induce transition mutations (*i.e.* purine ⇔ purine or pyrimidine ⇔ pyrimidine) in error-prone PCR.

dsDNA **Double stranded DNA**

epPCR **Error-prone polymerase chain reaction**
A method for randomly inserting nucleotide mutations into a template DNA sequence.

FACS **Fluorescence Activated Cell Sorting**
A specialised form of flow cytometry that allows a heterogeneous mixture of cells to be analysed and sorted into separate containers based on the specific light scattering or fluorescence properties of each cell.

Fmol **Femtomole**

Fwd **Forward**

g **Gramme**

g **Centrifugal acceleration**

GFP **Green fluorescent protein**

h **Hour**

Kb **Kilobases**

kcal/mol **Kilocalorie per mole**

l **Litre**

LB **Lysogeny broth**

LLB **Lennox lysogeny broth**

LV **Latent variable**
Variables that are not directly observed or measured, but are instead inferred by a mathematical or statistical model.

M **Molar**

MCS **Multiple cloning site**
A sequence of DNA containing multiple restriction enzyme recognition sites. Used to insert DNA fragments into a plasmid.

min **Minute**

ml **Millilitre**

mLB **Modified lysogeny broth**

mM **Millimolar**

ng **Nanogrammes**

NIPALS **Nonlinear iterative partial least squares**
An algorithmic variant of Partial Least Squares regression.

nm **Nanometer**

nM **Nanomolar**

OD **Optical density**

OFAT **One-factor-at-a-time**
An approach to experimental design in which a single experimental variable (or factor) is changed and the effect on the response(s) of interest is observed. All other factors are held constant. This procedure is repeated in turn for each of the factors of interest in a particular study.

Oligo **Oligonucleotide**

OMCL **Orthogonal Markov cluster**
A clustering algorithm used to identify homologous gene families from the genomes of multiple species of interest.

PBS **Phosphate buffered saline**

PCN **Plasmid Copy Number**
The average or expected number of a given plasmid per host cell

PCR **Polymerase chain reaction**

PLS **Partial least squares**
A specialised form of linear regression that can be used to infer the relationship between a matrix of predictor variables (*X*) and a matrix of response variables (*Y*). Latent variables are extracted from the original predictor and response matrices in a way that maximises the covariance between *X* and *Y*.

PRESS **Predicted residual sum of squares**
A summary statistic used in regression analysis as a measure of the fit of a model to a validation or test data set. Specifically, PRESS represents an

estimate of the squared prediction error between an empirically measured response value and the value predicted by the model.

PWM **Position weight matrix**
A method by which motifs in biological sequence data can be represented. A PWM for a given sequence has the dimension 4 x *L*, where *L* is the length of the DNA sequence and the four rows represent the four DNA nucleotides. The matrix is populated with the probabilities of the given nucleotides being present at the defined sequence position.

qPCR **Quantitative polymerase chain reaction**
A molecular biology technique for the quantification of nucleic acids. A fluorescent dye is used that intercalates with double-stranded DNA during the amplification phase of the PCR, resulting in a fluorescence signal that is proportional to the number of amplicons in a sample. Comparing this signal to a standard curve allows the concentration of the amplicon of interest to be calculated.

RBS **Ribosome binding site**
The DNA sequence immediately upstream of the start codon of an adjacent mRNA transcript to which the 16S subunit of the ribosome is recruited during translation initiation.

Rev **Reverse**

RFP **Red fluorescent protein**

rpm **Revolutions per minute**

SAM **Sequence alignment map**
A text-based bioinformatic file format. Used to store the alignment of biological sequence data to a reference sequence.

sec **Second**

SIMPLS **Statistically inspired modification of PLS**
An algorithmic variant of Partial Least Squares regression.

SMFR **Saturation mutagenesis of flanking regions**
A mutagenesis-based approach to the production of synthetic promoter libraries, in which promoter consensus regions are maintained while flanking sequences surrounding the core motifs are mutagenised.

TFBS **Transcription factor binding site**
The sequence of DNA within a promoter to which regulatory proteins (transcription factors) bind. Transcription factors can promote or block the recruitment of RNA polymerase to a given promoter, hence up- or down-

regulating the transcription rate of the downstream coding sequence.

TSS **Transcription start site**
The DNA sequence at the 5’ end of a gene, at which transcription of DNA to RNA begins.

U **Units**

UV **Ultraviolet**

V **Volts**

VIP **Variable importance in projection**
A summary statistic used in Partial Least Squares (PLS) modelling. Provides a measure of the contribution of a given predictor variable (*X*) to the observed variation in the response variable(s) (*Y*).

µg **Microgramme**

µl **Microlitre**

µm **Micrometre**

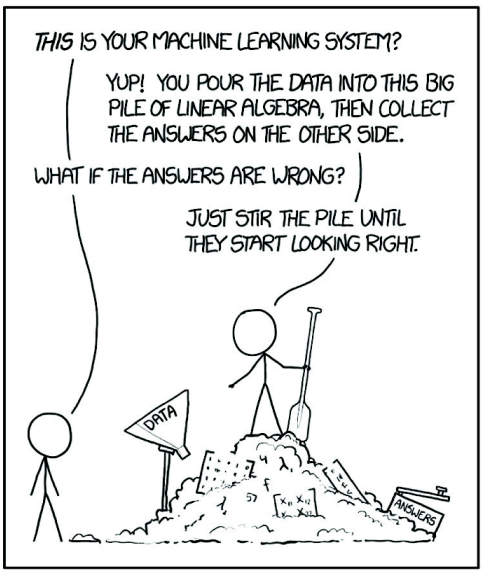
1G **First generation**

2G **Second generation**

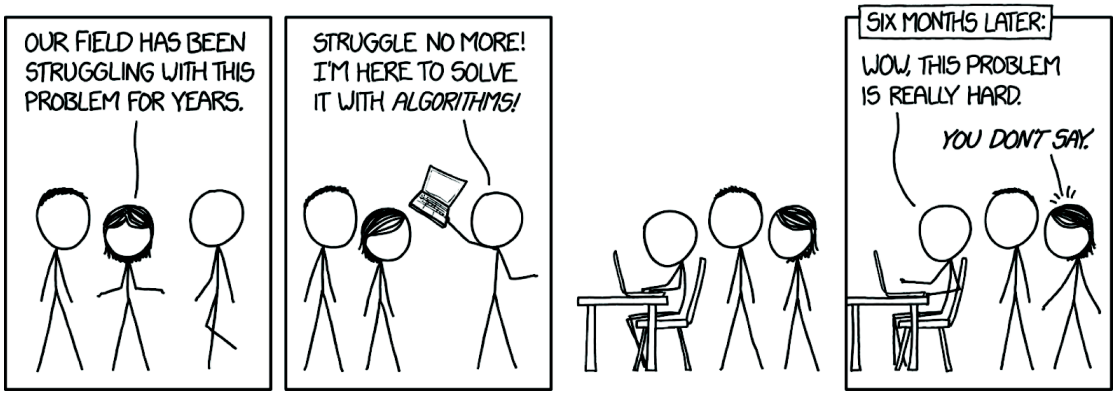
3G **Third generation**

4G **Fourth generation**

Visual Abstract



or,



xkcd.com

1 Introduction

Summary

The use of synthetic biology to produce biofuels that are chemically and structurally identical to the fossil fuels that they are intended to replace is of considerable industrial interest. Although model organisms are invaluable in proof-of-principle studies, they are not necessarily industrially applicable, partly due to the environmental extremes that can characterise industrial-scale bioproduction. However, engineering in non-model organisms such as the thermophilic bacterium *Geobacillus* is restricted by the limited availability of genus- or species-specific synthetic biology tools. For example, promoter sequences that allow varied and predictable control of transcription are a desirable feature of any synthetic biology toolkit. Promoters with the desired characteristics can be individually isolated from upstream of well-understood genes, or bioinformatically identified *en masse* using the genome or transcriptome of the organism of interest. Alternatively, pre-existing, well-understood promoters may be altered by mutation to generate synthetic sequences with novel activity. Statistical learning approaches that can decipher the effect of individual DNA bases or motifs on promoter output also have the potential to aid in promoter characterisation. Such models could be used to make predictions of promoter activity or *de novo* design synthetic promoter sequences. However, it is not clear which approach to promoter discovery and design is most applicable in an industrial context. This study therefore aims to provide a direct comparison between promoter discovery and design methods, and in particular to assess the applicability of statistical learning methods to promoter discovery and characterisation in *Geobacillus.*

1.1 Biofuels

Practical, political and environmental considerations have rendered current global fossil fuel consumption (and the resultant carbon dioxide emissions) unsustainable. The 2016 “Paris Agreement” resulted in a

commitment from 197 countries to limit global temperature increase to 1.5 °C above pre-industrial levels by the year 2100 (Kibria *et al.*, 2018), a target which requires a widespread reduction in emissions of greenhouse gasses. However, energy-related carbon dioxide emissions are projected to reach 43.2 billion metric tons by 2040, a 34% increase compared to the 2012 level (US Energy Information Administration, 2016). Given the current dependence of global economic activity on the combustion of fossil fuels (approximately 80% of all primary energy is derived from oil, coal or natural gas), the development of alternative energy sources that are renewable, sustainable and minimally disruptive to existing infrastructure is clearly necessary (International Energy Agency, 2017, Wojcik *et al.*, 2017).

The replacement of fossil fuels in the energy mix by biofuels (combustible fuels derived from biological material, primarily plant biomass) is an attractive proposition. Biofuels are of particular interest in the transport sector, where current infrastructure necessitates liquid, energy dense fuels (Shell International BV, 2016). The combustion of fossil fuels is deemed carbon-positive, as fuel combustion releases carbon into the atmosphere that was captured by ancient photosynthesis (Aro, 2015). In contrast, biofuels are considered carbon-neutral, as carbon dioxide that is removed from the atmosphere during plant growth is released during biofuel combustion; there is no net increase in atmospheric carbon (Mathews, 2008, Wojcik *et al.*, 2017). Although questions concerning the truly carbon-neutral nature of biofuels persist given the use of fossil fuels in the production process (Mathews, 2008, Aro, 2015), global demand for biofuel is predicted to reach 500 billion litres per year by 2040 (Cook *et al.*, 2017).

Biofuels are typically categorised as either first, second, third or fourth generation, depending on the origin of the feedstock and technological methods employed during their production, although these classifications have no legal or regulatory definition (Hoekman *et al.*, 2012). First generation (1G) biofuels are typically defined as those fuels that are produced by converting feedstocks that are primarily used as human food-sources, such as sugarcane, corn and wheat (Hoekman *et al.*, 2012, Wojcik *et al.*, 2017). Examples of 1G biofuel

include ethanol derived from the fermentation of plant sugars (bioethanol) and biodiesel derived from the transesterification of triglycerides from vegetable oils or animal fats (Hoekman *et al.*, 2012).

However, the diversion of food crops and the arable land used to grow them to biofuel production is controversial, and has led to the "Food vs. Fuel" debate (Tenenbaum, 2008, Bryant & Hughes, 2017). Second generation (2G) bioethanol and biodiesel aim to mitigate this issue by using lignocellulosic feedstock derived from the waste products of agriculture or forestry, or through the use of dedicated "energy crops" that do not impinge on human food supply and that are capable of growing on non-arable land (Carriquiry *et al.*, 2011). Both 1G and 2G bioethanol and biodiesel have been successfully commercialised, and are typically available to the consumer as petroleum/biofuel blends (Demirbas, 2009).

Despite the presence of 1G and 2G biofuels in the consumer energy mix, the use of such fuels is not without issues. For example, although the use of lignocellulosic feedstock in 2G biofuels mitigates "Food vs. Fuel" concerns, lignocellulose is highly recalcitrant to degradation and so typically requires energy- and chemical- intensive pre-treatment to liberate sugars that are subsequently fermented to produce bioethanol (Hess *et al.*, 2007, Wojcik *et al.*, 2017). Additionally, the hygroscopic nature of ethanol can cause corrosion to transportation infrastructure, an issue that can be compounded by the presence of contaminants such as halide or chloride ions (Howard, 2017). Biodiesel is also mildly hygroscopic and can form waxes at cold temperatures, restricting its use in cold climates or at high altitude (Brown *et al.*, 2018). The hygroscopicity of bioethanol and biodiesel also reduces their combustion temperature and energy density compared to fossil fuels (Oh *et al.*, 2018). As a result of these issues, 1G and 2G biofuels are typically restricted to a 10-15% blend with petroleum (Howard, 2017). This restriction is known as the blend wall (Tyner, 2015).

To circumvent the blend wall, so-called "advanced biofuels" make use of biogenic hydrocarbons that are chemically and structurally identical to the fossil

fuels that they are intended to replace (Brown *et al.*, 2018). In contrast to 1G and 2G biofuels, which use microbes as biocatalysts to convert feedstocks into fuel, third generation (3G) biofuels exploit oleaginous microbes as a source of naturally occurring hydrocarbons (Wojcik *et al.*, 2017). Algae (Cook *et al.*, 2017), cyanobacteria (Brown *et al.*, 2018), fungi (Leong *et al.*, 2018) and heterotrophic yeasts (Sargeant *et al.*, 2017) have all been identified as natural producers of hydrocarbons, and are of considerable academic and industrial interest (Cook *et al.*, 2017).

In contrast to the naturally occurring hydrocarbons that characterise 3G biofuels, fourth generation (4G) biofuels make use of synthetic biology for the design and construction of fuel-producing pathways (Howard *et al.*, 2013, Wojcik *et al.*, 2017). The synthetic biology approach can potentially facilitate optimisation of metabolic pathways for the production of tailored fuels or platform chemicals that are suited to existing infrastructure and engines (Howard, 2017), thus overcoming the blend wall. Additionally, microbes can potentially be engineered to degrade lignocellulosic biomass through the expression of heterologous saccharolytic enzymes (Bokinsky *et al.*, 2011), reducing the need for feedstock pre-treatment.

If 4G biofuels are to be produced on an industrial scale, careful consideration must be given to the choice of host organism. Although model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* are invaluable for laboratory-scale proof-of-principle studies (Bokinsky *et al.*, 2011, Howard *et al.*, 2013), these organisms may prove insufficiently robust for industrial-scale production. The development of alternative synthetic biology chassis organisms with the relevant characteristics is therefore of considerable scientific and industrial interest. Ideally, the chosen chassis organism should thrive under the potential environmental extremes (such extreme temperatures, pH and resource availability) that characterise large-scale industrial bioproduction (Adams, 2016). Additionally, certain organisms possess useful native metabolic pathways that are not found in model organisms (Lee & Kim, 2015). For example, the homologous expression of saccharolytic enzymes could facilitate host growth on lignocellulosic biomass, potentially reducing the

complexity of the heterologous pathways required to fabricate an industrially viable biofuel-producing microbe.

1.2 *Geobacillus*

The *Geobacillus* genus consists of Gram-positive, rod-shaped endospore forming bacteria that were initially classified as part of the genus *Bacillus* (Zeigler, 2014, Kananavičiūtė & Čitavičius, 2015). *Geobacillus* species are capable of growth at temperatures between 40 °C and 70 °C (Chen *et al.*, 2015) and appear to be globally ubiquitous, with representatives isolated from all seven continents (Zeigler, 2014). *Geobacillus* cultures have been isolated from environments as extreme as the Bolivian Andes and the Mariana trench, the deepest point in the world's oceans. *Geobacillus* species have also been extracted from subterranean oil fields and natural gas wells, although their presence in such locations may not be endogenous; the drilling process may lead to the introduction of bacterial species that are subsequently mistakenly identified as native (Struchtemeyer *et al.*, 2011).

It is hypothesised that such ubiquity may be a result of the spore-forming ability of *Geobacillus*, with life-cycles characterised by extensive growth in favourable environments, followed by sporulation and wide spread distribution (Hussein *et al.*, 2015). The most commonly reported natural sources of *Geobacillus*, however, reflect the thermophilic nature of the genus; compost piles, hot springs, geothermal soils and hydrothermal vents have all yielded isolates (Zeigler, 2014).

Geobacillus species have previously garnered interest for various biotechnological applications, typically as a source of thermostable enzymes for heterologous expression in mesophilic hosts. However, a number of *Geobacillus* species display key genomic and phenotypic attributes which render them potentially applicable for use as potential chassis organisms for industrial biotechnology or synthetic biology (Kananavičiūtė & Čitavičius, 2015).

The thermophilic nature of the genus is of particular interest, as large metabolic loads, and therefore high temperatures, are generated by fermentation at industrial scales. Increased temperatures also serve to reduce the risk of biotic contamination and increase the rate of feedstock conversion. Furthermore, cooling costs are reduced and the recovery of volatile products is simplified (Chen *et al.*, 2015). Additionally, a high growth rate, comparable to that of *Escherichia coli* or *Bacillus subtilis* (Suzuki *et al.*, 2013), and an ability to reach high cell densities mean that *Geobacillus* species could potentially produce large volumes of product from engineered pathways in a relatively short time (Kananavičiūtė & Čitavičius, 2015).

The catabolic versatility of *Geobacillus* also provides a potential advantage with regards to their application in large-scale industrial bio-production. Species in the genus have the reported ability to metabolise pentose and hexose sugars into ethanol, lactate, formate and acetate (Bezuidt *et al.*, 2015, Raita *et al.*, 2016, Zhou *et al.*, 2016). Furthermore, *G. thermoglucosidans* DSM2542 is able to utilise cellobiose and short-chain oligosaccharides such as xylan (Bartosiak-Jentys *et al.*, 2013). Expression of cellulases has additionally been reported in *Geobacillus* sp. T1 (Assareh *et al.*, 2012) and *Geobacillus* sp. R7 (Zambare *et al.*, 2011), and a xylanase has been isolated from *G. stearothermophilus* and engineered for improved thermostability (Zhang *et al.*, 2010).

The native expression of saccharolytic enzymes, coupled with the potential for the heterologous expression of engineered variants, raises the possibility of utilising *Geobacillus* species for the synthesis of desirable products, directly fuelled by the catabolism of a cheap, readily available lignocellulose-derived feedstock (Bhalla *et al.*, 2014).

The genetic tractability of *Geobacillus* is also advantageous, with successful engineering of the genus having previously been reported for a number of purposes. A highly cited example reports the production of enhanced ethanol yields by an engineered strain of *G. thermoglucosidans* (Cripps *et al.*, 2009). Ethanol yields were increased by 0.32 g per g of glucose substrate, as

compared to wild-type *G. thermoglucosidans*. Ethanol production was also reported at a yield of 0.47 g per g of cellobiose, as was successful fermentation of a mixed hexose and pentose feedstock. Subsequent metabolic flux analysis of the engineered strain suggested that yield could be increased in fed-batch fermentation to as much as 5.2 g ethanol per litre of culture per hour, based on growth media containing 12% weight/volume cellobiose (Niu *et al.*, 2015). Heterologous Isobutanol production has also been demonstrated in *G. thermoglucosidans*, with yields of 3.3 g per litre of culture from a feedstock containing glucose (Lin *et al.*, 2014a).

Engineering of *Geobacillus* has been rendered more practical by the development of a limited synthetic biology toolkit. One study, for example, reported two origins of replication, kanamycin and chloramphenicol selection markers, three reporter proteins, a 20 member synthetic promoter library (SPL) covering a 100-fold activity range and a four-member RBS library (Reeve *et al.*, 2016). Additional SPL and RBS libraries (Pogrebnyakov *et al.*, 2017), shuttle vectors (Taylor *et al.*, 2008, Bartosiak-Jentys *et al.*, 2013) and transformation methodologies (Kananavičiūtė & Čitavičius, 2015) can also be found in the literature, as can strategies for the development of gene knock-in and knock-out mutants (Cripps *et al.*, 2009, Sheng *et al.*, 2016), although off-target single-nucleotide polymorphisms and insertion/deletion mutations are not uncommon.

As a result of the characteristics and burgeoning genetic toolkit discussed above, *Geobacillus* was selected by the sponsor (Shell Research Ltd.) as the target host genus for the development of 4G biofuel (Howard *et al.*, 2013), translating the proof of principle study performed in *E. coli* to a more industrially relevant host.

The required pathways are complex, involving the coordinated expression of nine genes for alkane production. If additional sugar catabolism is required, more pathways must be expressed, increasing the requirement for different control systems such as a battery of promoter sequences that are significantly insulated from endogenous metabolism. Without an array of sufficiently characterised parts, the synthetic biology approach of combining

genetic modules to confer novel functionality to an organism cannot be applied in *Geobacillus*. Engineering in the strain would therefore remain the *ad hoc* process that has characterised biological engineering in the pre-synthetic biology era, with all of the caveats that process implies (Endy, 2005) and lacking the abstraction, standards and composition frameworks that define and expedite more mature engineering disciplines (Canton *et al.*, 2008). An expansion of the *Geobacillus* synthetic biology toolkit is therefore necessary to fully exploit the industrial potential shown by the genus.

1.3 Predictable system output for synthetic biology

Predictable output is a defining aspiration of synthetic biology. A number of factors affect the output from synthetic gene networks to a greater or lesser extent, including transgene copy number (Ajikumar *et al.*, 2010), integration into the genome or expression from plasmids (Tyo *et al.*, 2009), promoter activity (Blazeck & Alper, 2013), ribosome binding sites (Ravasi *et al.*, 2012, Lin *et al.*, 2014b, Markley *et al.*, 2015), codon bias of the host (Quax *et al.*, 2015), transcription rate and tRNA abundance (Angov, 2011), half-life of mRNA (Curran *et al.*, 2013), substrate and co-factor availability (Jones *et al.*, 2015), adjustment of enzyme kinetics (Bloom *et al.*, 2005), protein scaffolding (Dueber *et al.*, 2009) and sub-cellular localisation through the use of microcompartments (Parsons *et al.*, 2010, Boyle & Silver, 2012). The use of RNA as a control mechanism, either through the application of riboswitches (Mellin & Cossart, 2015) or toehold switches (Green *et al.*, 2014) has also emerged as a powerful tool for pathway control. Each of these aspects can be investigated and improved individually and subsequently be integrated by a model, a suite of experiments, or, ideally, using the combination of modelling and empiricism that defines synthetic biology.

Controlling transcription is often the simplest way to trigger expression of a transgene or synthetic pathway, and constitutive promoters with different and predictable activation characteristics are a desirable feature of any synthetic biology toolkit. Indeed, promoters with different and, most importantly, predictable effects on transcription may be used to regulate complex gene

circuits, balance engineered metabolic pathways and exploit new chassis for industrial-scale applications. However, in practice, promoter availability tends to be restricted to relatively few sequences (Lu *et al.*, 2009), which do not always perform as required and may not necessarily be transferrable to new microbial chassis. The fact that many promoters are characterised as merely "weak" or "strong" (Ellis *et al.*, 2009) highlights this issue; such definitions are hardly sufficient to allow adequate promoter selection for complex pathway engineering.

In lieu of a library of constitutive promoter elements with a broad range of activity, inducible promoter elements may appear superficially attractive as a possible alternative. By modulating the concentration of inducer, the desired level of protein production can, in theory, be achieved (Siegele & Hu, 1997). However, although the use of inducible promoter systems has been successful in some instances, in others it can prove inadequate. Promoter hypersensitivity to the inducer (Hammer *et al.*, 2006), the cost of adding large quantities of inducer to an industrial-scale fermenter (Jensen & Hammer, 1998b) or heterogeneous expression levels across a population (Khlebnikov *et al.*, 2001) all complicate the use of inducible promoters in industrial-scale cultures. Consequently, for large-scale production applications, constitutive promoters with "hard-wired", predictable properties are often preferred, and are therefore the focus of this study.

1.4 The structure of *cis*-regulatory elements

1.4.1 Prokaryotic cis-regulatory elements

A promoter can be broadly defined as a cis-regulatory element containing a suite of key sequence motifs that control the transcription of individual open reading frames (ORFs) or operons. In prokaryotes, the structure and organisation of natural promoter motifs is relatively well understood (Figure 1.1). Two conserved hexamers, located at approximately 10 and 35 base pairs (bp) upstream of the Transcription Start Site (TSS) serve as key binding regions for RNA polymerase (Kanhere & Bansal, 2005). No such conserved motifs have

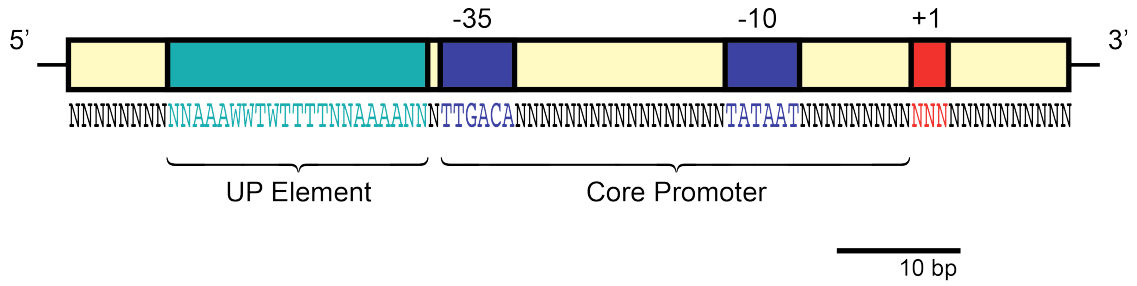


Figure 1.1: Schematic representation of a typical prokaryotic promoter element.

The Transcription Start Site (TSS) is shown in red. Two conserved hexamers, at approximately 10 and 35 bp upstream of the TSS are shown in blue. An upstream (UP) element is shown in turquoise. Such regions are typically rich in Adenine (A) and Thymine (T) residues, and may not be present in all promoter sequences. UP element consensus sequence is as derived by Estrem *et al.* (1998), and consensus regions within the core region are reproduced from Blazeck & Alper (2013) and Ross *et al.* (1998). N represents any deoxyribonucleotide. W represents A or T residues. G and C represent Guanine and Cytosine, respectively. The promoter represented in this figure is of arbitrary length. The length of promoters for synthetic biology applications is not rigidly defined. For example, short sequences such as the 35 bp Anderson promoters are available that contain only -10 and -35 motifs (iGEM, 2018). Alternatively, promoter sequences of 100 bp or greater containing insulator sequences or defined UP elements are also available (Davis *et al.,* 2011).

been identified in the region of sequence separating the two conserved regions, although a consensus length of 17 bp has been observed in some species (Nair & Kulkarni, 1994). Together, the consensus regions and the spacer DNA between them is often referred to as the core promoter.

In addition to these core promoter elements, an upstream region is present in some promoters. Typically rich in adenine and thymine residues, these UP elements boost transcription rate by facilitating binding with the C-terminal domain of the RNA polymerase α-subunit (Estrem *et al.*, 1998), and may also play a role in the subsequent process of transcription initiation, including open complex formation (Strainic *et al.*, 1998).

Once transcribed by the RNA polymerase, the mRNA transcript istranslated to protein. Interactions between the Ribosome Binding Site (RBS) and the ribosomal RNA are key for determining the efficiency of translation initiation (Laursen *et al.*, 2005). As such, the judicious choice of RBS represents one of the major synthetic biology control points (Reeve *et al*., 2014).

The role of the core promoter

The importance of consensus regions in conferring promoter activity to a sequence is well understood. Randomly generated, 103 bp sequences with no promoter activity have been shown to be only one or two mutations away from becoming active promoters, with such mutagenesis typically resulting in near-canonical -35 or -10 regions (Yona *et al.*, 2017).

Consensus regions are key to transcription initiation due to their interactions with RNA polymerase (Figure 1.2). The supply of RNA polymerase within a cell is one of the major limiting factors for the rate of translation initiation; the efficiency with which a given promoter or group of promoters can bind RNA polymerase is therefore a major determinant of promoter activity (Browning & Busby, 2004).

The RNA polymerase enzyme is comprised of subunits $\beta\beta'\alpha_2\omega$, with a temporary interaction between these core subunits and a σ factor being required for the formation of a holoenzyme capable of initiating transcription (Browning & Busby, 2004). It is the σ factor that serves to confer promoter specificity; by encoding multiple σ factors, bacteria are able to broadly globally modulate transcription patterns in response to environmental stimuli through up- or down-regulation of specific promoter families. Finer levels of control are subsequently achieved by activation or repression of specific transcripts through the action of transcription factors (Gruber & Gross, 2003).

Studies of *Thermus aquaticus* σ^A have shown that the anchoring of RNA polymerase to the promoter sequence is the result of extensive interactions between the protein and the phosphate backbone of the DNA. In particular,

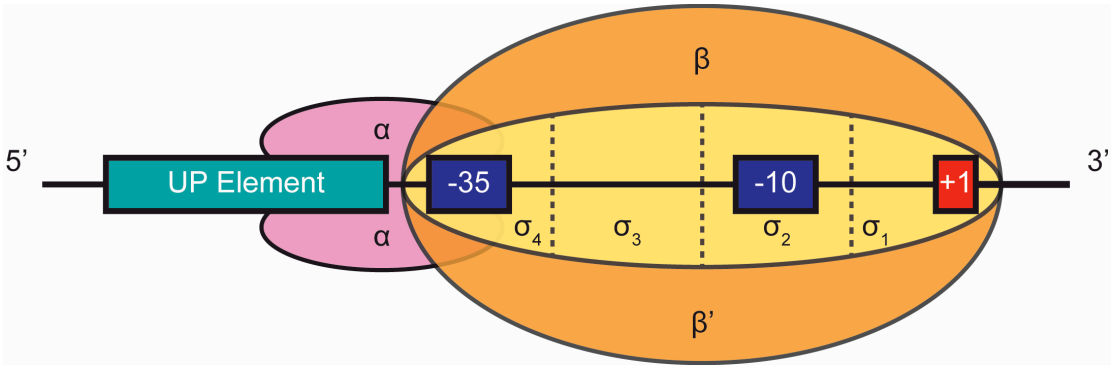


Figure 1.2: Schematic representation of the interaction between prokaryotic promoter sequence and RNA polymerase holoenzyme.

Promoter consensus regions are shown in dark blue, the upstream (UP) element is shown in turquoise and the Transcription Start Site (TSS) is shown in red. The RNA polymerase α subunits are shown in pink, β subunits are shown in orange and the bound σ factor is shown in yellow.

σ factor domains 2.4 and 4.2 bind the -10 and -35 elements respectively (Campbell *et al.*, 2002). In addition, consensus sequences may play a part in other aspects of transcription initiation. For example, the stabilisation of the open complex is a function of promoter elements within and downstream of the -10 region and their interaction with σ^1 (Ruff *et al.*, 2015).

Once the RNA polymerase has bound to the promoter, large scale conformational changes result in the DNA being moved into the active site cleft of the enzyme (Ruff *et al.*, 2015). Promoter DNA spanning from the -10 region to 2 bases downstream of the TSS is subsequently unwound to form an open complex (Gries *et al.*, 2010) in a process known as isomerisation. Thereafter, RNA synthesis initiates in an abortive, stochastic fashion, with the polymerase releasing short RNA transcripts (fewer than 10 nucleotides) and returning to the TSS. Abortive initiation continues until the nascent RNA molecule exceeds a critical length of approximately 11 nucleotides, at which stage the RNA polymerase successfully clears the promoter sequence (Saecker *et al.*, 2011).

The efficiency of promoter clearance is, in part, a function of the 20-nucleotide sequence immediately downstream of the TSS (Davis *et al.*, 2011).

Post-escape, the sigma factor is released and transcript elongation can proceed.

Although important in the initiation of prokaryotic transcription, it is by no means requisite for a promoter to have both consensus elements in order to be functional, and it is rare for an individual promoter sequence to have a fully canonical set of conserved domains (Browning & Busby, 2004). In some instances, for example, the presence of an extended -10 element is sufficient to offset the absence of σ domain 4, and therefore interactions between the σ factor and -35 consensus region, likely due to interactions between σ domain 3 and the extended -10 region (Brown *et al.*, 1997). The activity of a promoter with sub-optimal consensus sequences can also be increased through the actions of activating transcription factors, which can recruit RNA polymerase to the promoter (Browning & Busby, 2016).

Furthermore, it appears that certain positions within the consensus sequences are of greater importance than others; in the -35 region of *E. coli* promoters, for instance, -35T, -34T and -33G are the most heavily conserved residues (Lisser & Margalit, 1993). Also in *E. coli,* within the -10 region an Adenine residue at -11 is thought to play a key role in open complex formation (Cook & deHaseth, 2007), with a Thymine residue at -7 also thought to be key in stimulating promoter melting (Heyduk & Heyduk, 2014). It should be noted, however, that neither residue is mandatory, with similar initiation kinetics being observed when the -10 element as a whole is AT rich (Heyduk & Heyduk, 2014).

The DNA sequence between the two consensus regions does not typically contain any conserved motifs, with spacer sequence length instead regulating spatial alignment of the consensus regions and therefore facilitating interactions between promoter DNA and the RNA polymerase (Sztiller-Sikorska *et al.*, 2011). Optimal spacer length is therefore key in determining the efficiency of both RNA polymerase binding and open complex formation. It is hypothesised that the inefficiency of many natural promoter sequences is, at least in part, a result of sub-optimal spacer length. Such naturally inefficient

promoters in many cases therefore rely on additional promoter motifs, such as UP elements, or the activity of activator proteins, to boost transcription (Adhya *et al.*, 1993). In addition, the GC content of the spacer has also been shown to play a role in determining the expression level of the gene of interest (Deng *et al.*, 2018).

The role of upstream elements

In certain promoters, transcription levels are boosted by a region of DNA sequence located upstream of the -35 consensus region termed the UP element (Ross *et al.*, 1993). The most widely studied wild-type UP element, that of the *E. coli rrnB* P1 promoter, for example, spans positions -40 to -60, with two key motifs centred at -42 and -52, and has been shown to increase promoter activity by approximately 30-fold (Ross *et al.*, 1993).

Given the role of UP elements in boosting transcription and their apparent modularity (Rao *et al.*, 1994), the inclusion of such elements in synthetic promoter sequences is an interesting prospect. In one study, a synthetic consensus 24 bp UP element was placed immediately upstream of the -35 sequence of 17 constitutive *E. coli* promoters from the Anderson promoter collection, and the resulting synthetic promoters were characterised upstream of a GFP reporter. The presence of said UP element increased promoter activity in 15 out of the 17 characterised sequences by between approximately one- and 95-fold compared to the core promoter alone, with the percentage increase typically being greatest in promoters of moderate strength (Yan & Fong, 2017).

Interestingly, UP element insertion also appeared to reduce the stochastic fluctuation in promoter activity that can prove a hindrance when designing synthetic constructs and pathways (Yan & Fong, 2017). 70% of characterised promoter constructs displayed a statistically significant decrease in the coefficient of variation for GFP fluorescence when an UP element was placed upstream of a core promoter, as compared to the core promoter alone. Reduced expression variability may be a consequence of increased binding

affinity between RNA polymerase and promoter, and hence tighter control of mRNA levels (Yan & Fong, 2017).

However, a study of synthetic promoter sequences in *Pseudomonas putida* has shown that the presence of UP elements in a promoter is not a guarantee of increased protein expression. The integration of a rRNA promoter UP element either boosted observed fluorescence from the reporter protein mNeonGreen by up to five-fold, or reduced fluorescence by up to 23-fold, dependent on the promoter sequence into which the UP element was integrated (Elmore *et al.*, 2017). Additionally, the location at which UP elements are inserted into a synthetic promoter sequence must be carefully considered. Displacement of the *E. coli rrnB* P1 UP element by as little 5 bp is sufficient to abolish UP element dependant transcription (Meng *et al.*, 2001).

The role of the Ribosome Binding Site

Post-transcription, the control of translation initiation represents a major control point for synthetic biology. The Ribosome Binding Site (RBS), located upstream of the start codon in the 5' untranslated region of the mRNA transcript, facilitates the initiation of translation by recruiting ribosomes to the mRNA transcript, specifically through hybridisation of the RBS Shine-Dalgrano sequence to the 16S ribosomal RNA (Shine & Dalgarno, 1974, Boyle & Silver, 2012, Singh, 2014). By altering the RBS on a given mRNA transcript, the efficiency of translation initiation, and therefore protein output, can be modified (Ang *et al.*, 2013). Given that, in most instances, initiation is the rate limiting step in prokaryotic translation (Gualerzi & Pon, 1990, Laursen *et al.*, 2005), the judicious selection of RBS is a key consideration for regulating the output of synthetic pathways.

RBS sequences for synthetic biology applications can either be molecularly or bioinformatically identified from the species or genus of interest, or random mutation of well understood sequences can be used to generate libraries of RBS sequences with a range of activity levels (Anderson *et al.*, 2006, 2007). To facilitate the prudent selection of RBS for synthetic biology

projects, a number of computational models have also been developed that design custom RBS sequences that provide the desired translation initiation rate for a specified gene (Boyle & Silver, 2012, Reeve *et al.*, 2014). The most widely cited example, the RBS calculator developed by Salis *et al.*, is based on a thermodynamic model that quantifies the strength of interactions between the ribosomal RNA and the mRNA transcript and therefore predicts the translation rate of a given RBS sequence (Salis *et al.*, 2009).

1.4.2 Eukaryotic promoters

Eukaryotic promoters are more complex than their prokaryotic counterparts (Figure 1.3), with localisation of the transcriptional apparatus resulting from interactions between highly specific transcription factors, the promoter elements and co-activators (Hahn & Young, 2011). Broadly speaking, two regions are present: a core promoter element and an upstream enhancer (Blazeck & Alper, 2013). Both elements may be modified in order to modulate expression levels. The core region provides the basal sequence necessary for transcription initiation and may contain key motifs, the most widely studied of which is the TATA box, which typically occurs 40 to 120 bp upstream of the TSS (Hampsey, 1998). However, such motifs are by no means requisite for transcription initiation, as TATA boxes have been shown to appear in only 20% of *Saccharomyces cerevisiae* promoter elements (Basehoar *et al.*, 2004).

Upstream of the core promoter, the enhancer element serves to localise transcription factors, with the interactions between bound transcription factors and the transcriptional machinery serving as a determinant of promoter strength and control (Blazeck *et al.*, 2012). Transcription factor binding sites do not display uniform distribution across the enhancer element. The highest concentration of such binding motifs has been reported between 50 and 150 bp upstream of the TSS (Hughes *et al.*, 2000), although they may be present as much as 500 bp upstream of the TSS.

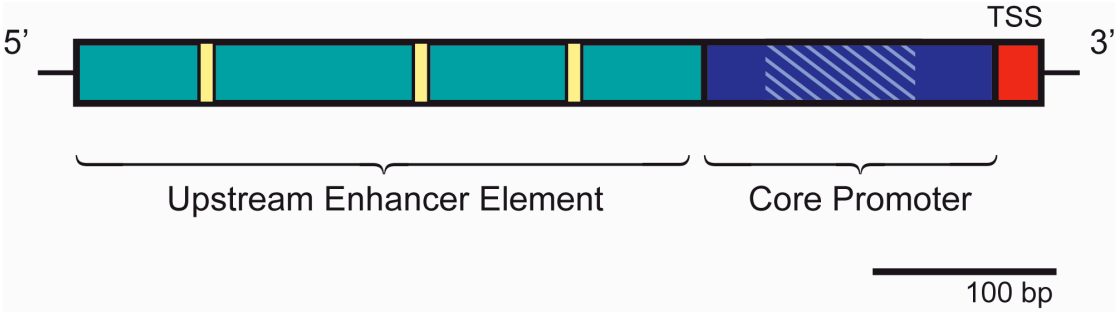


Figure 1.3: Schematic representation of a typical *Saccharomyces cerevisiae* promoter element.

The Transcription Start Site (TSS) is highlighted in red, and the core promoter element is shown in blue. Diagonal lines represent the area of the core promoter in which TATA boxes are most commonly found. The upstream enhancer is shown in turquoise, with transcription factor binding sites represented by yellow boxes in arbitrary positions.

1.5 Characteristics of *cis*-regulatory elements for synthetic biology

From an industrial perspective, it is preferable to have a production system that displays little variation, even if the overall output of that system is, on average, slightly less than that of an alternative that displays irregularities; synthetic biology aims to be boringly predictable rather than wonderfully complex.

Candidate *cis*-regulatory elements for synthetic biology must therefore be well characterised and yield homogeneous, consistent output, while also being insulated from background metabolisms and molecular control systems. However, consistency is often confounded by the inherently stochastic nature of gene expression, which subjects both *cis*-regulatory elements and any downstream proteins used in their characterisation to large degrees of noise (Rudge *et al.*, 2016). Expression noise can be broadly classified as either intrinsic (a consequence of the properties of the regulatory sequence and downstream gene) or extrinsic (variables which are not a direct result of the regulatory sequence but impact upon it, such as cellular concentrations of RNA polymerase or mRNA degradation) (Elowitz *et al.*, 2002). A useful example of

the stochasticity of gene expression is the all-or-nothing phenomenon (De Mey *et al.*, 2007), wherein expression reaches the expected level in a sub-set of the population, whilst the remaining cells display no expression. Well understood as a characteristic of inducible promoter systems (Siegele & Hu, 1997, Keasling, 1999, Morgan-Kiss *et al.*, 2002), all-or-nothing performance may also affect some constitutive promoter elements.

Multiple approaches are available for the identification of potentially applicable natural promoters, or for the development of synthetic constitutive promoter elements. These range from the more conventional PCR-based techniques and hybrid promoter engineering to the use of computational analysis and modelling for the *de novo* design of promoter elements with defined functionality.

1.6 Endogenous promoter sequences

The promoters available for use in synthetic systems have generally been limited to those endogenous elements isolated from model organisms, for instance, the *Escherichia coli lac* promoter and derivatives thereof (de Boer *et al.*, 1983, Makoff & Oxer, 1991, Jensen *et al.*, 1993, Terpe, 2006), and the arabinose-inducible P_{BAD} (Cagnon *et al.*, 1991, Guzman *et al.*, 1995, Wycuff & Matthews, 2000) promoter.

Phage genomes can also be used to generate novel promoters. For example, the p_L promoter, isolated from bacteriophage lambda, provides medium to high expression levels, and is tightly thermally-regulated by the cI repressor (Terpe, 2006, Valdez-Cruz *et al.*, 2010). p_L has been successfully employed to increase yield of various proteins in *E. coli* expression systems (Mellado & Salas, 1982, Simons *et al.*, 1984, Elvin *et al.*, 1990). Similarly, the T7 RNA Polymerase-based promoter system, also initially isolated from bacteriophage, has been widely adopted (Studier & Moffatt, 1986, Terpe, 2006).

Perhaps unsurprisingly, the native genome remains, to date, the most commonly mined source of promoter elements for engineering application in

Geobacillus. The *G. kaustophilus sigA* promoter, for example, has been used for the heterologous expression of an α-amylase and a β-galactosidase integrated into the *G. kaustophilus* HTA426 genome (Suzuki *et al.*, 2012).

Additionally, the promoter of the lactate dehydrogenase (*ldh)* gene has been isolated from both *G. thermodenitrificans* and *G. stearothermophilus* and utilised for heterologous expression. The *ldh* promoter from *G. thermodenitrificans*, for example, was used to facilitate heterologous production of Isobutanol in *G. thermoglucosidans*. Expression of the genes for a *Lactococcus lactis* ketoisovalerate decarboxylase, a *G. thermodenitrificans* ketol-acid reductoisomerase and a *Bacillus subtilis* acetolactate synthase as an operon under the control of the *G. thermodenitrificans ldh* promoter resulted in a yield of 3.3 g Isobutanol per litre of culture, when grown on 0.2 M glucose (Lin *et al.*, 2014a). Likewise, the *G. stearothermophilus* DSM2027 *ldh* promoter was used for the upregulation of the *G. thermoglucosidans* pyruvate dehydrogenase operon, resulting in increased ethanol yields (Cripps *et al.*, 2009).

Expression levels under the *G. stearothermophilus ldh* promoter have been shown to be highly dependent on culture aeration (Bartosiak-Jentys *et al.*, 2012), with oxygen dependence also evident in the *G. thermodenitrificans ldhA* promoter (Figure 1.4). Given the largely anaerobic nature of fermentation, the ability to induce enzymatic expression under oxygen limitation may be advantageous in certain scenarios (Kananavičiūtė & Čitavičius, 2015). However, the inherent variability of *ldh* promoter activity renders its use inadequate for more complex metabolic engineering, where constitutive, predictable output under a range of environmental conditions is required.

In addition to the constitutive *Geobacillus* promoter sequences discussed above, a number of inducible systems have also been identified in the genus, such as the temperature sensitive *G. stearothermophilus sgsE* promoter. Natively responsible for the induction of surface layer protein when culture incubation temperature is increased, the promoter was reported to upregulate expression of enhanced GFP (EGFP) in *B. subtilis* when temperature was increased from 28 °C to 45 °C (Novotny *et al.*, 2008).

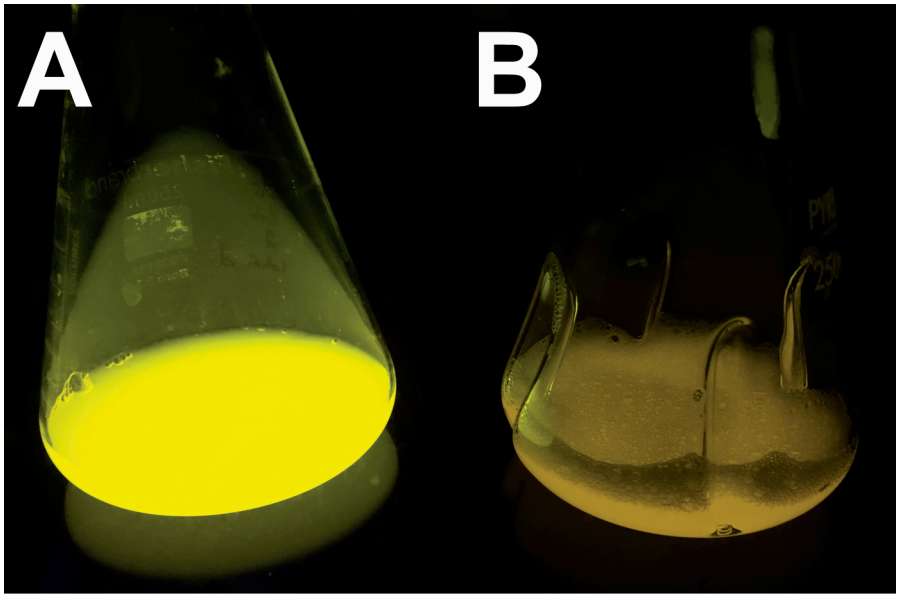


Figure 1.4: *G. thermoglucosidans* cultures expressing GFP under the control of the *G. thermodenitrificans ldhA* promoter.

G. thermoglucosidans was cultured in non-baffled (A) or baffled (B) 250 ml conical flasks. Cultures were incubated at 60 °C, with shaking at 220 rpm for 24 h, from an intial $OD_{600\ nm}$ of 0.1. Growth media was modified Lennox Broth (mLB: 10 g l^{-1} tryptone, 5 g l^{-1} NaCL, 5 g l^{-1} yeast extract, 1.05 mM $C_6H_6NO_6$, 0.91 mM $CaCl_2$., 0.59 mM $MgSO_4$ and 0.04 mM $FeSO_4$).

Baffles increase culture agitation, and therefore oxygen transfer. Increased GFP expression is clearly visible when cultures are grown in non-baffled flasks, highlighting the oxygen dependence of the *G. thermodenitrificans ldhA* promoter.

Ligand inducible promoter sequences in *Geobacillus* have also been reported, with a particular emphasis in the literature on sugar-inducible systems. Analysis of the *G. kaustophilus* HTA426 genome, for example, resulted in the isolation of promoters inducible through the addition of D-galactose, lactose, mannose and myoinositol (Suzuki *et al*., 2013). Similarly, the *G. stearothermophilus* NUB3621 *surP* promoter has been shown to be sucrose-inducible when cloned upstream of an α-galacosidase gene (Blanchard *et al*., 2014), although α-galacosidase activity was seen in the absence of inducer. Whilst the *surP* promoter may be therefore applicable in instances where

upregulation of a gene of interest is required, it may prove inadequate in situations where tight control is necessary.

A cellobiose-inducible promoter, Pβglu, isolated from upstream of the operon encoding the *G. thermoglucosidans* DSM2542 cellobiose-specific phosphotransferase system, has also been reported. Inducer-dependent promoter activity was confirmed by characterisation with the reporter gene *pheB,* with an approximately 600-fold up-regulation of catechol C2,3-dioxgenase activity in the presence of cellobiose observed in *G. thermoglucosidans* DSM2542. However, Pβglu provided inadequate levels of promoter activity when used to express a second gene, a *Thermotoga maritima* endoglucanase. A constitutive alternative isolated from upstream of the *G. thermoglucosidans* uracil phosphoribosyltransferase gene was therefore used in the place of Pβglu in a final engineered pathway (Bartosiak-Jentys *et al.*, 2013). This serves to highlight the risk in selecting a promoter based solely on characterisation upstream of a single reporter protein; the inherent context-dependency of promoter elements renders such an approach inadequate.

1.7 Molecular approaches to producing synthetic promoter libraries

Historically, identification of natural bacterial promoter elements has relied upon characterisation experiments consisting of labour intensive cloning of putative promoters upstream of a reporter gene (Zhou *et al.*, 2017). Whilst large numbers of native promoter elements have been identified in this manner, engineering projects have typically relied on promoters from the small subset of thoroughly characterised, widely employed sequences such as those discussed above. Whilst the success of projects employing these promoters cannot be doubted, the lack of diversity in promoter choice may be restrictive; just because a given promoter is readily available to the experimenter it does not necessarily follow that said promoter will be optimal in the context of interest.

Additionally, the inherent characteristics of natural promoter sequences render their use in synthetic biology applications potentially problematic. Natural promoter activity is often context-specific (Blazeck & Alper, 2013) and subject to

interaction with a multitude of regulatory proteins, complicating prediction of activity levels under varying conditions (Collado-Vides *et al.*, 1991). As a result of these inherent limitations, researchers have increasingly turned to libraries of Synthetic Promoter Libraries (SPLs) to meet their needs. The most commonly employed methods for SPL generation are discussed below, and summarised in Figure 1.5.

1.7.1 Saturation Mutagenesis of Flanking Regions

A key method of forming SPLs is based on the observation that the flanking regions surrounding consensus motifs within the promoter sequence have a role in determining activity (Jensen & Hammer, 1998a). Degenerate oligonucleotides allow known consensus motifs to be maintained while the flanking regions are mutagenised, leading to altered promoter activity. Promoter function is maintained in the synthetic sequences due to the preservation of the key consensus motifs, with altered expression levels likely being the result of minor changes in DNA confirmation within the randomised flanks (Jensen & Hammer, 1998a). For example, saturation mutagenesis of flanking regions (SMFR) was successfully used to produce a SPL with a 400-fold activity range in *Lactococcus lactis*, although some of the diversity in promoter activity levels was a result of synthesis errors in the consensus sequences and alteration to flank length (Jensen & Hammer, 1998a, 1998b).

However, the initial approach taken to saturation mutagenesis by Jensen and Hammer does not take into account the context-dependant nature of promoter activity (Jensen & Hammer, 1998a, 1998b). Consequently, current SPL generation uses a single PCR stage, with degenerate oligonucleotides coupled to either a full-length or truncated version of the gene that the promoter is intended to drive. This improvement allows for ectopic analysis or replacement of a wild-type promoter with a synthetic alternative, while maintaining the 5' mRNA of the target gene (Solem & Jensen, 2002, Hammer *et al.*, 2006).

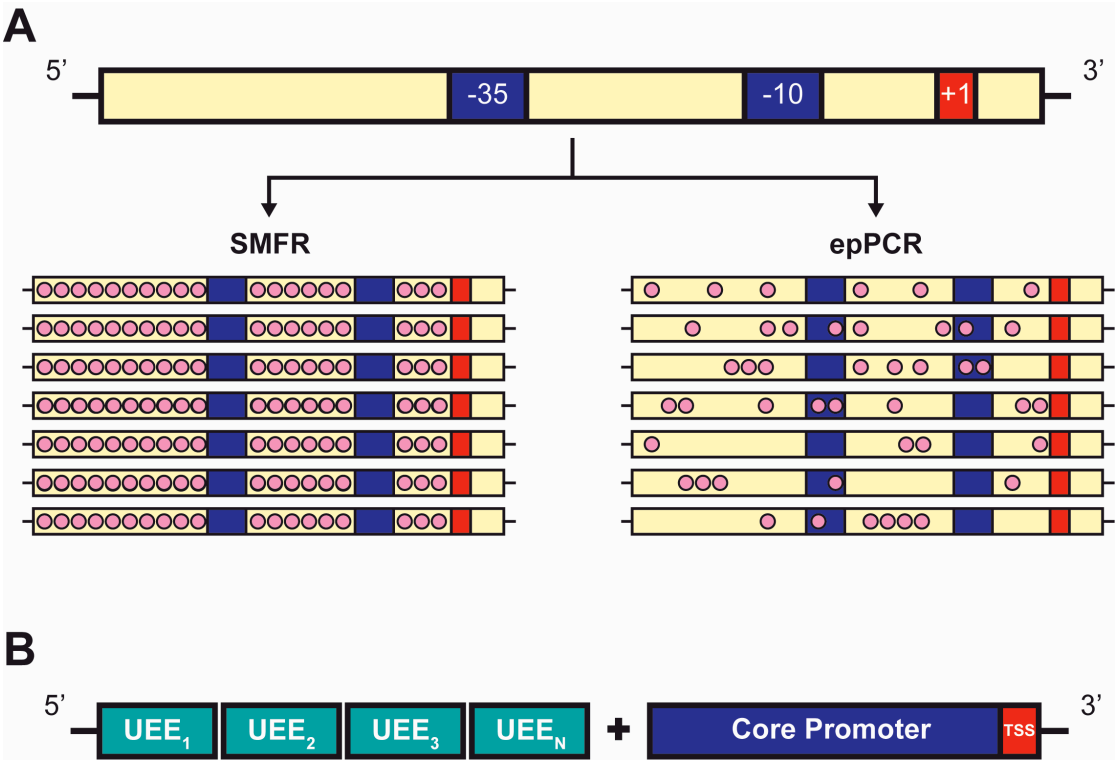


Figure 1.5: Summary of commonly employed molecular approaches for the production of synthetic promoter libraries in A) Prokaryotes and B) Eukaryotes.

A) Schematic representation of a typical prokaryotic promoter sequence shown, with consensus sequences highlighted in dark blue, and the Transcription Start Site in red. Saturation Mutagenesis of Flanking Regions (SMFR) employs degenerate oligonucleotides for the production of SPLs, wherein the consensus regions are maintained whilst the flanking regions are mutated (mutations are represented by pink circles). In contrast, error prone PCR (epPCR) results in the random introduction of mutations across the entire promoter sequence, including the consensus regions.

B) Schematic representation of a typical eukaryotic synthetic promoter. A core sequence (shown in dark blue) with known promoter activity is combined with multiple Upstream Enhancer Elements (UEEs) to achieve the desired strength. Figure adapted from Blazeck & Alper (2013).

SMFR has been successfully applied in a variety of prokaryotes and eukaryotes, including *Cornebacterium glutamicum* (Rytter *et al.,* 2014) and *Streptomyces coelicolor* (Sohoni *et al.*, 2014), yielding robust libraries with broad expression profiles. The methodology is also applicable to *S. cerevisiae,* wherein screening of an initial large library of colonies ultimately yielded 20 characterised promoters, displaying expression levels of yeast-enhanced green

fluorescent protein (GFP) that varied by approximately 22-fold (Ellis *et al.*, 2009).

In a separate study, a selection of constitutive promoters was initially isolated from the *S. cerevisiae* genome, and expression levels were subsequently characterised using expression profiles available from public databases. The promoter of the gene *PFY1* was chosen as a starting-point for its robust expression profile (Blount *et al.*, 2012). Knowledge of *PFY1* structure enabled identification of a rDNA enhancer-binding protein and a ploy-dT that were important for transcription initiation (Angermayr *et al.*, 2003). These regions were therefore held constant whilst a 48 base pair (bp) section of the promoter core was randomised, providing a library of 36 promoter elements with a broad range of expression levels.

It must be noted that none of the new sequences provided higher expression levels than the original *PFY1* promoter (Blount *et al.*, 2012). This inability to produce a synthetic promoter with higher expression levels than a natural alternative was also reported in *Francisella novicida* (McWhinnie & Nano, 2013).

A SMFR-type approach has also been successfully employed in the generation of a collection of synthetic promoter sequences with potential cross-genus applicability (Yang *et al.*, 2017c). The starting point was the strong synthetic minimal promoter P_{min}, from *S. cerevisiae.* Given that the -35 and -10 consensus sequences of *E. coli* and *B. subtilis* are identical, P_{min} required only four insertion mutations and one substitution to encode consensus regions for both species and a 17 bp spacer. The resulting promoter displayed comparable activity to that of P_{min} when characterised in *S. cerevisiae*, and had activity levels comparable to existing “strong” promoters in both *E. coli* and *B. subtilis.* Subsequent mutation of the flanking regions and library screening yielded three characterised promoter sequences, with a range of expression levels covering approximately an order of magnitude (Yang *et al.*, 2017). Crucially, all promoters maintained their relative activity levels across each of the three

species of interest, opening up possibilities for potential cross-host pathway optimisation.

SMFR has also been applied in *Geobacillus,* with the “strong” promoter from the *Geobacillus* sp. GHH01 *groESL* operon being mutagenised (Pogrebnyakov *et al.*, 2017). The region of the promoter responsible for temperature dependent expression of the operon was removed, and the flanking regions randomised, with selected bases immediately up and downstream of the consensus motifs being limited to Adenine or Thymine residues only. The *groESL* RBS was left unaltered. Promoter elements were cloned upstream of superfolder GFP (sfGFP) and characterised in *G. thermoglucosidans* strain C56-YS93, resulting in a library of 17 sequences with a reported 76-fold activity range.

A separate study also used the *Geobacillus* sp. GHH01 *groESL* promoter as the starting sequence for the generation of an SPL by SMFR (Jensen *et al.*, 2017). In this instance, the SPL was used for the expression of *G. stearothermophilus* β-galactosidase *BgaB* in *G. thermoglucosidans*. A five-fold variation in activity levels was observed across 28 characterised variants.

Although SMFR has successfully provided many new promoters, the technique requires labour intensive cloning and an *a priori* knowledge of promoter structure in the organism of interest, something that may not be immediately available in industrially relevant microbes. Furthermore, as many libraries use composite promoter scaffolds as a starting point, establishing a definitive wild-type reference expression baseline is impossible. Definitively stating whether SMFR will improve wild-type expression capability *pre hoc*, is therefore problematic (Blazeck & Alper, 2013). Additionally, by restricting mutagenesis to only the flanking regions, SMFR fails to take into account alterations to consensus sequences, which are known to play a significant role in modulating expression strength.

1.7.2 Error-prone PCR

Generating a SPL by applying error-prone PCR (epPCR) to an entire promoter sequence obviates any *a priori* knowledge of functional motif location and can potentially result in promoters with entirely new characteristics (Blazeck & Alper, 2013).

For example, the epPCR methodology was successfully used to mutagenise a bacteriophage P_L-λ promoter. Promoter mutants were cloned upstream of a *GFP* coding sequence and transformed into *E. coli*, resulting in a library containing approximately 9,000-12,000 functional clones (Alper *et al.*, 2005, Fischer *et al.*, 2006). Visual analysis of the transformants was initially used to identify 200 clones that covered "a wide range of fluorescence intensity". Subsequent cytometric analysis resulted in the identification of 27 clones, representing 22 discrete promoter sequences, which gave homogenous expression levels. Thorough characterisation of these 22 promoters resulted in a promoter library which was successfully employed to modulate levels of phosphoenolpyruvate carboxylase and lycopene production in *E. coli* (Alper *et al.*, 2005). The technique has also been successfully applied in yeast (Nevoigt *et al.*, 2006).

epPCR for promoter production has also been employed in *Geobacillus* (Reeve *et al.*, 2016). Here, the "strong", 245 bp, constitutive wild-type promoter pRpls from the *G. thermoglucosidans* genome was mutagenised, and amplicons were subsequently combined with the wild-type RBS sequence from the *G. thermoglucosidans PheB* gene. The complete *cis*-regulatory elements were characterised using sfGFP as a reporter. Screening of 100 colonies resulted in a characterised library of 20 promoter sequences, with a range of expression strengths of approximately 100-fold and a single promoter sequence of greater strength than wild-type pRpls. Further modulation of gene expression was subsequently achieved by replacing the wild-type RBS sequence with alternatives designed using the RBS calculator software tool developed by the Salis lab (Salis *et al.*, 2009).

However, although corrected for media autofluorescence, the data as reported by Reeve *et al.* lacked a true negative control such as would have been provided by transformants containing an empty vector. It is therefore difficult to ascertain whether several of the weaker members of the promoter library are truly active.

Additionally, although the rate at which mutations were incorporated into the promoter sequence was reported as approximately 10% (Reeve *et al.*, 2016), further analysis of the published promoter library revealed that a number of promoters were mutated at a much lower rate. For example, three promoter sequences displayed significantly different activity levels to the wild-type pRpls promoter, despite differing from the wild type by only one base pair at the sequence level. This result highlighted the complexity of the promoter design space, where relatively minor modifications at the sequence level can have significant impacts on the overall output of a system. High levels of sequence homology also increase the risk of homologous recombination between the mutated promoter sequence and the wild-type original, which remains within the *G. thermoglucosidans* genome (Pogrebnyakov *et al.*, 2017).

Despite these successful examples, the epPCR approach to SPL production has certain limitations: A reliance on a selection of a small subset of colonies for further analysis (Alper *et al.*, 2005, Yim *et al.*, 2013) renders discovery of a true optimum problematic. Moreover, the extensive screening required to isolate said subset should not be underestimated; it is typical for initial libraries of hundreds or thousands of bacterial colonies to ultimately yield relatively few fully characterised promoters. Both these problems become less of an issue if visual selection of colonies is replaced by high-throughput analytical techniques such as fluorescent assisted cell sorting and/or imaging cytometry, but they remain a non-trivial consideration.

1.7.3 Hybrid promoter engineering

In addition to the two mutagenic techniques discussed above, the generation of synthetic promoters through hybridisation of existing promoter

elements provides an alternative strategy for promoter genesis. More commonly employed in eukaryotes, hybrid promoter engineering combines minimal core promoter elements with various combinations of modular upstream activating sequences (UAS).

Blazeck *et al.* demonstrated the applicability of this approach to promoter design in *S. cerevisiae*, by showing that the addition of UAS to a core promoter resulted in increased expression levels compared to a wild-type baseline (Blazeck *et al.*, 2012). A roughly linear relationship was observed between the number of UAS modules added and promoter strength, with the addition of four such elements boosting expression of a weak constitutive promoter to levels comparable with the strongest endogenous promoter (Blazeck *et al.*, 2012). The magnitude of transcriptional increase was shown to depend both on the core element and UAS, but all core promoters were amenable to improvement (Blazeck *et al.*, 2012).

1.8 Computational methods for promoter discovery and design

In addition to the molecular methods for promoter discovery and SPL production discussed above, modern *in silico* techniques have broadened the promoter discovery pipeline. The proliferation of high quality “omics” data sets, for example, has served to expedite the discovery of endogenous promoters in a range of species, whilst mathematical abstraction of promoter sequences in a number of model organisms has expanded the promoter design space to allow for *de novo* design of synthetic promoter elements.

1.8.1 *In silico, high-throughput discovery of endogenous promoters*

Modern bioinformatics based approaches have rendered the discovery of large numbers of putative transcription control elements relatively trivial. BPROM software, for example, identifies putative promoter sequences regulated by *E. coli* sigma factor σ^{70}, based upon the presence and nucleotide

composition of functional motifs, with a claimed recognition accuracy of 80% (Solovyev & Salamov, 2011).

Although motif-based predictive models have proved successful for a limited number of model organisms, they may not always prove immediately applicable in non-model organisms, where understanding of species-specific regulatory motifs may be minimal (Umarov & Solovyev, 2017). Machine learning algorithms may present an alternative approach in such instances. Convolutional Neural Networks (CNN), for example, have been used to infer promoter activity in both prokaryotic and eukaryotic species at the sequence level (Umarov & Solovyev, 2017), thus obviating the need for detailed knowledge of motif structure and location. The resultant models were subsequently applied to classification of promoter and non-promoter sequences. Species-specific classification accuracy was observed, with error rates broadly comparable to those reported by BPROM (Solovyev & Salamov, 2011).

The proliferation of high-quality “omics” data sets has also served to broaden the promoter discovery pipeline. In *Streptomyces coelicolor*, for example, screening of transcriptome microarray data for genes whose expression profile remained constant under multiple growth conditions resulted in the identification of 166 putative global promoter elements in a species where the promoter toolbox was previously lacking (Li *et al.*, 2015). The same approach has been applied in *S. albus*, where 32 candidate promoters were identified, ten of which exhibited stronger activity than the strongest previously available *Streptomyces* promoters when characterised upstream of the reporter gene *xylE* (Luo *et al.*, 2015).

Transcriptome analysis has also been successfully applied for the identification of promoter elements in *E. coli* (Mendoza-Vargas *et al.*, 2009), where RNA-seq data aligned against the genome permitted robust identification of transcription start sites. Using the observation that the distance between promoter sequences and transcription start sites is highly conserved, this process permitted the identification of approximately 1,500 genome regions

likely to contain promoter sequences. These regions were subsequently analysed using a Position Weight Matrix (PWM) for the identification of probable promoter motifs and to determine which *E. coli* sigma factor was most likely to be responsible for their regulation. In total, approximately 800 putative promoter elements were reported (Mendoza-Vargas *et al.*, 2009).

Although the *in silico* approaches described above have certainly provided new promoters, they do not represent a systematic, theoretical examination of the promoter design space. If, for arguments sake, a promoter sequence is 100 bp in length, there are 4^{100} potential promoter sequences. Therefore, although the best sequence discovered by screening of natural promoter sequences may be sufficient for some experimental purposes, it is possible that other optima are present in a region of promoter design space which nature has not explored.

Mathematical models that are capable of deciphering the effect of individual DNA bases and motifs, or predicting promoter activity level in advance of *in vivo* characterisation have, in this context, considerable potential (Jensen *et al.*, 2006). Conventionally, the use of computational techniques in pathway design and optimisation has been limited to *post hoc* data analytics (Ellis *et al.*, 2009). However, computational modelling for the design and optimisation of biological systems is becoming more widespread, and a number of computational methodologies are available to facilitate the *de novo* design of synthetic *cis*-regulatory sequences. Given that the strength of the *cis*-regulatory sequence is one of the key determinants of system output, the potential ability of mathematical approaches to accurately determine promoter activity has broad reaching implications for the field of synthetic biology. The application of computational modelling to promoter design can potentially enhance and accelerate the design phase of the synthetic biology design-build-test cycle, and ultimately enhance our fundamental knowledge of genetic regulation in complex systems.

1.8.2 Position weight matrix models

Position weight matrix (PWM) models have been widely applied for the detection of transcription factor binding sites (Stromo, 2000, Sinha, 2006), and have also shown some promise in prediction promoter strength. By breaking promoter sequences down into constitutive motifs, PWM models were able to predict the strength of *E. coli* core promoter sequences recognised by sigma factor σ^E to a relatively high degree of accuracy (Rhodius & Mutalik, 2010). The core promoter PWM was subsequently combined with a score describing the activity of upstream elements to provide a model capable of predicting the strength of entire promoter sequences (Rhodius *et al.*, 2012). In addition to this predictive power, PWM models can potentially provide increased understanding of promoter structure, something that is often limited in novel microbial chassis.

Although PWM models have the potential to be applied to *de novo* sequence design, and their application for the *pre hoc* determination of strength in certain promoter families should not be overlooked, they are not without limitations. PWMs may prove inadequate for modelling in promoter families with a less conserved nature than those that interact with σ^E, as poorly conserved sequences require greater complexity within the model (Rhodius & Mutalik, 2010). Additionally, by assuming that the contribution of individual nucleotides to DNA-protein binding is independent and additive (Rhodius *et al.*, 2012), PWMs fail to account for the effect of interactions between positions.

Because of these potential limitations, the application of PWMs in novel microbial chassis, where knowledge of interactions between proteins and promoter sequences might be limited, is potentially challenging. PWMs should therefore be superseded in such situations by computational techniques that do not rely to such an extent on *a priori* knowledge of the system to which they are to be applied. Statistical models and machine learning approaches provide attractive alternatives in this instance, as they obviate the requirement for *a priori* understanding of the complex biological mechanisms that drive promoter activity (Bedbrook *et al.*, 2017).

1.8.3 Partial Least Squares regression

Partial Least Squares regression (PLS, alternatively referred to as Projection to Latent Structures) is a family of statistical methods that combine dimensionality reduction and linear regression to infer the relationship between multiple variables (Rosipal & Krämer, 2006). PLS has been successfully applied for modelling in a broad range of fields, including economics, genomics, political science and spectroscopy (SAS Institute Inc., 2016a). PLS models cope well with high dimensionality, and are particularly suited to dealing with "squat" data sets, where the predictors outnumber the observations. The ability to deal with high levels of correlation between predictors, known as multicollinearity, is also a feature of PLS algorithms. Experimental practicalities mean that the number of characterised *cis*-regulatory sequences that are available for model construction is unlikely to outnumber the number of nucleotides in a promoter sequence. Furthermore, the presence of conserved motifs within promoter sequences makes high levels of multicolinearity likely. The inherent characteristics of PLS therefore make it an attractive choice for the modelling of promoters.

The use of PLS modelling to quantitatively link DNA sequence to function is not a new concept (Jonsson *et al.*, 1993), although as a method for the generation of synthetic promoters it remains underutilised. In a pioneering study, 25 *E. coli* promoters were analysed using a Partial Least Squares (PLS) methodology, resulting in a statistical model that inferred the contribution of each individual nucleotide at any given position in the DNA sequence. In order to test the model, two synthetic sequences with high predicted activity levels were synthesised. The -35, -10 and +1 motifs were determined using the consensus sequence of the training set of 25 promoters, whilst the remainder of the synthetic sequences were determined using regression coefficients provided by the modelling process (Jonsson *et al.*, 1993).

In vivo characterisation of the synthetic promoters revealed activity levels within approximately 8% of the strength predicted by the model. Furthermore, the synthetic sequences were shown to provide higher expression

levels than any of those sequences found within the training set (Jonsson *et al.*, 1993). The application of PLS models for the design of synthetic promoter sequences can therefore be considered, in this respect, superior to molecular approaches to promoter production, where the activity levels of mutagenised sequences are rarely greater than the wild-type starting point.

Similar statistical methods were later applied to quantitatively link promoter DNA sequence and function for a library of synthetic *E. coli* promoters that were generated through SMFR (De Mey *et al.*, 2007). The generated PLS model was able to predict, with reasonable accuracy, the strength of promoter sequences that had not been used in the construction of the model (De Mey *et al.*, 2007).

In further validation of this computational technique, the predictive model was subsequently utilised to predict the strength of an endogenous *E. coli* promoter, that of the *ppc* gene (De Mey *et al.*, 2010). Based on this information, stronger promoters were selected from the previously characterised promoter library to fine tune *ppc* expression levels (De Mey *et al.*, 2007). This knock-in approach resulted in an increase in expression levels roughly in line with the model's predictions, with a three- to four-fold increase in mRNA levels seen at flask scale (De Mey *et al.*, 2010). Although the PLS regression doubtlessly aided in the optimisation process, it was not applied, in this instance to the *de novo* design of synthetic promoter sequences.

1.8.4 Artificial Neural Networks

The linear nature of PLS modelling is a potential drawback when applied to the analysis of promoter sequences, as the effects of any interactions between nucleotides may be confounded with the main effects for each individual nucleotide position (Jonsson *et al.*, 1993). PLS models therefore may not fully account for the complexity inherent in promoter structure, thereby increasing the probability of prediction errors and resulting in inadequate generality (Meng *et al.*, 2013). Indeed, many such models lack robust prediction accuracy (Meng & Wang, 2015), rendering their use in *de novo* sequence

design challenging.

Artificial Neural Networks (ANN) may provide a solution to these issues. Based upon a network of interconnected nodes designed to act as a rudimentary mimic of the brain, ANNs permit machine learning, as the order and force of connections between individual nodes may be altered (Buscema *et al.*, 2014).

By systematically altering node structure during the analysis of a training data-set, ANN models can potentially better represent the complex, non-linear interactions that occur within a promoter sequence than linear PLS models (Meng *et al.*, 2013). ANN modelling has previously proven successful for *de novo* promoter design (Meng *et al.*, 2013); using a set of synthetic promoters derived from the random mutagenesis of a wild-type *E. coli* promoter as a training set for an ANN model, strength predictions of sequences generated by *in silico* mutagenesis were used to select 16 synthetic sequences for *in vivo* characterisation (Meng *et al.*, 2013).

The expression levels predicted by the ANN displayed good correlation with the empirically measured *in vivo* activity of the 16 synthetic sequences, suggesting that such models are indeed applicable to synthetic promoter design.

Although ANNs have potential for predicting promoter activity and the *de novo* design of synthetic promoter sequences, their use in this context is not without limitations. The inherent mathematical complexity of such models renders ANNs something of a “black-box”, wherein promoter sequence information is transformed to a prediction of strength. The use of ANNs as explanatory models to improve understanding of promoter structure is therefore impossible. This is in direct contrast to linear regression techniques such as PLS, where the model coefficients provide a readily interpretable measure of the importance of a given nucleotide or motif to overall promoter strength. Careful consideration must therefore be given to the ultimate aim of the modelling process; a combination of the predictive power of ANNs with the

more readily interpretable output of PLS or PWM models may provide a useful trade-off between predictive and explanatory modelling power.

No matter which *in silico* approach is ultimately applied to *de novo* promoter design, the underlying principle remains, broadly speaking, the same. Models are trained on a known set of promoter sequences, using empirically gathered characterisation data to infer relationships between sequence and function. The accuracy of these inferences can subsequently be validated using a secondary group of sequences that were withheld from the training process. Once a model of sufficient predictive power has been obtained, it can then be applied to the *de novo* design of synthetic promoter sequences, or to the prediction of promoter strength.

1.9 The promoter discovery pipeline summarised

Industrial production systems that rely on biocatalysts require predictable outputs. As the complexity of engineered metabolic pathways increases, so too must the sophistication of the control systems which regulate them. For example, batteries of promoters that are orthogonal to endogenous metabolism and that provide predictable expression levels provide a potential technique for tuning synthetic pathways.

A multitude of different approaches are available for the identification and *de novo* design of promoter elements in prokaryotes (Figure 1.6). Endogenous promoters may be individually isolated from upstream of well-understood genes or operons, or identified en masse if genomic or transcriptomic data of sufficient quality is available in the species of interest. Once identified, natural promoters may be subjected to random mutagenesis for the production of SPLs, or mathematical abstraction of the promoter sequence can be used to model promoter strength or design *de novo* synthetic promoter elements with defined functionality.

The data that are obtained from the derivation of promoter libraries can potentially enhance our fundamental knowledge of genetic regulation in complex systems. For example, the effects of mutations in specific locations within the promoter sequence can be assessed, or mathematical models that infer the contribution to promoter activity of a given nucleotide at a given position could be use to gain insights into the interaction between DNA and protein. However, the extent to which such opportunities are exploited will depend on the objectives of specific projects. Stated alternatively, projects with an application-based objective might be more focused on “the recording and reporting of measurements and not on deeper mechanistic understanding” (Mutalik *et al.*, 2013b).

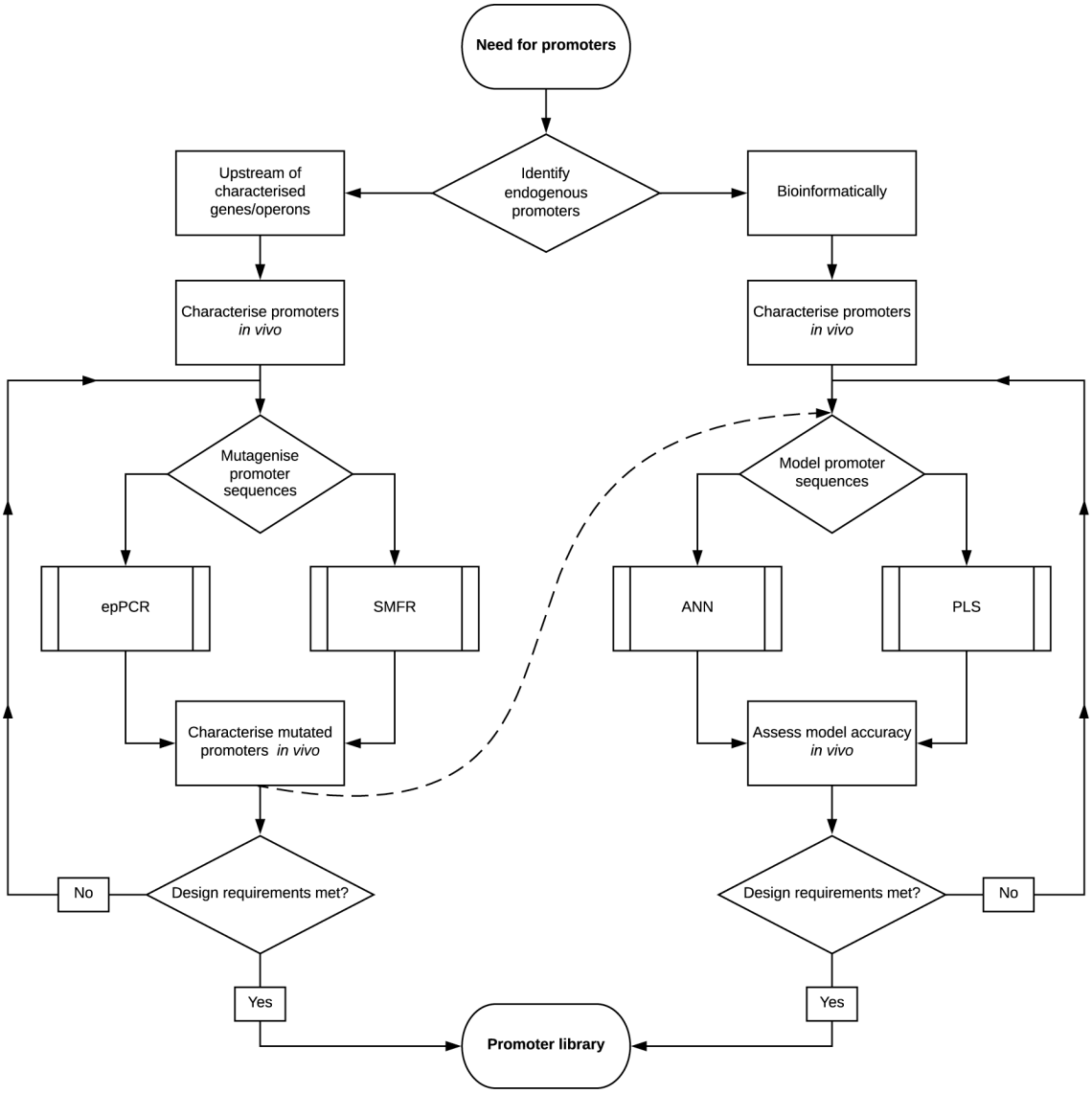


Figure 1.6: Workflow for promoter discovery in prokaryotes.

The discovery and design of prokaryotic promoters for synthetic biology applications typically begins with the isolation of endogenous regulatory sequences from the genus or species of interest. Putative regulatory sequences may either be individually isolated from upstream of previously characterised genes or operons, or bioinformatically identified *en masse*. The most commonly employed strategies for the design of synthetic promoters are mutagenesis-based (left-hand workflow). Thoroughly characterised, well-understood promoter sequences can be mutagenised using techniques like Error-prone PCR (epPCR) or Saturation Mutagenesis of Flanking Regions (SMFR). The resulting libraries of mutated sequences should subsequently be screened for promoter activity *in vivo.* Alternatively, the *in vivo* characterisation of large numbers of bioinformatically-identified putative promoters can be used to derive quantitative mathematical or statistical models of the relationship between promoter DNA sequence and function (right-hand workflow). Artificial Neural Networks (ANN) and Partial Least Squares (PLS) regression have both been employed by prior studies to model promoter activity in *E. coli*. Regardless of which modeling type is used, the accuracy of predictions of promoter activity should be assessed *in vivo*. Data derived from the *in vivo* characterisation of mutagenised promoter sequences can also potentially be used to train models of promoter function (dashed line). Multiple iterations of mutagenesis or mathematical modeling, with the associated *in vivo* characterisation, may be required to obtain a library of promoter sequences that meet the requirements of a given study. Flow chart rendered with LucidChart.

1.10 Hypothesis & Project aims

It is not clear which of the multitude of available approaches to promoter discovery and design is most appropriate in an industrial context. A direct comparison is necessary to identify which method represents the most practical approach to expanding the synthetic biology toolkit in non-model organisms. In particular, computational approaches to *de novo* promoter design remain underutilised aside from proof-of-principle studies in model organisms; **we hypothesise that statistical learning approaches to promoter discovery are applicable to non-model organisms, and will accelerate the discovery and characterisation of promoter libraries in *Geobacillus.***

To that end, the specific aims of this investigation are;

1. To rationally bioinformatically isolate constitutive *cis*-regulatory elements from the *Geobacillus* core genome (Chapter 3).

2. To thoroughly characterise the *in vivo* activity of these regulatory elements in a range of genetic and environmental contexts (Chapters 3, 4 & 5).

3. To use these characterisation data to develop robust statistical models that quantitatively link DNA sequence to promoter activity, and to subsequently assess the capability of these models to make accurate predictions of *in vivo* activity for synthetic and endogenous promoters (Chapters 3 & 4).

4. To use two commonly employed mutagenesis-based approaches, Saturation Mutagenesis of Flanking Regions and Error-prone PCR, to derive libraries of synthetic *Geobacillus* promoters (Chapter 6).

5. To identify which of the above methods is the most generally applicable in industrially relevant, non-model organisms (Chapter 7).

2 Materials & Methods

2.1 Materials

2.1.1 Media

Unless otherwise stated, all media were purchased as ready-made stocks from Becton Dickinson UK Limited (Berkshire, United Kingdom). All media were sterilised by autoclaving at 121 °C for 20 min before use.

Lysogeny Broth (LB) contained 10 g l^{-1} tryptone, 10 g l^{-1} NaCl and 5 g l^{-1} yeast extract. Lennox Lysogeny Broth (LLB) contained 10 g l^{-1} tryptone, 5 g l^{-1} yeast extract and 5 g l^{-1} NaCl.

For *Geobacillus* growth, modified Lysogeny Broth (mLB) used a basal 9of LLB. Once autoclaved, this was supplemented with 1.05 mM $C_6H_9NO_6$, 0.91 mM $CaCl_2$, 0.59 mM $MgSO_4$ and 0.04 mM $FeSO_4$ (Zeigler, 2001).

For all media types, agar was supplemented as required to 15 g l^{-1}. When required, ampicillin was added to a final concentration of 100 µg ml^{-1}, and kanamycin to a final concentration of 12.5 µg ml^{-1} for transgenic *Geobacillus* culture, and 50 µg ml^{-1} for transgenic *E. coli* culture.

2.1.2 Chemicals

Chemicals were purchased from Fisher Scientific UK Ltd (Loughborough, United Kingdom) and Sigma-Aldrich Company Ltd (Dorset, United Kingdom), unless otherwise stated.

2.2 Bioinformatic methods

2.2.1 Hardware

Bioinformatic analysis was performed using a local server containing 32 3.1 GHz CPUs and 256 GB RAM. The system was installed with Fedora version 2.1 Linux operating system.

2.2.2 Identification of putative cis-regulatory sequences from the Geobacillus core genome

The *Geobacillus* strains listed in Table 2-1 were sequenced *de novo* and genomes assembled by the Exeter Microbial Biofuels group prior to the start of the project.

	Size (bp)	Number of features
G. kaustophilus DSM7263	3,517,923	3,528
G. stearothermophilus DSM22	2,821,937	2,976
G. thermodenitrificans K1041	3,548,326	3,526
G. thermoglucosidans DSM2542	3,961,895	3,886

Table 2-1: *Geobacillus* species used in the prediction of a core *Geobacillus* genome.

The core genome of the four *Geobacillus* species was determined using the GET_HOMOLOGUES software package (Contreras-Moreira & Vinuesa, 2013). Three clustering algorithms were used to cluster homologous gene families: Bidirectional best blast hit (BDBH), COGtriangles (COG) and OrthoMCL (OMCL). In all cases, the default software parameters were used.

To isolate only those clusters which contained single-copy proteins present in all four *Geobacillus* species, the –t option was used. Single-copy

protein coding sequences were isolated as they were likely safer orthologues (Contreras-Moreira & Vinuesa, 2015).

Once isolated, single-copy protein coding sequences were extracted from the four genomes. Output files were parsed, reformatted to fasta format and imported into the Artemis programme (Rutherford *et al.*, 2000). For each entry, the 100 bp immediately upstream of the start codon was extracted. 100 bp sequences were isolated for analysis as the majority of elements that are known to affect transcription initiation in prokaryotes occur within 100 bp of the start codon (Mendoza-Vargas *et al.,* 2009, Davis *et al.,* 2011).

To identify putative promoter elements, extracted sequences were analysed using BPROM bacterial promoter prediction software (Solovyev & Salamov, 2011). To isolate promoter sequences that were likely to be orthogonal to endogenous regulatory pathways, those sequences containing known transcription factor binding sties (TFBS) were discarded.

Once identified, nucleotide sequences of all putative promoters were aligned using MUSCLE software (Edgar, 2004), with resultant alignments used to build a phylogenetic tree using FastTree software (Price *et al.*, 2009). Putative promoter sequences were subsequently manually clustered into 21 clades using FigTree (Rambaut, 2017).

Putative promoter sequences were selected for synthesis at random from each clade. True randomness was achieved by using a random number generator that converted atmospheric noise to numerical values (Haahr & Haahr, 1998). In early iterations, selected sequences were manually validated using Artemis to ensure that they did not overlap with any adjacent coding sequences (CDS). Later, to expedite this process, BEDTools intersect (Quinlan & Hall, 2010) was used to identify those putative promoter sequences which were non-overlapping.

Putative promoter sequences were aligned to transcripts of each of the four *Geobacillus* species using Bowtie 2 (Langmead & Salzberg, 2012). Indexes

of the genome files were prepared using the 'build' command. Putative promoters were then aligned to each genome using Bowtie 2, with the resultant alignments provided in Sequence Alignment Map (.sam) format. The alignment .sam files produced were converted to Binary Alignment Map (.bam) format, sorted and indexed using SAMtools (Li *et al.*, 2009).

The resultant alignments were compared against the four *Geobacillus* genomes using BEDTools intersect. The '–v' command was used to report only those putative promoter sequences which were non-overlapping with any annotated features in the genome transcripts. Output files were provided in .bam format, and were subsequently converted to fasta format using bam2fastx (Kim *et al.*, 2013).

2.2.3 Identification of putative cis-regulatory sequences from bacteriophage

The genomes of two bacteriophage, Thermus phage phi OH2 and *Geobacillus* phage GBSV1 (Liu *et al.*, 2009) were selected for analysis based on their ready availability from GenBank (National Centre for Biotechnology Information, https://www.ncbi.nlm.nih.gov/genbank/). The Artemis programme was used to identify intergenic regions of at least 100 bp length. The 100 bp nucleotide sequences immediately upstream of the adjacent CDS were subsequently extracted and analysed using BPROM to identify putative promoter elements.

2.3 General molecular genetic methods

2.3.1 Microbial strains

Geobacillus thermoglucosidans (type strain, DSM2542) was obtained from the DSMZ (Brunswick, Germany). Cultures were freeze-dried ampoules, and rehydrated as required following the DSMZ standard protocol.

NEB 5-alpha (New England Biolabs, Massachusetts, United States of America) chemically competent *Escherichia coli* strain (genotype: *fhuA2 D(argF-lacZ)U169 phoA glnV44 f80D(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*) was used for microbiological cloning, storage and amplification of plasmid vectors.

E. coli S17-1 was a gift from ZuvaSyntha Ltd. (Hertfordshire, United Kingdom) and was used as the mobilisation host for conjugal transformation of *Geobacillus* (genotype: *recA pro hsdRm RP4-Tc::Mu-Km::Tn7*). Transfer genes from the RP4 plasmid are integrated into the genome of *E. coli* S17-1, allowing for the conjugal transfer of plasmids containing the requisite mobilisation elements (Simon *et al.*, 1983).

2.3.2 Preparation of chemically competent Escherichia coli

E. coli NEB 5-alpha and S17-1 were made chemically competent using a modified version of the protocol described by Hanahan (Hanahan, 1985). Single *E. coli* colonies of the relevant strain were used to inoculate 5 ml LB broth and incubated at 37 °C for 16 h, with shaking at 220 rpm. 400 µl of the resulting stationary phase culture was used to inoculate 40 ml LB broth. Cultures were incubated at 37 °C, with shaking at 220 rpm, until an OD_{600} of 0.4-0.5 was reached.

Cells were harvested by centrifugation at 4,500 *g* for 8 min at 4 °C, and the supernatant discarded. The resultant pellet was re-suspended in 8 ml Transformation buffer 1 (TF1)**,** and incubated on ice for 15 min. (TF1: 150 g l^{-1} Glycerol; 30 ml l^{-1} 1 M CH_3CO_2K pH 7.5; 0.1 M KCl; 0.01 M $CaCl_2.2H_2O$. Adjusted to pH 6.4 with CH_3COOH, autoclaved, then supplemented with 50 ml l^{-1} filter sterilised 1 M $MnCl_2.4H_2O$).

Cells were subsequently harvested as before, the supernatant was removed and the pellet re-suspended in 4 ml Transformation buffer 2 (TF2: 150 g l^{-1} Glycerol; 0.075 M $CaCl_2.2H_2O$; 0.01 M KCl. Autoclaved, then supplemented with 20 ml l^{-1} filter sterilised 0.5 M MOPS-KOH pH 6.8). 100 µl

aliquots of competent cells were flash frozen in liquid nitrogen and stored at -80 °C until required.

2.3.3 Escherichia coli transformation

100-200 ng plasmid DNA was added to chemically competent *E. coli* of the relevant strain. Samples were incubated on ice for 40 min, then heat shocked at 42 °C for 2 min and incubated on ice for a further 5 min. 700 µl pre-warmed LB broth was subsequently added and the resulting samples were incubated at 37 °C with shaking at 220 rpm for 60 min.

After incubation, samples were centrifuged at 4300 *g* for 5 min, and 500 µl of the supernatant was removed. The cell pellet was re-suspended in the remaining supernatant, and 200 µl of the culture was plated out onto LB agar plates, with antibiotic selection as required. Plates were incubated at 37 °C for 16 h.

2.3.4 Geobacillus transformation

Chemically competent *E. coli* S17-1 was transformed as described in section 2.3.3.

Approximately 5 µl of transformed *E. coli* S17-1 was collected from a confluent plate-culture using a microbiological loop, suspended in 600 µl LLB broth and centrifuged at 4300 *g* for 5 min. The supernatant was removed and the resultant pellet re-suspended in a further 600 µl LLB broth. Approximately 10-15 µl of wild-type *Geobacillus* was collected from a confluent plate-culture using a microbiological loop, added to the *E. coli* suspension and re-suspended. The resulting bacterial mix was dispensed onto LLB agar plates, in drops of approximately 10 µl.

LLB plates were incubated at 37 °C for 7 h, followed by incubation at 60 °C for 1 h. The resulting biomass was re-suspended in 1 ml LLB broth, and used to create dilutions of 1:10 and 1:5 biomass to sterile LLB broth, 200 µl

aliquots of which were spread onto mLB agar plates containing 12.5 μg ml^{-1} kanamycin. Plates were incubated at 55 °C for approximately 65 h.

2.3.5 Culture & plasmid maintenance & storage

NEB 5-alpha *E. coli*, made chemically competent as described in 2.3.2, was used for microbiological storage and amplification of plasmid vectors. Single colonies were used to inoculate 5 ml LB broth, with antibiotic selection as required, and incubated at 37 °C, in the case of *E. coli* and 60 °C in the case of *Geobacillus.* All cultures were incubated with shaking at 220 rpm.

For long-term microbial storage, 500 μl stationary phase culture was added to 500 μl 50 % w/v sterile glycerol and thoroughly mixed by inversion. Samples were snap frozen in liquid nitrogen and stored at -80 °C.

2.3.6 Plasmid minipreps

Plasmid minipreps were performed from 5 ml overnight cultures. Extractions were performed using a GeneJet plasmid miniprep kit (Thermo Scientific, Loughborough, United Kingdom), according to the manufacturer's protocol. Eluted plasmid DNA was stored at -20 °C until required.

Eluted plasmid DNA was quantified using a Qubit™ Fluorometer (Thermo Scientific, Loughborough, United Kingdom) using the broad range double stranded DNA assay kit, according to the manufacturer's instructions.

2.3.7 Plasmid vectors

Plasmids were constructed, manipulated and visualised *in silico* using Clone Manager Professional edition version 9 (Scientific & Educational Software, Colorado, United States of America). All characterised putative promoter constructs used the pS797 backbone (Figure 2.1). A gift from ZuvaSyntha Ltd, (Hertfordshire, United Kingdom), pS797 contained some of the

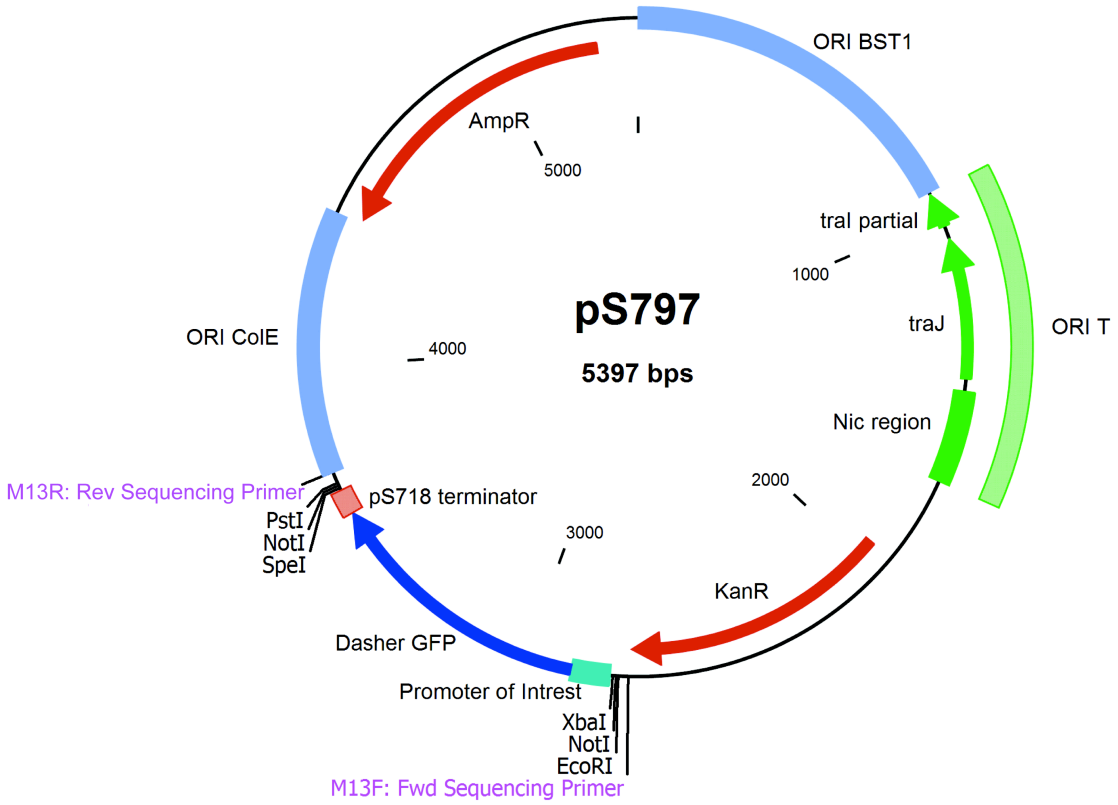


Figure 2.1: Plasmid map of pS797 expression vector used for characterisation of putative promoter elements.

The origin of transfer, ORI T, which contains the machinery necessary for mobilisation of the vector during conjugation, is shown in green. Antibiotic resistance genes are shown in red. Two origins of replication are present: ORI ColE for replication in *E. coli*, and ORI BST1, for replication in *Geobacillus.* Both are shown in blue. The binding sites of primers used for sequence verification are highlighted in purple.

The promoter of interest, shown in cyan, is located between multiple cloning sites (MCS) containing the listed restriction sequences. Also located between the MCS is the reporter protein (in this case Dasher *GFP*), highlighted in dark blue, and a terminator sequence, indicated by a red box.

required genes for the conjugal transformation of *Geobacillus*; the origin of transfer, ORI T, contains the Nic region and *traJ* gene from the conjugal plasmid RP4. The remaining genes required for conjugation were provided *trans* by *E. coli* S17-1.

pS797 also contained two origins of replication, ColE and BST1, to allow for replication in *E. coli* and *Geobacillus* respectively. Two antibiotic selection

markers were also present, allowing for selection by Ampicillin in *E. coli* and by Kanamycin in *Geobacillus.*

Unless otherwise stated, expression vectors were synthesised and the sequence verified by ATUM (previously DNA 2.0, California, United States of America).

2.3.8 "One pot" type IIS restriction cloning

The cloning methodology used for the production of expression vectors was adapted from Engler *et al.* (2008) and Kirchmaier *et al.* (2013). The methodology made use of the type II endonuclease BsaI, which cuts outside its DNA recognition site. The sequence between enzyme recognition and cleavage sites can be user defined, resulting in unique post-digestion overhangs. By matching the 3' overhang of one part to the 5' overhang of the part which was to be immediately downstream in the final construct, digested fragments were only able to ligate in a defined manner (Figure 2.2) (Engler *et al.*, 2008).

All enzymes used for cloning were purchased from Thermo Scientific (Loughborough, United Kingdom), from either the FastDigest or Anza brands, unless otherwise specified.

For DNA parts to be utilised in the cloning strategy, prefixes and suffixes containing the requisite restriction endonuclease recognition sites were added to the parts of interest (Figure 2.3). Bases highlighted in red varied depending on the DNA part to which the affix sequence was attached to permit ligation of parts in the specified order. Part-specific sequences are summarised in Table 2-2.

Cloning affixes were added *in silico* to the part sequence, and the composite DNA parts were synthesised by ATUM (previously DNA 2.0, California, United States of America). Parts were synthesised in the ATUM cloning vector pJ201, a high copy number plasmid encoding kanamycin resistance.

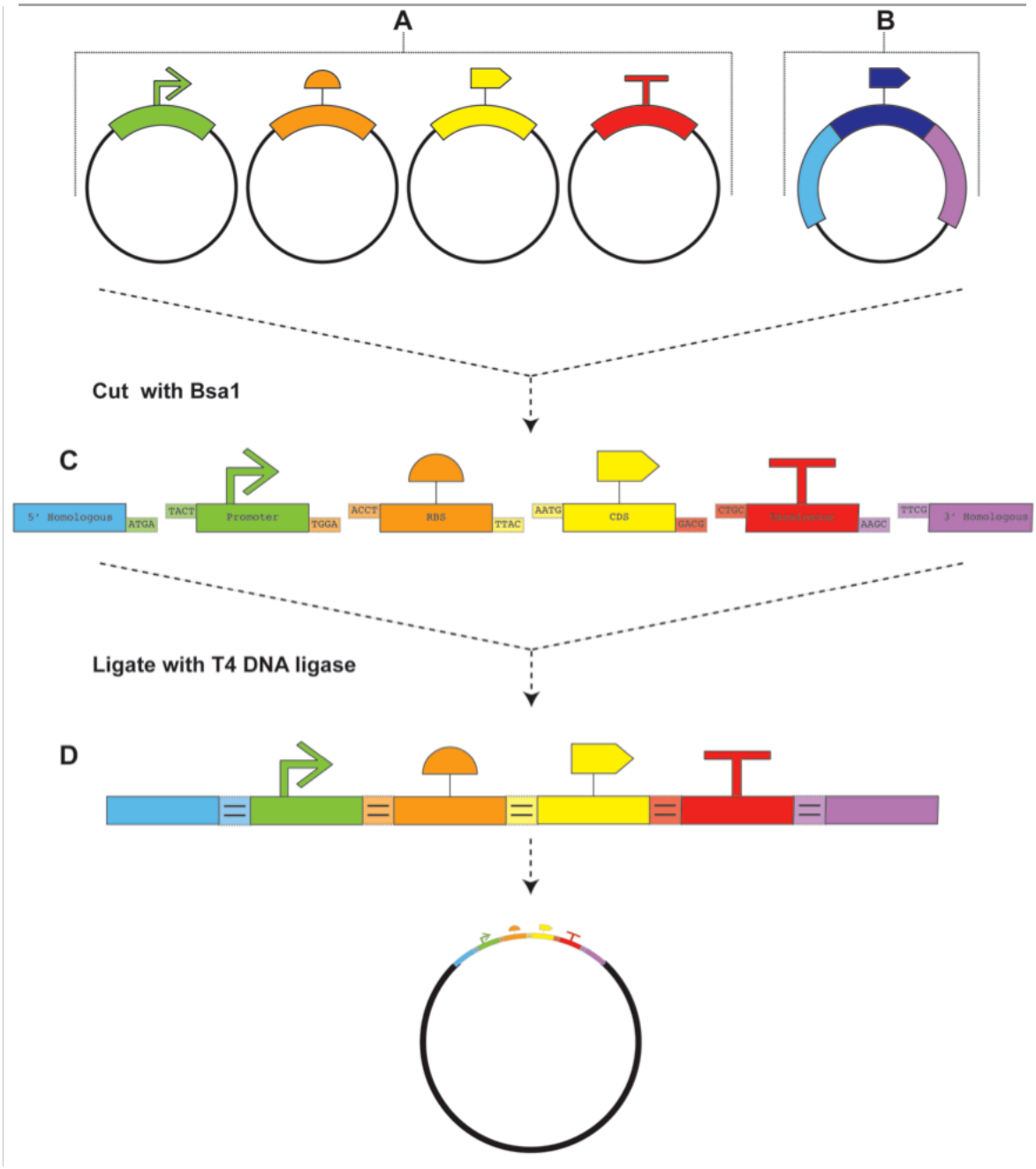


Figure 2.2: Schematic representation of one-pot cloning protocol.

DNA parts to be combined were flanked with unique five base pair sequences and BsaI restriction sites. Prior to cloning, parts were typically inserted into pEX1C3 entry vectors (A). Alternatively, parts were used in the form of linear double stranded DNA fragments produced by PCR or DNA synthesis. The destination vector (B) contained the gene encoding DNA toxin ccdB (dark blue), to act as a negative selection marker, flanked by BsaI sites and unique five base pair sequences. As a further selection mechanism, destination and entry vectors carried different antibiotic resistance genes.

Digestion with type II endonuclease BsaI resulted in parts with specific overhangs (C). The 5' overhang of each part was only compatible with the 3' overhang of the part immediately upstream, and vice versa. Ligation with T4 DNA ligase resulted in a completed plasmid (D), with 4 bp scar sequences between the parts.

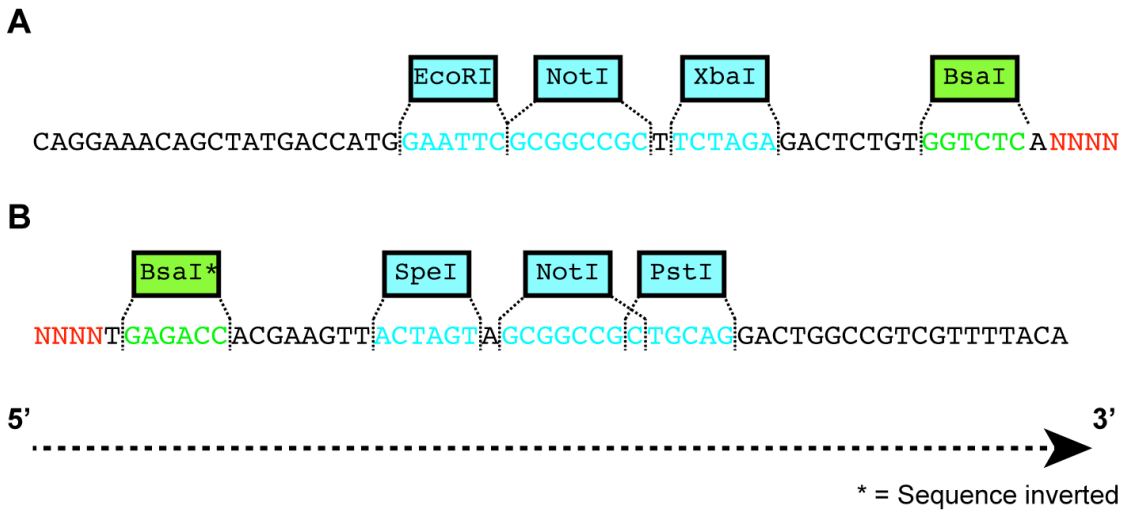


Figure 2.3: DNA sequences of A) prefixes and B) suffixes added to DNA parts for use in one-pot cloning.

BsaI restriction sites shown in green. Biobrick restriction sites are shown in blue, and allowed for the insertion of DNA parts into entry vectors. Sequences that allowed for the production of unique overhangs post digestion with BsaI are highlighted in red. Bases coloured black were present to act as spacers.

DNA Part	5' Overhang (5' → 3')	3' Overhang (5' → 3')
Promoter	TACT	TGGA
RBS	ACCT	TTAC
CDS	AATG	GACG
Terminator	CTGC	AAGC
Destination vector	TTCG	ATGA

Table 2-2: DNA sequences of unique overhangs that resulted from the digestion of DNA parts with BsaI.

For reasons of cost and turn-around time, putative promoter sequences to be cloned were synthesised as linear double stranded DNA fragments (gBlocks) by Integrated DNA Technologies (Iowa, United States of America). Linear fragments were either used directly in cloning reactions, or were inserted

into the pEX1C3 entry vector (Figure 2.4) by digestion with EcoR1 and Pst1 and ligation with T4 DNA ligase, all following the manufacturer's standard protocol.

pEX1C3 was adapted from the pSB1C3 sequence available from the iGEM registry of standard biological parts. BsaI restriction sites were removed from the Ampicillin resistance gene, and BsaI sites were inserted into the Multiple Cloning Sites (MCS). Additionally, the RFP coding sequence was replaced with the gene encoding DNA gyrase toxin ccdB to act as a negative selection marker during cloning (Huang *et al.*, 2010). All alterations were made using a QuikChange Lightning site-directed mutagenesis kit (Agilent Technologies, California, United States of America), according to the manufacturer's protocol.

The bioinformatically determined 100 bp *Geobacillus* putative *cis*-regulatory sequences were assumed to contain both promoter and RBS elements. Consequently, the scar sequence that would have resulted from cloning disparate promoter and RBS sequences together (ACCT) was inserted into the putative promoter sequences *in silico*. The promoter-RBS boundary was determined by alignment of putative *cis*-regulatory sequences to identify conserved regions using WebLogo version 2.8.2 (Crooks *et al.*, 2004). A highly conserved region of 15 base pairs at the 3' terminus of the 100 bp sequence space was identified as the putative RBS. The cloning scar sequence ACCT was therefore inserted between these 15 bases and the remaining 85 bp of putative promoter elements. The composite promoter-RBS sequences with cloning affixes were subsequently synthesised as single parts.

A modified version of the pS797 plasmid was used as a destination vector (Figure 2.5). To make pS797 compatible with the cloning methodology, BsaI restriction sites were removed from the ampicillin resistance gene by point mutation. Additionally, sequences containing BsaI sites and overhangs compatible with the 5' overhang of the promoter block and the 3' overhang of the terminator block were inserted into the MCS. Finally, the gene for the DNA gyrase toxin ccdB was inserted between the BsaI sites to act as a negative selection marker (Huang *et al.*, 2010). All modifications were made using a

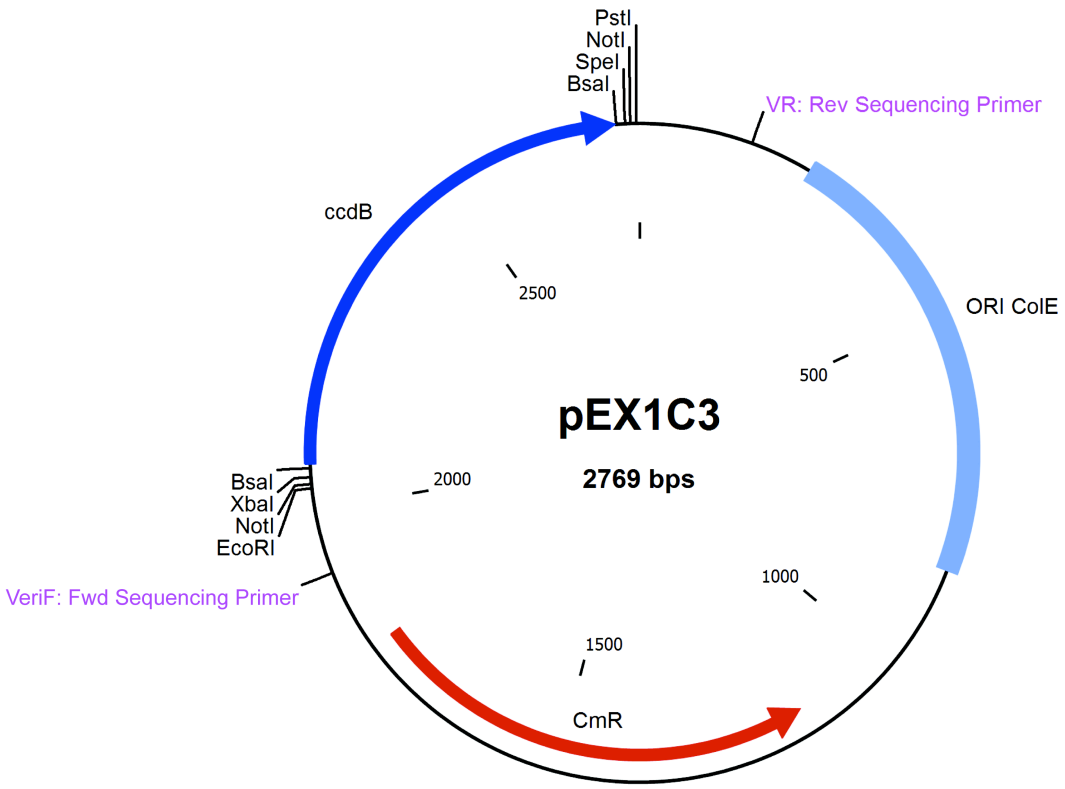


Figure 2.4: Plasmid map of pEX1C3, used as an entry vector for one-pot cloning.

The origin of replication, ORI ColE is shown in light blue. The chloramphenicol resistance cassette is shown in red, and the binding sites of primers used for sequence verification are highlighted in purple. The negative selection marker *ccdB* is shown in dark blue and is located between the multiple cloning sites (MCS) that contain the listed restriction sequences.

QuikChange Lightning site-directed mutagenesis kit (Agilent Technologies, California, United States of America), according to the manufacturer's protocol.

Cloning reactions consisted of 20 fmol of each entry vector or linear DNA fragment and 20 fmol of the destination vector, with 10 U Eco31I restriction endonuclease (a BsaI isoschizomer) and 1 U T4 DNA ligase in 2 µl ligation buffer (10x Thermo Scientific FastDigest buffer supplemented with 0.5 mM ATP) to a final reaction volume of 20 µl with ddH_2O (Kirchmaier *et al.*, 2013).

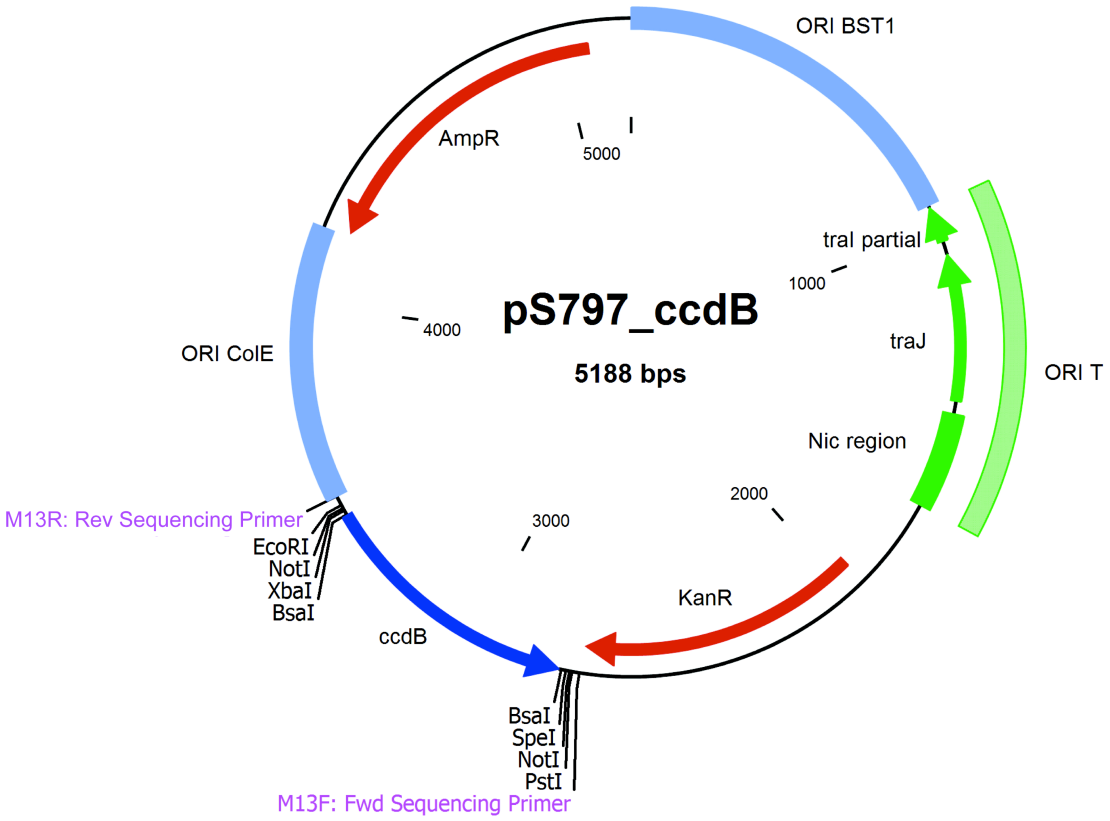


Figure 2.5: Plasmid pS797 as used as a destination vector for "one pot" cloning reactions.

The origin of transfer, ORI T, which contains the machinery necessary for mobilisation of the vector during conjugation, is shown in green. Antibiotic resistance cassettes are shown in red. Two origins of replication were present: ORI ColE for replication in *E. coli*, and ORI BST1, for replication in *Geobacillus.* Both are shown in blue. The binding sites of primers used for sequence verification are highlighted in purple.

A negative selection marker, the gene encoding the DNA gyrase toxin ccdB, is shown in dark blue and is located between multiple cloning sites (MCS) containing the listed restriction sequences.

Reactions were incubated for 50 cycles of 37 °C for 2 min then 20 °C for 5 min. This was followed by final incubation steps of 50 °C for 5 min then 80 °C for 5 min.

10 µl of the incubated cloning reaction mix was used to transform chemically competent NEB 5-alpha *E. coli*, following the protocol described in section 2.3.3. Constructs were selected for using ampicillin, as ampicillin

resistance was carried by none of the entry vectors. Further selection pressure was provided by the presence of the DNA gyrase toxin ccdB in the un-digested destination vector

Plasmid construction was verified by diagnostic digest and visualisation by gel electrophoresis as described in sections 2.3.9 and 2.3.10 respectively. Constructs were also verified by sequencing, as described in section 2.3.11.

2.3.9 Diagnostic digests

Restriction endonucleases were purchased from Thermo Scientific (Loughborough, United Kingdom), from either the FastDigest or Anza brands. Digests were preformed using one of the four restriction endonucleases with a single restriction site in the pS797 MCS (Figure 2.5). All digests were performed according to the manufacturer's protocol.

2.3.10 Gel electrophoresis

Gel electrophoresis was performing using a 1 % w/v agarose gel (10 mg ml^{-1}). Gels were formed of broad separation grade agarose (Fisher Scientific, Loughborough, United Kingdom) and TAE electrophoresis buffer (TAE buffer: 40 mM Tris; 20 mM CH_3COO-; 1 mM EDTA in ddH_2O) to a final volume of 50 ml or 100 ml, dependent on the number of samples to be run. Gels were stained with either 10 µg ml^{-1} ethidium bromide or 1 µl ml^{-1} 10, 000 X SYBR™ safe DNA gel stain (Thermo Scientific, Loughborough, United Kingdom).

5 µl HyperLadder™ 1kb (Bioline Reagents Ltd., London, United Kingdom) was loaded to each gel as a size reference marker, and gels were run at either 80 V (50 ml gels) or 120 V (100 ml gels) for 1 h.

DNA bands were visualised and photographed using a UV transilluminator.

2.3.11 DNA sequencing

Where required, constructs were verified by DNA sequencing (Sanger *et al.*, 1977), as performed by Source BioScience (Nottingham, United Kingdom). Primer sequences utilised are listed in Table 2-3.

	Sequence (5' → 3')	Notes
M13F	TGT AAA ACG ACG GCC AGT	Provides forward strand sequence. Used to sequence verify constructs in pS797.
M13R	CAG GAA ACA GCT ATG ACC	Provides reverse strand sequence. Used to sequence verify constructs in pS797.
VF2	TGC CAC CTG ACG TCT AAG AA	Provides forward strand sequence. Used to sequence verify constructs in pEX1C3.
VR	ATT ACC GCC TTT GAG TGA GC	Provides reverse strand sequence. Used to sequence verify constructs in pEX1C3.

Table 2-3: Primers used for sequence verification of plasmid DNA.

Results were evaluated for coverage and sequence using a combination of SnapGene Viewer version 3.1.4 (SnapGene, Illinois, United States of America) and Clone Manager Professional edition version 9 (Scientific & Educational Software, Colorado, United States of America).

2.3.12 Determination of Plasmid Copy Number (PCN) by quantitative PCR (qPCR)

Cultures for which PCN was to be determined were aliquoted and harvested by centrifugation at 4300 *g* for 5 min. Two aliquots were stored per culture, one of 10 µl and one of 400 µl. The supernatant was removed and pellets were flash frozen in liquid N_2 and stored at -80 °C until required.

Immediately before analysis, sample pellets were thawed on ice and re-suspended in 0.2 culture volume of 10 mM Tris-HCL pH 8.5. Suspensions were incubated at 95 °C for 5 min to lyse cells. qPCR reactions were set up using a Corbett Robotics CAS-1200 (Qiagen, Netherlands).

qPCR reactions consisted of 10 µl DyNAmo Flash SYBR green qPCR 2x master mix (Thermo Scientific, Loughborough, United Kingdom), 2 µl sample and 1 pM each of the forward & reverse primers, to a final volume of 20 µl with ddH_20. Disparate reactions were performed for the amplification of plasmid and genome amplicons, using the PCR primers listed in Table 2-4. Each reaction was performed in triplicate.

	Sequence (5' → 3')	Amplicon size (bp)	Notes
Plas_F	CTA TGT GGC GCG GTA TTA TC	177	Amplified region of ampicillin resistance gene in pS797.
Plas_R	CGC AGT GTT ATC ACT CAT GG	177	
Gen_F	GCT GGC GTT CTC TTA GTA CC	177	Amplified unique region of *G. thermoglucosidans* genome.
Gen_R	GCT GAG ACG GCT GTT ATC AC	177	

Table 2-4: Primers used in determination of plasmid copy number by qPCR.

Reactions were incubated using a Corbett Research Rotor Gene 6000 (Qiagen, Netherlands) using the following conditions: 95 °C for 7 min, followed by 40 cycles of 95 °C for 10 s then 60 °C for 20 s. Data were acquired during the 60 °C incubation step, using an excitation and emission wavelength of 470 nm and 510 nm respectively. The gain of the instrument was set at 5.

Melt analysis was performed immediately after the final extension cycle. Samples were held at 50 °C for 30 s, followed by pre-melt conditioning of 50 °C for 90 s. Temperature was subsequently ramped to 99 °C in 1 °C increments, at a rate of 5 s per 1 °C.

Plasmid and genomic standard curves were included with each qPCR run. Plasmid standard curves were generated using pS797 extracted by miniprep from NEB 5-alpha *E. coli* as described in section 2.3.6. Genomic standard curves were generated using *G. thermoglucosidans* genomic DNA extracted from 10 ml overnight cultures using a GeneElute Bacterial Genomic DNA kit (Sigma-Aldrich Company Ltd, Dorset, United Kingdom), according to the manufacturer's protocol.

Standard curves contained plasmid or genomic DNA diluted with ddH_2O to fall within the range 2-7 ng μl^{-1}. From these initial concentrations, four ten-fold serial dilutions with ddH_2O were produced and also served as standards.

If a qPCR reaction for a given sample failed to provide a quantifiable result, the secondary pellet from that culture was thawed on ice and re-suspended in 0.1 culture volume of 10 mM Tris-HCL pH 8.5. Cell lysis and qPCR analysis was then performed as before.

All data analysis was performed using Rotor Gene 6000 Series software version 1.7 (Qiagen, Netherlands). Threshold cycle (C_t) values were calculated using automatic baseline adjustment. DNA concentrations were reported in ng μl^{-1}.

To convert DNA concentrations to a measure of plasmid copy number, the following formula was used (Integrated DNA Technologies, 2016):

$$Number\ of\ copies\ per\ \mu l = \frac{x * (6.0221 * 10^{23})}{(N * 660) * (1 * 10^{9})}$$

where x is the amount of amplicon in ng and N is the length of the dsDNA amplicon in bp. PCN was then calculated by the formula:

$$PCN = \frac{Number\ of\ copies\ per\ \mu l\ (Plasmid)}{Number\ of\ copies\ per\ \mu l\ (Genome)}$$

2.4 Molecular methods for synthetic promoter production

2.4.1 Generating promoter libraries by Saturation Mutagenesis of Flanking Regions (SMFR)

To identify consensus motifs within *Geobacillus* promoter sequences, 34 sequences that had been previously shown to have promoter activity were aligned and the resulting alignment visualised using WebLogo version 2.8.2 (Crooks *et al.*, 2004). This alignment led to the identification of three conserved motifs, which were hypothesised to be putative RBS, -10 and -35 regions. Relative to the start codon of the upstream CDS, these regions spanned from -1 to -18 inclusive (putative RBS), from -25 to -31 inclusive (putative -10 motif) and from -37 to -43 inclusive (putative -35 motif).

Degenerate oligonucleotides were designed which encoded the sequence of the *G. thermodenitrificans ldhA* promoter in the three putative motif locations. The remainder of the 150 bp sequence was degenerated, with requisite cloning affixes at the 5' and 3' termini to permit use in the type IIS cloning strategy. At each position where degeneracy was specified in the oligonucleotide sequence, all four nucleobases had an equal probability of occurring.

The length of the final promoter sequence and the requisite cloning affixes prohibited synthesis as a single oligonucleotide. As such, two oligonucleotides were designed (Figure 2.6). The first encoded the antisense strand of the final promoter sequence, up to and including the putative -35 motif, as well as a BsaI site and the requisite overhang sequence to allow for use in one-pot cloning. The second oligonucleotide encoded the sense strand of the remaining promoter sequence and cloning suffix, including the putative -35 motif, to allow for annealing of the two oligonucleotides.

Degenerate oligonucleotides were synthesised by Eurofins Genomics (Ebersberg, Germany). A complete list of oligonucleotides and the DNA primers used for their conversion to dsDNA is provided in Table 2-5.

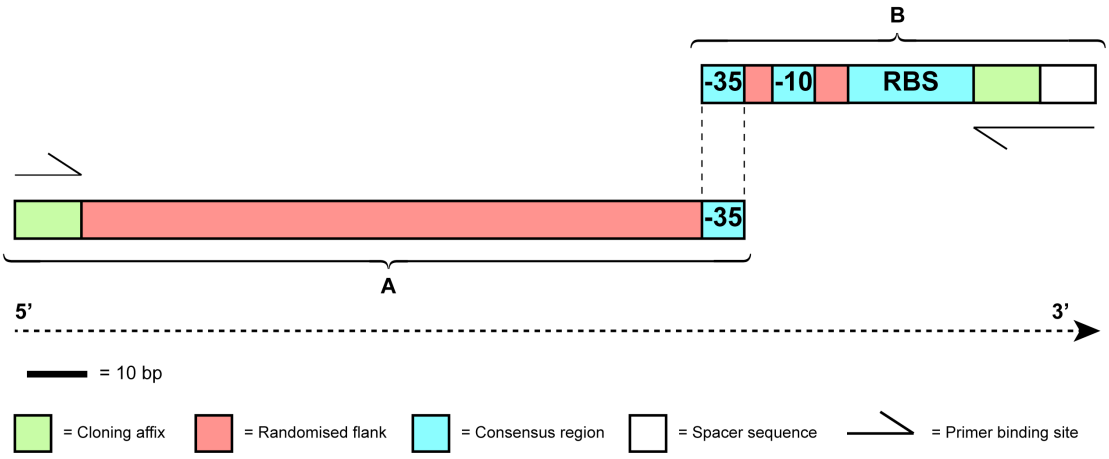


Figure 2.6: Schematic representation of degenerate oligonucleotides used in synthetic promoter production by Saturation Mutagenesis of Flanking Regions.

Cloning affixes are highlighted in green. Degenerate regions are shown in red, with consensus regions shown in blue. The putative -35 consensus motif was present on both oligonucleotides to permit the two sequences to be annealed together. Primer binding sequences to allow for conversion to dsDNA are indicated.

	Sequence (5' → 3')
Degenerate Oligo 1	CCAGAGTATGAN$_{103}$ATTTTA
Degenerate Oligo 2	TAAAATAN$_5$TGAATGTN$_6$ACCTATAAGAAGGGAGAATAGTAATG TGAGACCACGAAGTTA
PCR Primer (Fwd)	GGT CTC ATA CT
PCR Primer (Rev)	TAA CTT CGT GGT CTC ACA TT

Table 2-5: Oligonucleotide and primer sequences used for synthetic promoter production by Saturation Mutagenesis of Flanking Regions.

Degenerate oligonucleotides were annealed using a modified version of the protocol described by Cronn *et al.* (2008); oligonucleotides were suspended at 200 µM in Annealing buffer (10 mM Tris-Cl pH 8.0, 1 mM EDTA, 50 mM NaCl). Equal volumes of each degenerate oligonucleotide were combined, and the resultant mix was heated at 95 °C for 2 min and then cooled to 30 °C at a

rate of 1 °C min^{-1}. Annealed oligonucleotides were snap cooled to 4 °C, diluted to 15 µM in 10 mM Tris-HCL pH 8.5 and stored at -20 °C until required.

SMFR PCR reactions used Phusion High-Fidelity DNA polymerase (Thermo Scientific, Loughborough, United Kingdom). Reactions comprised 5x Phusion HF buffer combined with 1 µl annealed oligonucleotides, 200 µM each of dNTPs, 0.5 µM each of the forward and reverse primers and 0.02 U $µl^{-1}$ Phusion polymerase, made up to a total volume of 20 µl with ddH_2O.

Reactions were incubated using the following cycling conditions: 47.6 °C for 30 s, 72 °C for 10 s, followed by 30 cycles of 98 °C for 10 s, 47.6 °C for 30 s then 72 °C for 10 s. This series was followed by a final incubation at 72 °C for 10 min.

The resultant mutated promoter sequences were cloned into pS797 with the *GFP* reporter sequence and the pS718 terminator. The cloning protocol was performed as described in section 2.3.8. Promoter and RBS entry vectors were replaced in the cloning master mix by 1 µl PCR product from the above reaction.

Finally, 10 µl of the incubated cloning reaction mix was used to transform chemically competent *E. coli* NEB 5-alpha, according to the protocol described in section 2.3.3.

2.4.2 Generating synthetic promoter libraries by error prone Polymerase Chain Reaction (epPCR)

Error-prone Polymerase Chain Reaction (epPCR) was performed using the *G. thermodenitrificans ldhA* promoter sequence as a template. The scar sequence `ACCT` was inserted *in silico* between the promoter and putative RBS sequence to mimic the effect of combining promoter and RBS by molecular methods. The requisite cloning prefix and suffix were added to the *ldhA* promoter sequence *in silico* to allow for use in the cloning strategy detailed in section 2.3.8.

The *ldhA* promoter sequence with cloning affixes was synthesised as a linear gBlock gene fragment by Integrated DNA Technologies (Iowa, United States of America). PCR primers were designed to allow error-prone amplification of the complete *ldhA* promoter sequence whilst leaving the BsaI restriction sites from the cloning affixes unaltered. The sequences of the promoter, cloning affixes and PCR primers are summarised in Table 2-6.

The synthesised *ldhA* promoter gene fragment was cloned into the pEX1C3 entry vector by restriction enzyme digest with FastDigest EcoRI and PstI and ligation with T4 DNA ligase (Thermo Scientific, Loughborough, United Kingdom), according to the manufacturer's standard protocol. Entry vector construction was verified by restriction enzyme digest and gel electrophoresis as described in sections 2.3.9 and 2.3.10 respectively.

	Sequence (5' → 3')
G. thermodenitrificans ldhA promoter	CTGCCTCGTCCATTTTTTGCTTAATGGAGGTTGTCATGAAAA TGACAAACAACGTCCAAACAATTGCCATAATCGTTTACGCATA GTTTCGATTTCATCGCGTAAAATAATTTGTGAATGTATTCACA CCTATAAGAAGGGAGAATAGT
Cloning prefix	CAGGAAACAGCTATGACCATGGAATTCGCGGCCGCTTCTAGAG ACTCTGTGGTCTCATACT
Cloning suffix	AATGTGAGACCACGAAGTTACTAGTAGCGGCCGCTGCAGGACT GGCCGTCGTTTTACA
PCR Primer (Fwd)	AGA CTC TGT GGT CTC ATA CT
PCR Primer (Rev)	TAA CTT CGT GGT CTC ACA TT

Table 2-6: DNA sequences and primers used for synthetic promoter production by error-prone PCR.

Mutagenic PCR reactions utilised a commercially available mix of 8-oxo-dGTP and dPTP to promote transversion and transition mutations respectively (Zaccolo *et al.*, 1996, Paul *et al.*, 2013). Reactions comprised 5x GoTaq reaction buffer (Promega, Wisconsin, United States of America), 2 mM $MgCl_2$, 0.5 mM Mutagenesis dNTP mix (TriLink Biotechnologies, California, United States of America), 500 nM of each of the forward and reverse primers and

1.25 U GoTaq DNA polymerase (Promega, Wisconsin, United States of America). The final reaction volume was made up to 25 µl with ddH_2O.

Reactions were incubated using cycling conditions adapted from Reeve *et al.* (Reeve *et al.*, 2016): 95 °C for 2 min , followed by cycles of 92 °C for 1 min, 55 °C for 1.5 min, 72 °C for 5 min for 20 cycles, and a final incubation of 72 °C for 10 min.

To digest any remaining template DNA, the resulting reactions were supplemented with 10x FastDigest Buffer and 1 µl FastDigest Dpn1 (both Thermo Scientific, Loughborough, United Kingdom) to a final reaction volume of 30 µl with ddH_20. Reactions were incubated at 37 °C for 10 min, followed by heat inactivation of the Dpn1 enzyme at 80 °C for 5 min.

Any remaining non-natural dNTPs were removed from the mutated sequences by a further PCR cycle using Phusion High-Fidelity DNA polymerase (Thermo Scientific, Loughborough, United Kingdom). 5x Phusion HF buffer was combined with 1 µl of the product from the above reaction, 200 µM each of dNTPs, 0.5 µM each of the forward and reverse primers and 0.02 U μl^{-1} Phusion polymerase, made up to a total volume of 20 µl with ddH_2O.

Reactions were incubated using the following cycling conditions: 98 °C for 30 s, followed by 30 cycles of 98 °C for 10 s, 52.3 °C for 30 s then 72 °C for 10 s. This series was followed by a final incubation at 72 °C for 10 min.

The resultant mutated promoter sequences were cloned into pS797 upstream of the *GFP* reporter sequence and the pS718 terminator. The cloning protocol was performed as described in section 2.3.8. Promoter and RBS entry vectors were replaced in the cloning master mix by 1 µl PCR product from the above reaction.

Finally, 10 µl of the incubated cloning reaction mix was used to transform chemically competent *E. coli* NEB 5-alpha, according to the protocol described in section 2.3.3.

2.4.3 Initial screening of synthetic promoter libraries

Prior to characterisation in *Geobacillus,* synthetic promoter libraries were pre-screened in *E. coli* NEB 5-alpha to identify active promoter elements. The transformation plates that resulted from both SMFR and epPCR library production methods were visually examined using a blue-light transilluminator and amber filter (Labtech International Ltd, East Sussex, United Kingdom). Visibly fluorescing individual colonies were picked and used to inoculate 5 ml LB broth, laced with 100 µg ml^{-1} Ampicillin.

Cultures were incubated at 37 °C with shaking at 220 rpm for 16 h, whereupon plasmids were extracted by miniprep and constructs verified by diagnostic digest, gel electrophoresis and DNA sequencing, as described in sections 2.3.6, 2.3.9, 2.3.10 and 2.3.11 respectively.

Once visibly fluorescing colonies had been picked from the *E. coli* transformation plates, the remaining colonies were resuspended in 1 ml LB broth and used to inoculate 250 ml conical flasks containing 40 ml LB broth laced with 100 µg ml^{-1} Ampicillin. Cultures were subsequently incubated at 37 °C, with shaking at 220 rpm, for 16 h.

100 µl of the resulting stationary phase culture was diluted 10-fold with sterile phosphate buffered saline (PBS; 0.01 mM $Na_2PO_4.7H_2O$, 3mM KCl, 140 mM NaCl, pH 7.4) for analysis by BD FACS Aria II Fluorescence Activated Cell Sorter (FACS) (BD Biosciences, California, United States of America). The cytometer was fitted with a 100 nm nozzle, and a sheath fluid of PBS was used. Culture fluorescence was excited at 488 nm and fluorescence intensity was recorded using the 530/30 nm detector.

A negative control, *E. coli* NEB 5-alpha transformed to contain an empty pS797 vector, was used to provide baseline fluorescence. Any cell events with fluorescence values falling within a gate created around the control population were ignored; a lack of fluorescence indicated either a failed cloning reaction or

the mutated promoter sequence within the analysed cell having no promoter activity.

To isolate mutated promoter sequences with a range of activity levels, the remaining fluorescence range was divided into eight gates of equal size. Single cell events from each gate were individually sorted directly into a well of a 96-well microplate containing 200 µl LB broth containing 100 µg ml^{-1} Ampicillin. “Sweet Spot” monitoring was used to ensure droplet formation efficiency, and sorting purity was set to “Single Cell”.

Microplates were subsequently incubated for 16 h at 37 °C, with shaking at 800 rpm, using PHMP Thermoshakers (Grant Instruments, Cambridgeshire, United Kingdom). After incubation, culture fluorescence and absorbance was analysed by a Tecan plate reader, as described in section 2.5.4. Culture fluorescence was normalised to absorbance, and any cultures with a resulting fluorescence value greater than that of the pS797 negative control were plated onto LB agar plates with the relevant antibiotic selection, and incubated at 37 °C for 16 h.

Subsequently, single colonies were subsequently picked and grown overnight in 5 ml LB broth. Putative promoter constructs were extracted by miniprep and verified by diagnostic digest, gel electrophoresis and Sanger sequencing, as described in sections 2.3.6, 2.3.9, 2.3.10 and 2.3.11 respectively.

Once the DNA sequence of the putative promoters was verified, the sequences were characterised for activity in *G. thermoglucosidans* and *E. coli* NEB 5-alpha, as described in section 2.5.

2.5 Characterisation of putative promoter activity

2.5.1 Starter culture preparation

Fluorescent reporter proteins were used for promoter characterisation. In the case of putative promoters that were bioinformatically identified from the

Geobacillus core genome, both *GFP* and *mOrange* reporters were used. Synthetic promoters generated by epPCR and SMFR mutagenesis were characterised upstream of *GFP*.

In the case of cultures expressing GFP, colonies were manually pre-screened for fluorescence using a blue-light transilluminator with an amber filter (Labtech International Ltd, East Sussex, United Kingdom). When both fluorescing and non-fluorescing colonies were observed on the same petri dish, only those colonies that were visibly fluorescing were picked for further analysis.

For characterisation of promoter activity in *E. coli* strains NEB 5-alpha and S17-1, single transformed colonies were picked and used to prepare overnight cultures as described in section 2.3.5.

For characterisation in *Geobacillus,* transformed colonies were picked and restreaked on mLB agar plates, with antibiotic selection as required. Plates were incubated at 55 °C for 16 h. The resulting biomass was subsequently re-suspended in 5 ml mLB broth.

2.5.2 Culture growth in 250 ml Conical flasks

Starter cultures, prepared as described in section 2.5.1, were used to inoculate 60 ml of the relevant media, with antibiotic selection as required, to an OD_{600} of 0.1. Flasks were incubated at either 37 °C in the case of *E. coli* cultures or 60 °C in the case of *Geobacillus* cultures, with shaking in all cases at 220 rpm.

At the required time points, 200 µl sample aliquots were loaded to 96-well plates for analysis of culture absorbance and reporter fluorescence by a Tecan plate reader, as described in section 2.5.4.

2.5.3 Culture growth in 96-well plates

Sterile, black 96-well microplates with clear, flat bottom wells with lids were purchased from Corning (Flintshire, United Kingdom). Starter cultures, prepared as described in section 2.5.1, were used to inoculate 1 ml of the relevant media, with antibiotic selection as required, to an OD_{600} of 0.1. 200 µl sample aliquots were loaded onto 96-well plates using either a Corbett Robotics CAS-1200 (Qiagen, Netherlands) or a Gilson Pipetemax 268 (Gilson Inc, Wisconsin, United States of America).

To minimise the effect of position dependent bias, to which assays performed in 96-well plate format can be susceptible (Liang *et al.,* 2013), sample aliquots were loaded in a Latin rectangle design (Figure 2.7); no starter culture was represented more than once on any given row or column (Falcón, 2015). Starter cultures were allocated position groups at random, with aliquots from either 10 (Figure 2.7A) or 20 (Figure 2.7B) separate starter cultures loaded per plate.

96-well plates with lid covers have been shown to suffer from significant loss of culture in the outermost wells through evaporation (Chavez et al., 2016). To account for such edge effects, wells at the plate periphery were filled with 200 µl aliquots of sterile mLB broth.

Microplates were incubated using PHMP Thermoshakers (Grant Instruments, Cambridgeshire, United Kingdom). Incubation was at 37 °C in the case of E. coli cultures or 60 °C in the case of Geobacillus cultures, with shaking in all instances at 800 rpm.

2.5.4 Quantification of putative promoter activity

Tecan plate reader

Population level absorbance and fluorescence measurements were taken using a Tecan Infinite 200 PRO microplate reader (Tecan, Switzerland).

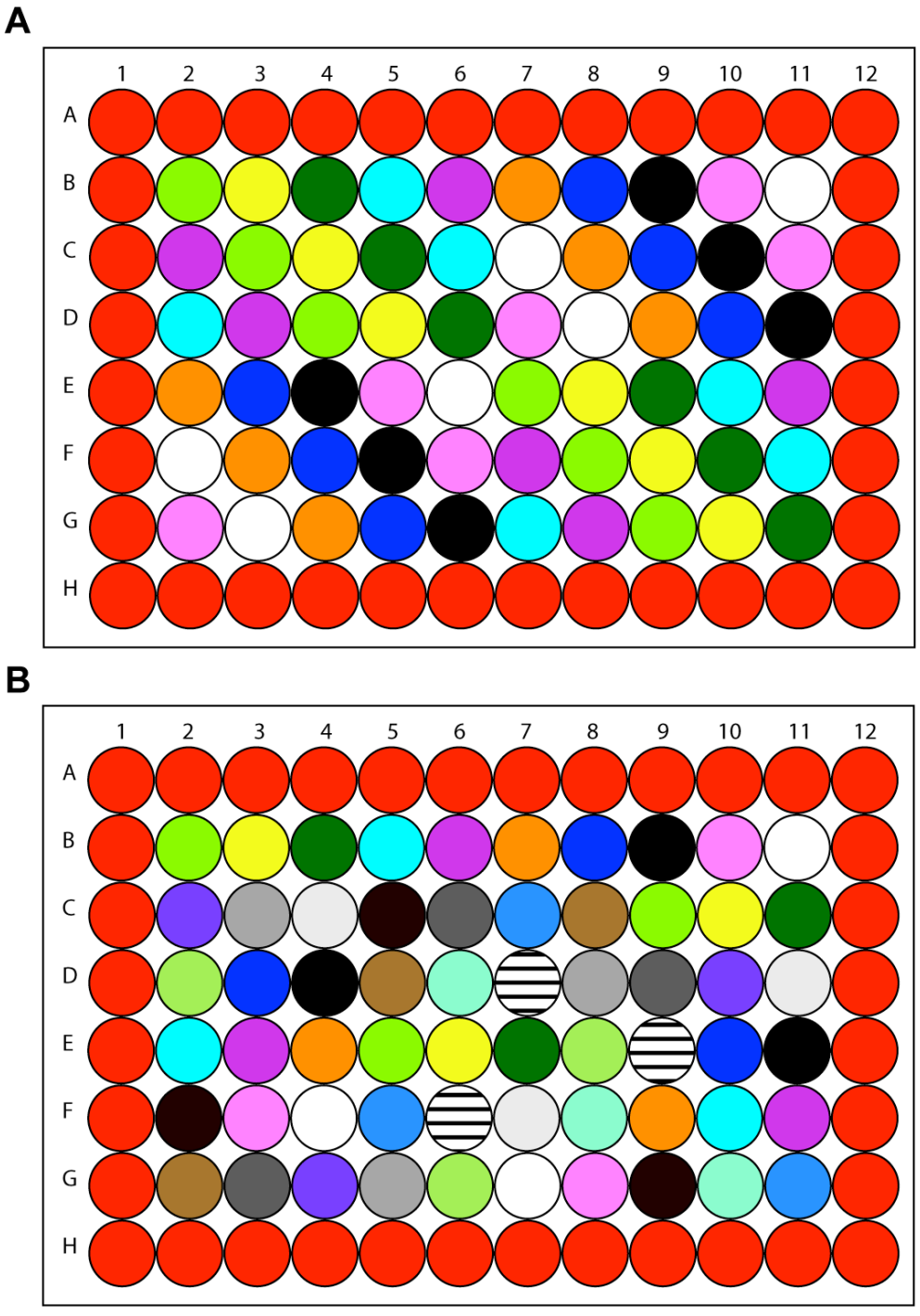


Figure 2.7: Schematic representation of Latin rectangle 96-well plate design.

96-well plates containing culture aliquots from A) 10 or B) 20 starter cultures. Disparate colours and patterns represent aliquots taken from the same starter culture. In both cases the outermost wells, highlighted in red, contain sterile growth media to reduce edge effects.

For measurements of GFP activity, fluorescence excitation and emission values were 477 nm and 515 nm respectively. For measurements of mOrange activity, excitation and emission values were 546 nm and 576 nm respectively. In both cases, gain was set to 56. Absorbance of all cultures was measured at 600 nm.

In all instances, well fluorescence or absorbance was reported as the mean of five individual reads: one read from the well centre, and one read from each of the four cardinal points of the well. A border between the edge of the well and the edge of the read zones was 300 μm in the case of absorbance measurements and 500 μm in the case of fluorescence measurements.

Flow cytometry

Flow cytometry was performed using a BD FACS Aria II Fluorescence Activated Cell Sorter (FACS), equipped with a 100 μm nozzle (BD Biosciences, California, United States of America). Cell fluorescence was excited at 488 nm and fluorescence intensity was recorded using the appropriate detectors: 530/30 nm in the case of cultures expressing GFP and 585/42 nm in the case of cultures expressing mOrange.

A sheath fluid of phosphate buffered saline (PBS) (0.01 mM $Na_2PO_4.7H_2O$, 3mM KCl, 140 mM NaCl, pH 7.4) was used for analysis of all samples. Cultures were diluted 10-fold with sterile PBS immediately prior to analysis. In the case of cultures grown in 96-well plate format, culture aliquots from the same starter culture were recombined and briefly vortexed to mix prior to dilution with PBS.

Unless otherwise stated, 100,000 events were recorded per population. Flow cytometry data were analysed using FlowJo version 10.2 (FlowJo LLC, Oregon, United States of America).

2.6 Experimental design, statistical analysis and predictive modelling

All statistical analysis and predictive modelling was performed using a combination of Prism versions 6 & 7 (GraphPad Software Inc., California, United States of America) and JMP pro versions 12 & 13 (SAS Institute Inc., North Carolina, United States of America).

All Design of Experiments (DoE) was performed using JMP pro versions 12 & 13. DoE is an empirical statistical approach to the design and analysis of experiments that allows the simultaneous alteration of multiple variables (Lendrem *et al.,* 2015b, Brown *et al.,* 2018b). Statistical models are subsequently applied to model the experimental response surface and to identify the optimal settings for each variable (Kumar *et al,* 2013). DoE aims to avoid the pitfalls of more classical One Factor at a Time (OFAT) experimentation, in which the final result may vary depending on the starting point used for each variable, and in which the true optimum combination of variables may be missed (Tye, 2004, Lendrem *et al.,* 2015a).

3 Bioinformatic identification of putative promoters & their characterisation in *G. thermodenitrificans*

Summary

Species from the genus *Geobacillus* have potential as microbial chassis for large-scale industrial bio-production. However, their widespread application is hampered by the relative lack of species-specific synthetic biology parts like reliable promoter sequences. To expand the number of regulatory sequences available for use in the genus, putative promoters were bioinformatically identified from upstream of genes that were shown to have homologues in four *Geobacillus* species. The genomes of two bacteriophage were also analysed to identify putative promoters. A rationally selected group of the putative regulatory sequences was subsequently characterised *in vivo* in *G. thermodenitrificans*, using GFP fluorescence as a measure of promoter activity. The *in vivo* characterisation of libraries of candidate promoters using reporter proteins is commonplace. However, *in silico* approaches to accurately determine the output of previously uncharacterised promoters, or to aid in the *de novo* design of synthetic promoter sequences, can potentially enhance and accelerate the design phase of the synthetic biology design-build-test cycle. Data derived from the experimental characterisation of *Geobacillus* promoters were therefore used to train a Partial Least Squares model to quantifiably link promoter DNA sequence to GFP output. This model was subsequently used to make predictions of activity for 12 novel synthetic putative promoter sequences. The accuracy of these predictions was assessed *in vivo.*

3.1 Introduction

Despite the potential of *Geobacillus* species as microbial chassis for large-scale industrial bio-production (Kananavičiūtė & Čitavičius, 2015), their widespread application is hampered by the lack of a diverse toolkit of robust genetic parts to expedite a synthetic biology approach to engineering in the genus. In particular, the ability to select a reliable promoter of known activity is

of paramount importance. Promoters that have previously been used for metabolic engineering in *Geobacillus* species, such as the *G. kaustophilus sigA* promoter (Suzuki, 2012), were typically isolated from the genome. The lactate dehydrogenase promoters from both *G. stearothermophilus* and *G. thermoglucosidans* have also both been applied for metabolic engineering in the genus (Cripps *et al.*, 2009, Lin *et al.*, 2014), although the oxygen dependent nature of the *ldh* promoter is well documented (Bartosiak-Jentys *et al.*, 2012). Alternatively, mutagenesis of previously characterised *Geobacillus* promoters has yielded libraries of synthetic promoter sequences. An approach based on Saturation Mutagenesis of Flanking Regions (SMFR), for example, yielded a library of 17 sequences, covering a reported green fluorescent protein (GFP) expression range of 76-fold (Pogrebnyakov *et al.*, 2017). Additionally, a Synthetic Promoter Library (SPL) generated using epPCR contained 20 promoter sequences, covering a 100-fold range of GFP expression strengths (Reeve *et al.*, 2016).

To expand the library of natural promoter sequences available for use in *Geobacillus*, the core genome of four species (*G. kaustophilus* DSM7263, *G. stearothermophilus* DSM22, *G. thermodenitrificans* K1041 and *G. thermoglucosidans* DSM2542) was bioinformatically identified. 100 bp putative promoters were subsequently isolated from immediately upstream of the start codon of coding sequences (CDS) in the core genome. 100 bp sequences were chosen as the majority of elements that are known to affect transcription initiation in prokaryotes occur within 100 bp of the start codon (Mendoza-Vargas *et al.,* 2009, Davis *et al.,* 2011). Putative promoter sequences were also isolated from two species of bacteriophage. 31 putative promoter sequences were synthesised upstream of *GFP* in the pS797 vector and characterised *in vivo* in *G. thermodenitrificans*, which was selected as the host organism by the sponsor, Shell Research Ltd.

Data derived from the experimental characterisation of these 31 putative promoters were then used to train a Partial Least Squares (PLS) model to quantifiably link promoter DNA sequence to GFP output.

3.1.1 Partial Least Squares modelling

Partial Least Squares (PLS, alternatively referred to as “Projection to Latent Structures”) models are used to infer the relationship between matrix of predictor variables, *X* (in this instance promoter DNA sequence) and a matrix of empirically measured responses, *Y* (in this instance GFP fluorescence) (Wold *et al.*, 2009).

PLS can be viewed as a specialised form of multivariate linear regression. Two features of the PLS algorithm render it particularly suited to modelling promoters. Firstly, PLS was designed to accurately model high-dimensional data sets, and is particularly adapted to scenarios where variables outnumber responses (Wold *et al.*, 2001). Given that each promoter sequence consisted of 100 independently modelled nucleotides (giving a total of 4^{100} potential sequences), hyper-dimensionality was an inherent characteristic of the promoter design space.

Secondly, PLS has been shown to produce accurate predictive models in instances where correlation between *X* variables, also known as multicolinearity, is high (Palermo *et al.*, 2009). Given the presence of DNA sequence motifs within promoters, multicolinearity was also likely to be a feature of the promoter data set. A PLS modelling approach was therefore selected to infer the relationship between promoter DNA sequence and function.

Mathematical principles of PLS

A number of algorithmic variants of PLS are available, but the underlying mathematical principles remain generally the same. PLS resolves the issues that result from high dimensional data by applying the assumption that the relationship between *X* and *Y* can be accurately inferred through a smaller number of underlying, or latent, variables (LVs) which are not directly observed or measured (Rosipal & Krämer, 2006). In brief, a matrix of individual *x* values, *X,* is decomposed to two secondary matrices. *T* (also called the *X*-scores)

consists of LVs of *X*, whilst *P'* (the *X*-loadings, alternatively termed coefficients) is used to relate *T* to *X* (Formula 1). A matrix of errors, *E*, is also calculated, and represents the residual information remaining in *X* once the LVs have been extracted (Gowen *et al.*, 2010). *Y* is also decomposed to comparable matrices of latent variables, coefficients (*Q'*) and errors (*F*) (Formula 2). Throughout the process, LVs are calculated in a manner that attempts to minimise error whilst maximising the covariance between the LVs of *X* and *Y*.

$$X = TP' + E \quad (1)$$

$$Y = TQ' + F \quad (2)$$

T is calculated through a linear transformation of *X* via a matrix of weights, *W*:

$$T = XW \quad (3)$$

Once *T* has been calculated, it can be used to generate an equation for general predictions of *Y* by substituting equation (3) into equation (2) (Gowen *et al.*, 2010):

$$Y = (XW)Q' + F \quad (4)$$

Model validation and interpretation

The optimum number of LVs to extract from *X* must be carefully considered. Models containing large numbers of LVs risk being overfit to the training data, providing an accurate description of the relationship between *X* and *Y*, but performing poorly when applied to making predictions based on observations that were not present in the training data set. Equally, extracting too few LVs will result in a model with insufficient statistical power (Gowen *et al.*, 2010). To optimise the number of LVs extracted and therefore calculate the most parsimonious model, the van der Voet T^2 test is applied (van der Voet, 1994). Multiple models are constructed, with a different number of LVs being extracted from the original data set by each model, up to a given maximum. The optimum model is judged to be the one with the smallest number of LVs, whose

prediction error is not statistically significantly greater than the model with the minimum error (Tobias, 1995).

In addition to optimising the number of LVs in candidate models, predictive power must also be assessed. To prevent the model being overfit to the training data set, candidate models should be used to predict *Y* for an independent validation data set, and the resulting errors assessed. In data sets where the number of individual *x* values is small, withholding large quantities of data from model training to use for model validation is likely to result in poor statistical power, reducing the probability of extracting useful information from the data (Button *et al.*, 2013). The selection of cross-validation (CV) methods must therefore be carefully considered to maximise the statistical power of candidate models whilst maintaining predictive accuracy.

In *K*Fold CV, the data set is divided at random into *K* portions, ideally of equal size. A model is trained on $K - 1$ parts of the data set and subsequently evaluated on the withheld data. Each of the *K* parts of the data set is iteratively used for model validation (Jung & Hu, 2015). A total of *K* models are therefore fit, and model performance is subsequently evaluated using the root mean PRESS statistic (Predicted Residual Sum of Squares). Root mean PRESS provides a measure of the squared prediction error between the observed values for a given *K* and the values predicted by the model; a lower root mean PRESS is indicative of more accurate prediction. The model with the lowest root mean PRESS is typically selected for more detailed interrogation. The underlying assumption of *K*Fold CV is that models constructed on a proportion of the data set are not statistically significantly different to those models constructed on the complete set of data (Beleites & Salzer, 2008).

Once models have been satisfactorily trained and validated, interpretation of the model parameters can provide useful insights to the system being analysed. Both weights and coefficients can be interpreted as quantitative measures of the impact of a given predictor on the measured response. In this instance, weights and coefficients can be used to determine whether a given

nucleotide at a given position within the promoter DNA sequence is increasing or decreasing GFP fluorescence.

The PLS algorithm also calculates a summary statistic to determine the importance of a given variable in determining the model prediction. The Variable Importance score (VIP) is calculated as the weighted sum of squares of the PLS weights, and takes into account the amount of explained *y* variance in each extracted LV (Wold *et al.*, 2001, Farrés *et al.*, 2015). A threshold VIP of 0.8 is commonly accepted, below which variables are judged to have an statistically insignificant impact on model output (Eriksson *et al.*, 2006). Variables with a small VIP and a small model coefficient are candidates for removal from subsequent models to increase model parsimony (SAS Institute Inc., 2016). Conversely, *x* values with a high VIP, large weights and large coefficients are judged to have statistically significant impacts on model output.

Generating synthetic promoter sequences

Once a PLS model is trained and validated, it can be used to make predictions of *Y* for novel *x* values. In this instance, a PLS model was applied to making predictions of GFP fluorescence, *Y*, for a group of putative synthetic promoter sequences, *X*. The accuracy of these predictions was subsequently assessed *in vivo*.

3.2 Results

3.2.1 Bioinformatic identification of putative promoters

Identification of putative promoters from the Geobacillus core genome

Four *Geobacillus* species, *G. kaustophilus* DSM7263, *G. stearothermophilus* DSM22, *G. thermodenitrificans* K1041 and *G. thermoglucosidans* DSM2542 were sequenced *de novo* and their genomes assembled. To identify proteins common to all four *Geobacillus* species, single-copy proteins were clustered into homologous gene families using the GET_HOMOLOGUES software package (Contreras-Moreira & Vinuesa, 2013). To increase prediction robustness, three separate clustering algorithms were used and the resulting gene families compared. Bidirectional best blast hit (BDBH), COG triangles (COG) and OrthoMCL (OMCL) algorithms returned 1,924, 1,914 and 1,902 gene clusters respectively, with 1,886 clusters being identified by all three algorithms (Figure 3.1). The core genome of the four *Geobacillus* species of interest was therefore shown to contain 1,886 genes. Given that homologues of these core genes were shown to be present in each of the four *Geobacillus* species, a total of 7,544 core genes were therefore identified.

The 100 bp immediately upstream of the start codon was extracted from each of the 7,544 core genes, and BPROM software was used to identify putative *cis*-regulatory sequences. DNA sequences were scored on the basis of the presence and nucleotide composition of functional motifs, with sequences being identified as putative promoters if they scored higher than a pre-determined threshold (Solovyev & Salamov, 2011). To isolate promoter sequences that were likely to be orthogonal to endogenous regulatory pathways, those sequences containing known transcription factor binding sites (TFBS) were discarded.

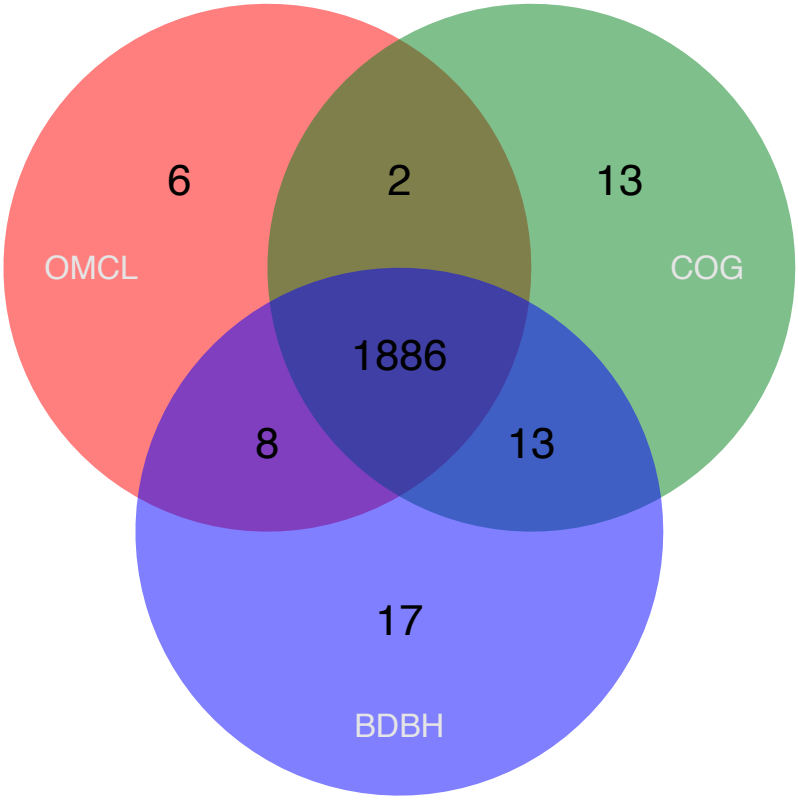


Figure 3.1: Venn diagram showing the number of homologous gene families identified in the genomes of the four *Geobacillus* species of interest by Bidirectional best blast hit (BDBH), COG triangles (COG) & OrthoMCL (OMCL) clustering algorithms.

Rendered using GET_HOMOLOGUES software.

1,489 putative 100 bp promoters[1] that did not contain TFBS were identified. The number of sequences isolated from each of the four *Geobacillus* species of interest is summarised in Table 3-1.

[1] Throughout the remainder of this thesis, the 100 bp, bioinformatically identified *cis*-regulatory elements are referred to as promoters, unless otherwise specified. Given that the 100 bp sequences were isolated from immediately upstream of the start codon of the adjacent CDS, the *cis*-regulatory sequences contained both promoter and RBS putative sequences, but the term promoter is used as shorthand.

	Abbreviation	Putative promoters
G. kaustophilus DSM7263	GKAU	403
G. stearothermophilus DSM22	GSTEA	370
G. thermodenitrificans K1041	GTDN	345
G. thermoglucosidans DSM2542	GTGNS	371

Table 3-1: Number of putative promoters isolated from the four *Geobacillus* species of interest.

A phylogeny consisting of 21 clades was constructed from the identified putative promoters (Figure 3.2A). This phylogeny was then used to select sequences at random for *in vivo* characterisation. A compromise was required between a desire to maximise the sequence diversity of the characterised promoters and a need to ensure that *in vivo* characterisation of the number of sequences selected was experimentally feasible. Two putative promoters were therefore selected at random from each of the 13 clades of the phylogeny that contained more than 50 sequences (Figure 3.2B).

Once selected, putative promoters were manually checked to ensure that they did not overlap with any adjacent CDS. If a putative regulatory sequence was found to overlap with a CDS, the promoter was discarded and a replacement sequence was selected at random from the same clade as the original. In total, 26 putative promoter sequences were selected to be synthesised upstream of *GFP* in the pS797 vector. The bioinformatic pipeline used for putative promoter discovery is summarised in Figure 3.3.

Identification of putative promoter sequences from bacteriophage

Intergenic regions of at least 100 bp were identified in two bacteriophage, *Thermus* phage phi OH2 and *Geobacillus* phage GBSV1. From these intergenic regions, the 100 bp sequences immediately upstream of the start codon of the adjacent CDS were extracted. The extracted sequences were subsequently screened to identify putative promoter sequences using BPROM.

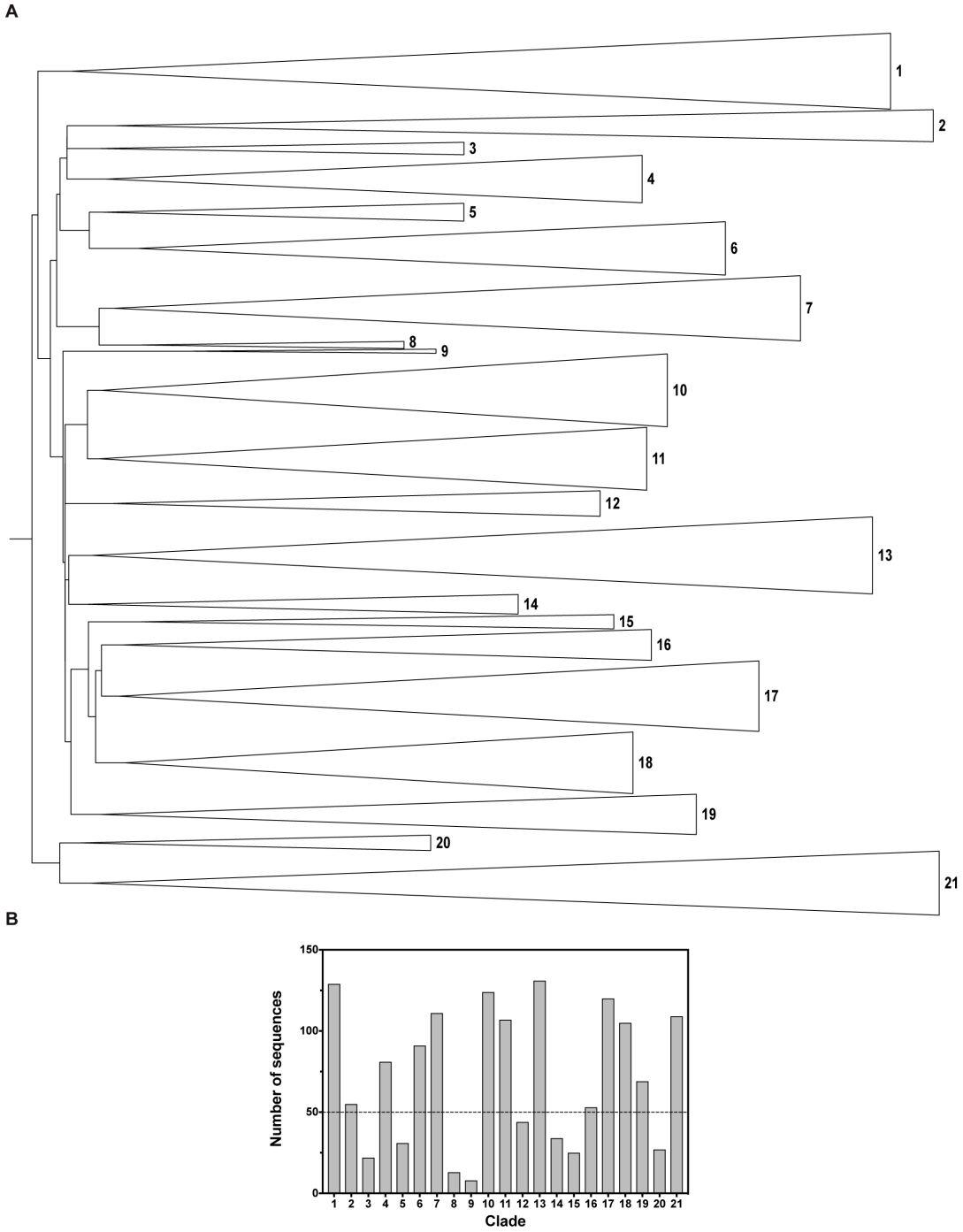


Figure 3.2: Phylogeny of putative promoter sequences.

A) Sequences were aligned using MUSCLE. The phylogeny was rendered using Figtree software and is rooted at the midpoint.

B) Bar chart showing the number of putative promoter sequences in each of the 21 clades of the promoter phylogeny. The dashed line shows the threshold of 50 sequences. Clades were discarded if they contained fewer than 50 sequences.

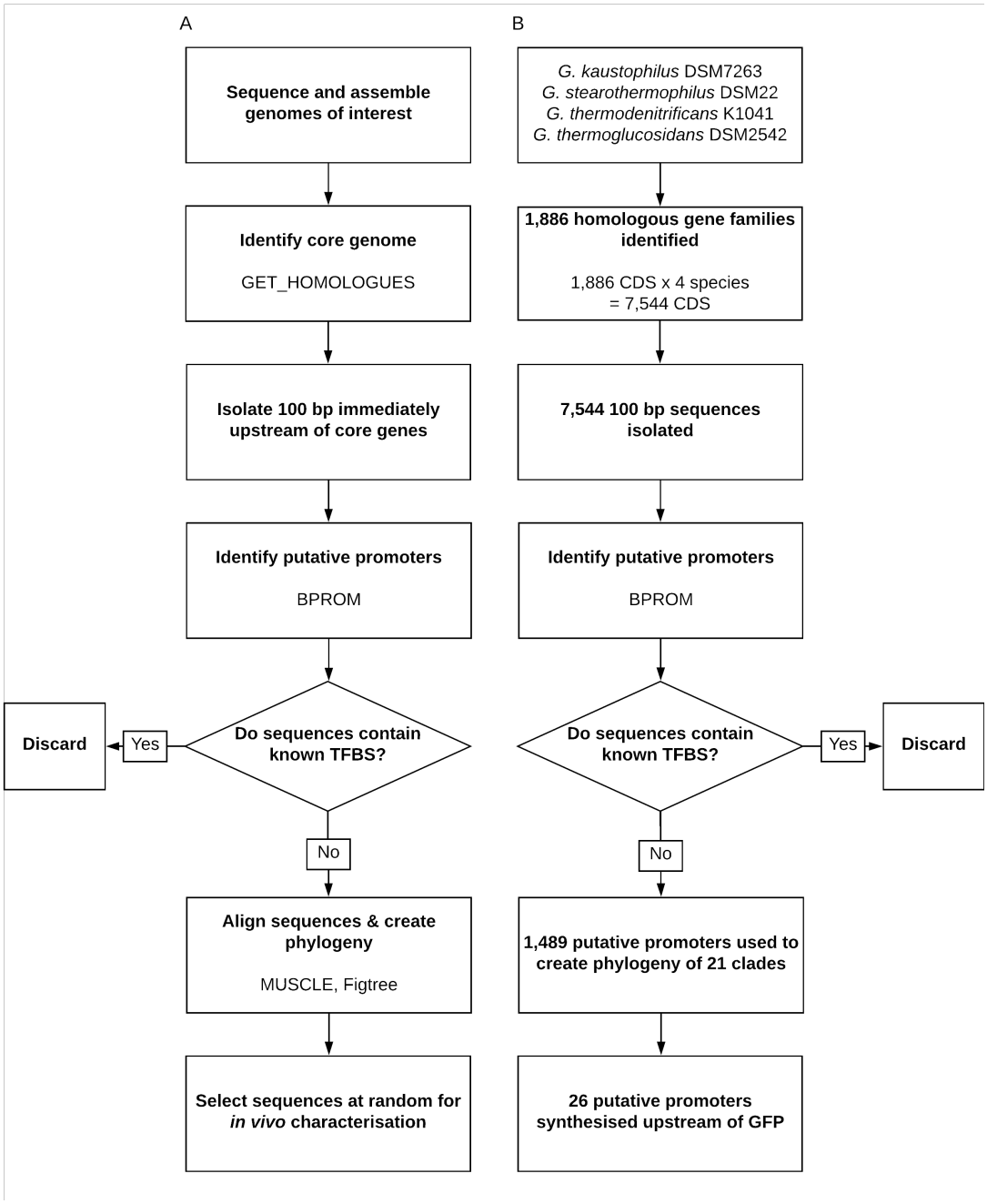


Figure 3.3: Bioinformatic pipeline for the identification of putative promoters.

A) shows the generalised bioinformatic pipeline, with relevant software named. B) shows the results of the pipeline when applied to four *Geobacillus* species, as discussed in the text. Rendered using Lucidchart.

Any putative promoters containing known TFBS were discarded. The number of intergenic regions and putative promoters identified in each phage is summarised in Table 3-2. Two putative promoters were selected at random from each phage for *in vivo* characterisation upstream of *GFP* in pS797.

	Intergenic regions ≥100 bp	Putative promoters
Geobacillus phage GBSV1	9	9
Thermus phage phi OH2	12	7

Table 3-2: Number of intergenic regions and putative promoters identified in the two phage of interest.

3.2.2 Putative promoter characterisation in G. thermodenitrificans

Wild-type Geobacillus growth characteristics

To ascertain the time points at which promoter characterisation measurements should be taken, wild-type *G. thermodenitrificans* and *G. thermoglucosidans* were cultured in both 96-well plate and 250 ml flask growth formats (Figure 3.4). Both *G. thermodenitrificans* and *G. thermoglucosidans* displayed typical bacterial exponential growth, with both species reaching stationary phase after between 7 and 8 h incubation in either culture format. Maximum growth rate was achieved between 3 h and 4 h incubation in 96-well plate format, and between 2 h and 3 h incubation in 250 ml flask format. For both species, a higher final optical density at 600 nm ($OD_{600\ nm}$) was reached by cultures grown in 96-well plates than those cultures in flasks, although analysis of the data by ordinary two-way ANOVA revealed no significant difference in final $OD_{600\ nm}$ between growth formats or species.

For promoter characterisation experiments, measurements of culture absorbance and fluorescence were therefore taken after 2.5 h (mid-log phase), 7 h (early stationary phase) and 24 h (late stationary phase) incubation.

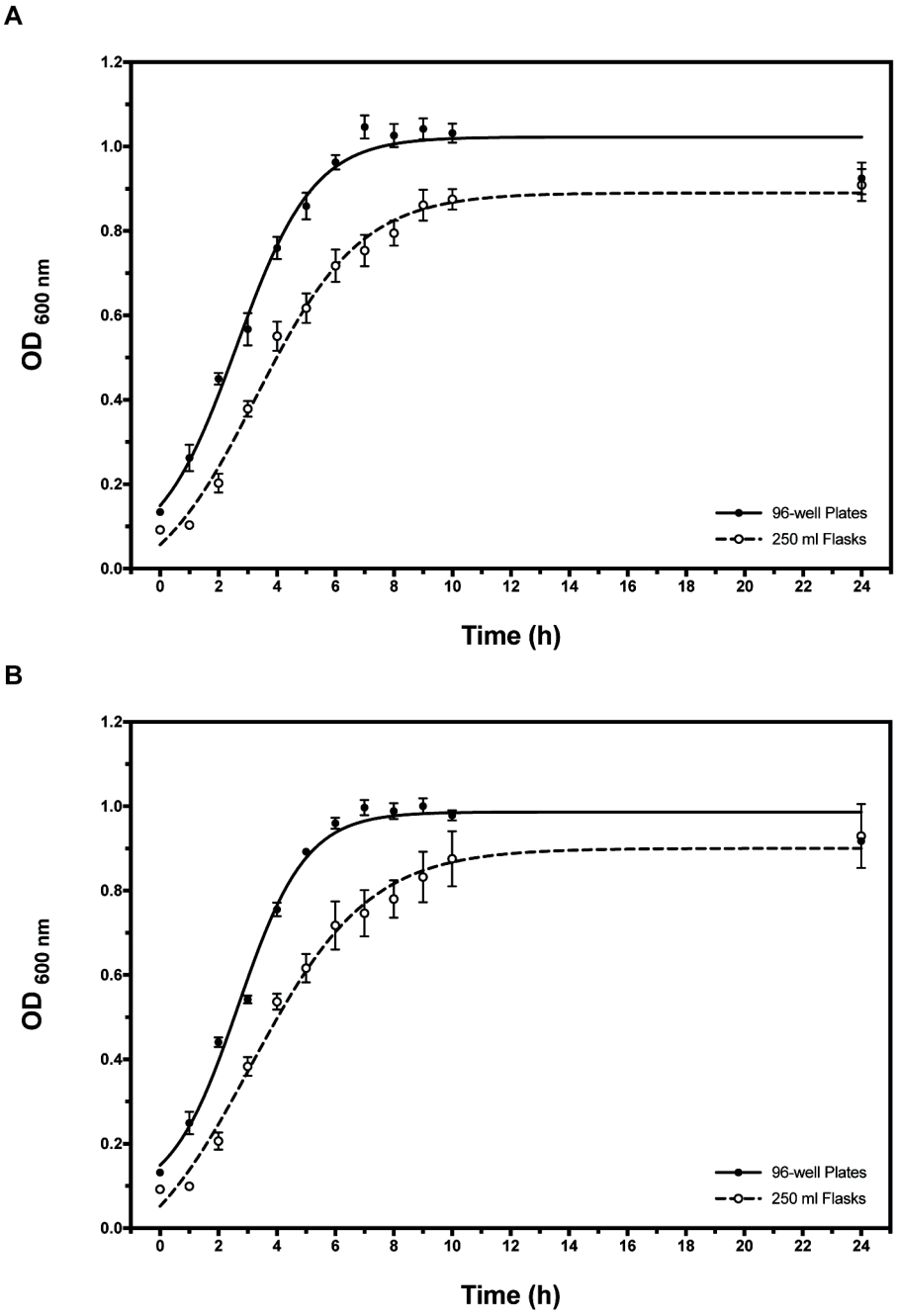


Figure 3.4: Growth curves of wild-type A) *G. thermodenitrificans* & B) *G. thermoglucosidans* in 96-well plate and 250 ml conical flask growth formats.

Geobacillus were cultured in mLB broth. To prepare starter cultures, single *Geobacillus* colonies were restreaked on mLB agar plates and incubated overnight at 55 °C. The resultant biomass was resuspended in 5 ml mLB broth and used to inoculate cultures to an initial $OD_{600\ nm}$ of 0.1. Incubation was at 60 °C, with shaking at 220 rpm in the case of cultures grown in flasks, and 800 rpm in the case of cultures grown in 96-well plates. Points represent the mean growth of three starter cultures arising from independent colonies, with standard deviation error bars shown, unless hidden by the point. The curves represent the best fit of the data using a four parameter logistic sigmoidal equation, rendered using Prism software.

For promoter characterisation experiments, measurements of culture absorbance and fluorescence were therefore taken after 2.5 h (mid-log phase), 7 h (early stationary phase) and 24 h (late stationary phase) incubation.

Characterisation of putative promoters

G. thermodenitrificans was chosen as the host species for *in vivo* characterisation by the industrial sponsor (Shell Research Ltd.). Experimental practicalities restricted the number of bioinformatically identified putative promoter sequences that could be feasibly characterised *in vivo*. High-throughput approaches to promoter characterisation, combining flow cytometry with multiplexed DNA and RNA sequencing, have previously been successfully employed to characterise libraries of thousands of regulatory sequences in species including *Bacillus subtilis* and *E. coli* (Kosuri *et al.*, 2013, Johns *et al.*, 2018). However, these cytometric methods require sufficiently large numbers of transformants: approximately 50-fold coverage of the promoter library being characterised is required for accurate characterisation. Low transformation efficiencies in *G. thermodenitrificans* precluded such high-throughput screening, as obtaining the required number of transformants was impractical.

In lieu of characterising all 1,489 of the bioinformatically identified putative *Geobacillus* promoters, a subset of sequences was selected. In order to maximise the sequence diversity of the chosen promoters, two putative promoters from each of the clades of the *Geobacillus* phylogeny that contained more than 50 sequences (Figure 3.2B) were selected at random. Additionally, four putative phage promoters were selected at random. Characterisation of the resulting group of 30 sequences was hypothesised to allow empirical exploration of a sufficient portion of the promoter design space whilst being experimentally feasible.

Of the 30 selected putative promoter sequences, one could not be immediately synthesised by ATUM (previously DNA 2.0, California, United States of America). As a result, an additional sequence was chosen at random for synthesis. Ultimately however, the problematic sequence was successfully

synthesised. The initial promoter library therefore contained 31 characterised sequences.

Comparing relative fluorescence of cultures grown in 96-well plate & 250 ml conical flasks

Growth of wild-type *G. thermodenitrificans* revealed no significant difference in culture $OD_{600\,nm}$ between cultures grown in 96-well plate and 250 ml conical flask format (Figure 3.4). To investigate any potential difference in fluorescence of cultures grown in the two growth formats, and hence to determine a growth format for subsequent promoter characterisation experiments, the initial group of putative promoters was characterised in *G. thermodenitrificans* cultured in both 250 ml flask and 96-well plate format (Figure 3.5).

A linear regression of the data returned a R^2 value of 0.626, indicating a positive correlation in culture fluorescence between the two growth formats. Additionally, multiple t-tests were used to compare the relative fluorescence of each *promoter::GFP* fusion in each growth format, corrected for multiple comparisons using the Holm-Šidák method. Only the promoter GSTEA_02162 was found to cause significantly different GFP expression between the two growth formats, with a higher mean fluorescence being recorded when cultures were incubated in 250 ml flasks. Given the lack of significant difference in fluorescence between growth formats for the majority of the characterised promoter sequences and the increased throughput afforded by growth in 96-well plates, subsequent characterisation experiments were performed in 96-well plate format only.

Promoter activity when cultured in 96-well plate format.

Of the 31 putative promoter sequences, only four (GSTEA_02364, GTDN_00966, N352_gp54 and GTGNS_00505) resulted in GFP expression that was statistically significantly different from the negative control,

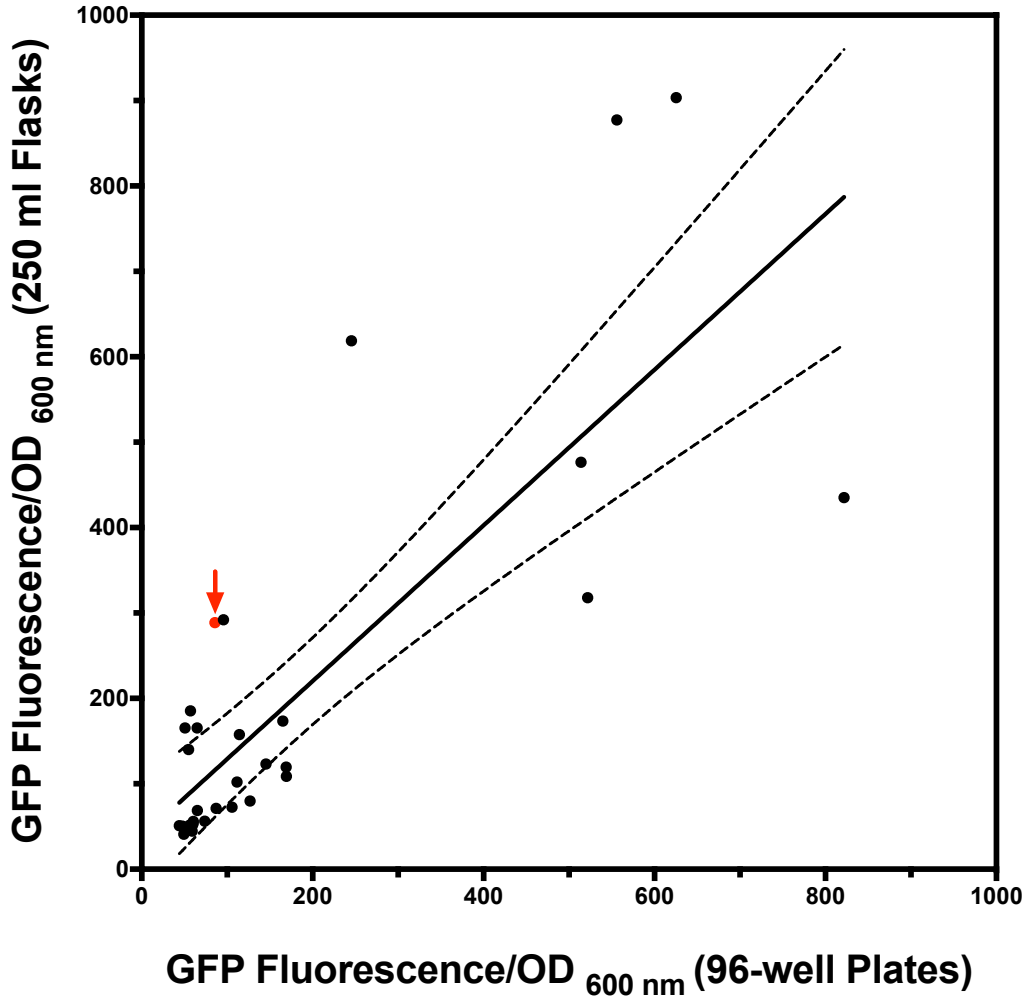


Figure 3.5: Comparing fluorescence output of promoters in *G. thermodenitrificans* cultured in 96-well plate and 250 ml flask growth formats.

Fluorescence and absorbance measurements after 24 h incubation. Points represent the mean fluorescence output of each promoter, from three starter cultures arising from independent transformants in each format. The solid line represents a linear regression of the data, with 95% confidence limits shown by the dashed lines. Fluorescence output for each *promoter::GFP* fusion in each growth format were compared using multiple t-tests, using the Holm-Šidák method to correct for multiple comparisons and a significance level of 0.05. The point coloured red indicates the promoter GSTEA_02162, which showed a statistically significant difference in GFP fluorescence between the two growth formats.

G. thermodenitrificans transformed with the empty pS797 vector (Figure 3.6). Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test and a significance level of 0.05. Only one promoter, GTGNS_00505, resulted in mean GFP expression that was greater than that of the positive control; this difference was not statistically significant. It could therefore be argued that only 13% of the characterised sequences were active promoters.

In total, the promoter library covered a 18.6-fold range of expression strengths, although this range dropped to only 1.6-fold if only the four "active" promoters were considered.

3.2.3 Modelling the relationship between promoter sequence and function

To account for any batch effects introduced by technical sources of variation between replicates, all promoter fluorescence[2] measurements were normalised to a measurement of fluorescence from the positive control, the *G. thermodenitrificans ldhA* promoter, cultured on the same 96-well plate as the promoter of interest. The PLS platform of the JMP software (SAS Institute, North Carolina, United States of America) was used to model the relationship between promoter DNA sequence (*X*) and GFP fluorescence output in *G. thermodenitrificans* (*Y*). Each of the 100 nucleotide positions in the promoter DNA sequence was modelled as an individual *x* variable. *Promoter::GFP* fluorescence output was as measured after 24 h incubation (Figure 3.6). For each of the 31 characterised promoter sequences, measurements from three starter cultures arising from independent transformants were included in the modelled data set, giving a total of 93 *y* values. The NIPALS (nonlinear iterative PLS) algorithm was used, with a maximum of 15 LVs permitted. *K*Fold CV was used, with *K* = 7, (*i.e.* the default value recommended by the JMP software).

[2] The term "promoter fluorescence" is used throughout this thesis to refer to the fluorescence activity of *promoter::reporter* fusions. Obviously, it is the reporter, not the promoter, from which fluorescence arises. However, the term "promoter fluorescence" provides useful shorthand.

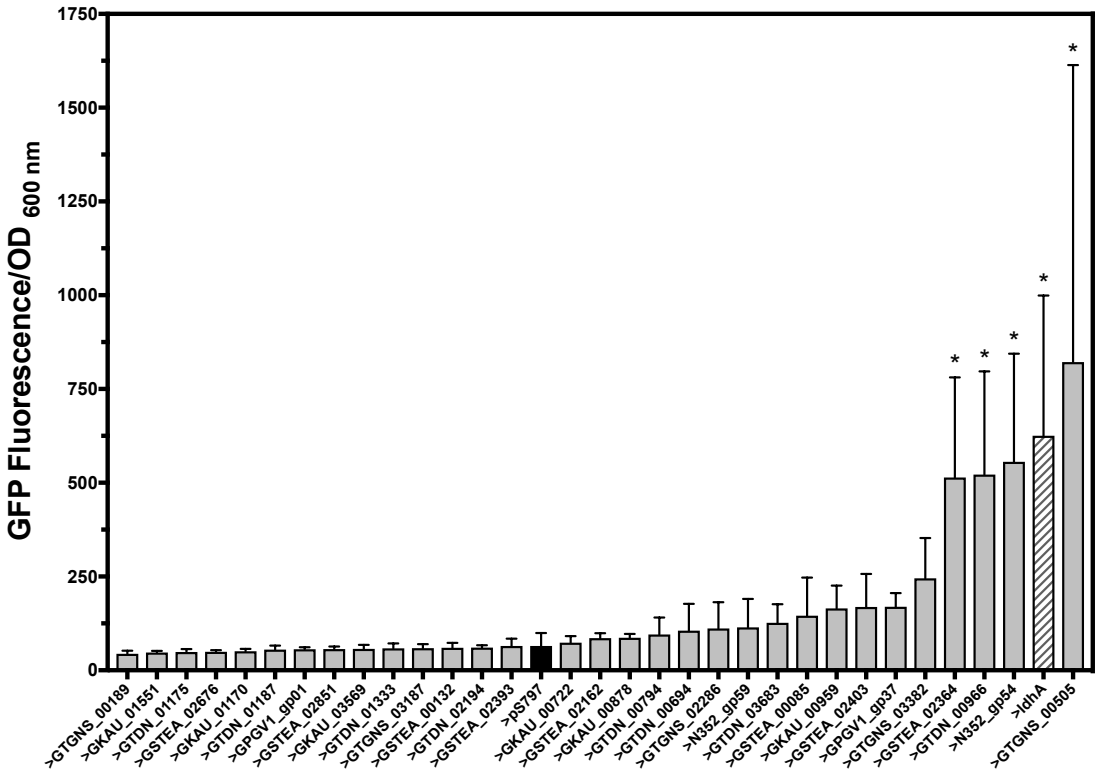


Figure 3.6: 31 putative promoters characterised upstream of GFP in *G. thermodenitrificans* cultured in 96-well plate format.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the hatched bar. The negative control, *G. thermodenitrificans* transformed with an empty pS797 vector, is highlighted in black. Bars represent the mean of n = 3 independent starter cultures, except in the case of the two controls, where n = 12. Standard deviation error bars shown, unless hidden by the bar. Promoter sequences with mean relative fluorescence output that was statistically significantly different to the negative control are labeled with an asterisk. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, using a significance level of 0.05.

Within the characterised promoter library, all four DNA nucleotides were represented at all 100 positions within the promoter sequence, with one exception. The -11 position (relative to the start codon of the upstream CDS) lacked a cytosine residue (Figure 3.7). The modelled promoter sequence space therefore contained a total of 399 *x* variables. In theory, the PLS model should have consequently been able to determine the contribution to GFP fluorescence of (almost) any nucleotide at any position within the promoter sequence.

The most parsimonious PLS model that was obtained used two LVs to model the relationship between *X* and *Y*, and explained 8.573% of the variation observed in *X* and 80.251% of the variation observed in *Y*. 69.844% of the variation observed in *Y* was explained by the first LV alone.

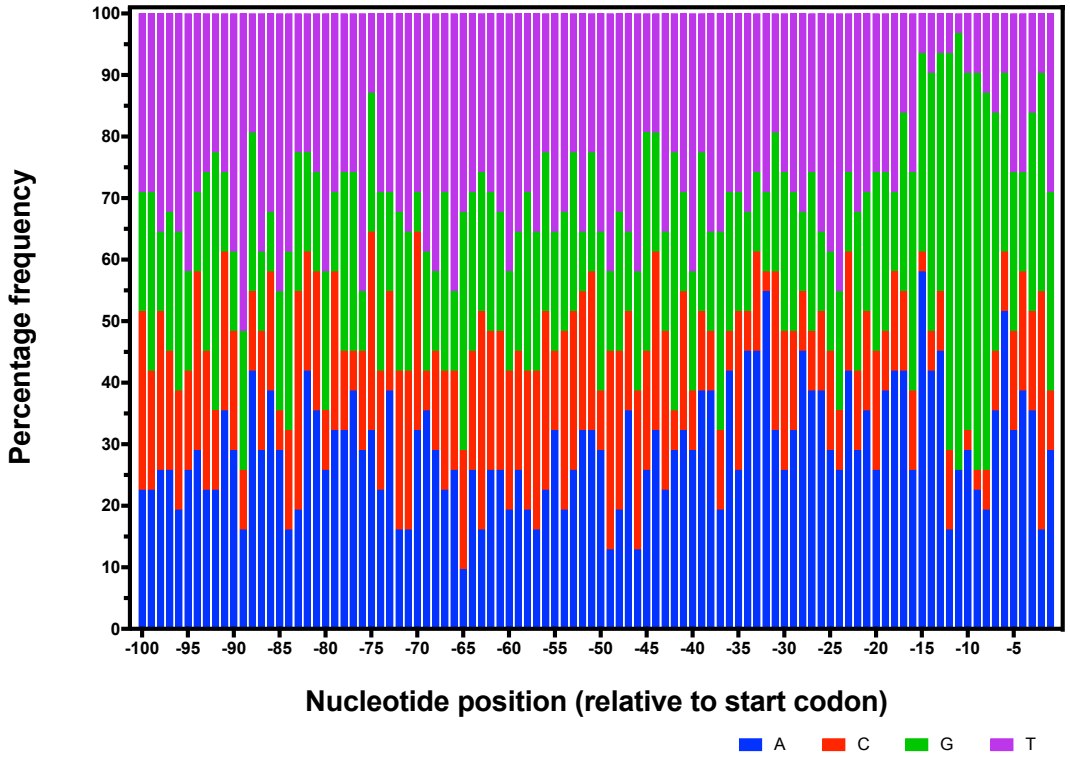


Figure 3.7: Percentage frequency of nucleotides at all positions within the set of 31 characterised putative promoters.

Diagnostics plots of the model performance when applied to the CV data set appeared indicative of good predictive power. A strong positive correlation was seen between empirically measured fluorescence values and values predicted by the model (Figure 3.8A). A linear regression of the data returned an R^2 value of 0.803. However, analysis of the model residuals (the difference between empirically measured and predicted *y* values) questioned the model's predictive power. The model residuals were clearly heteroscedastic – residuals increased in magnitude as the predicted GFP fluorescence increased. (Figure 3.8B). This result suggested that the predictive accuracy of the model deteriorated as promoter strength increased. Additionally, model residuals were shown to not be normally distributed by a Shapiro-Wilk W test, at a significance level of 0.05 (Figure 3.8C). The W value returned by the analysis was 0.780 (3sf), and Prob<W was <0.0001. PLS, like all conventional multivariate linear regression models, operates under the assumption that model residuals will be uncorrelated, have a mean of 0 and will be normally distributed. The lack of normality in the prediction residuals was therefore potentially indicative of underlying structural biases within the model (Eck, 2018, Schmidt & Finan, 2018).

Inaccurate prediction of fluorescence output for the stronger promoters was potentially the result of the skewed nature of the training data set. Of the 31 characterised sequences, only four promoters displayed mean activity levels that were statistically significantly greater than the negative control (Figure 3.6). As such, strong promoters were potentially under-represented in the training data set, reducing the predictive power of the model.

3.2.4 Generating putative synthetic promoters

The simulator function of the JMP software was used to generate 5,000 synthetic putative promoter sequences. To generate a synthetic promoter, nucleotides were selected at random at each of the 100 sequence positions. To increase the probability of including any key consensus motifs from the original training set of 31 wild-type promoters in the synthetic sequences, the probability of including any key consensus motifs from the original training set of 31

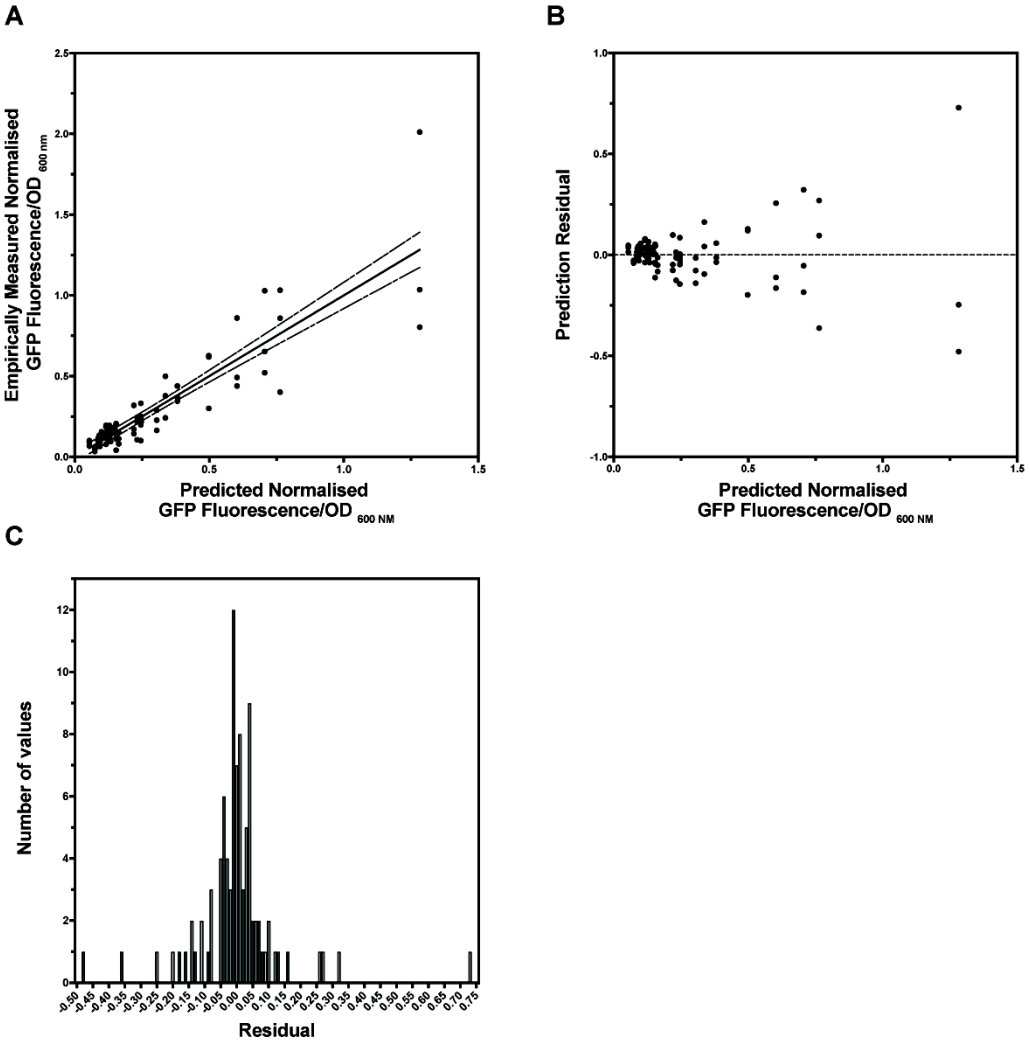


Figure 3.8: Partial Least Squares model diagnostics.

A) Empirically measured *promoter::GFP* fluorescence output, normalised to the positive control, the *G. thermodenitrificans ldhA* promoter, plotted against the normalised fluorescence as predicted by the PLS model. The solid line represents a linear regression of the data, with 95% confidence limits represented by the dashed lines.

B) Normalised fluorescence as predicted by the PLS model, plotted against the prediction residual (the difference between empirically measured and predicted fluorescence values). The dashed line is shown at the point where prediction residual is equal to 0.

C) Histogram of model residual distribution. Analysis of the data by Shapiro-Wilk W test at the 0.05 significance level revealed that the data were not normally distributed. The W value returned was 0.780 (3sf), and Prob<W was <0.0001.

wild-type promoters in the synthetic sequences, the probability of a nucleotide being assigned to a given sequence position was weighted based on the distribution of nucleotides found in the 31 characterised promoters (Figure 3.7). The optimum obtained PLS model was subsequently used to make a prediction of GFP fluorescence for the synthetic sequences.

12 synthetic sequences, with predicted normalised fluorescence output in the range 0.525-0.570 (*i.e.* approximately half the fluorescence output of the *G. thermodenitrificans ldhA* promoter) were selected for *in vivo* characterisation. Given the observed weakness of the PLS model at making predictions of fluorescence output for strong promoters (Figure 3.8), only those synthetic putative promoters with moderate predicted strength were selected for *in vivo* characterisation. BLAST queries raised against each of the 12 synthetic sequences returned no hits, indicating an absence of similar sequences in the GenBank database.

The 12 synthetic putative promoter sequences were synthesised upstream of *GFP* in the pS797 vector by ATUM. Empirical measurements of GFP fluorescence after 24 h incubation of *G. thermodenitrificans* transformants showed minimal correlation between the predicted and empirically measured GFP fluorescence values (Figure 3.9). Of the 12 characterised synthetic putative promoters, none resulted in normalised GFP expression that was significantly different to the pS797 negative control, as determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test. Only one synthetic putative promoter, GSYN_00011, had an empirical normalised GFP output that fell within one standard deviation of the predicted value.

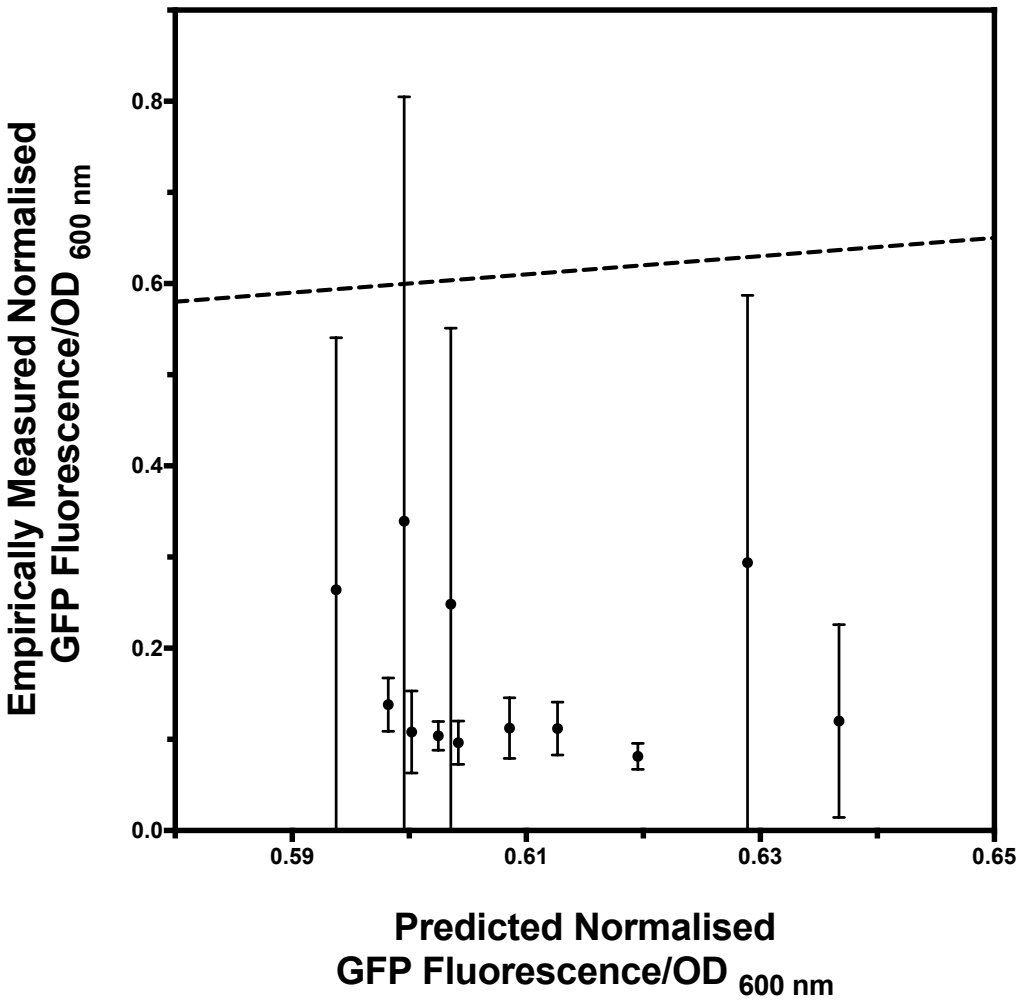


Figure 3.9: Empirically measured fluorescence output of putative synthetic promoter sequences, plotted against fluorescence as predicted by the Partial Least Squares model.

Points represent individual putative synthetic promoter sequences. Empirical measurements are the mean of three starter cultures arising from independent *G. thermodenitrificans* transformants, characterised after 24 h incubation. All fluorescence measurements were normalised to the positive control, the *G. thermodenitrificans ldhA* promoter. Standard deviation error bars are shown. The dashed line represents the point at which empirically measured and predicted fluorescence values are equal.

3.3 Discussion

Bioinformatic analysis of four species of *Geobacillus* and two species of bacteriophage resulted in the identification of 1,515 putative promoter sequences. Of these putative regulatory sequences, 31 were characterised upstream of *GFP* in *G. thermodenitrificans* (Figure 3.6). *GSTEA_02364::GFP*, *GTDN_00966::GFP*, *N352_gp54::GFP* and *GTGNS_00505::GFP* were the only *promoter::reporter* fusions for which mean fluorescence output was significantly greater than that of the negative control, *G. thermodenitrificans* transformed to contain the empty vector pS797. It could therefore be argued that only four of the 31 promoters were truly "active" in *G. thermodenitrificans*, giving BPROM a specificity of only 13% for identifying *Geobacillus* promoter sequences. This result compared poorly with BPROM's claimed specificity of 80% in *E. coli* (Solovyev & Salamov, 2011).

BPROM's lack of specificity for *Geobacillus* promoters was perhaps not surprising. The promoter classification algorithm described putative functional motifs and scored their oligonucleotide composition on the basis of a training set of *E. coli* sigma70 promoters (Solovyev & Salamov, 2011). However, transcription regulatory motifs that are active in *E. coli* may not be representative of the regulatory mechanisms of *Geobacillus* (Cardinale & Arkin, 2012).

Alternative *in silico* methods for prokaryotic promoter identification have been posited, including machine learning techniques such as Hidden Markov-models (Mann *et al.*, 2006) and artificial neural networks (Umarov & Solovyev, 2017). Application of these *in silico* techniques could potentially have increased the accuracy of putative promoter identification in *Geobacillus* as compared to BPROM. However, *de novo* motif identification approaches can be time consuming, and sophisticated machine learning techniques for promoter identification require prior understanding of the underlying statistical and biological characteristics of the sampled DNA sequences to provide a robust training set (Song, 2011). Such machine learning techniques are therefore not

always immediately applicable in non-model organisms, such as *Geobacillus*, for which promoter motifs may not have been previously described.

The relative inaccuracy of BPROM in identifying *Geobacillus* promoters did not necessarily preclude future application of the approach. BPROM can provide a “quick and dirty” screening technique to isolate a large set of putative promoter elements in a genus of interest. Given the ease and relatively low cost of DNA synthesis, *in vivo* characterisation can subsequently be used to further refine promoter selection, as was shown in this chapter.

The 31 putative promoter sequences that were characterised in *G. thermodenitrificans* displayed a total range of GFP expression of 70-fold (Figure 3.6). This result compared favourably with libraries of *Geobacillus* promoters described in the literature, which covered 100- (Reeve *et al.*, 2016) and 76-fold (Pogrebnyakov *et al.*, 2017) ranges of expression levels when characterised upstream of GFP. However, whilst the two published libraries contained promoter sequences with expression levels spanning the entirety of their stated range, the library described in this chapter was mostly comprised of sequences with no statistically significant promoter activity.

This discrepancy in the number of active sequences in the published promoter libraries and the promoters described in this chapter was potentially a result of differences in the way in which the libraries were conceived. The two published libraries used an *a posteriori* approach, in which previously characterised promoter sequences were mutagenised. By maintaining known consensus regions (Pogrebnyakov *et al.*, 2017) or keeping the rate at which mutations were incorporated by epPCR low (an average error rate of 10% was reported by Reeve *et al.,* although three of the 245 bp promoter sequences differed from the wild-type starting promoter by only one nucleotide) (Reeve *et al.*, 2016), the probability of maintaining promoter activity in mutant sequences was likely increased.

The failure of the initial PLS model to accurately predict the *in vivo* activity of 12 synthetic putative promoter sequences (Figure 3.9) may have

been the result of the sparsity of promoter activity levels in the training data set. Deficiencies in model predictive power resulting from sparse data sets are well understood (Beleites & Salzer, 2008). The heteroscedastic nature of the model residuals (Figure 3.8B) was also possibly the result of an over representation of weak or inactive promoters in the characterised promoter sequences; the training set only contained a limited amount of information regarding active promoter sequences, thus increasing prediction variance and model instability (Beleites & Salzer, 2008).

Prediction accuracy might have been improved by increasing the number LVs that were extracted from the data to reduce any underlying systematic biases; the optimal obtained PLS model extracted only two LVs from the training data, from a maximum permitted number of 15 LVs. However, such an increase in model complexity carried the risk of overfitting the model to the training data (Deng *et al.*, 2015). *In lieu* of increasing model complexity, an expansion of the training data set was deemed necessary to increase model predictive power. Theoretically, the more times a given nucleotide was observed at a given position within the promoter sequence, the more accurately the effect of that nucleotide on promoter output could be calculated (Liao *et al.*, 2007). The *in vivo* characterisation of a greater number of putative promoter sequences was also thought likely to increase the range of expression levels in the characterised library of *Geobacillus* regulatory sequences.

The need for a training data set containing a greater number of active promoter sequences was corroborated by studies that had previously used PLS to model promoter function. De Mey *et al.*, for example, required a training set of 42 *E. coli* promoter sequences, of 57 bp length, to accurately infer the relationship between DNA sequence and promoter function (De Mey *et al.*, 2007). Additionally, Jonsson *et al.* required 25 68 bp promoter sequences, all of which showed *in vivo* activity, to obtain an accurate model of *E. coli* promoter function (Jonsson *et al.*, 1993). Both of the published studies required a greater number of active promoters than were described in this chapter to model a design space of comparatively reduced dimensionality (*i.e.* sequences that were 57 bp or 68 bp long, rather than the 100 bp sequences that were modelled

in this investigation). The size of the training data set used in this investigation was therefore hypothesised to be inadequate to accurately infer the relationship between promoter sequence and function.

3.4 Summary

Bioinformatic analysis of the core genome of four *Geobacillus* species, (*G. kaustophilus* DSM7263, *G. stearothermophilus* DSM22, *G. thermodenitrificans* K1041 and *G. thermoglucosidans* DSM2542) and the genomes of two bacteriophage (*Thermus* phage phi OH2 and *Geobacillus* phage GBSV1) resulted in the identification of 1,515 putative promoter sequences. Data derived from the experimental characterisation of 31 putative promoters in *G. thermodenitrificans* were used to train a PLS model that inferred the relationship between DNA sequence and promoter function. Despite providing an accurate fit of the training data, the PLS model was unable to accurately predict the *in vivo* regulatory activity of 12 synthetic putative promoter sequences.

It was hypothesised that the number of *in vivo* characterised promoter sequences was inadequate to provide a significantly robust training data set: previous studies that had successfully used PLS to model promoter function were trained on a greater number of promoter sequences than were characterised in this chapter. *In vivo* characterisation of additional putative promoter sequences was therefore deemed necessary.

4 Modelling promoter activity as a function of nucleotide sequence in *G. thermoglucosidans*

Summary

The lack of predictive power shown by the Partial Least Squares (PLS) model discussed in Chapter 3 was hypothesised to be a result of the small size of the training data set as compared to the scale of the promoter design space. Three progressively larger sets of bioinformatically identified putative promoter sequences were therefore characterised *in vivo,* and the resulting characterisation data were used to derive PLS models of the relationship between promoter DNA sequence and function.

The linear nature of PLS modelling was also hypothesised to have contributed to the inaccuracy of the model discussed in Chapter 3. This linearity may have rendered PLS models unable to accurately account for any non-linearity in the promoter design space, increasing the probability of prediction errors. The address this possible deficiency, Artificial Neural Networks (ANNs) with non-linear activation functions were applied to training promoter sequence-function models. ANNs have previously been shown to perform poorly when the system under investigation is complex and the number of observations in the training data set is small. Partition modelling was therefore used to identify those positions within the promoter sequence that were predicted to be having the largest impact on promoter output. Downstream ANNs, derived from the same training data sets as the PLS models, were restricted to modelling promoter activity as a function of only those sequence positions that were identified as important by the partition models, thereby reducing the dimensionality of the promoter design space.

ANN and PLS models trained on each of the training data sets were used to predict *pre hoc* the activity of either putative synthetic promoter sequences or bioinformatically identified putative promoters. In all instances, the accuracy of these predictions was subsequently assessed *in vivo*.

4.1 Introduction

Chapter 3 discussed the *in vivo* characterisation of 31 putative promoter sequences in *G. thermodenitrificans*. The data that were derived from this experimental characterisation were subsequently used to fit a Partial Least Squares (PLS) model that attempted to infer the relationship between promoter DNA sequence and function.

Analysis of the optimal PLS model obtained suggested an accurate fit of the training data. However, when tested using 12 predicted synthetic promoter sequences, the model was unable to accurately predict *in vivo* promoter activity. The lack of predictive power was hypothesised to be a result of the small size of the training data set as compared to the scale of the promoter design space (4^{100} potential 100 bp sequences) and an over-representation of inactive promoter sequences in the training data set.

In addition to the size of the training data set, the personality of the model used to infer the relationship between promoter sequence and function may also have contributed to the lack of predictive power of the models discussed in Chapter 3. PLS is a linear regression method which assumes a linear relationship between the *X* and *Y* matrices (Wold *et al.*, 2009) and in which the model weights and coefficients are calculated as linear combinations of the original *x* and *y* variables. This linearity may have confounded the effects of any interactions between nucleotides within the promoter sequence with the main effects for each individual nucleotide position (Jonsson *et al.*, 1993). The mathematical abstraction of the relationship between promoter sequence and function afforded by PLS models might therefore have been too simplistic to accurately account for the complexity inherent in promoter structure, increasing the probability of prediction errors (Meng & Wang, 2015). To address this possible deficiency, non-linear Artificial Neural Networks (ANNs) were applied to training promoter sequence-function models.

4.1.1 Artificial Neural Networks and Partition Modelling

Artificial Neural Networks (ANNs) provide an alternative strategy for training promoter sequence-function models that can avoid the potential issues caused by the linearity of PLS models. The term “ANN” is somewhat of a catch-all for a family of algorithmic variants, although the underlying principles remain generally the same (Buscema *et al.*, 2014). Based around a rudimentary model of a mammalian brain, ANNs feed linear combinations of input matrices, *X,* into hidden layers consisting of multiple nodes (Figure 4.1).

At each node of the hidden layer, non-linear functions (known as activation functions) are applied to the input data. A linear combination of the hidden nodes is subsequently used to map the data to either additional hidden layers, or to an output layer containing the response matrix, *Y* (SAS Institute Inc, 2016b). During this training phase, the strength of the connection between nodes, known as the connection weight, is calculated in order to obtain a model which best describes the design space of interest (Prieto *et al.*, 2016). In addition to optimising model weights, the number of nodes and hidden layers can be varied during the training process, as can the activation function personality. Training therefore results in a learning process that maps the input data to predictions of response for the system under investigation (Buscema *et al.*, 2014).

The JMP software fits ANNs using a multilayer perceptron algorithm and backpropagation. During training, the model weights are initially assigned normally distributed random starting values. Predictions of *Y* are made from this initial network, and cross-validation (CV) statistics are calculated on a validation data set. The model parameters are then systematically altered by gradient descent in order to optimise model performance. When altering the model weights no longer improves the CV statistics (*i.e.* when altering the model weights no longer reduces prediction error), model fitting is stopped (Gotwalt, 2011). Finding a combination of weights that globally minimises prediction error is likely to result in an ANN that is overfit to the training data (Hastie *et al.*, 2009). A penalty term is therefore applied. Penalty terms aim to constrain the

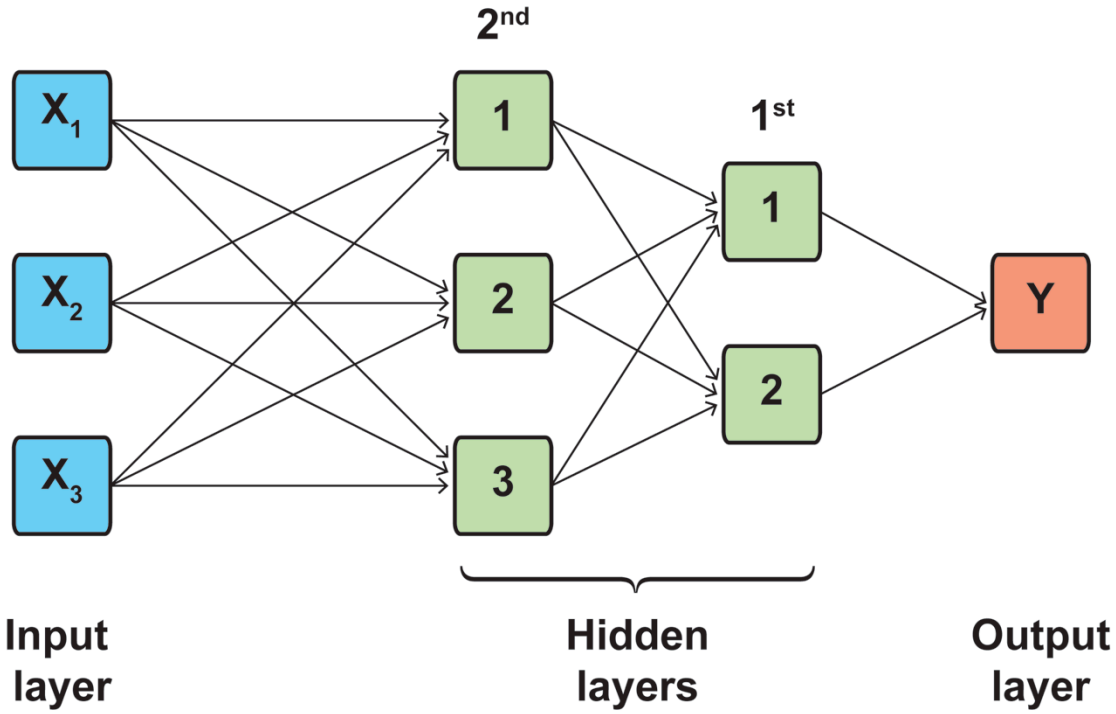


Figure 4.1: Schematic representation of an Artificial Neural Network.

Each layer consists of multiple nodes. Nodes are represented by the individual boxes. Hidden layers are numbered, with the 1st hidden layer being closest to the output layer.

Input data matrices (shown in blue) are linearly combined and fed into the 2nd hidden layer. At each node in the hidden layer (shown in green) a non-linear transformation of the input data is applied. Once calculated, the results of these transformations are linearly combined and fed into the 1st hidden layer, where more non-linear transformations are applied. The results from the 1st hidden layer are then linearly combined and fed into the output layer (shown in red), which gives a prediction of *y* from the values of *x* given in the input matrix.

In this example, two hidden layers and a single *y* value are shown, but this is arbitrary; in theory, any number of hidden layers with any number of nodes can be combined to map the response surface of interest.

model fit process so that weights do not converge to excessively large values (Setiono, 1997). Once the learning process is complete, a test set of data that is completely unknown to the ANN is applied in order to quantify network predictive performance (Pasini, 2015).

The structure and inherent flexibility of ANNs potentially affords a high degree of predictive power. Theoretically, ANNs have universal approximation capability; given sufficient training data, hidden layers and nodes, any response can be predicted to any accuracy (Hornik, 1989, SAS Institute Inc, 2016b).

However, whilst the non-linearity of ANNs can potentially provide a more accurate mathematical abstraction of the promoter design space than a PLS modelling approach, interpretation of the model output is confounded by the inherent complexity of the ANN hidden layers. Whereas the weights and coefficients of PLS models can be readily interpreted to provide a measure of the contribution of individual nucleotides to promoter activity, analysis of the underlying structure of ANNs does not provide readily interpretable information about the system being modelled (Sjöberg *et al.*, 1995). As such, the potential increase in predictive power afforded by ANNs comes at the expense of a reduced insight into promoter structure and the relationship between DNA sequence and function, as can potentially be provided by PLS models.

4.1.2 A Design of Experiments approach to ANN design.

Backpropagation and CV maximise the performance of individual networks, not network structure; activation function personality, the number of hidden layers and the number of nodes in each hidden layer must be defined. The JMP software has the capability of fitting ANNs with either one or two hidden layers, each containing any number of nodes. The nodes can run one of three activation functions: Gaussian, Linear or TanH. Furthermore, three penalty functions (Absolute, Squared or Weight Decay) are also available (SAS Institute Inc, 2016b).

Given the multi-factorial nature of the optimisation problem posed by ANN design, a statistical Design of Experiments (DoE) approach was applied. DoE systematically alters multiple variables simultaneously, allowing multi-dimensional design spaces to be efficiently explored (Lendrem *et al.*, 2015b). Statistical models are subsequently applied to model the experimental response surface to identify the optimal settings for each variable (Kumar *et al.*, 2013). DoE aims to avoid the pitfalls of more classical One Factor at a Time (OFAT) experimentation, in which the final result may vary depending on the starting point used for each variable, and in which the true optimum combination of variables may be missed (Tye, 2004, Lendrem *et al.*, 2015a).

In this instance, ANN parameters were defined as variables in the DoE, with the R^2 and Root Average Squared Error (RASE) values that were returned when ANNs were applied to a test data set serving as responses. In each instance where DoE was applied to ANN design, the DoE custom design platform in the JMP software was used to define a group of 20-30 ANN architectures. The results of these initial architectures were then analysed using standard least squares and partial least squares statistical modelling to identify the combination of variables that was predicted to maximise R^2 value and minimise the RASE value returned when candidate models were applied to the test data.

4.1.3 Dimensionality reduction

Although ANNs can potentially provide a more precise mathematical abstraction of the promoter design space than PLS models, the amount of data required to provide a robust training data set for neural models is possibly restrictive. When the system under investigation is complex, ANNs typically perform poorly if the training data set contains a small number of observations (Bataineh & Marler, 2017). If robust ANN models of *Geobacillus* promoters were to be obtained, the training data set therefore needed to be either expanded to contain more observations (*i.e*. more characterised promoter sequences) or reduced in terms of number of *x* variables (*i.e*. the number of nucleotide positions within promoter sequence being modelled).

In the case of PLS models, high-dimensional design spaces (*i.e*. those with many *x* variables) are modelled by incorporating regression and dimensionality reduction through the extraction of latent variables (LVs) from the initial training data (Boulesteix & Strimmer, 2006). However, ANNs contain no such inherent mechanism for reducing the dimensionality of the design space (Tobias, 1995). Feature selection techniques were therefore required to identify sub-regions of the promoter design space which contained the most relevant information, thereby maximising the predictive power of downstream sequence-function models, and reducing the risk of overfitting (Liu *et al.*, 2017).

Partition, or decision tree, modelling was the method by which the dimensionality of the promoter design space was reduced. By applying algorithms that fit binary decision trees through recursive partitioning, partition models provide a powerful classification technique that can also be applied to data discovery (SAS Institute Inc, 2016b). Partitioning of the training data set allows the relationship between a response variable and a set of factors to be described without the use of a mathematical model (Figure 4.2) (Baltagi & Kussener, 2014).

During partition model training, a randomly selected portion of the data set is split into groups that differ maximally in terms of the response of interest. For example, the maximum difference in fluorescence output from two groups of promoters might be obtained by splitting the training set into a group of promoters which contain guanine residues at the -15 position, and another group where adenine, cytosine or thymine residues are present at -15 (Figure 4.2B).

The resulting sub-groups can be further split, resulting in the formation of a tree-like structure (Figure 4.2C). The process is then repeated multiple times on different randomly selected portions of the original training data set, so that a "forest" (Ho, 1995) of decision trees is formed (Figure 4.2D-E). Across the entire forest, the more times a given factor causes a split in the data set, the better that factor is predicted to be at explaining variation in the response of interest.

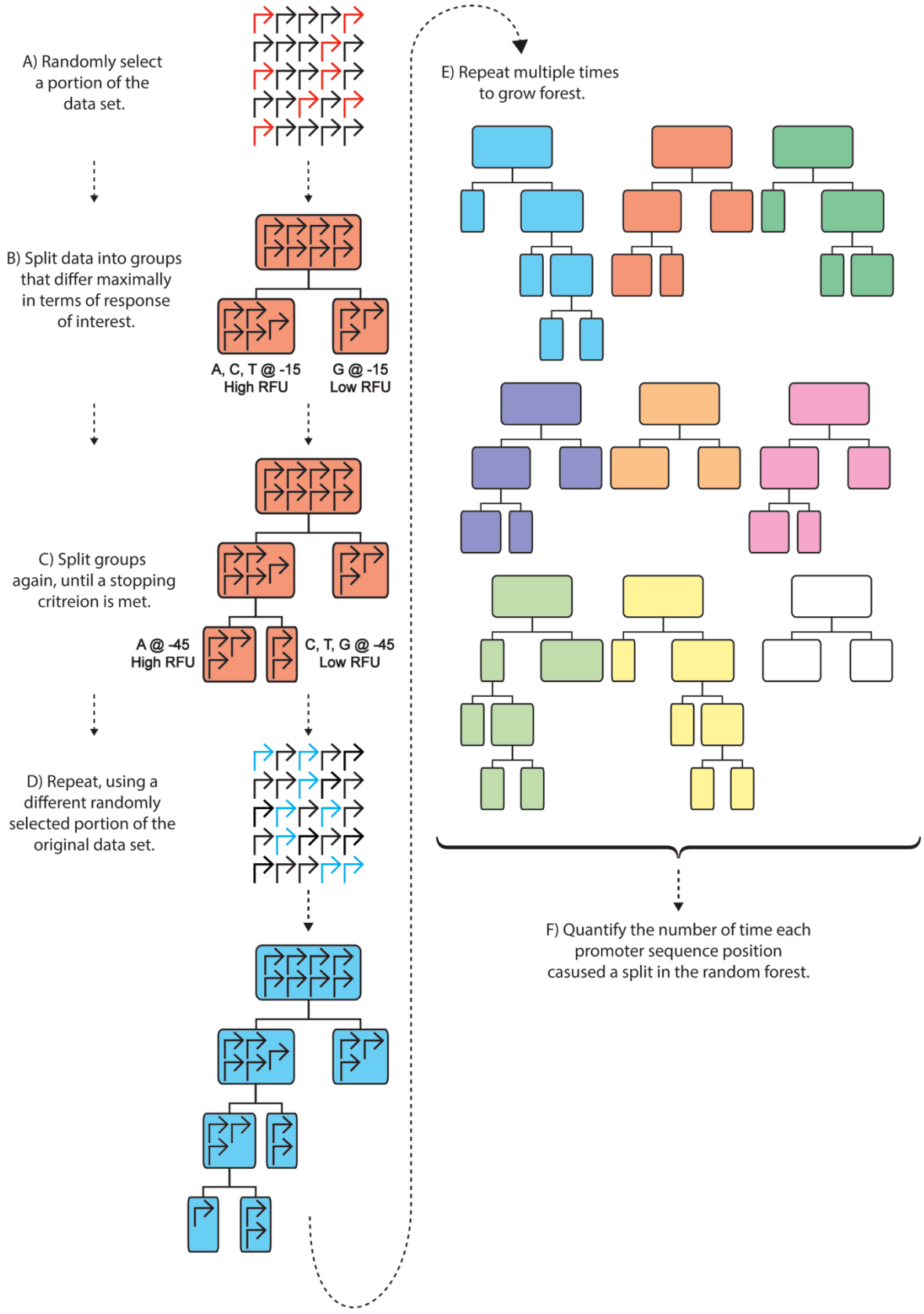


Figure 4.2: Schematic representation of a random forest partition model, as applied to promoter sequences.

By training multiple trees that are all randomly different from one another, the correlation between individual tree predictions is reduced. Individual trees risk producing splits that are overfit to the training data and which therefore display inadequate generality and poor explanatory power when applied to novel data points. However, given the Law of Large Numbers, increasing the total number of trees and averaging their predictions reduces the likelihood of overfitting (Breiman, 2001), resulting in improved generality and increased robustness, especially if the input data is noisy (Criminisi *et al.*, 2012).

Random forest algorithms were suitable for application to the promoter data set as the models are non-parametric and make no assumptions about the underlying distribution of the data being analysed. Additionally, random forests can be applied in instances when the number of *x* variables (in this case promoter sequence position and nucleotide) is greater than the number of observations, *y* (GFP fluorescence) (Manilich *et al.*, 2011). Promoter sequence positions causing large numbers of splits in a random forest were predicted to explain a greater amount of the observed variance in fluorescence output than those positions causing fewer splits.

Downstream promoter sequence-function models, using either ANN or PLS algorithms, were subsequently restricted to modelling only those sequence positions that were identified as important by partition models. In this way, promoter strength was modelled on a design space of reduced dimensionality, potentially allowing models of improved predictive power to be obtained.

The strategy of training models based on a reduced portion of the 100 bp promoter DNA sequence was supported by the results of the PLS modelling discussed in Chapter 3. The optimum PLS model that was obtained suffered from a lack of predictive power when applied to synthetic promoter sequences that were not part of the training data (Figure 3.9), but did return a good fit of the training data set. 80.251% of the variation observed in the empirically measured GFP fluorescence levels (*Y*) was explained using only 8.573% of the variation in the promoter sequence matrix (*X*). This result suggested that only a

small fraction of the total promoter sequence was responsible for the majority of the observed fluctuation in GFP output.

Provided the partition models accurately identified key sequence positions, sequence-function models that did not use the entire 100 bp promoter should therefore have been capable of satisfactorily explaining the majority of empirically observed fluctuations in GFP activity.

4.1.4 Model Averaging

An additional strategy that was applied to improve model predictive power was model averaging. Although single "best" performing models are most often presented in support of conclusions, this approach falsely assumes that only one model explains the data (Clyde, 2002). Model averaging is analogous to calculating a mean value of a continuous measurement of interest from biological replicates in order to obtain a measure with less variance to the true or ideal value. As long as incorrect predictions are in the minority and model stability (*i.e.* the tendency for predicted values to change based on alterations to the training data) is not unduly affected, model aggregation should improve the accuracy of the final predicted value (Beleites & Salzer, 2008).

When applied to ANNs, the process of model averaging by combining multiple individual networks is known as "ensembling" (Hansen & Salamon, 1990). By training a series of networks on the same task and then combining the outputs of these ANNs, ensembling attempts to exploit information about the design space that was captured by ANNs that might otherwise have been judged redundant if only a single best performing ANN were isolated (Sharkey, 1996). Theoretically, individual members of the ensemble can counteract deficiencies in other members, thereby improving the generalisation performance of the ensemble as compared to the performance of the individual constituent networks (Yang *et al.*, 2013).

Famously, “all models are wrong but some are useful” (Box & Draper, 1986); by combining multiple “useful” models, aggregation aims to decrease the degree to which model predictions are “wrong”.

4.1.5 Sponsor mandated change in host organism

The promoter characterisation experiments described in Chapter 3 utilised *G. thermodenitrificans* as the host organism. However, the transformation efficiency of *G. thermodenitrificans* was low, hindering the high-throughput screening of potential regulatory sequences and other genetic parts of interest. Research performed by a collaborator, the industrial biotechnology company ZuvaSyntha Ltd. (Hertfordshire, United Kingdom), suggested that *G. thermoglucosidans*, a close relative to *G. thermodenitrificans*, was more amenable to transformation than *G. thermodenitrificans*. A change in host organism was therefore mandated by the industrial sponsor (Shell Research Ltd.) As such, *G. thermoglucosidans* was used as the chassis organism for all subsequent promoter characterisation.

4.1.6 Deriving sequence-function models with improved predictive power

To expand the training data set beyond that reported in Chapter 3, three progressively larger sets of putative promoter sequences, termed A, B and C, were characterised in *G. thermoglucosidans*. Empirical data derived from these characterisation experiments were used to derive sequence-function models, using ANN, PLS and random forest models. Once trained and validated, these models were applied to the generation of synthetic promoter sequences and to predicting *pre hoc* the promoter activity of bioinformatically identified putative *Geobacillus* promoter sequences that had not been previously characterised. The accuracy of these predictions was subsequently assessed *in vivo*.

4.2 Results

4.2.1 Characterisation and modelling of data set A

Characterisation of putative promoter sequences in G. thermoglucosidans

To comply with the change in host organism mandated by the industrial sponsor (Shell Research Ltd.), the 31 putative promoter sequences that were characterised in *G. thermodenitrificans* and discussed in Chapter 3 were re-characterised in *G. thermoglucosidans*.

Additionally, to increase the proportion of the promoter design space explored by the empirical data set, putative promoters that had not previously been characterised in *G. thermodenitrificans* were selected for synthesis. 100 bp sequences were randomly selected, one from each of the 13 promoter phylogeny clades (Figure 3.2), so that a total of three putative promoters from each clade of the *Geobacillus* promoter phylogeny were empirically characterised. The exception was clade seven, as three sequences from this clade had already been selected in the first modelling iteration. Of the 12 sequences that were selected, one putative promoter could not be synthesised. A total of 11 previously uncharacterised putative promoters were therefore added to the *Geobacillus* promoter library, giving a total of 42 sequences.

Three starter cultures arising from independent transformation events for each *promoter::GFP* fusion were initially characterised. If the mean fluorescence of a given promoter was greater than that of the positive control, the *G. thermodenitrificans ldhA* promoter, a further six independent transformants of that promoter were characterised.

As measured by GFP fluorescence, the 42 characterised sequences displayed a total expression range of 126-fold (Figure 4.3). 14 sequences had a significantly greater mean fluorescence than the negative control, as measured by ordinary one-way ANOVA with Dunnett's multiple comparisons test. These 14 sequences covered a range of expression of 5-fold.

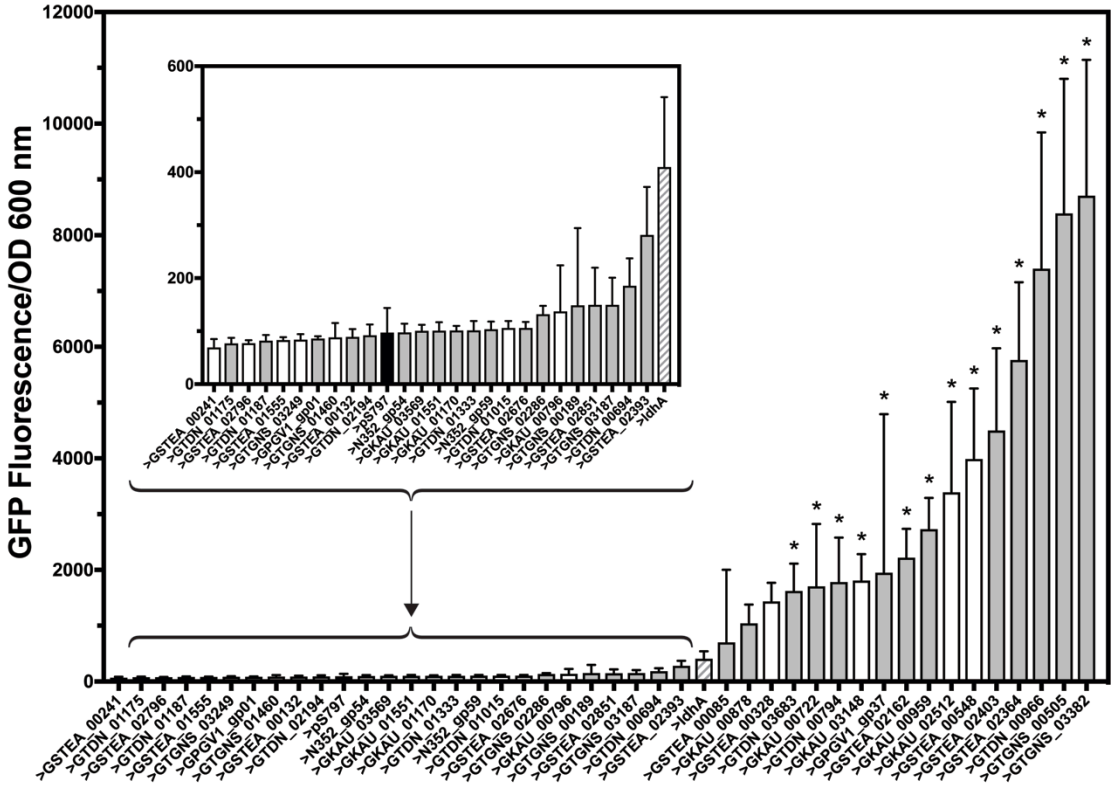


Figure 4.3: Putative promoters characterised upstream of GFP in *G. thermoglucosidans.*

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. Promoters selected during the first iteration of characterisation performed in *G. thermodenitrificans* are shown in grey. Promoters selected at random for the second iteration of characterisation experiments are shown in white. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the hatched bar. The negative control, *G. thermoglucosidans* transformed with an empty pS797 vector, is highlighted in black. Bars represent the mean of $3 \leq n \leq 9$ starter cultures arising from independent transformants, except in the case of the two controls, where $n = 23$. Standard deviation error bars are shown, unless hidden by the bar. Promoter sequences with mean relative fluorescence values that were statistically significantly different from the negative control are labelled with an asterisk. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, using a significance level of 0.05.

In Chapter 3, promoters were defined as "active" if their mean fluorescence output was statistically significantly greater than that of the negative control. However, this threshold was considered too harsh for partitioning active and non-active sequences in *G. thermoglucosidans*. Indeed, when characterised in *G. thermoglucosidans*, the *G. thermodenitrificans ldhA* promoter was judged to not be statistically significantly different from the negative control (adjusted P-value = 0.999).

However, the *ldhA* promoter was clearly active; *G. thermoglucosidans* cultures expressing GFP under the control of the *ldhA* promoter were fluorescent when observed using a blue light transilluminator with an amber filter (Figure 4.4). Using the *ldhA* promoter as the lower threshold for determining active promoters resulted in 17 sequences, covering a 22-fold range of expression levels, being defined as active.

In addition to the library of bioinformatically identified putative promoters, the 12 synthetic putative promoter sequences that were discussed in Chapter 3 were also re-characterised in *G. thermoglucosidans*. None of the 12 sequences showed fluorescence output that was statistically significantly different from the negative control. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test at the 0.05 significance level.

Modelling the relationship between promoter DNA sequence and function

The lack of predictive power of the first PLS model iteration (Chapter 3) was hypothesised to be, in part, a result of an over-representation of inactive promoters in the training data set. PLS models were therefore trained on only those promoter sequences with a mean fluorescence output that was greater than that of the *G. thermodenitrificans ldhA* promoter. Four promoter sequences with a lower mean fluorescence output than the *ldhA* promoter, GSTEA_02851, GSTEA_02393, GTGNS_00189 and GTGNS_03187, were also included in the training data set to provide examples of DNA sequences with no acceptable levels of promoter activity. Within the training data set all four DNA nucleotides were represented at 95% of the promoter sequence positions. The -9, -10, -11,

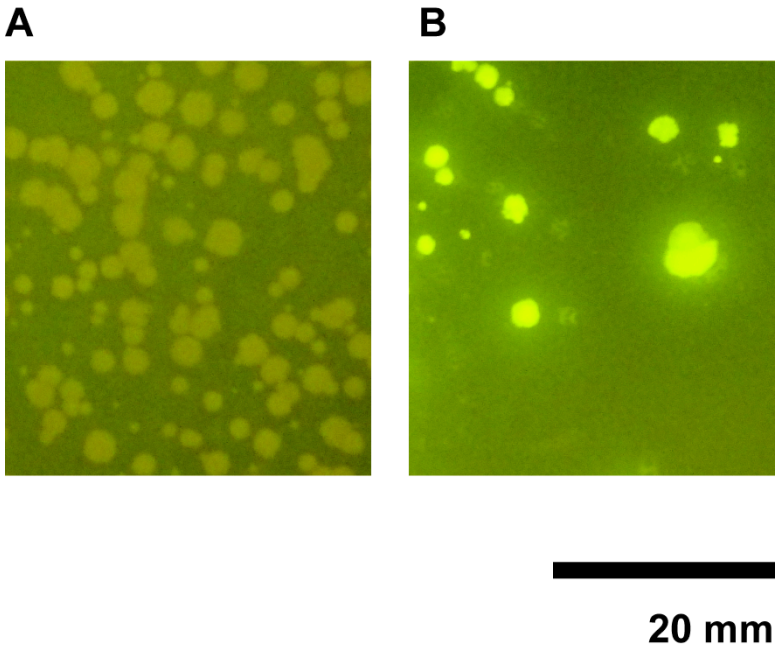


Figure 4.4: *G. thermoglucosidans* transformants cultured on mLB agar.

Panel A shows *G. thermoglucosidans* transformants containing the empty vector pS797 (no visible fluorescence). Panel B shows *G. thermoglucosidans* transformants expressing GFP under the control of the *G. thermodenitrificans ldhA* promoter (visible fluorescence). *G. thermoglucosidans* was incubated at 55 °C for 72 h. Cultures were illuminated using a blue light transilluminator with an amber filter. Images were taken using a Panasonic DMC T235, with ISO 800 sensitivity. F-stop was set to 4.5. Exposure time was 0.25 s.

-34 and -36 positions lacked a cytosine residue, and the -11 position also lacked a thymine residue. The training data set therefore contained 394 *x* variables.

For the first PLS model (Chapter 3), measures of GFP activity when fused to each promoter were normalised to a measurement of fluorescence of GFP under the control of the *G. thermodenitrificans ldhA* promoter, cultured on the same 96-well plate as the promoter of interest. Normalisation was intended to account for any batch effects introduced by technical sources of variation between measurements of biological replicates. However, the normalisation process may instead have resulted in an inaccurate quantification of promoter activity.

As an example, Figure 4.5 shows replicate measurements of both the *ldhA* promoter and the promoter GSTEA_02393, grown from starter cultures arising from independent transformation events in two separate 96-well plates. The fluorescence activity of the *ldhA::GFP* fusion differed by approximately two-fold between biological replicates, whilst expression from *GSTEA_02393::ldhA* remained broadly consistent. Normalisation therefore resulted in a quantification of GSTEA_02393 promoter activity that varied by approximately 3.6-fold between biological replicates, a difference that was not representative of the empirical fluctuation in GSTEA_02403 activity. All subsequent promoter sequence-function models were therefore trained on un-normalised fluorescence data.

PLS models have previously been shown to return improved predictive performance when the modelled data set contains few outliers, and when the modelled data have a somewhat symmetrical distribution (Cox & Gaudard, 2013). Therefore, to remove any outlying points from the training data set, any fluorescence measurements for a given promoter that fell outside of one standard deviation of the mean for that promoter were excluded (Figure 4.6A). Additionally, to provide a symmetrically distributed training data set, the data were logarithmically transformed (Figure 4.6B & C), resulting in a training data set that was bi-normally distributed.

A total of 136 *y* values were available for model training. 20% of these *y* values were randomly selected and withheld from model training to serve as an independent test set for determining model predictive accuracy. A total of seven PLS models were constructed. The models varied in terms of the number of putative promoters included in the training data set, which of the PLS algorithms and validation methodologies were applied, and the number of Latent Variables (LVs) that were extracted.

An initial PLS model trained on the *G. thermoglucosidans* data set (hereafter referred to as PLS_iteration_A_1) utilised the NIPALS algorithm with holdback cross-validation. 80% of the available data were used for model

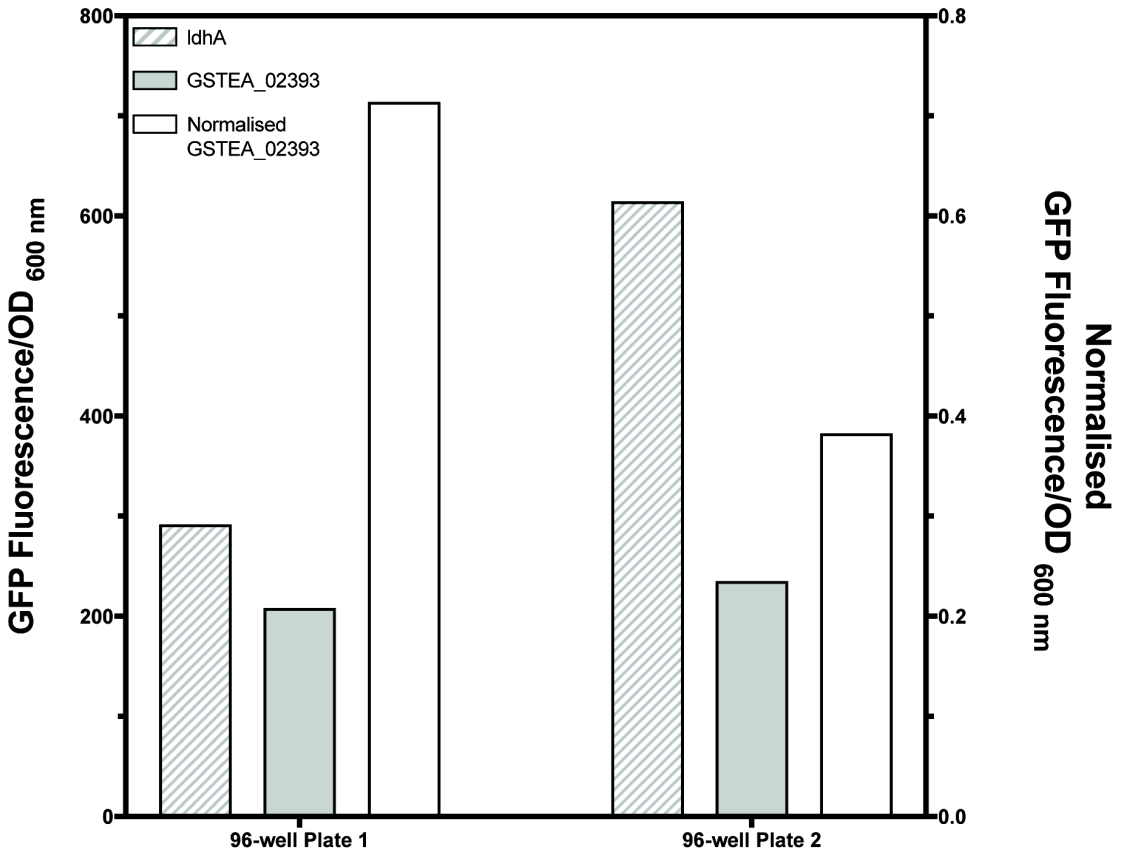


Figure 4.5: The effect of normalising promoter activity measurements to the *G. thermodenitrificans ldhA* promoter.

Fluorescence and absorbance measurements after 24 h incubation. The hatched bars represent the *ldhA* promoter. The grey bars represent the raw fluorescence output of GFP under the control of the promoter GSTEA_02393. The white bars, plotted on the right-hand y-axis, represent the fluorescence output of *GSTEA_02393::GFP*, normalised to the fluorescence output of *ldhA::GFP*.

training, with the remaining 20% used for validation. The resulting model extracted 15 LVs from the data, and explained 76.971% of the variation observed in promoter sequence (*X*) and 97.612% of the observed variation in promoter fluorescence output (*Y*). 225 out of the 394 *x* variables exceeded the VIP threshold of 0.8 (Eriksson *et al.*, 2006), suggesting that approximately half of the *x* variables were having a significant effect on model output. In an attempt to increase model parsimony, a second PLS model was trained using only the 225 statistically significant *x* variables. However, the resulting PLS model

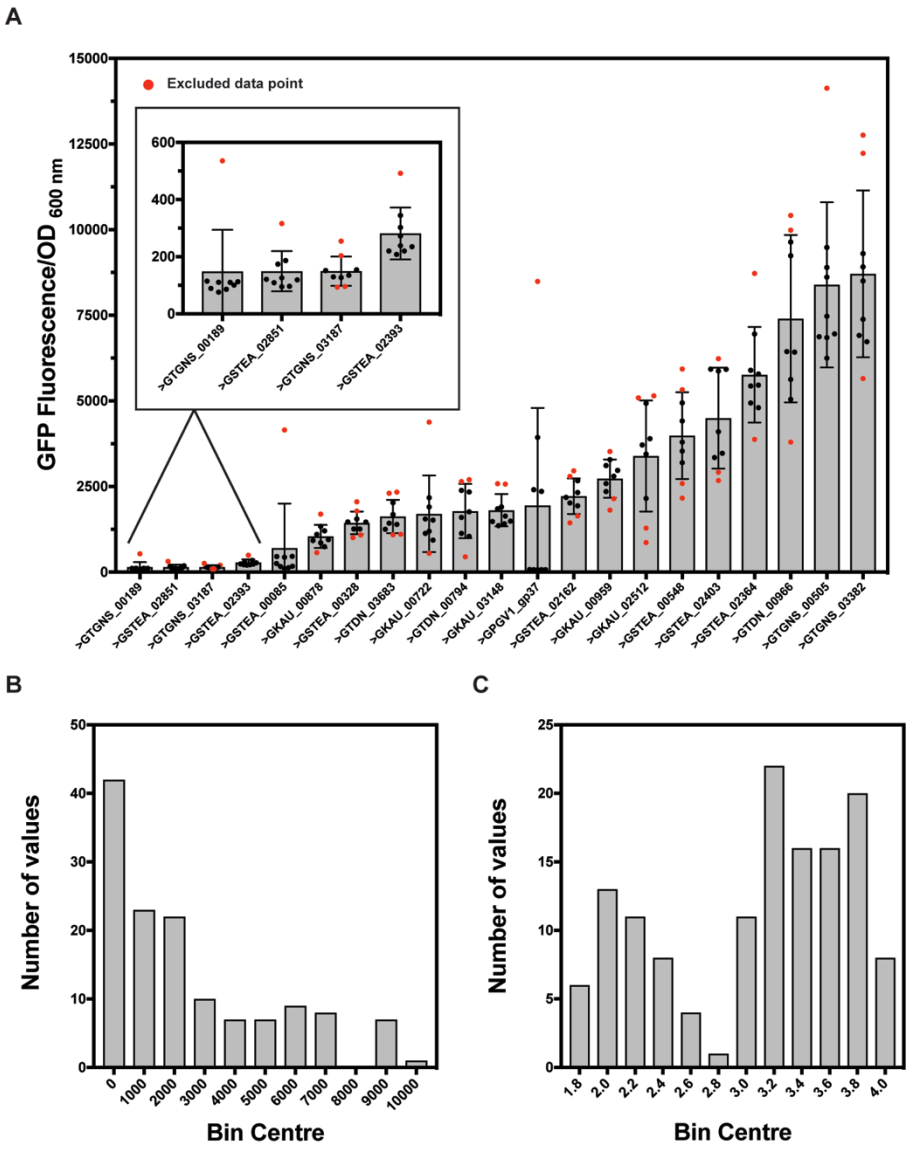


Figure 4.6: Data transformation for the first iteration of Partial Least Squares modelling in *G. thermoglucosidans*.

A)) GFP expression levels of the promoter sequences included in the PLS training data set. Bars represent the mean fluorescence of 9 starter cultures arising from independent transformation events, with standard deviation error bars shown unless hidden by the bar. Points represent fluorescence measurements of individual starter cultures. Points are highlighted red when they fall outside of 1 standard deviation of the mean for the given promoter. Data points that are coloured red were excluded from the PLS training data set.

Histograms show the distribution of expression levels in the PLS training data set. B) shows the distribution of raw fluorescence data. C) shows the distribution of the data once a log(10) transformation had been performed.

extracted the same number of LVs as PLS_iteration_A_1 and did not show an improvement in the explained proportion of variation in promoter fluorescence output (97.611%, compared to 97.612% from PLS_iteration_A_1).

A PLS model was also trained using the SIMPLS variant of the PLS algorithm, again using holdback cross-validation with an 80%-20% split of training to validation data. However, the SIMPLS model did not return an improvement in statistical power as compared to the model trained using the NIPALS algorithm. This was consistent with previous observations that there is no significant difference between the performance of the two algorithms when *Y* is univariate (de Jong, 1993).

Although PLS_iteration_A_1 was shown to provide an optimal fit to the training data set as shown in Figure 4.6, analysis of the model suggested that predictive power could be further improved by removing the promoter sequence GPGV1_gp37 from the training data set (Figure 4.7). One measure of the overall impact a particular variable has on model performance is the Euclidean distance of that value from the model origin. If particular *y* values are outliers in comparison to the majority of *y*, then the outlying values may be having an unduly large effect on model output (Cox & Gaudard, 2013).

Figure 4.7A showed that measurements of GFP fluorescence under GPGV1_gp37 did not cluster with the majority of the training data set in terms of distance from the Y model. Additionally, the comparatively large residual values observed for GPGV1_gp37 as compared to the rest of the training data set (Figure 4.7B) were indicative of poor predictive accuracy for this promoter sequence.

The model PLS_iteration_A_1 may have assigned undue weight to promoter sequence GPGV1_gp37 due to the large variation in GFP expression that was observed between biological replicates expressing this promoter (Figure 4.7C). Of the eight biological replicates included in the model, five displayed no GFP expression, with the remaining three replicates resulting in expression levels between six and ten times greater than that of the

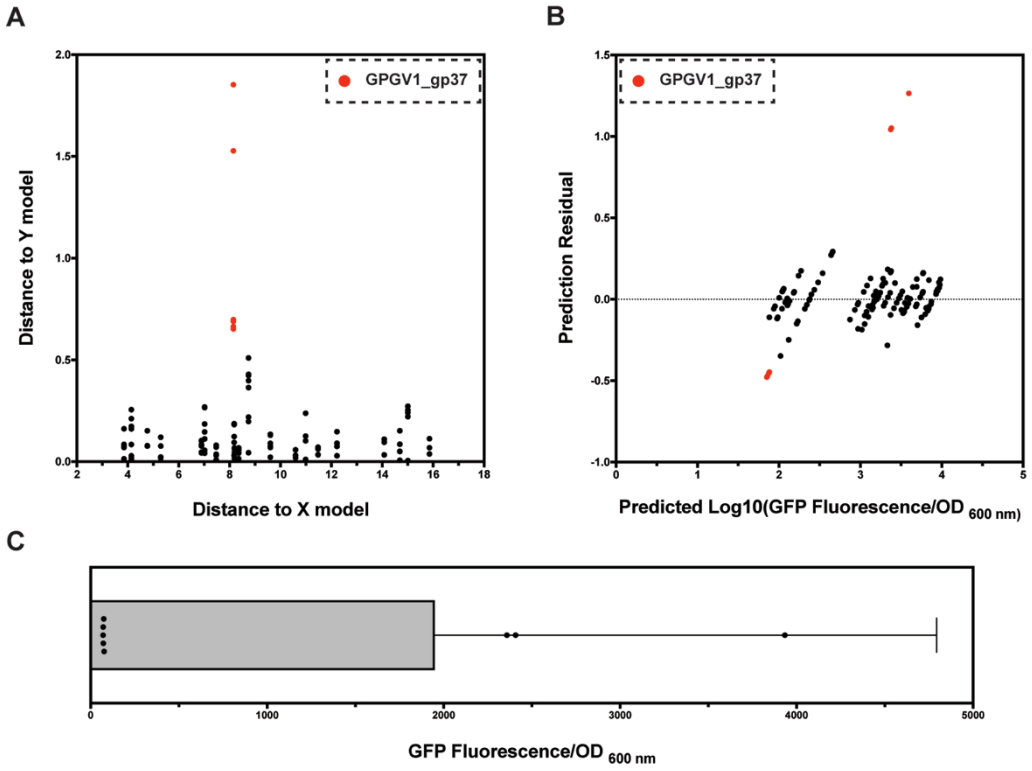


Figure 4.7: The effect of including promoter sequence GPGV1_gp37 on the model PLS_iteration_A_1.

A) the Euclidean distance of individual measurements of promoter activity from the *X* and *Y* models . B) Fluorescence output predicted by the PLS model plotted against prediction residual (the difference between empirically measured and predicted fluorescence values). Dashed line shown at the point where model residual is equal to 0. In both A & B, measurements of GFP output from GPGV1_gp37 are highlighted in red.

C) Empirically measured fluorescence output of GFP under the control of GPGV1_gp37. Points represent individual biological replicates from individual starter cultures, the bar represents the mean of n = 8 replicates, with standard deviation error bar shown.

G. thermodenitrificans ldhA promoter. Across the remainder of the training data set, whilst promoters did display variation in GFP expression between biological replicates, (Figure 4.6), the apparent Boolean “on” or “off” nature of GPGV1_gp37 was not observed from other promoter sequences. Subsequent PLS models were therefore trained on a training data set from which all measurements of GPGV1_gp37 had been removed.

PLS models were trained on the data set from which GPGV1_gp37 had been removed using the NIPALS algorithm and *K*Fold cross-validation. To observe the effect of *K* size on model output, three individual PLS models were initially trained, using *K* values of 3, 7 and 10 respectively. Of the three PLS models obtained, the model that explained the greatest amount of the observed variation in promoter fluorescence output used *K*= 7. That model is hereafter referred to as PLS_iteration_A_2. 14 LVs were extracted from the data, and the model explained 75.201% of the variation observed in promoter sequence (*X*) and 93.404% of the observed variation in promoter fluorescence output (*i.e.* activity; *Y*). 235 *x* values returned VIP values that exceeded the 0.8 threshold. A model trained using only those 235 *x* values explained 6.187% more of the variation observed in *X* and 0.01% more of the variation observed in *Y* than PLS_iteration_A_2, but required 15 LVs to do so. Given the relatively minor differences in explanatory performance between the two models and the more parsimonious nature of PLS_iteration_A_2, the model PLS_iteration_A_2 was selected for further interrogation.

When applied to the CV data set, PLS_iteration_A_2 displayed good predictive power. A strong positive correlation was seen between empirically measured fluorescence values and values predicted by the model; a linear regression of the data had an R^2 value of 0.964 (Figure 4.8A). Additionally, no significant patterns were observed in the model residuals (Figure 4.8B), which was suggestive of there being no underlying structural biases in the model (Cherkasov *et al.*, 2014). Of the 27 *y* values in the independent test set, 59% had positive residuals, indicating that PLS_iteration_A_2 tended slightly to over-predicting fluorescence values.

Analysis of the distribution of the model residuals by Shapiro-Wilk W test showed that there was insufficient evidence to reject the null hypothesis that the underlying distribution of the residuals was normal, at a significance level of 0.05 (W = 0.948, Prob<W = 0.187) (Figure 4.8C). However, visual inspection of a histogram of the model residuals questioned this conclusion, as the residuals did not appear normally distributed, with a visible skew towards larger residuals (Figure 4.8C). Previous studies have shown that the Shapiro-Wilk test can be

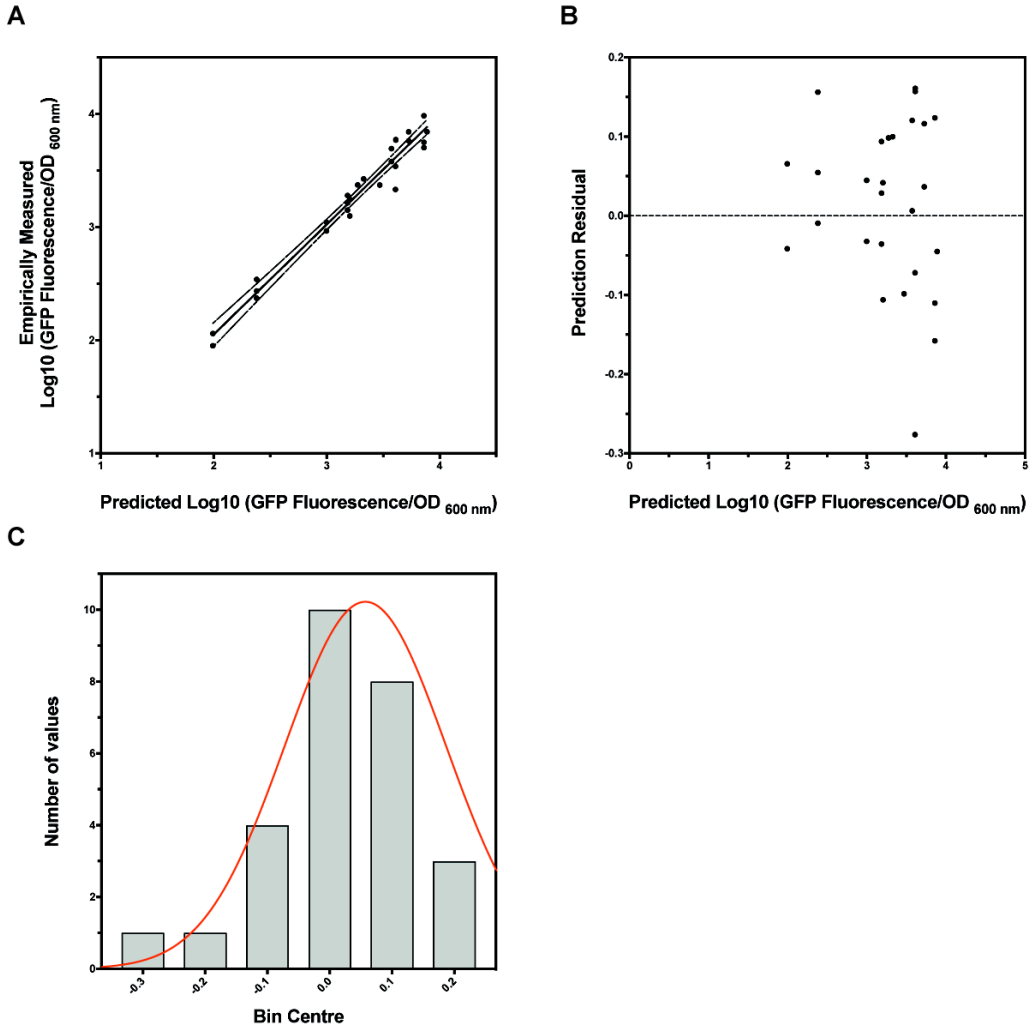


Figure 4.8: Partial Least Squares model PLS_iteration_A_2 diagnostics.

A) Empirically measured, Log10 transformed promoter fluorescence output, plotted against the Log10 transformed fluorescence predicted by the model. The solid line represents a linear regression of the data, with 95% confidence limits shown by dashed lines. The R^2 value of the linear regression was 0.964.

B) Log10 transformed fluorescence predicted by the model plotted against the prediction residual. The dashed line is shown at the point where the prediction residual is equal to 0.

C) Histogram of model residual distribution. A Shapiro-Wilk W test showed that there was insufficient evidence to reject the null hypothesis that the underlying distribution was normal (W = 0.948, Prob<W = 0.187). The red line represents a Gaussian distribution of the data, as rendered by Prism software. The R^2 value of the curve was 0.9807.

prone to type II errors (*i.e.* fail to reject a false null hypothesis) when the sample size is small (*e.g.* n < 50) (Razali & Wah, 2011, Le Boedec, 2016). In the case of the residuals for model PLS_iteration_A_2, n was equal to 27, raising the possibility that the failure of the Shapiro-Wilk test to reject the null hypothesis was erroneous.

Generating synthetic putative promoters

Once trained and validated, PLS_iteration_A_2 was applied to the generation of synthetic putative promoter sequences. As with the synthetic sequences discussed in Chapter 3, the simulator function of the JMP software was used to generate 100 bp putative promoter sequences. The probability of a nucleotide being assigned to a given sequence position was weighted based on the distribution of nucleotides found in the training data set. Five putative synthetic promoter sequences were selected for *in vivo* characterisation. BLAST queries raised against each of the five sequences returned no hits, indicating an absence of similar sequences in the GenBank database. The GFP expression level predicted for these five sequences fell within the range of expression levels observed in the training data set (Figure 4.9A). Whilst PLS models are in theory capable of extrapolation beyond the training data set (Sanderson *et al.*, 2008), it was decided to first validate the model through interpolation.

The five putative synthetic promoter sequences were synthesised upstream of *GFP* in the pS797 vector. Additionally, to expand the library of characterised *Geobacillus* promoter sequences, five bioinformatically identified putative promoters that had not been previously characterised were selected for synthesis at random.

Measurements of fluorescence after 24 h incubation of *G. thermoglucosidans* transformants showed that none of the 5 synthetic putative promoter sequences had any *in vivo* promoter activity (Figure 4.9B); there was no significant difference between cultures expressing GFP under the control of the five putative promoter sequences and the pS797 negative control.

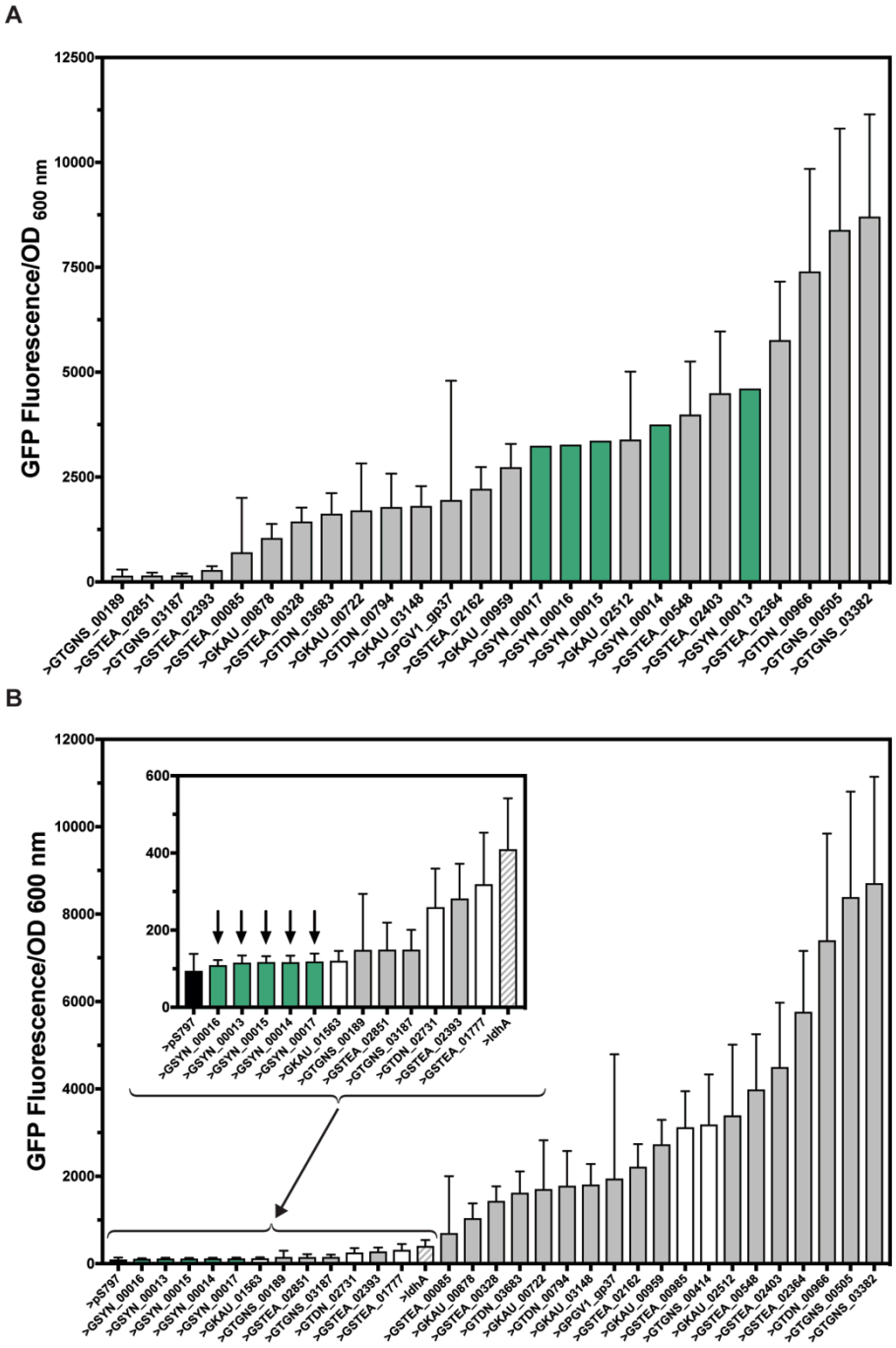


Figure 4.9: Fluorescence output of GFP under the control of synthetic putative promoters as A) predicted by the Partial Least Squares model PLS_iteration_A_2 & B) as empirically measured.

A) Grey bars represent the fluorescence output of GFP under the control of promoters as used in model training. Bars represent the mean of n = 9 starter cultures arising from individual transformation events, with standard deviation error bars shown. Predicted fluorescence outputs of GFP under the control of putative synthetic promoter sequences are shown in green.

B) The negative control, *G. thermoglucosidans* transformed with empty pS797 vector, is shown in black. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the hatched bar. Bars represent the mean of $3 \leq n \leq 20$ starter cultures arising from individual transformation events. Standard deviation error bars are shown. Synthetic putative promoters are shown in green. Previously uncharacterised natural *Geobacillus* promoters are shown in white.

Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, at a significance level of 0.05.

Sequence analysis of putative synthetic promoters

Of the 17 putative, synthetic promoters that were characterised *in vivo* (the five sequences modelled by PLS_iteration_A_2 and the 12 sequences discussed in Chapter 3), none displayed any *in vivo* promoter activity. To evaluate possible causes, a sequence logo of the 17 "active" promoters used in the training of PLS_iteration_A_2 (Figure 4.10A) was compared to a sequence logo of the 25 "inactive" characterised *Geobacillus* putative promoters, whose *in vivo* GFP output was less than that of the *ldhA* promoter (Figure 4.10B), and a sequence logo of the 17 synthetic putative promoters (Figure 4.10C). Promoter DNA sequences were aligned and visualised using WebLogo version 2.8.2 (Crooks *et al.*, 2004).

All three sequence logos showed a heavily conserved region of adenine- and guanine-rich sequence, located between 15 and 7 bp upstream of the start codon of the adjacent GFP CDS. Given the similarities of both the location and the sequence of these conserved regions to the canonical Shine-Dalgarno sequence (Shine & Dalgarno, 1974), they were hypothesised to be putative Ribosome binding sites.

Conserved regions upstream of the putative RBS were also observed in the aligned "active" promoter sequences (Figure 4.10A). Sequence motifs spanning from -36 to -32 and from -51 to -46 were hypothesised to be -10 and -35 regions, respectively. Crucially, whilst these putative consensus regions were also present in the synthetic putative promoter sequences (Figure 4.10C), they were not observed in the alignment of "inactive" promoter sequences (Figure 4.10B). Likewise, regions spanning from -71 to -65 and from approximately -90 to -80 are more heavily conserved in the "active" and synthetic putative promoter sequences than in the "inactive" sequences.

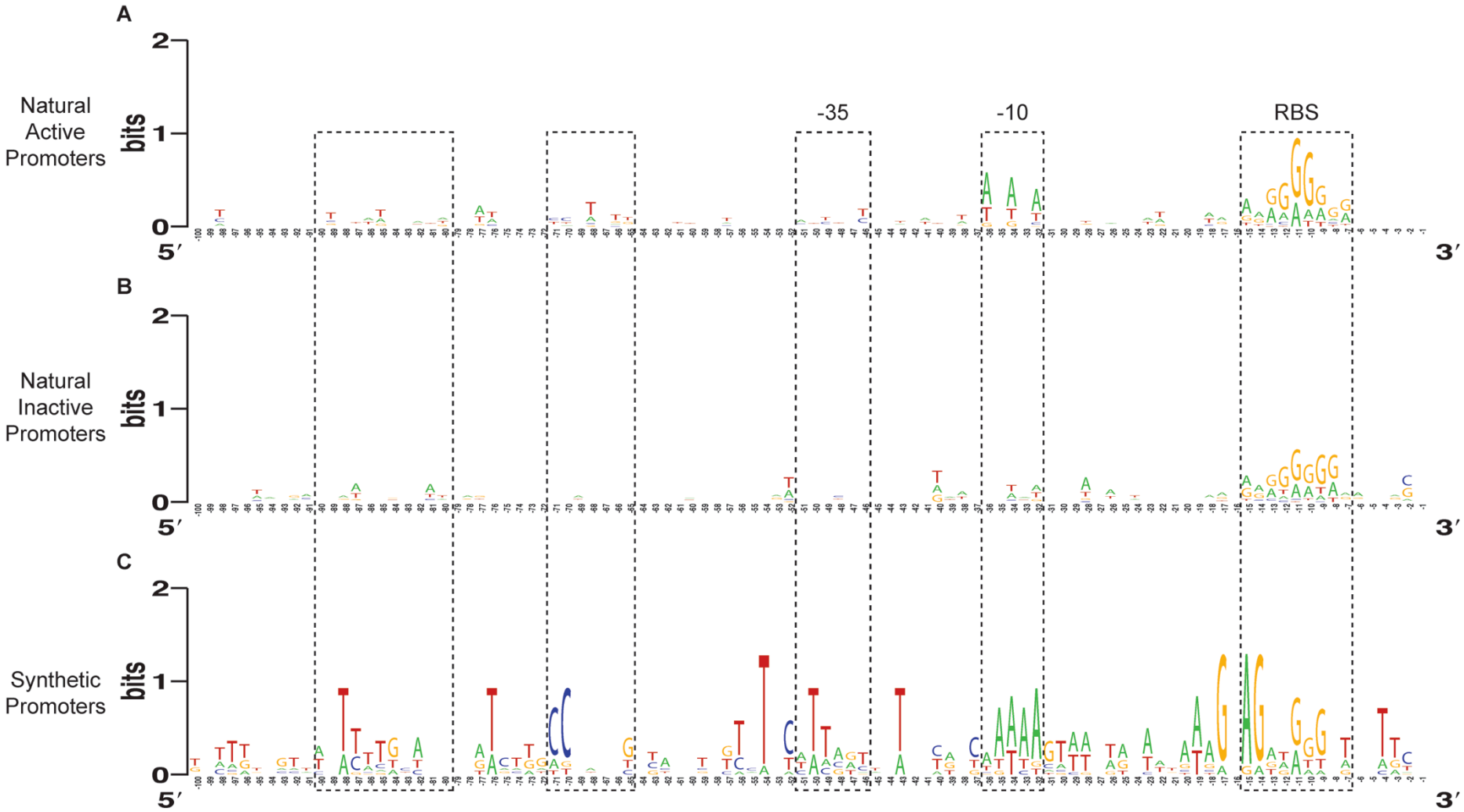


Figure 4.10: Visualisation of sequence alignments of putative promoters.

Sequence logos show promoter sequence alignments of A) "Active" B) "Inactive" and C) synthetic putative promoter sequences. The overall height of individual stacks indicates the degree of sequence conservation at a given position, and the height of nucleotide symbols indicates the conservation of each nucleic acid at that position. Position numbering is relative to the start codon of the upstream CDS. The regions of sequence discussed in the text are bounded by the dashed lines, with putative motif identification shown above. Rendered using WebLogo version 2.8.2 (Crooks *et al.*, 2004).

Motifs and individual nucleotides that were conserved in the training data set were expected to be conserved in the synthetic promoter sequences. During the generation of the synthetic putative promoter sequences, the probability of nucleotides being assigned to each position within the sequence was weighted based on the distribution of nucleotides found in the training data set. As a group, the synthetic promoter sequences were also much more heavily conserved than either of the two groups of natural putative promoter sequences (as shown by greater stack height and nucleotide symbol size in Figure 4.10C as compared to Figure 4.10 A or B).

The lack of promoter activity observed in the synthetic sequences did not therefore appear to be caused by key motifs such as the RBS being omitted from the DNA sequences. Instead, it was hypothesised that the optimum PLS model that was obtained was unable to accurately infer the contribution of a given nucleotide or motif at a given sequence position, resulting in inaccurate predictions of synthetic promoter activity.

4.2.2 Characterisation and modelling of data set B

Given the lack of predictive power of the model PLS_iteration_A_2 when applied synthetic promoter sequences, an expansion of the training data set was deemed necessary. A further 10 previously uncharacterised, bioinformatically identified putative promoter sequences were therefore selected for *in vivo* characterisation.

To exploit the structure of the promoter design space revealed by previous *in vivo* characterisation (Caschera *et al.*, 2010), additional promoters were selected at random from clades of the promoter phylogeny that had previously yielded high-performing promoters. The eight promoters that resulted in the highest GFP expression in data set A (Figure 4.6A) represented five clades of the *Geobacillus* promoter phylogeny (Table 4-1). An additional two putative promoter sequences were therefore selected at random from each of these five clades.

Clade	Sequences characterised in data set A
6	GKAU_00959, GSTEA_02403, GTDN_00966
11	GKAU_02512
13	GTGNS_00505
18	GSTEA_02364, GTGNS_03382
21	GSTEA_02403

Table 4-1: Clades of the *Geobacillus* promoter phylogeny containing strong promoter sequences, as characterised in data set A.

The 10 previously uncharacterised promoters were combined with the 26 promoter sequences shown in Figure 4.9B to form data set B. In addition to the inclusion of the additional promoter sequences, four further alterations were made to the experimental design for promoter characterisation between data sets A and B. Firstly, to assess the functional reliability of the *Geobacillus* putative promoters, sequences were also characterised upstream of a second reporter protein coding sequence, the *RFP* derivative *mOrange*. Secondly, a type IIS restriction cloning system was implemented to facilitate the routine use of characterised putative promoter elements for the control of alternate CDS. Application of the cloning strategy resulted in a 4 bp scar being introduced between putative promoter and RBS sequences. The total length of the characterised *Geobacillus* regulatory elements therefore increased from 100 bp to 104 bp. The implications of both altered genetic context and the introduction of restriction cloning scar sequences are discussed in Chapter 5.

Alterations were also made to the way in which measurements of promoter activity were entered into the training data set for data set B. For both the PLS models derived from *G. thermoglucosidans* data set A and the models discussed in Chapter 3, the training data sets contained multiple measurements of fluorescence output for each promoter sequence. Given the relatively small number of characterised promoter sequences, fluorescence measurements from biological replicates were included to increase the amount of data that was available for model training.

However, during the validation process, the data set was partitioned at random into training and CV sets. As such, identical promoter sequences were present in both the training and CV sets. Known as data twinning (Forman & Scholz, 2010), this duplication of data points may have artificially inflated the model's predictive accuracy (Raccuglia *et al.*, 2016), as the entry in the CV set could be fitted more accurately that it could have been if it were not also present in the training set (Clarke *et al.*, 2009). To avoid data twinning in subsequent modelling iterations, single mean measures of activity for each promoter sequence were modelled.

Characterisation of putative promoter sequences in G. thermoglucosidans

Of the 36 putative promoters that were initially included in data set B, two (GKAU_01563 and GSTEA_02851) contained restriction sites that were incompatible with the type IIS cloning strategy. 34 putative promoter sequences were therefore available for *in vivo* characterisation. When used to express GFP, the promoter library covered a 113-fold range of expression levels (Figure 4.11A). 22 promoters, covering a 30-fold expression range, resulted in mean GFP expression that was higher than that of the *G. thermodenitrificans ldhA* promoter. 14 promoters resulted in mean GFP fluorescence that was statistically significantly greater than the pS797 negative control at the 0.05 significance level, as determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test.

When cloned upstream of *mOrange*, two sequences, GSTEA_00342 and GTGNS_02755, could not be transformed into *G. thermoglucosidans*. The resulting promoter library therefore contained 32 sequences, and covered an activity range of 47-fold (Figure 4.11B). 14 sequences resulted in mOrange fluorescence levels that were statistically significantly different to the negative control. These 14 sequences covered an expression range of three-fold.

Little correlation was observed in promoter activity levels between the two reporter proteins (Figure 4.12). A linear regression of the data returned an R^2 value of 0.267. Eight promoter sequences, including the *ldhA* promoter, fell

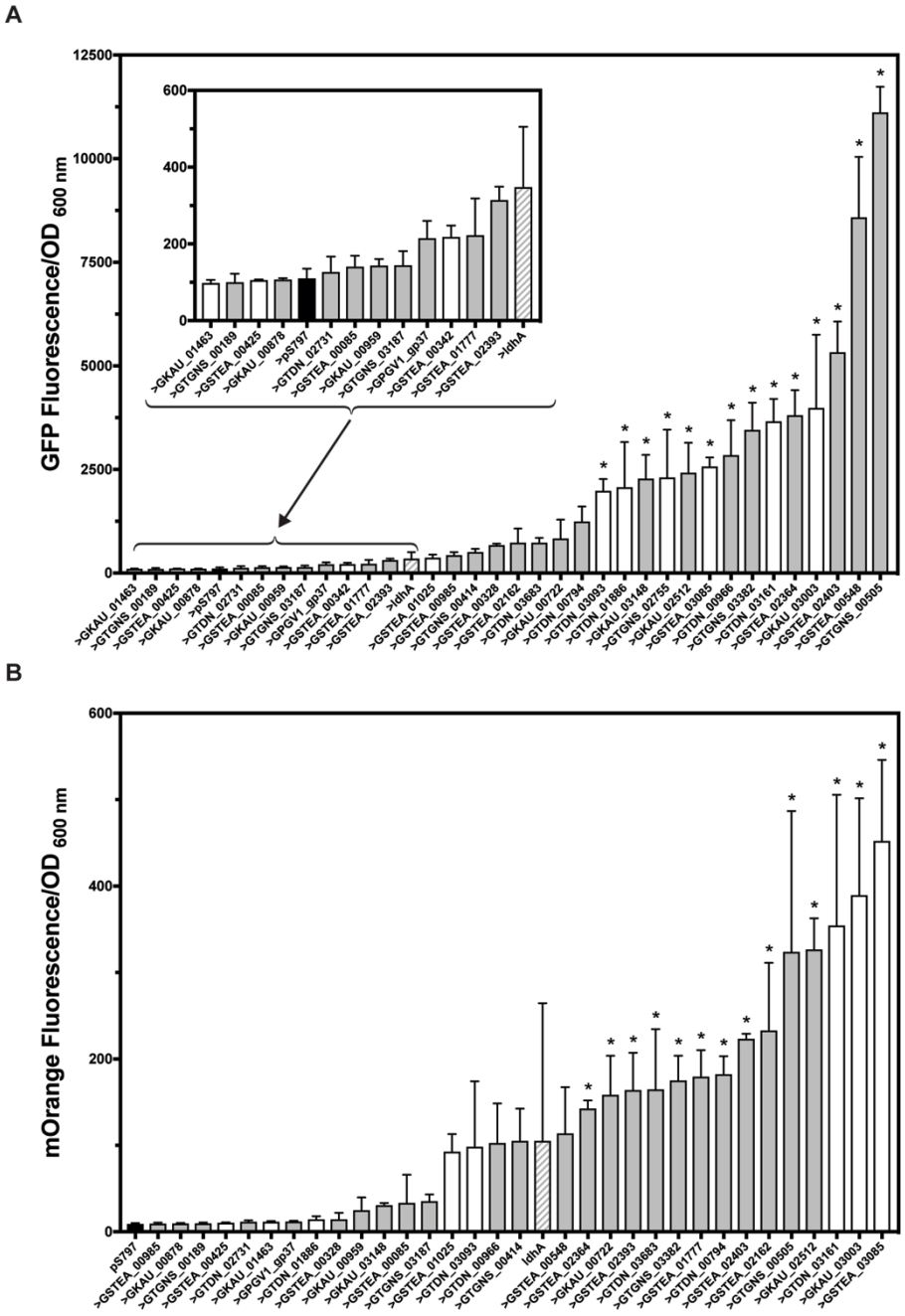


Figure 4.11: Putative promoters characterised upstream of A) GFP & B) mOrange in *G. thermoglucosidans*.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. Previously uncharacterised promoters are shown in white. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the hatched bar. The negative control, *G. thermoglucosidans* transformed with an empty pS797 vector, is represented by the black bar. Bars represent the mean of n = 3 starter cultures arising from independent transformation events, except in the case of the two controls, where $3 \leq n \leq 9$. Standard deviation error bars are shown, unless hidden by the bar. Promoter sequences with mean fluorescence values that were significantly different from the negative control are labelled with an asterisk. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, using a significance level of 0.05.

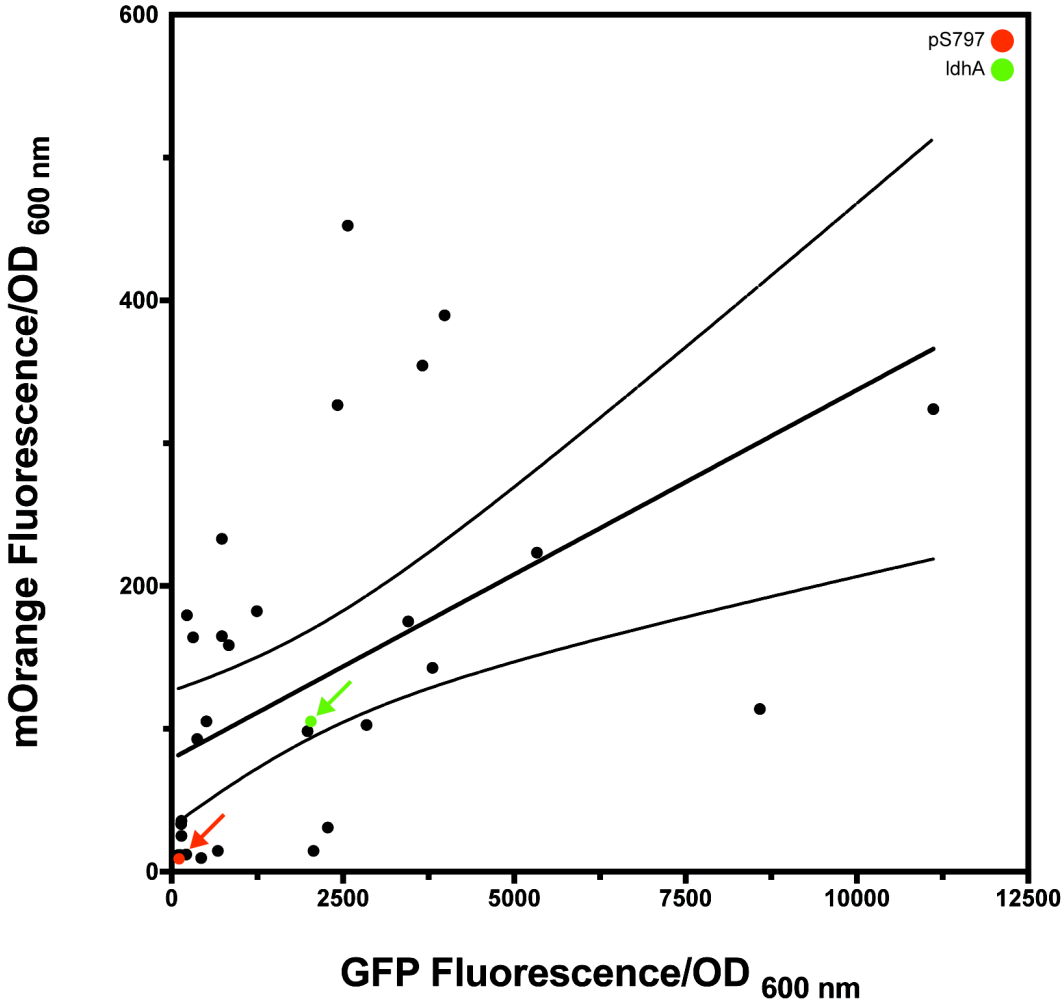


Figure 4.12: Fluorescence output of GFP and mOrange under the control of putative promoter sequences.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. Points represent the mean fluorescence output of reporter proteins under the control of individual promoter sequences from 3 ≤ n ≤ 9 starter cultures arising from independent transformation events. The negative control, *G. thermoglucosidans* transformed to contain empty pS797 vector, is coloured red. The positive control, the *G. thermoglucosidans ldhA* promoter, is coloured green. The solid line represents a linear regression of the data, with 95% confidence limits represented by the dashed lines.

within the 95% confidence limits of the linear regression, indicating a reasonable correlation in activity levels between the two reporters for these sequences.

Partition modelling

100 random forest models were generated for each of the *promoter::GFP* and *promoter::mOrange* fluorescence data sets. In all instances, 20% of the available promoter sequences were randomly selected and withheld from model training to serve as a validation set. Each random forest contained a maximum of 100 decision trees, with early stopping if the addition of further trees to the forest did not improve the validation statistic (SAS Institute Inc, 2016b). Each tree was trained on a data set containing 26 randomly selected promoter sequence positions, drawn with replacement.

Once partitioning was complete, the number of times each promoter sequence position caused a split in all 100 random forests was quantified (Figure 4.13A). In the case of the GFP reporter, the modelling suggested that the sequence positions with the biggest impact on promoter performance were distal to the CDS. Of the 10 sequences positions that resulted in the greatest number of splits, 7 were located in the 5' half of the promoter sequence. The exceptions were the -49, -14 and -5 positions, which returned the 2nd, 8th and 9th most splits, respectively. This result was somewhat counterintuitive, given that the regions of promoter sequence that are canonically hypothesised to contribute the most the promoter activity (the RBS, -10 and -35 motifs) are located in the proximal end of the promoter sequence.

The results from the random forests trained using the mOrange data set showed minimal correlation to the GFP results (Figure 4.13B). For the mOrange data set, 7 out of the 10 sequence positions that resulted in the greatest number of splits were located in the 3' half of the promoter sequence. The lack of correlation between the partitioning results for the two reporter proteins was likely a result of the lack of correlation observed in promoter activity for the two reporters (Figure 4.12). Given that fluorescence was used as the response variable in random forest construction, the difference in promoter activity between the two reporters would have resulted in different partition results.

For both reporter CDS, the promoter sequence positions that were identified as important in determining promoter activity showed minimal

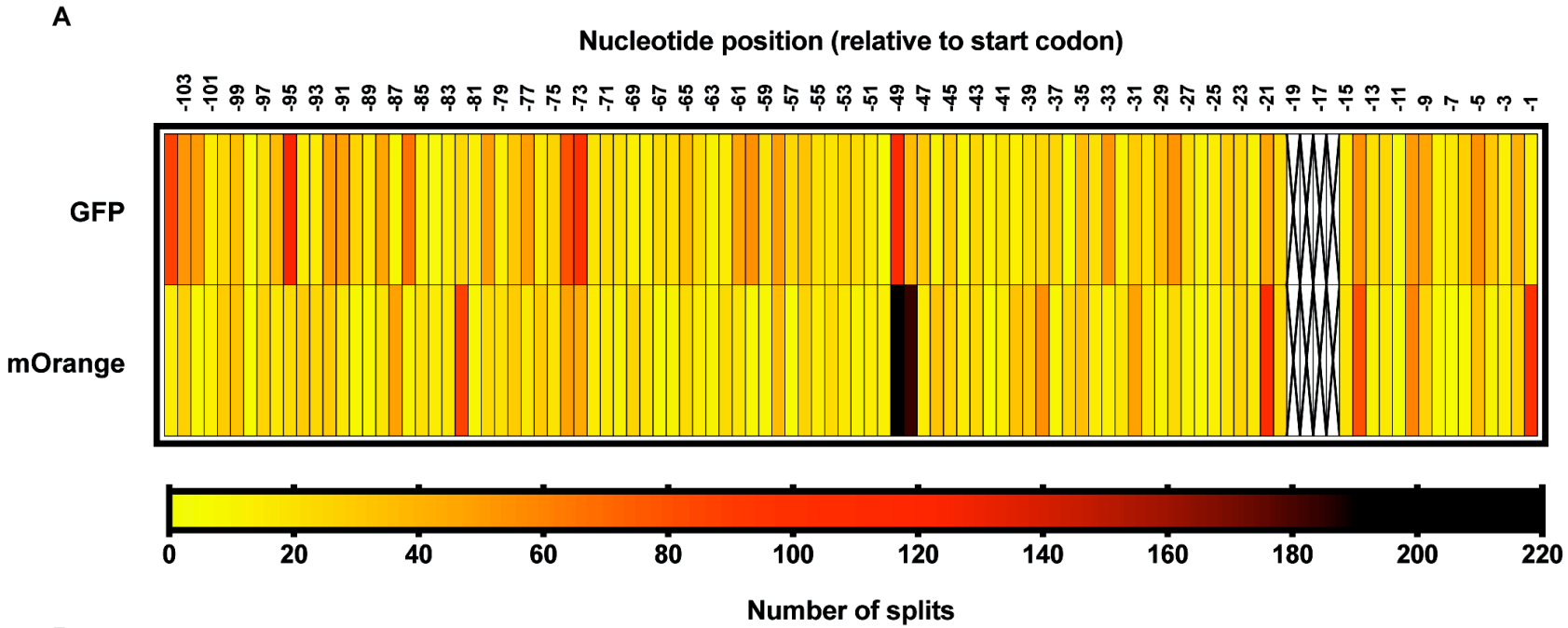


Figure 4.13: Data set B partition modelling results.

A) Heat map showing the number of splits caused by individual promoter sequence positions when GFP or mOrange fluorescence data was used for model construction. The hatched region represents the ACCT cloning scar between promoter & RBS. As all promoter sequences were identical in this region, these 4 positions were not included in the partition model.

B) Comparing the importance of promoter sequence positions between the 2 reporter proteins. Points represent individual sequence positions. The dashed line is shown at the point where position rank for both reporters is equal. The lower the rank of a position, the more times that position caused splits in the boosted forest, and therefore the more important that position was judged to be for determining promoter output by the model. Sequence positions that were important for determining promoter activity under both reporters should therefore have clustered in the bottom left hand corner of the graph.

sequence conservation (Figure 4.14). Sequence positions with high degrees of homology across all 34 promoter sequences were likely to have little statistical power with regards to determining differences in promoter output, as promoters with significantly different activity levels could have the same nucleotide present at conserved positions. This explanation was proposed as the reason for the partition modelling identifying sequence positions outside of the canonical consensus regions as important in determining promoter activity. The partition modelling results therefore highlighted the importance of considering entire promoter sequences instead of only consensus regions when designing synthetic promoters *de novo*, as non-consensus regions were shown to have a significant impact on promoter activity.

Partial Least Squares sequence-function models

The lack of correlation between the GFP and mOrange partition results precluded the construction of multivariate promoter sequence-function models that made simultaneous predictions of fluorescence output for both GFP and mOrange. Initial sequence-function models were therefore trained using only the GFP data set as a proof-of principle.

The predictive performance of sequence-function models derived from data set A was judged based on prediction error when models were applied to the CV data set (Figure 4.8). However, previous studies have shown that high predictive accuracy when applied to CV data does not necessarily correlate to adequate generality (Golbraikh & Tropsha, 2002). This lack of correlation was readily apparent in the models derived from data set A; the model PLS_iteration_A_2 returned an R^2 value of 0.964 when applied to the CV data, but displayed poor predictive power when applied to making predictions of synthetic promoter activity (Figure 4.8). An external test set was therefore required to accurately quantify the predictive power of promoter sequence-function models (Sheridan, 2013).

Five promoter sequences were selected to provide an independent test set on which to measure predictive power. So that the test set

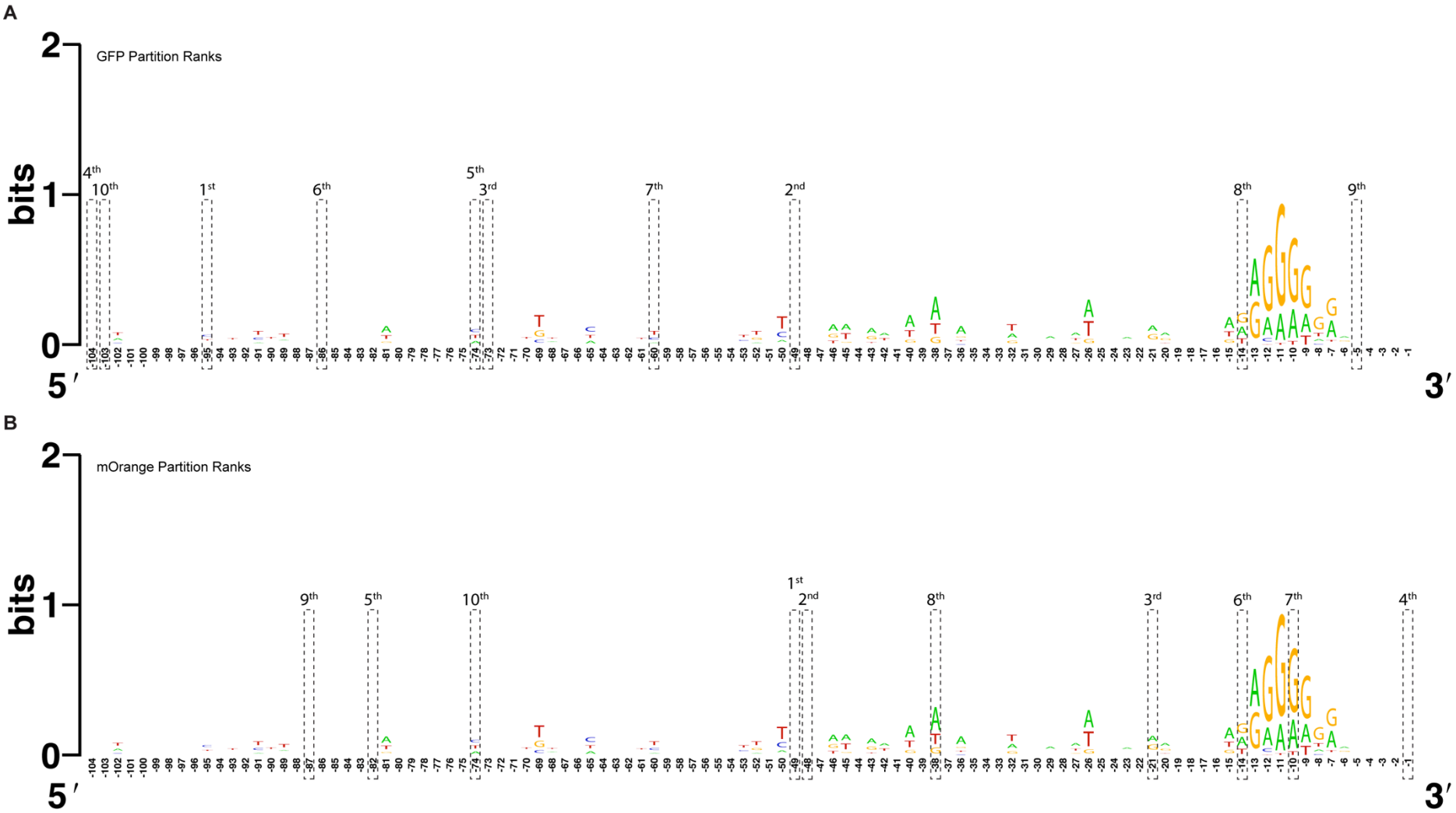


Figure 4.14: Visualisation of a sequence alignment of the 34 putative promoter sequences used in partition modelling of data set B.

The highlighted sequence positions are those which caused the highest number of splits in 100 random forest partition models, trained using A) GFP and B) mOrange fluorescence data as response variables. Rendered using WebLogo version 2.8.2 (Crooks *et al.*, 2004).

contained promoter sequences with a range of activity levels, the distribution of GFP fluorescence levels in data set B (Figure 4.11A) was analysed. One promoter sequence was chosen at random from each of the 1st and 3rd distribution quartiles, and three sequences were chosen at random from the interquartile range. The chosen sequences (ordered from strongest to weakest) were GSTEA_02403, GTDN_01886, GTDN_03093, GSTEA_00328 and GTGNS_00189.

Outlier analysis was performed to identify any *promoter::GFP* fusions that might have negatively impacted upon PLS model performance (Cox & Gaudard, 2013). Quantile range analysis with a Q value of 3 showed that none of the measurements of *promoter::GFP* activity were outliers. However, Huber M-Estimation with a K value of 4 returned two outlying measurements of fluorescence activity; the mean GFP output of the two strongest promoters, GSTEA_00548 and GTGNS_00505 was outlying with regards to the remaining 32 promoter sequences. However, both sequences were included in the training data set, as the fact that there were certain promoters that resulted in extremely high GFP expression levels was deemed biologically relevant; any final sequence-activity model needed to be able to accurately predict promoter activity at all biologically feasible levels of activity.

By indicating which nucleotides within the promoter sequences were likely to contribute most to promoter strength, the partition modelling results provided a useful tool for dimensionality reduction. Downstream PLS models could be trained on data sets from which redundant sequence positions had been removed. However, the random forest results provided no indication as to how many sequence positions should be used as *x* variables to maximise the predictive power of the final sequence-function models.

PLS models were therefore trained using 3, 5, 7, 9, 11, 13 or 15 sequence positions as *x* variables. Sequence positions were included in the PLS models in descending order of the number of splits caused in the 100 random forest partition models (Figure 4.13). For each number of *x* variables that were considered, eight PLS models were fit, using the settings summarised

in Table 4-2. A total of 56 different PLS models were therefore trained. The optimal model was the one that returned the highest R^2 value when applied to the test set of five promoter sequences.

Model	PLS algorithm	Validation methodology
1	SIMPLS	*K*Fold, *K* = 7
2	NIPALS	*K*Fold, *K* = 7
3	SIMPLS	*K*Fold, *K* = 5
4	NIPALS	*K*Fold, *K* = 5
5	SIMPLS	*K*Fold, *K* = 3
6	NIPALS	*K*Fold, *K* = 3
7	SIMPLS	Holdback, proportion 0.33
8	NIPALS	Holdback, proportion 0.33

Table 4-2: Summary of settings used in Partial Least Squares model construction.

The optimal PLS model that was obtained (hereafter referred to as PLS_iteration_B_1) modelled GFP fluorescence as a function of five promoter sequence positions. The sequence positions included in the model were -49, -73, -74 -95 and -104. The model used the SIMPLS PLS algorithm and holdback CV, and used one LV to explain 68.224% of the variation in GFP fluorescence observed in the training set.

When applied to the test data set, PLS_iteration_B_1 returned an R^2 value of 0.793 (Figure 4.15). The empirically measured GFP output from two of the five test promoter sequences fell within one standard deviation of the mean of the value predicted by the model. These results suggested that PLS_iteration_B_1 had reasonable predictive power when applied to previously unseen data. During analysis of the model, the mean GFP fluorescence caused by the promoter GTGNS_00505 was observed to be outlying in terms of Euclidean distance to the PLS model centre. However, removal of GTGNS_00505 from the training data set did not yield a model of improved predictive accuracy as compared to PLS_iteration_B_1.

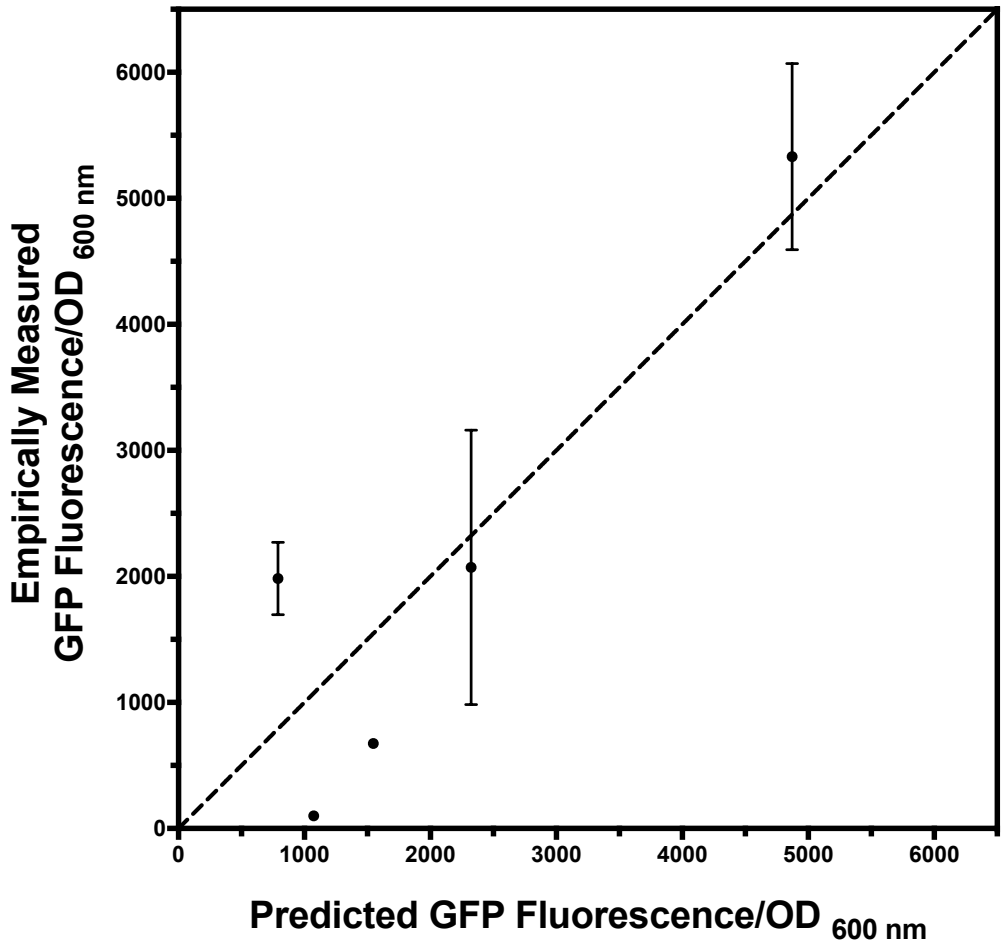


Figure 4.15: Empirically measured GFP fluorescence levels plotted against GFP fluorescence levels as predicted by the Partial Least Squares model PLS_iteration_B_1.

Points represent the activity levels of individual promoter sequences. Empirically measured fluorescence and absorbance after 24 h incubation in 96-well plate format, Empirical values are the mean of n = 3 starter cultures arising from independent transformants, with standard deviation error bars shown, unless hidden by the points. The dashed line represents the point at which empirically measured and predicted fluorescence values are equal.

The second best performing PLS model obtained, PLS_iteration_B_2, was trained using 10 promoter sequence positions, the SIMPLS algorithm and *K*Fold CV with *K* = 7. When applied to the test set, PLS_iteration_B_2 returned a R^2 value of 0.680.

In attempt to further increase prediction accuracy, PLS_iteration_B_1 and PLS_iteration_B_2 were aggregated. However, the R^2 value that was

returned when the aggregate model was applied to the test set was 0.7819, *i.e.* smaller than the R2 value returned by PLS_iteration_B_1. This result suggested that aggregating the two PLS models did not increase predictive accuracy.

Artificial Neural Network sequence-function models

ANNs were also applied to train promoter sequence-function models using the data derived from the empirical characterisation of promoter sequences in data set B.

The custom design DoE platform of the JMP software was used to define 20 ANN architectures. All architectures consisted of a single hidden layer and used the squared penalty method. The number of nodes in the hidden layer, the number *x* variables (promoter sequence positions) modelled and the activation function personality were the factors included in the experimental design. As with the PLS models, sequence positions were included in the ANNs in descending order of the number of splits caused in the 100 random forest partition models (Figure 4.13). The performance of each of the 20 network designs was quantified using the R^2 and the Root Average Squared Error (RASE) values that were returned when the ANNs were applied to an independent test data set (SAS Institute Inc, 2016b).

So that the test set contained promoter sequences with a range of activity levels, the distribution of GFP fluorescence levels in data set B (Figure 4.11A) was analysed. One promoter sequence was chosen at random from each of the 1st and 3rd distribution quartiles, and three sequences were chosen at random from the interquartile range. The chosen sequences (ordered from strongest to weakest) were GTGNS_00505, GTDN_01886, GTDN_03093, GSTEA_00328 and GTGNS_00189. The remaining 29 promoter sequences were randomly split 70:30 into training and validation sets. The same training, validation and test sets were used to train all ANNs.

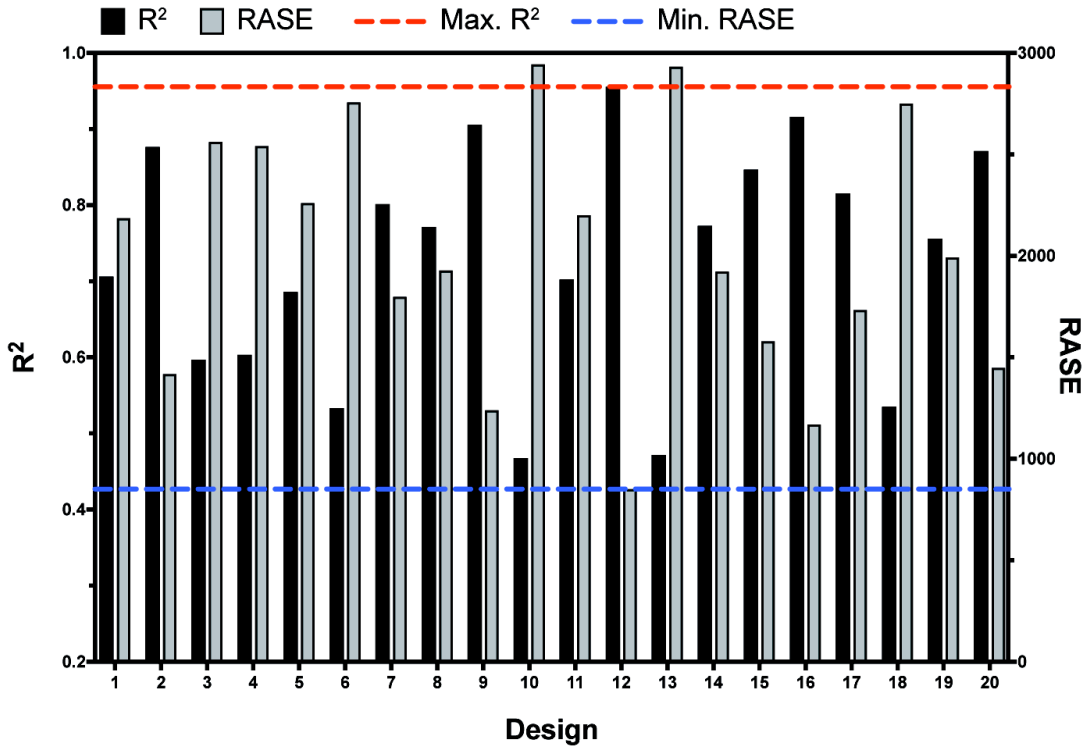
Given the random attribution of values to the ANN weights at the start of the learning process, the final solution to which an individual ANN converges is

dependent on the random seed used to generate the starting weight values (Hastie *et al.*, 2009). 1,000 ANNs were therefore fit for each of the 20 network architectures, with each model using a unique, defined random seed. For each network architecture, an ensemble network was created. Ensembles contained the five ANNs of a given architecture that returned the highest R^2 value when applied to the test set. The R^2 and RASE values that were returned by the 20 ensemble networks are shown in Figure 4.16**.**

The optimal ANN obtained modelled GFP fluorescence as a function of nine promoter sequence positions, using 11 nodes and linear activation functions. When applied to the test set, this ANN returned an R^2 value of 0.955 and a RASE value of 850.84. That model is hereafter referred to as ANN_1. The apparent high-performance of a linear network was somewhat surprising; when linear activation functions were used, ANNs decomposed to simple linear interpolators (Laudani *et al.*, 2015). Completely linear ANNs therefore did not contain the non-linearity that was one of the major the reasons for their application to promoter sequence-function analysis.

The results from each of the 20 ANN architectures were analysed by standard least squares model with effect screening emphasis. The resulting analysis predicted that the R^2 value of the test set would be maximised by an ANN that modelled eight promoter sequence positions using 11 hidden nodes and linear activation functions. However, this model architecture returned an R^2 value of 0.9346 when applied to the test set, which was lower than the R^2 value returned by ANN_1 (0.9555).

In the 21 single layer ANNs discussed above, the dimensions of the promoter design space were reduced by predicting fluorescence output as a function of only those sequence positions that were predicted by the partition modelling to have strong predictive power. However, ANNs of increased complexity could potentially model design spaces of large dimensionality, without the need for partitioning. For example, ANNs with multiple hidden layers could potentially accurately model highly dimensional design spaces by using the 2nd layer (the layer closest to the original *x* variables) to identify key



Design	Activation function personality	N (Nodes)	N(*x*)
1	Linear	3	6
2	Linear	11	10
3	Gaussian	5	10
4	Linear	3	5
5	Tan	5	6
6	Gaussian	5	5
7	Gaussian	11	6
8	Gaussian	3	9
9	Linear	9	6
10	Tan	11	5

Design	Activation function personality	N (Nodes)	N(*x*)
11	Gaussian	3	8
12	Linear	11	9
13	Tan	3	10
14	Gaussian	9	5
15	Tan	9	9
16	Tan	11	7
17	Gaussian	11	10
18	Tan	3	5
19	Tan	9	10
20	Linear	5	9

Figure 4.16: R^2 and Root Absolute Squared Error (RASE) values returned by 20 single layer Artificial Neural Network architectures when applied to a test data set.

Bars representing R^2 values are shown in black, and are plotted on the left hand y-axis. Bars representing RASE values are shown in grey, and are plotted on the right-hand y-axis. The dashed red and blue lines represent the R^2 and RASE values returned by the optimal obtained ANN, which was trained using design 12. For each ANN design, the table shows the activation function personality, the number of nodes in the hidden layer (N(Nodes)) and the number of promoter sequence positions (N(*x*)) modelled.

promoter sequence positions, in a role analogous to the dimension-reduction performed by the partition models (SAS Institute Inc, 2016b).

DoE-guided ANN optimisation was therefore applied to two-layer ANN design. The custom design platform in the JMP software was used to define 30 ANN architectures. Five variables were included in the DoE: the number of nodes in the first hidden layer, the number of nodes in the second hidden layer, the activation function personality used in the hidden layers and the penalty function personality. All 30 network architectures modelled GFP activity as a function of the complete 104 bp promoter sequence, and all 30 network architectures were run 1,000 times, with each model using a unique defined random seed. For each network architecture, the five ANNs that returned the highest R^2 value when applied to the test set were aggregated.

The R^2 and RASE values that were returned when the 30 aggregated models were applied to the test set were analysed using a standard least squares model with effect screening emphasis. The R^2 value of the test set was predicted to be maximised by an ANN in which both hidden layers contained seven nodes and used Gaussian activation functions. The optimal penalty method was predicted to be Weight Decay. The predicted optimal model architecture was fit 1,000 times using 1,000 different random seeds, and the five models that returned the highest R^2 value when applied to the test data set were aggregated. The resulting ensemble model returned an R^2 value of 0.9765 when applied to the test set, and is hereafter referred to as ANN_2.

The training set R^2 value returned by ANN_2 was higher than that returned by all of the two-layer network designs specified by the initial DoE design. The next best performing two-layer model, hereafter referred to as ANN_3, returned an R^2 value of 0.948 when applied to the test set. ANN_3 also used the Weight Decay penalty term, but used a 13-node first hidden layer with Gaussian activation functions, and a three-node second hidden layer with TanH activation functions.

ANN_1, ANN_2 and ANN_3 (Figure 4.17) all returned R^2 values that were greater than 0.9 when applied to independent test data. This result suggested that all three of these ANN models had good predictive power. To further test the predictive power of the obtained models, ANN_1, ANN_2, ANN_3 and the PLS model PLS_iteration_B_1 were subsequently used to make predictions of activity for all of bioinformatically identified putative promoter sequences that had not been characterised *in vivo*.

In both data sets A and B and in Chapter 3, bioinformatically identified putative promoter sequences were examined manually to ensure that they did not overlap with any adjacent CDS. Whilst manual inspection was sufficient when the number of sequences being analysed was small, this approach was inadequate when *in silico* predictions of activity were required for all 1,489 putative promoters. BEDTools intersect (Quinlan & Hall, 2010) was therefore used to isolate non-overlapping putative promoter sequences. In total, 636 of the 1,489 putative *Geobacillus* promoters were shown to be non-overlapping. The number of promoters identified in each of the 4 *Geobacillus* species of interest is summarised in Table 4-3.

	Putative promoters	Non-overlapping Putative promoters
G. kaustophilus DSM7263	403	176
G. stearothermophilus DSM22	370	187
G. thermodenitrificans K1041	345	130
G. thermoglucosidans DSM2542	371	143

Table 4-3: Number of non-overlapping putative promoters isolated from each of the four *Geobacillus* species of interest.

Despite the three ANNs returning equally accurate predictions when applied to the test set of five promoter sequences (Figure 4.17), the predictions that were returned when the three ANNs and the PLS model PLS_iteration_B_1 were applied to putative promoter sequences that had not been characterised *in vivo* showed little correlation (Figure 4.18). The two models that returned the

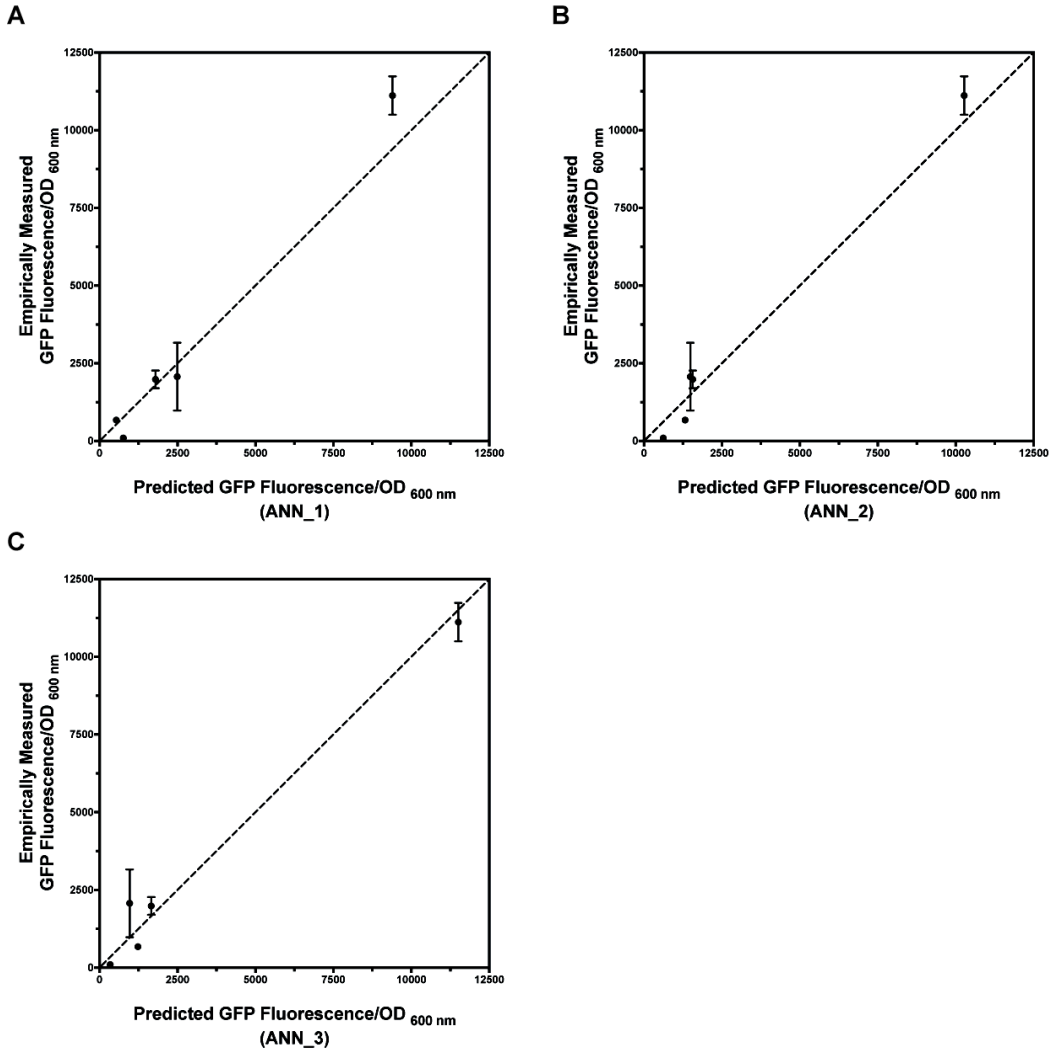


Figure 4.17: Empirically measured fluorescence output of GFP under the control of the five promoters from the test data set, plotted against fluorescence as predicted by the three optimal Artificial Neural Network models obtained.

Points represent individual promoter sequences. Empirical measurements were taken after 24 h growth in 96-well plate format and are the mean of n = 3 starter cultures, arising from independent transformation events. The dashed lines represent the point at which empirical and predicted values are equal.

most similar predictions of activity were ANN_1 and PLS_iteration_B_1, although a linear regression of the two sets of predictions returned an R^2 value of only 0.2463.

In the case of ANN_3, prediction clusters were also apparent, with putative promoter sequences predicted to either result in high or minimal GFP

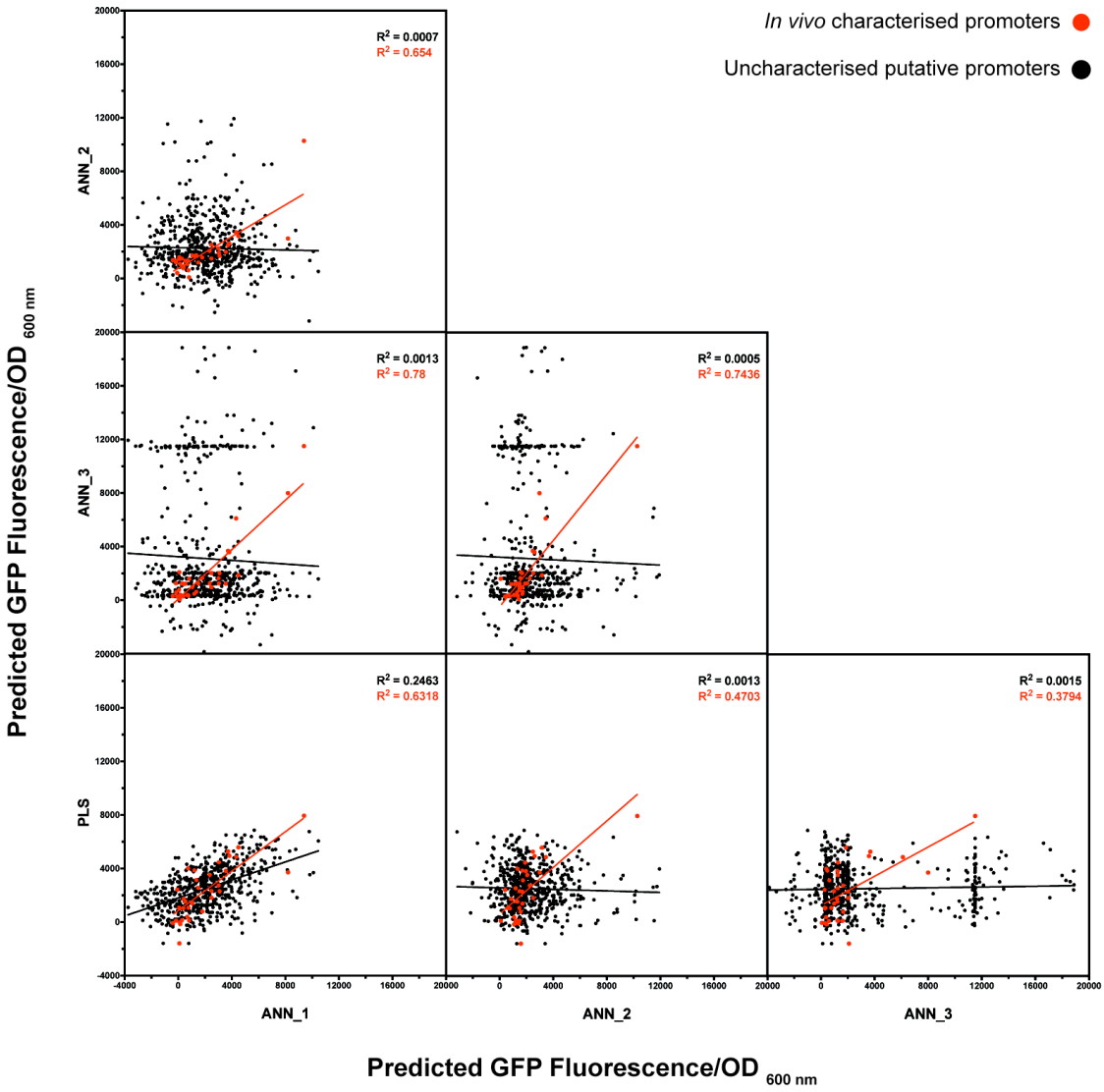


Figure 4.18: Scatterplot matrix showing GFP fluorescence output of putative promoter sequences as predicted by high performing Artificial Neural Network & Partial Least Squares models.

Points represent individual promoter sequences. Predictions of fluorescence output are as calculated by neural networks ANN_1, ANN_2 and ANN_3 and PLS_iteration_B_1**.** Red points represent promoter sequences used in the model training process. Black points represent bioinformatically identified putative *Geobacillus* promoters that were not characterised *in vivo*. The solid lines represent linear regressions of the data.

expression, with few sequences predicted to be of intermediate strength.

The lack of correlation between the predicted activity levels for promoter sequences that had not been characterised *in vivo* suggested that the predictive power of the four models was not as comparable as was implied by the similar R^2 values that were returned by the test set. Had all four of the models truly had comparable predictive accuracy, strong positive correlations would have been expected between predictions of activity for uncharacterised putative promoters.

To ascertain which, if any, of the four promoter sequence-function models showed the greatest predictive accuracy, 14 previously uncharacterised putative promoter sequences were selected at random from across the *Geobacillus* promoter phylogeny for *in vivo* characterisation. Three promoters could not be successfully cloned upstream of GFP using the type IIS restriction cloning strategy. A secondary test set of 11 promoters was therefore available (Figure 4.19A) to assess the predictive accuracy of the four putatively powerful models.

None of the four models were able to accurately predict the *in vivo* activity levels of the 11 promoter sequences (Figure 4.19B-E). In particular, the models showed a tendency to over-predict the activity of sequences that had no *in vivo* activity, and under-predict the activity of functioning promoters. Of the 11 promoters in the secondary test set, eight resulted in GFP fluorescence that was not statistically significantly greater than the negative control (Figure 4.19A). However, all four of the models predicted that the majority of these eight sequences would have *in vivo* promoter activity. In contrast, two of the three promoters that did show *in vivo* activity (GKAU_03578 and GSTEA_01279) were stronger than predicted. The activity of the third active promoter, GTGNS_02828, was weaker than predicted by models ANN_1 & PLS_iteration_B_1, and stronger than predicted by models ANN_2 and ANN_3. The inability of the models to differentiate between active and inactive promoters was hypothesised to be the result of the models being unable to identify critical nucleotides or motifs that were not present in the training data.

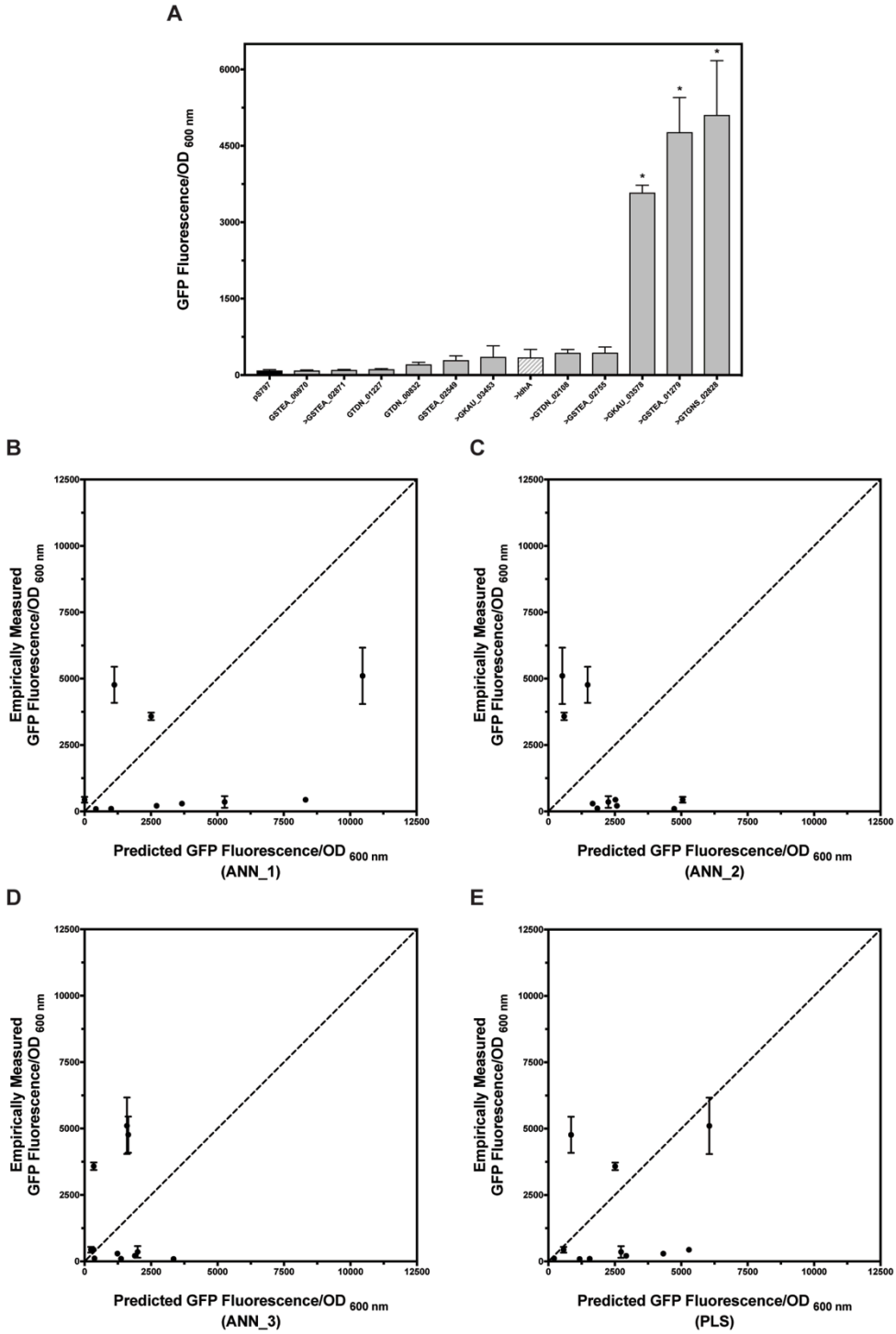


Figure 4.19: Empirically measured promoter activity of a secondary test set of 11 putative promoters, as compared to the activity levels predicted by Artificial Neural Network & Partial Least Squares models derived from data set B.

Points and bars represent the activity levels of individual promoter sequences. Empirical measurements of promoter activity taken after 24 h growth in 96-well plate format. In all cases, bars or points represent the mean of $3 \leq n \leq 9$ independent starter cultures, with standard deviation error bars shown, unless hidden by the bar or point. A) Empirically measured activity of promoters in the secondary test set. The hatched bar represents the positive control & the black bar represents the negative control. Promoters for which mean fluorescence output was statistically significantly different to the negative control are indicated by an asterisk. Significance was determined by one-way ANOVA with Dunnett's multiple comparisons test at a significance level of 0.05. B-E) Empirically measured promoter activity plotted against activity as predicted by models B) ANN_1, C) ANN_2, D) ANN_3 & E) PLS_iteration_B_1. The dashed lines represent the points at which empirically measured and predicted fluorescence values are equal.

4.2.3 Characterisation and modelling of data set C

The lack of predictive power shown by the PLS and ANN promoter sequence-function models derived from data set B was hypothesised to be a result of the size of the training data set. ANNs theoretically have universal approximation capability (Hornik, 1989), but such capability requires a significantly large training set. When the design space being modelled is complex, ANNs cannot provide an accurate abstraction unless the training data set explores a sufficiently large proportion of said space (Bataineh & Marler, 2017). ANNs trained on small data sets are therefore likely to provide inadequate generality when applied to novel data.

A final expansion of the training data set was therefore performed. Including the 11 putative promoter sequences that were selected for model validation, a total of 45 putative promoter sequences had been characterised *in vivo* in data set B. 52 additional promoter sequences were selected at random from across all clades of the promoter phylogeny and synthesised upstream of *GFP* and *mOrange* in the pS797 vector by ATUM (previously DNA 2.0, California, United States of America). The three promoter sequences from data set B that could not be cloned upstream of *GFP* using the type IIS restriction cloning strategy were also synthesised by ATUM. Data set C therefore contained a total of 100 putative promoter sequences.

In total, 95 promoter sequences were characterised upstream of GFP in *G. thermoglucosidans* (Figure 4.20). Two *promoter::GFP* fusions could not be synthesised by ATUM, and three sequences could not be transformed into *G. thermoglucosidans*. The promoter library covered a total expression range of 148-fold in steady increments. Of the 95 characterised sequences, 31, covering an expression range of 6.8-fold resulted in mean GFP expression levels that were statistically significantly greater than the negative control. The weakest promoter to exceed this threshold was the *ldhA* promoter.

During analysis of data set B, the *ldhA* promoter was used as the threshold for defining active sequences. 22 sequences, covering a 30-fold range in GFP expression levels, exceeded this threshold in data set B.

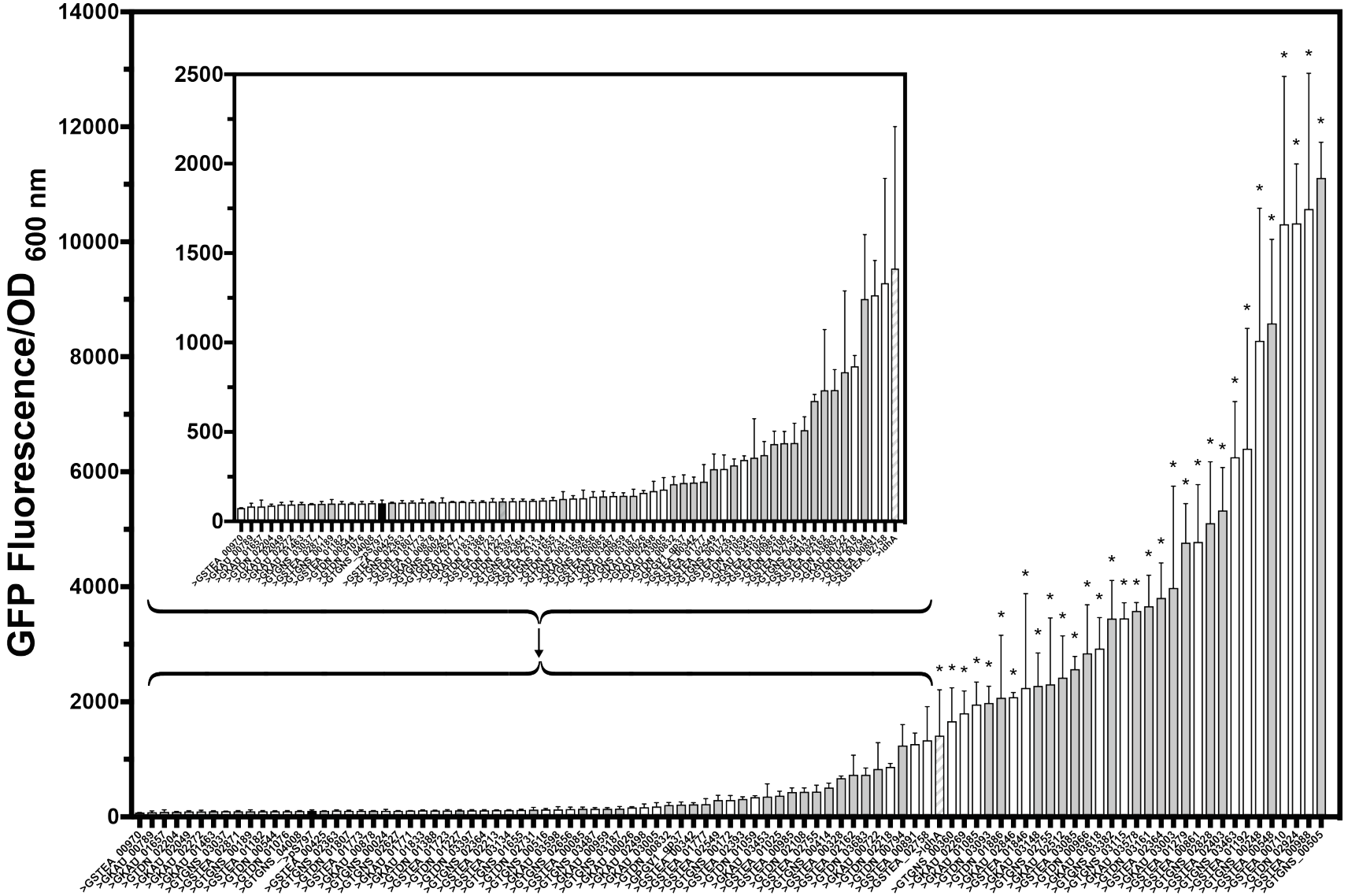


Figure 4.20: Putative promoters characterised upstream of GFP in *G. thermoglucosidans*.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the hatched bar. The negative control, the empty pS797 vector, is shown in black. Grey bars represent previously characterised promoters, white bars represent previously uncharacterised sequences. Bars represent the mean of $3 \leq n \leq 9$ starter cultures, arising from independent transformants. Standard deviation error bars are shown, unless hidden by the bar. Promoter sequences that resulted in mean fluorescence that was statistically significantly different from the negative control are labelled with an asterisk.

However, repeated measurements of *ldhA* promoter activity in data set C returned a higher mean GFP expression level for the *ldhA* promoter than was observed in data set B. (371.8 AU in data set B, 1414.3 AU in data set C). The *ldhA* promoter therefore provided a harsher threshold for defining promoter activity in data set C than in data set B. Using the mean fluorescence for the *ldhA* promoter from data set B as the cut-off for defining promoter activity in the data set C resulted in 44 promoters, covering a GFP expression range of 30-fold, being defined as active. Increasing the number of characterised sequences in the *Geobacillus* promoter library did not therefore yield a significant increase in GFP expression range as compared to data set B, although the number of "active" sequences was doubled.

82 sequences, covering an expression range of 107-fold were successfully characterised upstream of mOrange in *G. thermoglucosidans* (Figure 4.21). Nine sequences could not be synthesised upstream of mOrange, and 11 *promoter::mOrange* fusions could not be transformed into *G. thermoglucosidans*. Of the 82 characterised sequences, 32, covering an expression range of 8.4-fold, resulted in mOrange expression that was statistically significantly greater than the negative control. The mean fluorescence of *ldhA::mOrange* was also statistically significantly greater than the negative control. The range of mOrange activity levels displayed by the "active" sequences in data set C was 2.8-fold greater than that observed in data set B.

Homogeneity of expression

Ideally, *cis*-regulatory elements for synthetic biology applications should yield homogeneous, predictable expression of the protein of interest at the single-cell level (Gasser *et al.*, 2015). However, the promoter characterisation data discussed up to this point were obtained at the population-level, using a Tecan Infinite 200 PRO microplate reader. These data did not therefore account for variation in promoter behaviour between individual cells (Beal *et al.*, 2012). Flow cytometry was therefore used to analyse the intra-population variation in fluorescence activity displayed by the characterised *promoter::reporter* fusions.

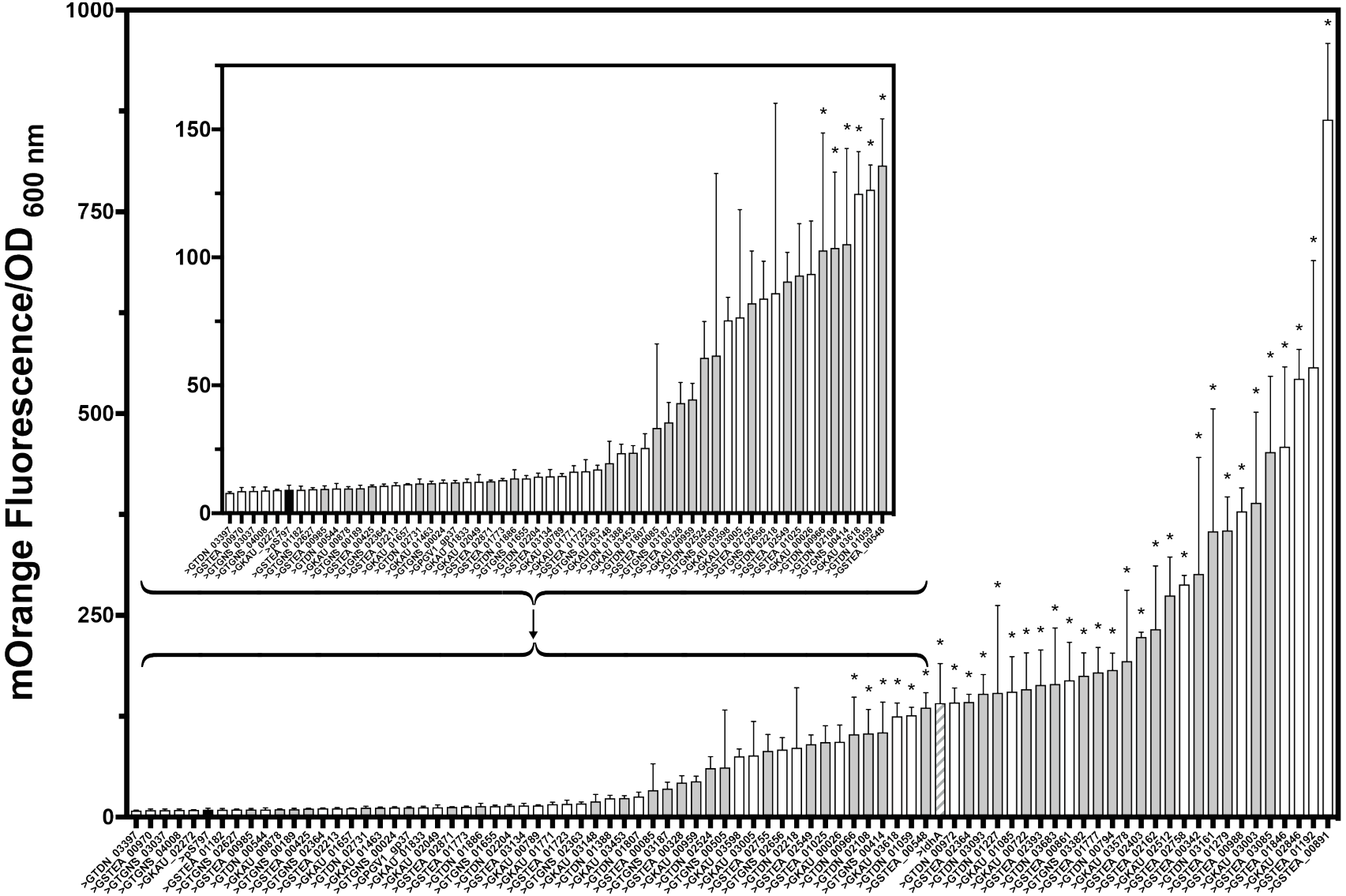


Figure 4.21: Putative promoters characterised upstream of mOrange in *G. thermoglucosidans*.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the hatched bar. The negative control, the empty pS797 vector, is shown in black. Grey bars represent previously characterised promoters, white bars represent previously uncharacterised sequences. Bars represent the mean of 3 ≤ n ≤ 9 starter cultures, arising from independent transformants. Standard deviation error bars are shown, unless hidden by the bar. Promoter sequences that resulted in mean fluorescence that was statistically significantly different from the negative control are labelled with an asterisk.

Expression of GFP and mOrange under the control of *G. thermodenitrificans ldhA* promoter was shown to be highly heterogeneic. Analysis of 11 starter cultures arising from independent transformation events for each reporter protein returned fluorescence intensities that covered a 3-log range for both *ldhA::GFP* (Figure 4.22C) and *ldhA::mOrange* (Figure 4.23C). Therefore, despite the comparatively widespread use of the sequence to control the expression of heterologous proteins in *Geobacillus* (Cripps *et al.*, 2009, Bartosiak-Jentys *et al.*, 2012, Lin *et al.*, 2014, Kananavičiūtė & Čitavičius, 2015), the *ldhA* promoter did not satisfy the synthetic biology requirement of homogeneous, predictable expression.

In contrast to the expression heterogeneity displayed by the *ldhA* promoter, the majority of the putative promoters that were identified in this study offered homogeneous expression at the single-cell level. To assess the intra-population variation in expression levels, 100,000 events from each of three cultures arising from independent transformation events were combined to form a single “meta” population for each of the characterised *promoter::reporter* fusions. Of the 95 analysed *promoter::GFP* fusions, only two, GKAU_03003 and GTDN_01059 returned a robust Coefficient of Variance (CVar) that was greater than that returned by *ldhA::GFP* (Figure 4.22A). Of the 82 characterised *promoter::mOrange* fusions, 60 returned a CVar value lower than that returned by *ldhA::mOrange* (Figure 4.23A).

Five exemplar promoter sequences (GPGV1_gp37, GSTEA_00891, GSTEA_02364, GSTEA_02755 and GSTEA_03085) were identified that, when characterised upstream of GFP, cumulatively covered the same range of expression as the *ldhA* promoter, increasing in steady increments (Figure 4.22B and C). When characterised upstream of mOrange, the five promoters did not display the same rank order of expression levels, but did cumulatively cover the same expression range as the *ldhA* promoter, again increasing in steady increments (Figure 4.23B & C).

The characterised promoter library therefore contained promoter sequences that afforded tight control of protein expression across a three-log

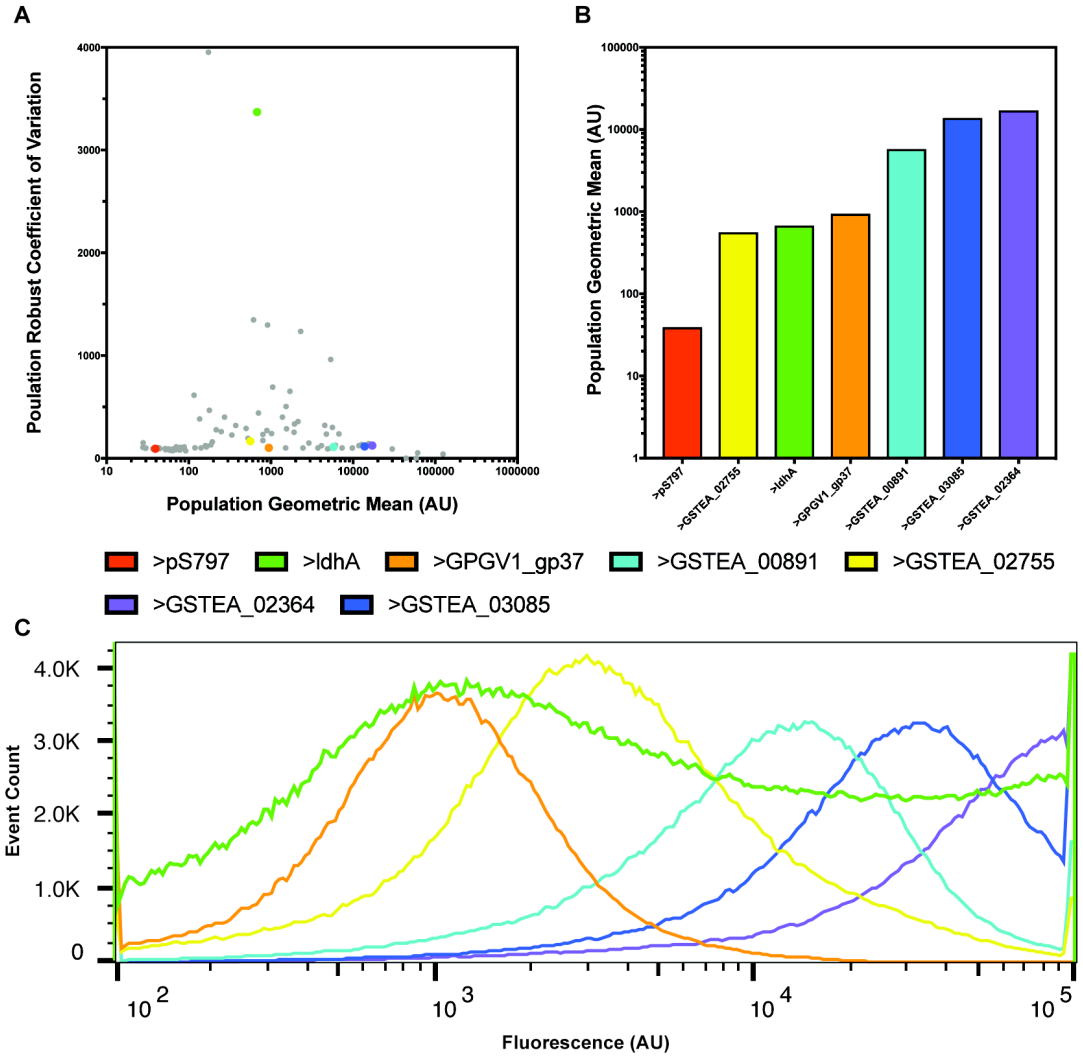


Figure 4.22: FACS analysis of *G. thermoglucosidans* cultures expressing GFP.

Bars, histograms and points represent individual promoter sequences. For each *promoter::GFP* fusion, 100,000 events from each of three starter cultures arising from independent transformation events were combined to form a single "meta" population of 300,000 events. The exceptions were the two controls; bars, histograms and points representing pS797 and the *ldhA* promoter represent 15 and 11 starter cultures, respectively. Cultures were excited at 488 nm and fluorescence intensity was recorded using a 530/30 nm detector.

range of fluorescence intensity. Members of the library were therefore potentially broadly applicable to synthetic biology and metabolic engineering projects in *Geobacillus* where expression homogeneity is required, and represented a significant improvement as compared to the *ldhA* promoter, which has previously been applied to pathway engineering in *Geobacillus*.

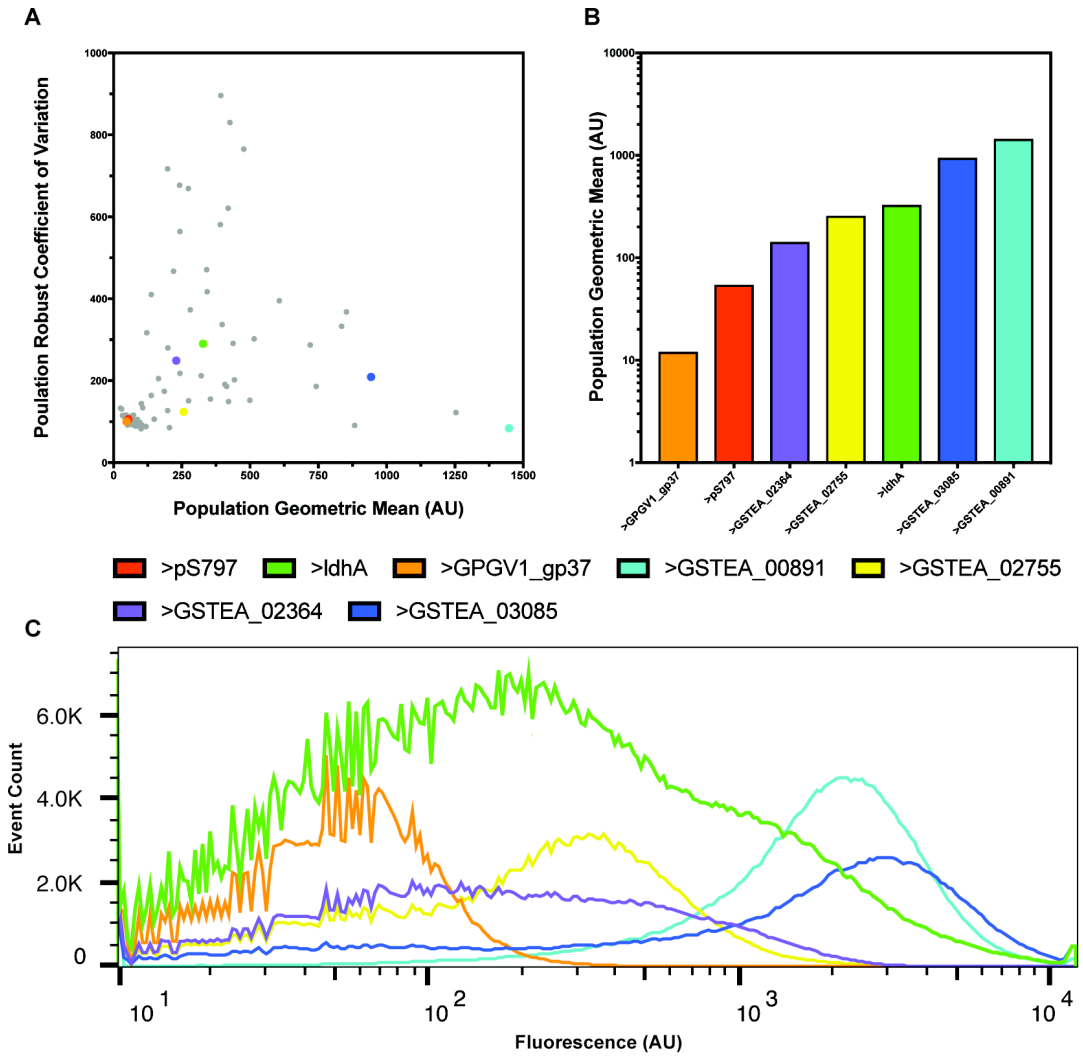


Figure 4.23: FACS analysis of *G. thermoglucosidans* cultures expressing mOrange.

Bars, histograms and points represent individual promoter sequences. For each *promoter::mOrange* fusion, 100,000 events from each of three starter cultures arising from independent transformation events were combined to form a single "meta" population of 300,000 events. The exceptions were the two controls; bars, histograms and points representing pS797 and the *ldhA* promoter represent 15 and 11 starter cultures, respectively. Cultures were excited at 488 nm and fluorescence intensity was recorded using a 585/42 nm detector.

Partition Modelling

100 random forests were fit to both the *promoter::GFP* and *promoter::mOrange* fluorescence data sets, using the same settings as applied the data from data set B, and the number of times each position caused splits in the data sets across all 100 forests was quantified. The results of the partition modelling are shown in Figure 4.24.

For both reporter proteins, the sequence positions that caused the greatest numbers of splits in the 100 random forests were spread across the entire promoter sequence. In the *promoter::GFP* data set, for example, the five promoter sequence positions that caused the greatest number of splits were, in descending order, -83, -88, -20, -95 and -5. In the *promoter::mOrange* data set, the five positions causing the most splits were -73, -43, -2, -20 and -25. This result again showed the importance of considering nucleotides and motifs outside of the canonical consensus regions when *de novo* designing synthetic promoter sequences or training sequence-function models.

The partition results also validated the use of 100 bp promoter sequences in *Geobacillus.* Although it may seem trivial to say so, the simplest way to reduce the dimensionality of the promoter sequence-function models would have been to characterise shorter promoters. Instead of bioinformatically isolating the 100 bp upstream of *Geobacillus* CDS as putative promoter elements, for example, 50 bp putative promoters could have been isolated. 50 bp sequences would have been of sufficient length to contain the putative RBS, -10 and -35 consensus regions identified in Figure 4.10, and may therefore have displayed promoter activity.

However, the partition results (Figure 4.24) showed that sequence positions upstream of the -50 position were likely to be important in determining promoter activity. Putative promoter elements of reduced length would therefore not have contained vital upstream nucleotides or motifs, and may have therefore shown reduced promoter activity as compared to the 100 bp sequences.

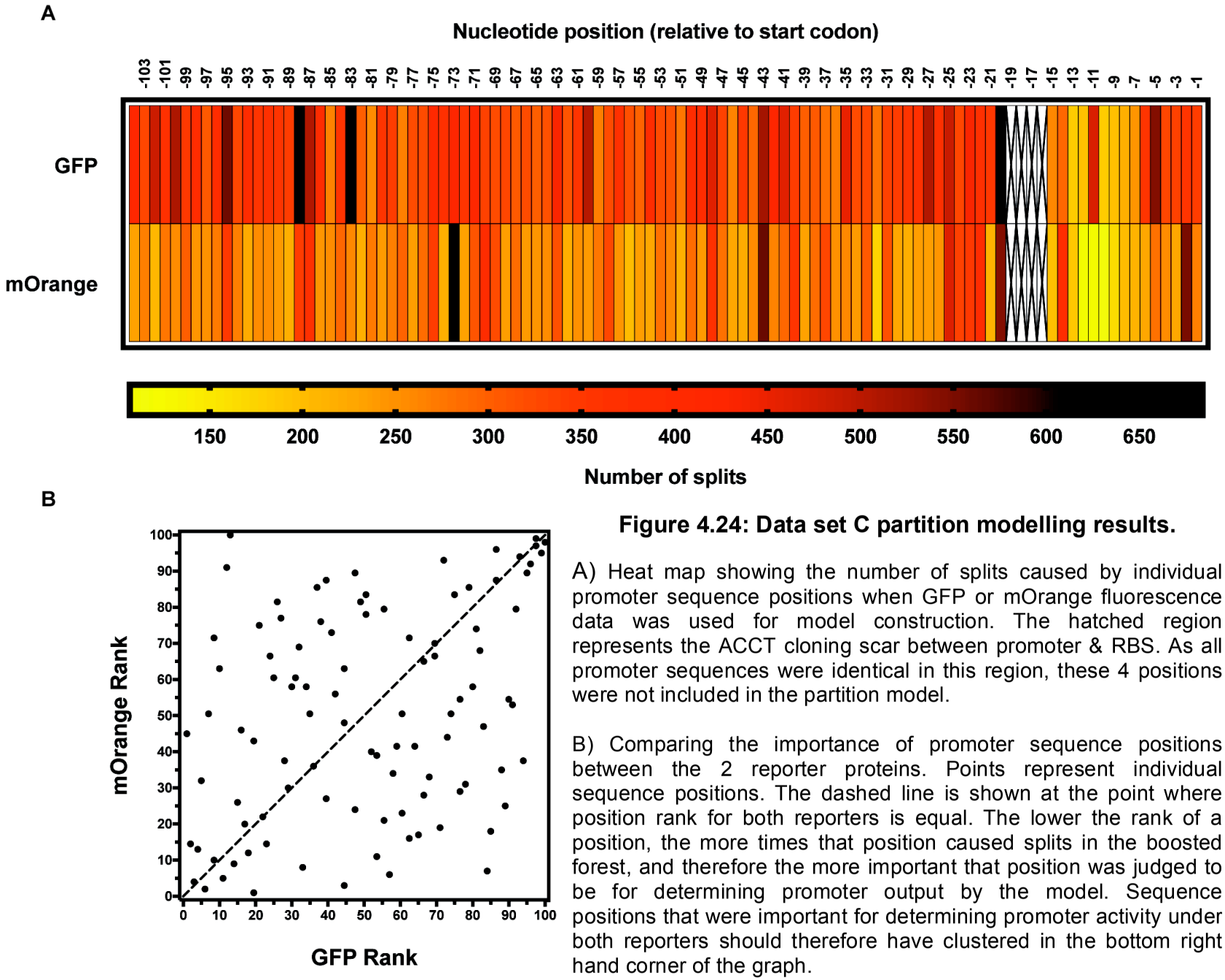


Figure 4.24: Data set C partition modelling results.

A) Heat map showing the number of splits caused by individual promoter sequence positions when GFP or mOrange fluorescence data was used for model construction. The hatched region represents the ACCT cloning scar between promoter & RBS. As all promoter sequences were identical in this region, these 4 positions were not included in the partition model.

B) Comparing the importance of promoter sequence positions between the 2 reporter proteins. Points represent individual sequence positions. The dashed line is shown at the point where position rank for both reporters is equal. The lower the rank of a position, the more times that position caused splits in the boosted forest, and therefore the more important that position was judged to be for determining promoter output by the model. Sequence positions that were important for determining promoter activity under both reporters should therefore have clustered in the bottom right hand corner of the graph.

Promoter sequence positions were included in downstream ANN & PLS sequence-function models in descending order of the number of splits caused in the 100 random forest partition models. As with data set B, the lack of correlation between the GFP and mOrange partition results (Figure 4.24B) precluded the construction of multivariate promoter sequence-function models that made simultaneous predictions of fluorescence output for both GFP and mOrange.

Partial Least Squares sequence-function models

10 promoter sequences were selected to form an independent test set on which to measure model predictive power. So that the test set contained promoter sequences with a range of activity levels, the distribution of GFP fluorescence levels in data set C (Figure 4.20) was analysed. Two sequences were selected at random from the 1st distribution quartile, five promoters were selected from the interquartile range, and three sequences were selected from the 4th quartile.

Outlier analysis was performed to identify any *promoter::GFP* fusions that might have negatively impacted upon PLS model performance (Cox & Gaudard, 2013). Quantile range analysis with a Q value of 3 showed that none of the measurements of *promoter::GFP* activity were outliers. However, Huber M-Estimation with a K value of 4 returned four outlying measurements of fluorescence activity. *GSTEA_00710::GFP*, *GSTEA_00988::GFP*, *GTDN_02924::GFP* and *GTGNS_00505::GFP*, which returned the four highest mean fluorescence measurements in data set C, were calculated to be outliers with respect to the remaining 90 *promoter::GFP* constructs. As with the strong outlying promoter sequences discussed in data set B, the four outlying promoter sequences were included in the training set for sequence-function models, on the basis that their inclusion provided the models with biologically relevant data.

PLS models were trained that modelled GFP fluorescence (y) as a function of varying numbers of promoter sequence positions (x). The number of sequence positions modelled was systematically increased from 10 to 50 in

increments of five. PLS models that fit GFP fluorescence as a function of the complete 104 bp promoter sequence were also analysed. For each of the 10 potential groups of *x* variables, two types of Cross Validation (CV) were applied. Four models were fit using *K*Fold CV, using *K* values of 4, 5, 7 or 10. Models were also fit with holdback CV. For each of the 10 groups of *x* variables, 2,000 models were fit using holdback CV, with half of the models holding back 20% of the training data for model validation and the other half holding back 33%.

Given that *Y* (GFP fluorescence) was univariate, all models used the NIPALS PLS algorithm. CV was used to determine the optimum number of LVs to extract from the data, with a maximum of 10 LVs permitted per model. For each of the 10 groups of sequence positions, the single optimum model obtained using *K*Fold CV and the single optimum model obtained using holdback CV was identified (Figure 4.25).

For both CV methodologies, the R^2 value of the test set was highest when 20 promoter sequence positions were included in the PLS models (Figure 4.25A), with prediction accuracy decreasing as the number of nucleotides included was increased beyond this point. At all model complexities, holdback CV provided more accurate predictions of GFP fluorescence output from the independent test set than *K*Fold CV. As well as returning the highest observed R^2 values, models trained on 20 nucleotides returned the lowest observed RASE values (Figure 4.25B).

The optimum PLS model that was obtained (hereafter referred to as PLS_iteration_C_1) modelled GFP fluorescence as a function of 20 promoter sequence positions, and held-back 33% of the training data for CV. PLS_iteration_C_1 returned an R^2 value of 0.6024 when applied to the training and validation sets, and an R^2 value of 0.8901 (Figure 4.26A) when applied to the test set.

No correlation was apparent in the model residuals (Figure 4.26b). Analysis by Shapiro-Wilk W test showed insufficient evidence at the 0.05 significance level to reject the null hypothesis that the underlying distribution of

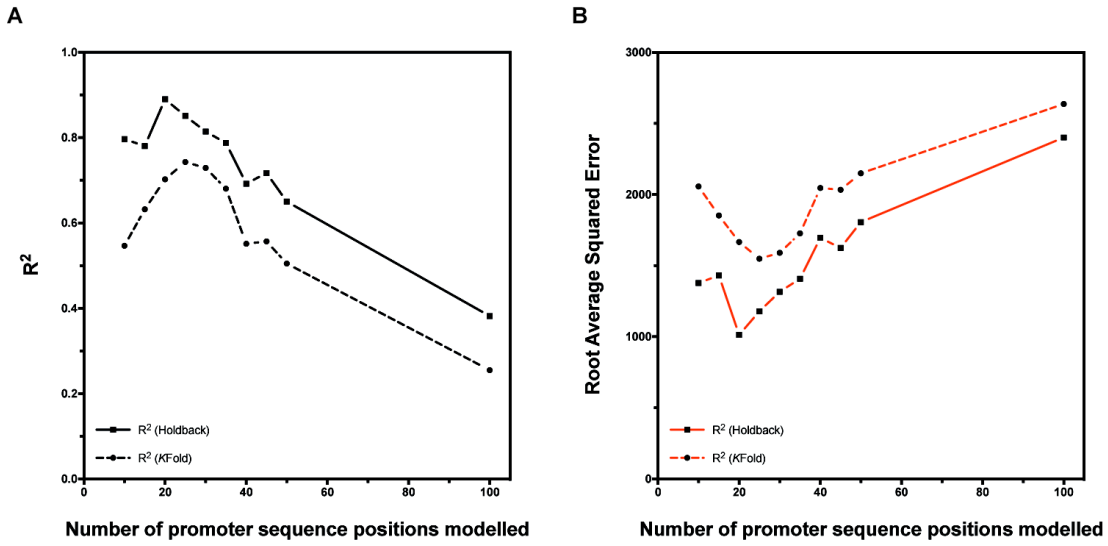


Figure 4.25: R^2 and Root Average Squared Error (RASE) values returned by Partial Least Squares (PLS) models when applied to a test data set.

PLS models were trained using the NIPALS algorithm, with a maximum of 10 Latent Variables (LVs) permitted per model. For each number of sequence positions modelled, PLS models were fit using *K*Fold CV with *K* values of 4, 5, 7 and 10, and holdback CV, with 20% or 33% of the training data set withheld to act as a validation set. Points represent the model that returned the highest R^2 value for the given number of promoter sequence positions.

The square points and solid lines represent models trained using holdback Cross Validation (CV). The circular points and dashed lines represent models trained using *K*Fold CV.

the model residuals was normal (Figure 4.26C) (W = 0.993, Prob<W = 0.380). However, as was the case with the residuals for model PLS_iteration_A_2 (Figure 4.8), visual analysis of a histogram of the model residuals questioned this conclusion, as the residuals were clearly not normally distributed (Figure 4.26C). Given the small sample size (n = 10 in the test data set) and the previously reported poor power of the Shapiro-Wilk test when the sample size being assessed is small (Razali & Wah 2011, Le Boedec, 2016), the result of the Shapiro-Wilk test was judged inadequate for determining normality in the residuals of model PLS_iteration_C_1. Despite the potential lack of normality, the lack of correlation in the residuals (Figure 4.26B) suggested that PLS_iteration_C_1 did not contain significant underlying biases. Additionally,

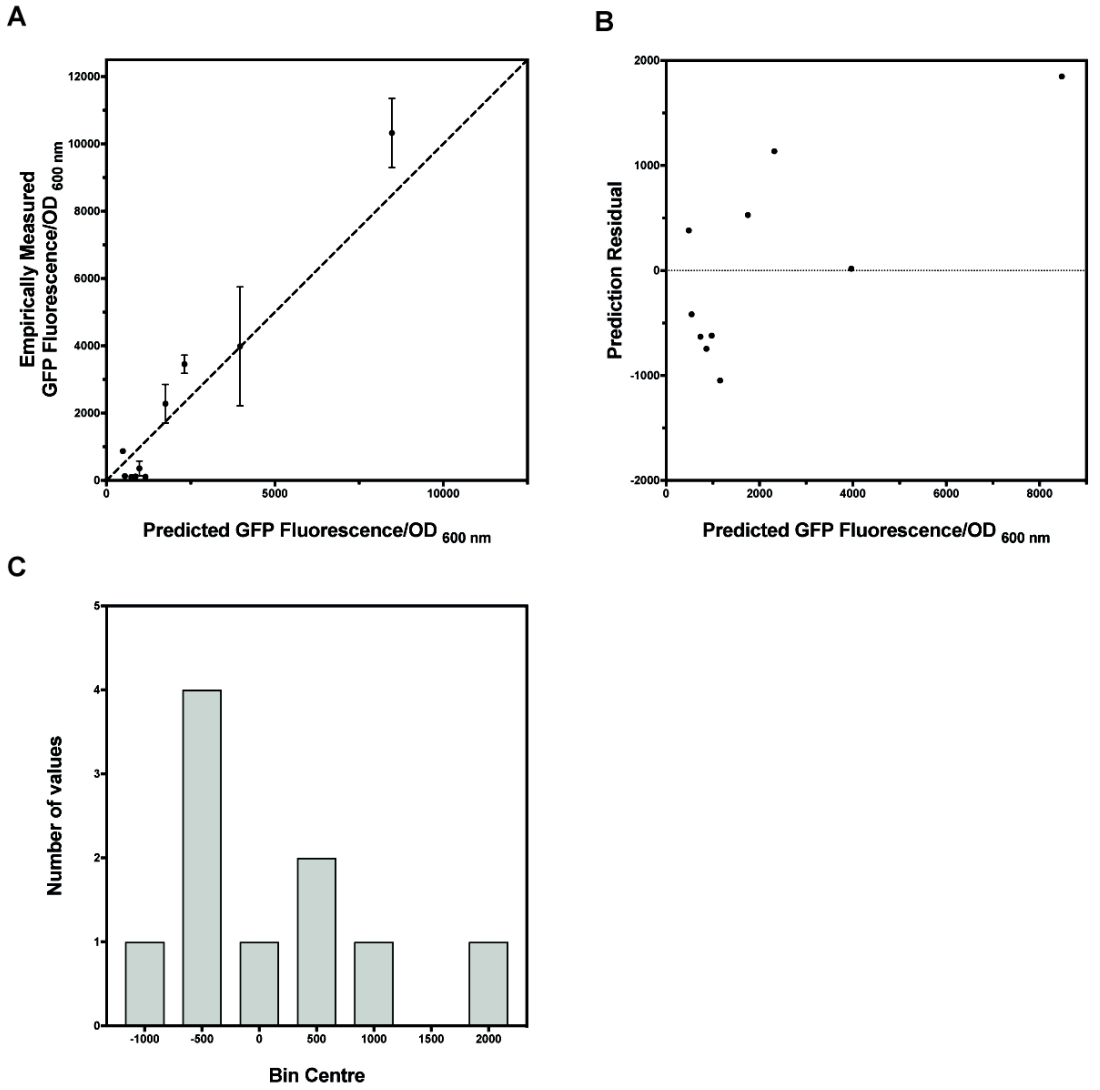


Figure 4.26: Model diagnostics for optimal obtained Partial Least Squares model of data set C, PLS_iteration_C_1.

A) Empirically measured GFP fluorescence, plotted against the GFP fluorescence as predicted by the model. Points represent the activity levels of individual promoter sequences. Empirically measured fluorescence and absorbance after 24 h incubation in 96-well plate format, Empirical values are the mean of n = 3 starter cultures arising from independent transformants, with standard deviation error bars shown, unless hidden by the points. The dashed line represents the point at which empirically measured and predicted fluorescence values are equal

B) GFP fluorescence predicted by the model, plotted against prediction residual. Dashed line shown at the point where the prediction residual is equal to 0.

C) Histogram of model residual distribution. A Shapiro-Wilk W test returned W = 0.993 and Prob<W = 0.380. There was therefore insufficient evidence to reject the null hypothesis that the underlying distribution of the model residuals was normal at the 0.05 significance level. However, visual analysis of the histogram questioned this conclusion, as the data were clearly not normally distributed.

the model provided a reasonable fit of the training data and had good predictive power when applied to previously unseen data.

Artificial Neural Network sequence-function models

As with the ANNs that were trained on data set B, ANNs were trained using a statistical Design of Experiments (DoE) approach. ANNs were trained using only a single hidden layer. Whilst two-layer networks could potentially have been used to map the promoter response surface, the resulting increase in model complexity carried the risk of increasing model variance and overfitting the model to the training data (Hastie *et al.*, 2009). Additionally, a single-layer network containing 19 nodes had previously been shown to be sufficiently complex to describe 224 bp regulatory sequences in *E. coli* (Meng *et al.*, 2013).

A screening design was used to identify which of the ANN parameters were having the greatest impact on model performance. The ANN parameters that were included in the screening design are summarised in Table 4-4. The specified parameters were combined in a full-factorial manner, which resulted in 81 ANN architectures being specified. Each of the 81 network architectures was fit 500 times, with each fit using a unique, specified random seed. All models used the squared penalty method. For each of the 81 ANN architectures, the single ANN that returned the highest R^2 when applied to the test set was identified.

Parameter	Levels specified
Activation function personality	Gaussian, Linear or TanH
Number of nodes in the hidden layer	3, 5 or 7
Cross Validation (CV) methodology	*K*Fold, with *K* = 4, *K* = 5 or *K* = 8
Number of promoter sequence positions modelled	10, 20 or 100

Table 4-4: Artificial Neural Network parameters included in screening experiment, and the values specified for each parameter.

When applied to the test data set, the 81 ANNs returned R^2 values ranging from 0.435 to 0.934. The optimum ANN obtained modelled GFP fluorescence as a function of 20 promoter sequence positions, using a five-node hidden layer, the TanH activation personality and *K* = 5 CV. Of the 10 models that returned the highest R^2 values, eight used the TanH activation function personality, suggesting that a sigmoidal activation function provided the most accurate mathematical abstraction of the promoter design space. This result was concurrent with the literature, as sigmoidal activation functions had previously been used to train promoter sequence-function models in *E. coli* (Meng *et al.*, 2013).

The results of the screening experiment were subjected to statistical analysis. A standard least squares model with effect screening emphasis showed that both the number of promoter sequence positions included in the ANN and the personality of the activation function were having a statistically significant impact on model performance at a significance level of 0.05 (number of sequence positions LogWorth = 7.632, P = 0.000, activation function personality LogWorth = 6.298, P = 0.000). The number of nodes in the hidden layer and the *K* value used in CV did not return statistically significant results (P = 0.582 and P = 0.671, respectively).

The screening results were also analysed by PLS model, using *K*Fold CV where K = 7. The test set R^2 value was used as the *y* variable and the parameters summarised in Table 4-4 were used as *x* variables. The resulting model extracted a single LV from the data, and was capable of explaining 12.5% of the cumulative variation in *X* and 53.525 % of the cumulative variation in *Y.*

Four factors exceeded the VIP threshold value of 0.8 (Eriksson *et al.*, 2006), which suggested that these factors were having a statistically significant effect on ANN predictive power. The Linear and TanH activation functions both exceeded the VIP threshold, as did the effect of including 20 and 100 promoter sequence positions in the ANNs (Figure 4.27A). Analysis of the model coefficients showed that the TanH activation function was predicted to positively

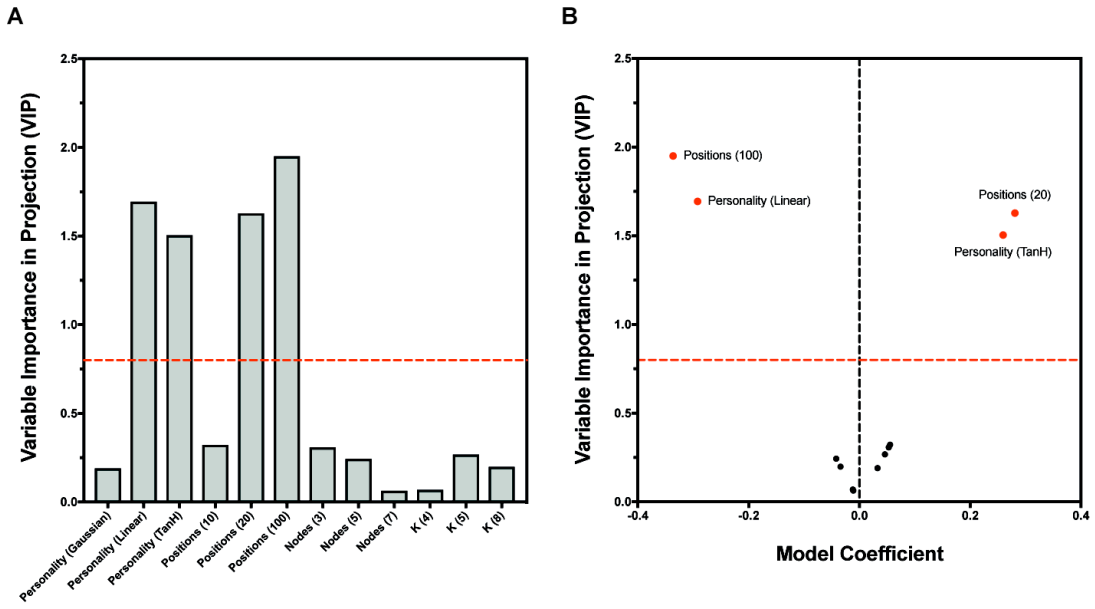


Figure 4.27: Assessing the contribution of Artificial Neural Network model parameters to determining predictive power using a Partial Least Squares model.

A) Variable Importance Plot (VIP) and B) VIP v Coefficient plot.

In both panels the dashed red line represents the VIP threshold value of 0.8, above which *x* variables are predicted to have a statistically significant impact on model output.

VIP and coefficient values were returned by a PLS model that modelled the R^2 value returned by ANNs when applied to a test data set, as a function of activation function personality (Personality), the number of promoter sequence positions included in the ANN design (Positions), the number of nodes included in the hidden layer (Nodes) and the value of *K* used in ANN Cross Validation (*K*).

contribute to ANN predictive power, whereas the Linear activation function was predicted to negatively impact predictive power (Figure 4.27B).

Taken together, the results of the screening design suggested that the TanH activation function and the number of sequence positions included in theANN were likely to be vital in determining ANN predictive power. A second iteration of ANN design was therefore undertaken using only TanH activation functions. All ANNs used a single hidden layer and the squared penalty function. The number of promoter sequence positions that were included in the ANNs ranged from 10 to 100, increasing in increments of 10 positions, and the

number of nodes included in the hidden layers was between 3 and 15, increasing in increments of 2. The parameters were combined in a full-factorial manner. As such, a total of 70 ANN architectures were generated. Each of the 70 ANN architectures were fit 1,000 times, using unique, specified random seeds. 500 runs of each network design used *K*Fold CV where *K* = 4, and the remaining 500 runs used *K*Fold CV where *K* = 5.

The individual ANN that returned the highest R^2 value when applied to the test set was identified for each of the 70 network architectures (Figure 4.28). The best performing ANN that was obtained returned an R^2 value of 0.9304 when applied to the test set, and modelled GFP fluorescence as a function of 20 promoter sequence positions, with 5 nodes in the hidden layer. However, the ANN design space did not contain an obvious single local optimum. Instead, ANNs returning high R^2 values (> 0.8) were returned by at least 1 network design for each of the 10 groups of sequence positions that were analysed.

Model predictive power was shown to decrease when either 90 or 100 promoter sequence positions were modelled (Figure 4.28). In the case of ANNs that modelled GFP fluorescence as a function of complete promoter sequences, the optimum network obtained used 9 nodes in the hidden layer and returned an R^2 value of 0.8199 when applied to the test set. To test if ANNs of increased complexity could better model complete promoter sequences, single layer models were trained using 17, 19 or 21 nodes. Models containing more than 21 nodes in the hidden layer were not trained, as highly complex networks (*i.e.* those with many nodes) are known to result in models which have high variance and that are overfit to the training data (Hastie *et al.*, 2009). 1,000 ANNs were trained for each of the 17-, 19- and 21- node architectures, but none performed better than the nine-node model when applied to the test set, either in terms of R^2 or RASE (Figure 4.29).

Model ensembling was applied to the ANNs in an attempt to improve predictive power. The ensembling strategy is summarised in Figure 4.30. For each number of promoter sequence positions that were modelled, the network architecture that returned the single ANN with the highest test set R^2 value was

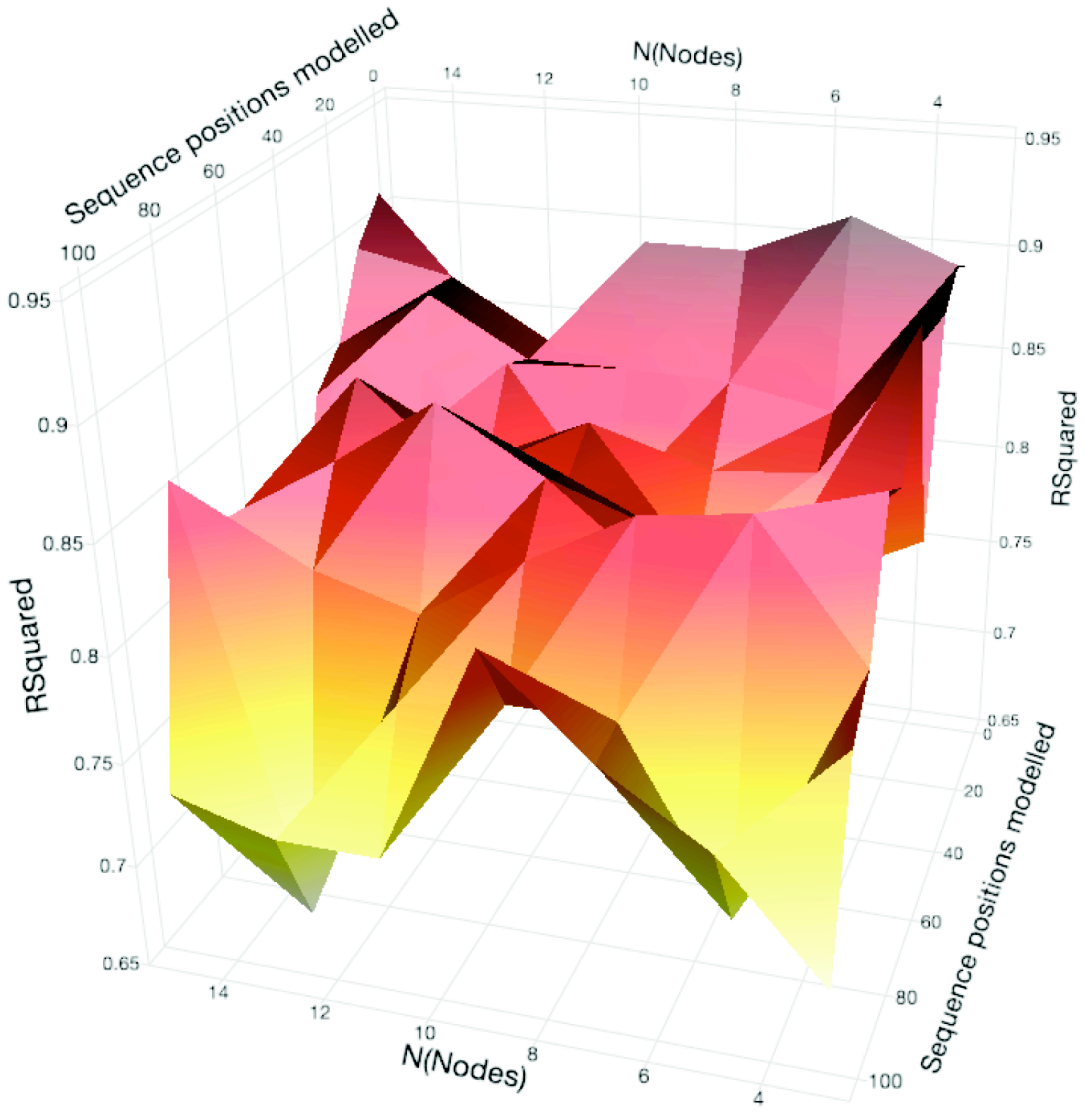


Figure 4.28: Response surface showing the R^2 values returned by Artificial Neural Networks using the TanH activation function when applied to a test data set.

ANNs varied in terms of the number of promoter sequence positions included in the model and the number of nodes in the ANN hidden layer. A total of 70 ANN architectures were generated. Each of the 70 ANN architectures was fit 1,000 times, using unique, specified random seeds. 500 runs of each of the 70 network architectures used *K*Fold CV where *K* = 4, and the remaining 500 runs used *K*Fold CV where *K* = 5. For each of the 70 architectures, the single model that returned the highest R^2 value when applied to the test set was included in the response surface.

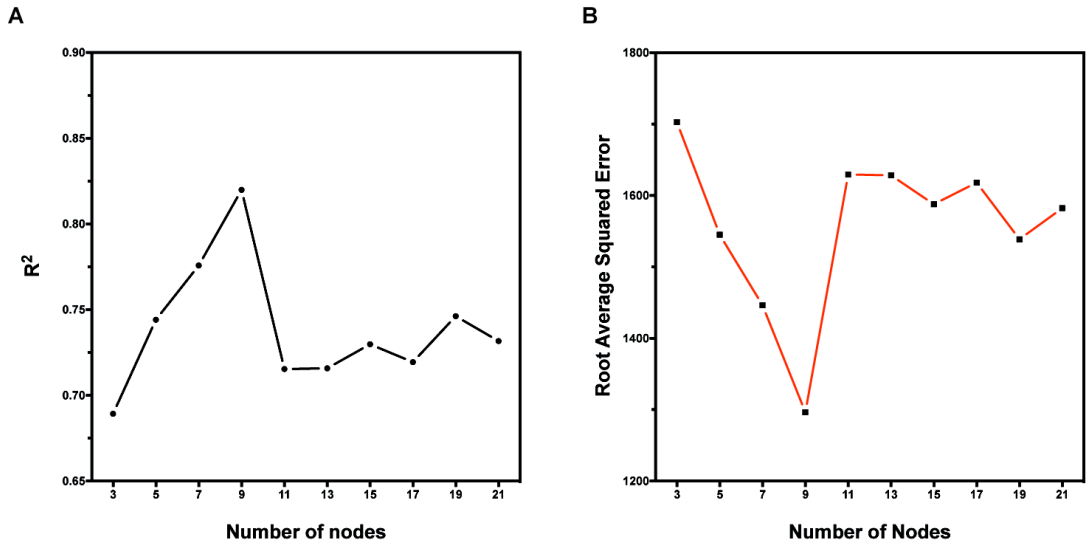


Figure 4.29: Model performance statistics for single layer Artificial Neural Networks modelling GFP fluorescence as a function of complete promoter sequences.

R^2 and Root Average Squared Error (RASE) values were returned when ANNs were applied to an independent test data set of 10 promoter sequences. For each number of nodes specified 1,000 ANNs were fit, with each fit having a unique, specified random seed. For each number of nodes, 500 models were fit using *K*Fold CV with $K = 4$ and 500 models were fit using *K*Fold CV with $K = 5$. Points represent the R^2 and RASE values returned by the single highest performing model for each number of nodes.

identified. Once the optimal network architecture was identified, ensemble models were created. The top 10 highest performing ANNs with the chosen architecture (as judged by the test set R^2 value) were used to create nine progressively larger ensembles, with ANNs being included in the ensemble in descending order of their test set R^2 value. Once formed, the ensemble models were applied to the test set of promoter sequences to quantify their predictive power.

Ensemble models did not always return higher R^2 values than the best performing constituent model for the given number of promoter sequence positions. For example, the individual best performing ANN that modelled GFP fluorescence as a function of 30 promoter sequence positions returned an R^2 value of 0.8995 when applied to the test data set. The best performing

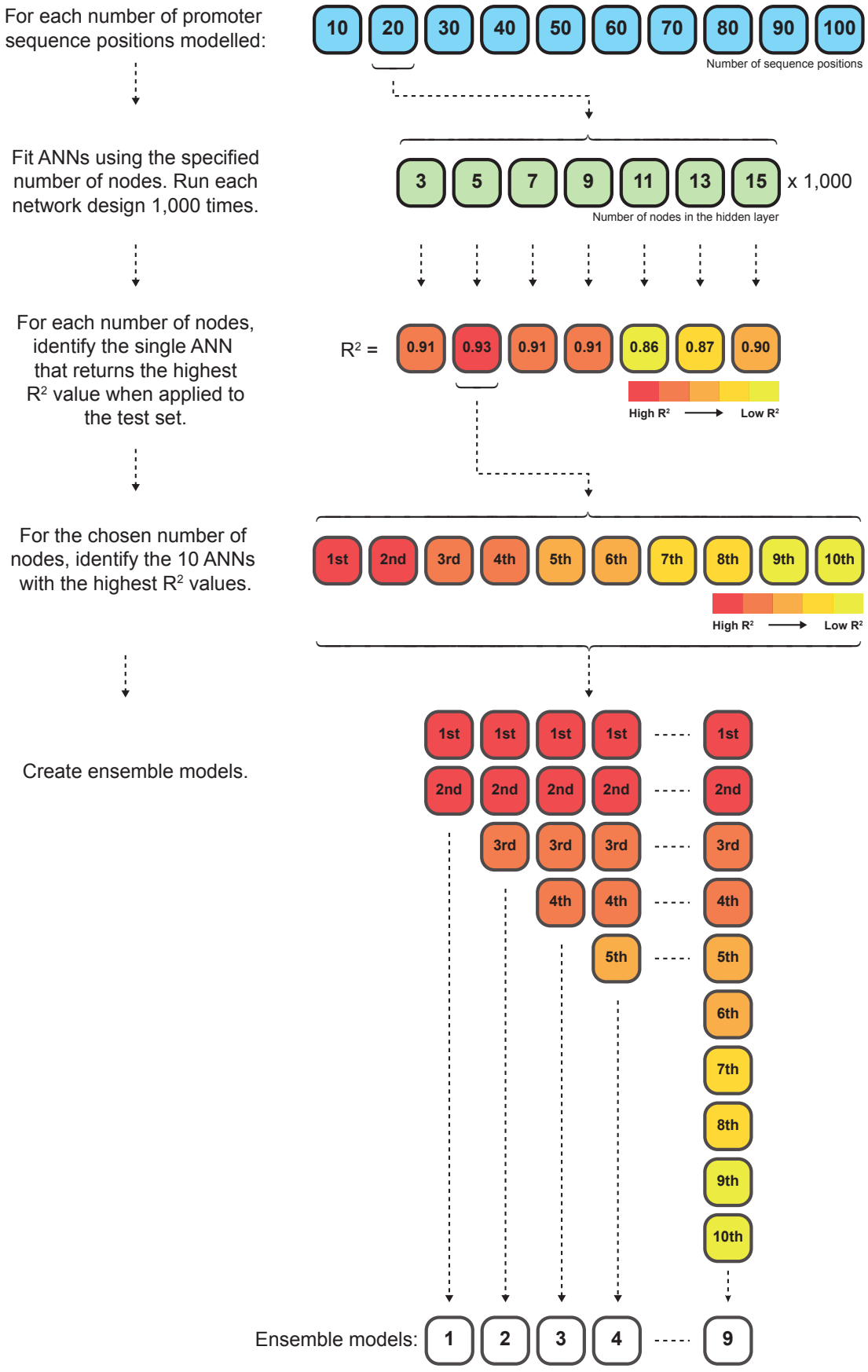


Figure 4.30: Schematic representation of the Artificial Neural Network ensembling strategy.

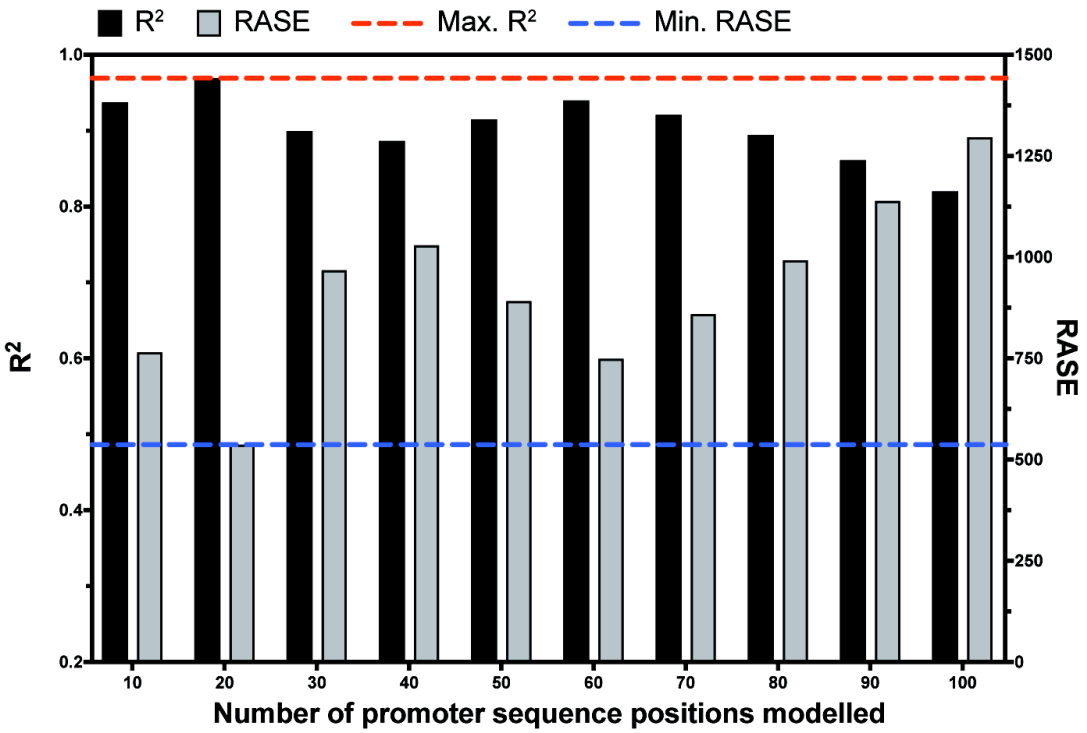
ensemble, however, returned an R^2 value of 0.8766. Figure 4.31 summarises the test set performance of the optimal model, whether individual ANN or ensemble, that was obtained for each of the groups of promoter sequence positions that were analysed.

In total, five models were identified that returned an R^2 value greater than 0.9 when applied to the test data set (Table 4-5). All five of the models appeared to have good predictive power when applied to the test data set, and also returned high R^2 values when applied to the training and validation test sets. These results suggested that the models provided both a good fit of the training data and performed well when applied to previously unseen data.

Model	Sequence positions (*x*)	N(Nodes)	R^2 (Training & Validation)	R^2 (Test)
10_Tan15	10	15	0.9251	0.9373
20_Tan5	20	5	0.9746	0.9691
50_Tan3	50	3	0.8979	0.9149
60_Tan13	60	13	0.9753	0.9399
70_Tan11	70	11	0.9930	0.9209

Table 4-5: A summary of the architecture and performance of five high-performing Artificial Neural Network models.

Modelling GFP fluorescence as a function of small numbers of promoter sequence positions carried the potential risk of data twinning. In the case of ANNs predicting GFP fluorescence as a function of 10 promoter sequence positions, for example, it was theoretically possible that the nucleotides at those 10 sequence positions in a promoter in the test set would have been identical to the nucleotides at those 10 positions in a promoter in the training and validation sets. Predictive power would therefore have been artificially inflated (Clarke *et al.*, 2009, Raccuglia *et al.*, 2016). However, analysis of the promoter sequences showed that none of the promoters in the test set were an exact match for any of the promoters in the training and validation sets when the 10, 20, 50, 60 or 70 sequence positions included in the high-performing ANNs were considered.



Sequence Positions (*x*)	N (Nodes)	Ensemble model? (Y/N)	Number of constituent models	Model name
10	15	Y	2	10_Tan15
20	5	Y	2	20_Tan5
30	3	N	-	30_Tan3
40	11	Y	2	40_Tan11
50	3	N	-	50_Tan3
60	13	Y	2	60_Tan13
70	11	Y	3	70_Tan11
80	5	Y	7	80_Tan5
90	15	N	-	90_Tan15
100	9	N	-	100_Tan9

Figure 4.31: R^2 and Root Absolute Squared Error (RASE) values returned by 10 single layer Artificial Neural Network designs when applied to a test data set.

Bars representing R^2 values are shown in black, and are plotted on the left hand y-axis. Bars representing RASE values are shown in grey, and are plotted on the right-hand y-axis. The dashed red and blue lines represent the R^2 and RASE values returned by the optimal obtained ANN, which was trained on 20 promoter sequence positions and had 5 nodes in the hidden layer. For each number of promoter sequence positions (*x*) modelled, the table shows number of nodes in the hidden layer (N(Nodes)) and whether or not multiple ANNs were aggregated to reach the final test statistics. If models were aggregated, the number of constituent models that went into the final prediction is shown.

The high R^2 values that were returned by the 5 high-performing models shown in Table 4-5 did not therefore appear to have been an artefact of data-twinning, as was the case with the models that were trained on data set A.

A secondary test set was identified to further test the predictive power of the five high-performing ANNs and the PLS model PLS_iteration_C_1. 10 putative promoter sequences were selected at random from across the *Geobacillus* promoter phylogeny and synthesised upstream of *GFP* by ATUM (previously DNA 2.0, California, United States of America).

When applied to the secondary test set (Figure 4.32), none of the isolated models showed the same level of predictive accuracy as had been observed in the primary test set (Figure 4.31). The model that returned the most accurate predictions was 10_Tan15, which returned an R^2 value of 0.5211. (Figure 4.32C). The second highest R^2 value was returned by the PLS model PLS_iteration_C_1 (Figure 4.32B, R^2 = 0.3595). All six of the tested models performed particularly poorly when the empirically measured fluorescence output of a *promoter::GFP* fusion was low.

The best performing ANN of the complete promoter sequence, 100_Tan9, did not show greater predictive accuracy than 10_Tan15 when applied to the secondary test set; 100_Tan9, returned an R^2 value of -0.068. This result validated the use of partition modelling to identify key promoter sequence positions for inclusion in downstream ANN and PLS models. The partition models were capable of accurately identifying promoter sequence positions that were key in determining promoter output, thereby reducing the dimensionality of the design-space modelled by the downstream ANN and PLS models, boosting predictive power.

The model averaging approach that was initially applied to the ANNs derived from data set C could be defined homogeneous; the constituent ANNs for each ensemble used the same model architecture (*i.e.* they all had the same number of nodes in the hidden layer) but applied different model weights. However, heterogeneous ensembles, in which the constituent models use

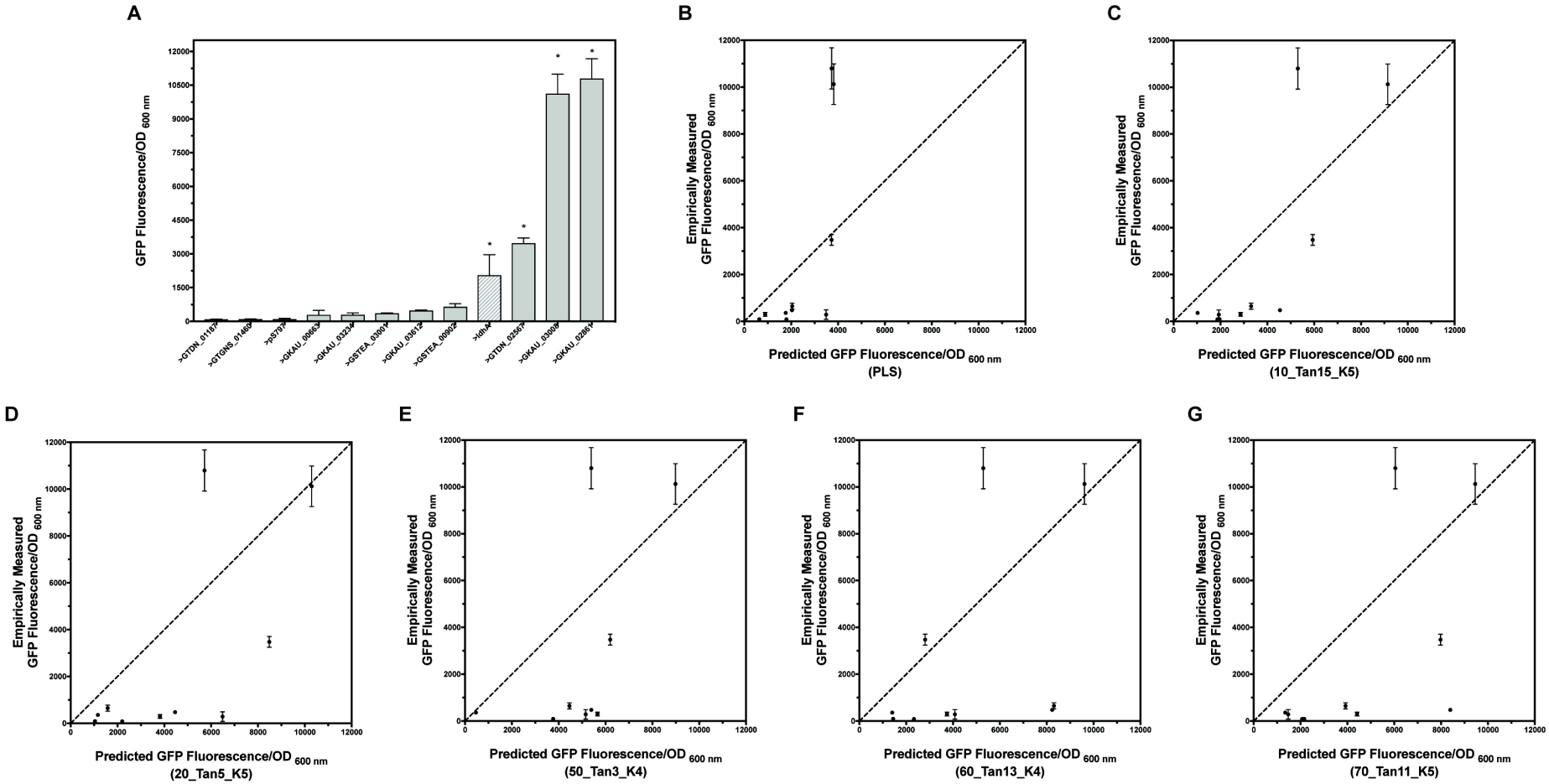


Figure 4.32: Empirically measured promoter activity of a secondary test set of 10 putative promoters, as compared to the activity levels predicted by Artificial Neural Network & Partial Least Squares models derived from data set C.

Empirical measurements of promoter activity taken after 24 h growth in 96-well plate format. Bars and points represent the mean of n = 3 starter cultures arising from independent transformants, with standard deviation error bars shown, unless hidden by the bar or point. Panel A shows empirically measured fluorescence. The hatched bar represents the positive control, the *ldhA* promoter and the black bar represents the negative control, the empty vector pS797. Promoters for which mean GFP fluorescence was statistically significantly different to the negative control are indicated by an asterisk. Panels B-G show empirically measured fluorescence as predicted by B) the PLS model PLS_iteration_C_1, C) 10_Tan15, D) 20_Tan5, E) 50_Tan3, F) 60_Tan13 & G) 70_Tan11.

different architectures and/or data sets (Yang *et al.*, 2013) were not initially used.

Ensemble networks have previously been shown to be most useful when the constituent models return predictions that are as accurate but diverse as possible (Granitto *et al.*, 2005). High levels of prediction diversity between the five high-performing ANNs and PLS_iteration_C_1 were apparent when the models were applied to bioinformatically identified putative *Geobacillus* promoters that had not been characterised *in vivo* (Figure 4.33). The predictions of promoter activity returned by even the two most similar models (60_Tan13 and 70_Tan11) showed, at very best, minimal positive correlation. A linear regression of the two sets of predicted values returned an R^2 of 0.4437. Linear regressions comparing the other 14 possible combinations of the six models returned R^2 values ranging from 0.052 (PLS_iteration_C_1 v 50_Tan3) to 0.3813 (PLS_iteration_C_1 v 20_Tan5). The extremely low R^2 values that were returned by the linear regressions indicated that the predicted activity level of each previously uncharacterised putative promoter varied significantly depending on which model was used to make the prediction. Although such prediction diversity is known to be required for effective ensembling, it does not guarantee that an ensemble model will out-perform individual constituent models (Johansson *et al.,* 2007).

To test if a heterogeneous ensemble could out-perform the optimal homogenous ensemble (the model 10_Tan15), all possible combinations of the five high-performing ANNs (Table 4-5) and PLS_iteration_C_1 were generated. The resulting 57 ensembles were then applied to the secondary test set. The best performing heterogeneous ensemble combined PLS_iteration_C_1 and the ANNs 10_Tan15 and 50_Tan3. When applied to the training and test sets, the ensemble returned R^2 values of 0.8873 and 0.9402, respectively. However, when applied to the secondary test set the ensemble returned an R^2 value of 0.4779, and therefore did not show improved predictive power as compared to 10_Tan15, which returned an R^2 value of 0.5211.

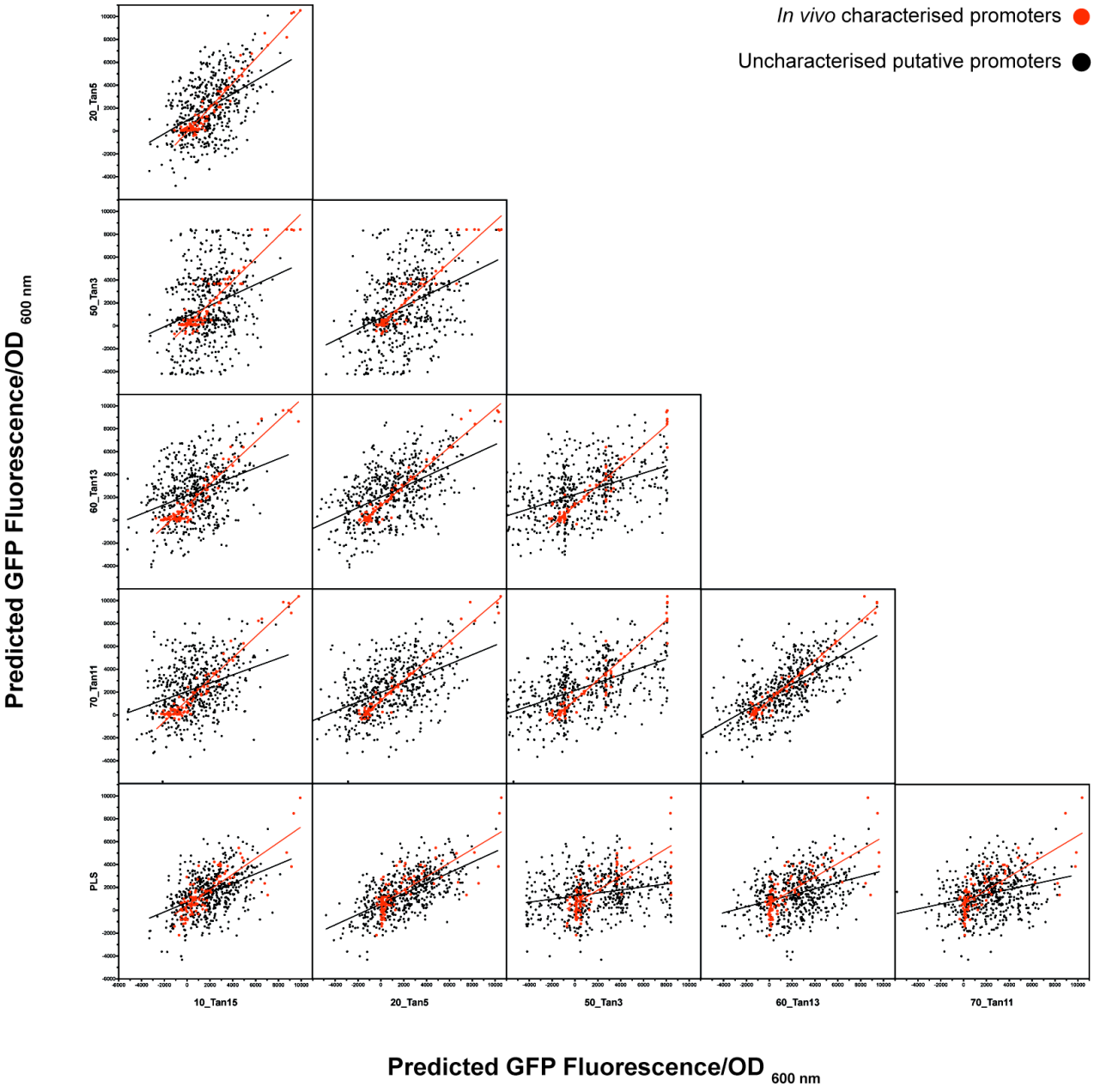


Figure 4.33: Scatterplot matrix showing GFP fluorescence output of putative promoter sequences as predicted by high performing Artificial Neural Network and Partial Least Squares models.

Points represent individual promoter sequences. Predictions of fluorescence output are as calculated by the ANNs 10_Tan15, 20_Tan5 50_Tan3, 60_Tan13 and 70_Tan11, and by PLS_iteration_C_1. Red points represent promoter sequences that were included in the model training and validation process. Black points represent bioinformatically identified putative promoter sequences that were not characterised *in vivo*. The solid lines represent linear regressions of the data.

4.3 Discussion

4.3.1 Identification and characterisation of putative promoters

Three progressively larger sets of bioinformatically identified putative promoter sequences were characterised in *G. thermoglucosidans*. In total, 105 *promoter::GFP* fusions, covering a total expression range of 148-fold, were characterised (Figure 4.20*,* Figure 4.32A). Of these sequences, 47, covering an expression range of 30-fold, were defined as active promoters. 82 *promoter::mOrange* fusions, covering a total expression range of 107-fold were also characterised (Figure 4.21), of which 32, covering an expression range of 8.4-fold, were defined as active promoters. As discussed in Chapter 5, context effects meant that correlation in promoter activity between the two reporter proteins was minimal (a linear regression of the two data sets returned an R^2 value of 0.199).

At the start of this investigation, the constitutive promoters that had been applied to the control heterologous expression in the genus *Geobacillus* were three endogenous regulatory elements; the *G. kaustophilus sigA* promoter (Suzuki *et al.*, 2012), the oxygen-dependent *ldhA* promoter, isolated from *G. thermodenitrificans* or *G. stearothermophilus* (Cripps *et al.*, 2009, Bartosiak-Jentys *et al.*, 2012, Lin *et al.*, 2014, Kananavičiūtė & Čitavičius, 2015) and the *G. stearothermophilus* ribonuclease HIII promoter (Blanchard *et al.*, 2014). The endogenous *Geobacillus* promoters that were characterised in this investigation therefore represented a major expansion in the number of constitutive regulatory sequences available for synthetic biology and metabolic engineering projects in this genus of industrially relevant thermophiles.

In total, 58 *promoter::GFP* fusions and 50 *promoter::mOrange* fusions were classed as inactive. Furthermore, the mean fluorescence activity of 23% of *promoter::GFP* fusions and 17% of *promoter::mOrange* fusions fell within one standard deviation of the mean fluorescence of the negative control.

Alternative methods for promoter identification might have shown greater specificity in *Geobacillus*. “Omics” approaches, for example, (Mendoza-Vargas

et al., 2009, Li *et al.*, 2015, Luo *et al.*, 2015) could potentially have resulted in more accurate determination of transcription start sites. Alternatively, machine learning techniques could have been used to screen the *Geobacillus* genome for putative promoter elements (Mann *et al.*, 2006, Umarov & Solovyev, 2017). However, these approaches can be time- and resource-intensive, and, in the case of the machine-learning approaches, require prior understanding of the biological and statistical characteristics of the sampled DNA sequences (Song *et al.*, 2016). Such information is not always readily available in non-model, industrially relevant organisms.

In contrast, the bioinformatic approach to promoter discovery that was applied in this investigation allowed putative promoter sequences with a broad range of *in vivo* activity levels to be identified and characterised relatively quickly and easily. This approach is therefore potentially broadly applicable to a wide range of industrially relevant organisms.

4.3.2 Promoter sequence-function modelling

Partition modelling revealed that regions outside of the canonical consensus regions (*i.e.* -10, -35 and RBS motifs) were key for determining promoter activity in *Geobacillus* (Figure 4.24). When promoter sequences were characterised upstream of GFP, sequence positions -83, -88 and -95 caused the 1st, 2nd and 4th most splits in 100 random forests, respectively. When characterised upstream of mOrange, positions -73, -70 and -87 caused the 1st, 6th and 10th most splits, respectively. Additionally, a sequence alignment of 21 promoter sequences that showed promoter activity *in vivo* showed regions of AT-rich conserved sequence towards the 5' terminus of the promoter sequence that were not as heavily conserved in sequences that did not show promoter activity (Figure 4.10).

In *E. coli*, studies have shown that the canonical UP-element is typically AT-rich, and boosts transcription initiation through interactions with the C-terminal domain of the RNA polymerase alpha subunit (Ross *et al.*, 1993, Aiyar *et al.*, 1998, Estrem *et al.*, 1998). UP-elements have also been reported to

increase transcription activation in *Bacillus subtilis* (Meijer & Salas, 2004, Phan *et al.*, 2012) The results of the partition models and sequence alignments suggested that promoter regions upstream of the -35 box also play an important role in boosting transcription initiation in *Geobacillus*.

The use of ANN and PLS models to infer a quantitative relationship between the DNA sequence of a promoter and *in vivo* function was only moderately successful. ANN and PLS models were trained using data derived from the characterisation of the three progressively larger sets of putative promoters. In data set C, 95 characterised *promoter::GFP* fusions were available for division into training, validation and initial test sets. To reduce the dimensionality of the promoter design space and thereby potentially increase the predictive power of the final models, GFP fluorescence was modelled as a function of varying number of promoter sequence positions. Five putatively high-performing ANNs were identified (Table 4-5), respectively using 10, 20, 50, 60 or 70 promoter sequence positions to infer promoter activity. A putatively high-performing PLS model was also obtained (Figure 4.26).

When applied to training, validation and primary test sets the putatively high-performing ANN and PLS models all returned high (> 0.89) R^2 values, which was suggestive of strong predictive capability. However, these measures of predictive power proved overly optimistic. When applied to secondary test sets, the six putatively high-performing models displayed inadequate generality (Figure 4.31). The best-performing ANN, which modelled GFP fluorescence as a function of 10 promoter sequence positions using the TanH activation function and 15 nodes in a single hidden layer, returned an R^2 of 0.5211 when applied to a secondary test set. Prediction accuracy was particularly poor when a DNA sequence showed no *in vivo* promoter activity.

Although novel in *Geobacillus*, a small number of *E. coli* promoter sequence-function models that applied either ANNs or PLS and displayed good predictive power are described in the literature. For example, the PLS model described by De Mey *et al.* was successfully applied to pathway engineering in *E. coli*. Once trained, the model was used to make a prediction of strength for

the wild-type *ppc* promoter, which was not in the training data set. The wild-type promoter was subsequently replaced *in vivo* with promoter sequences that were predicted by the PLS model to have greater strength. The resulting upregulation of the *ppc* gene was broadly as predicted by the model (De Mey *et al.*, 2010).

Successful derivation of accurate PLS models of the relationship between promoter sequence and function was also reported by Jonsson *et al.*, who were able to accurately predict the *in vivo* expression strength of two synthetic promoters. The synthetic sequences were derived by selecting nucleotides at each sequence position based on which of the four nucleotides had the largest positive PLS model coefficient at that position (Jonsson *et al.*, 1993). Finally, the ANN derived by Meng *et al.* was used to design synthetic promoter sequences that were subsequently used to successfully upregulate expression of a small peptide toxin, BmK1, in *E. coli* (Meng *et al.*, 2013).

The predictive power of the published sequence-function models was in contrast to the relative lack of predictive power shown by the models derived in this investigation (Figure 4.31). This discrepancy may have been a result of the way that the promoter libraries on which the models were trained were generated.

In this investigation, promoter sequences were selected for *in vivo* characterisation in a way that required no *pre hoc* understanding of *Geobacillus* regulatory elements. By selecting promoters from across the *Geobacillus* promoter phylogeny, this investigation aimed to maximise the sequence diversity of the characterised promoters, and therefore maximise the proportion of the promoter design space that was empirically explored. In contrast, the published studies employed an *a posteriori* approach, in which the characterised promoters either contained defined motifs or were highly homologous.

The PLS model reported by De Mey *et al.*, for example, applied a training set derived from Saturation Mutagenesis of Flanking Regions (SMFR). *E. coli* -10 and -35 consensus regions were flanked with 12 semi-conserved and 20

fully degenerate nucleotides, resulting in a library of 49 characterised 57 bp sequences (De Mey *et al.*, 2007). 42 of these sequences were used for model training. The remaining seven sequences were used as a test set, and the resulting PLS model was able to accurately predict promoter activity for six out of the seven sequences. The DNA sequences of the synthetic promoter sequences used for model training were not published. However, given the conserved nature of the degenerate oligonucleotide used for promoter derivation, significant sequence similarities were likely between the training and test data sets.

The neural network described by Meng *et al,* was also trained on a mutagenesis-derived promoter library. Mutations were introduced into the 224 bp sequence of the *E. coli* wild-type *Trc* promoter using epPCR, with the error-rate reported to have “reached up to about 20%” (Meng *et al.*, 2013). However, analysis of the published promoter sequences revealed an average error rate of only 10%, with 18 sequences containing fewer than 10 mutations.

Finally, although the training set of 25 promoters that was applied by Jonsson *et al.* was not derived by mutagenesis, sequences were selected for inclusion in the training data only if they contained a 17 bp spacer between the -35 and -10 regions and a 7 bp spacer between the -10 region and the transcription start site (Jonsson *et al.*, 1993). Four nucleotide positions from the 68 bp sequences were also identical across all 25 promoters.

By maintaining consensus regions (De Mey *et al.*, 2007), using epPCR with a low error rate to mutate a single promoter sequence (Meng *et al.*, 2013), or by keeping spacing between key regions constant (Jonsson *et al.*, 1993), the studies discussed above trained sequence-function models on promoter libraries with considerable sequence homology. Such homology ensured that key promoter sequence motifs, such as consensus and spacer sequences, were either identical or broadly consistent in promoters within the training data sets. Individual sequence positions within different promoter sequences were therefore likely to be more directly comparable than if such structures were not conserved (Jonsson *et al.*, 1993).

By characterising promoter sequences that were potentially not consistent with regards to consensus motif sequence or location, the *Geobacillus* promoter library that was discussed in this investigation likely lacked the underlying structure that was present in the published libraries. The ANN and PLS models of the *Geobacillus* promoter design space therefore had to account for changes in both promoter DNA sequence and motif structure, potentially resulting in reduced predictive power.

Promoter sequence-activity models trained on data sets containing large amounts of sequence homology potentially ignore regions of promoter design space that encode promoters with desirable characteristics (Jonsson *et al.*, 1991). However, it would appear that DNA sequence data alone is not sufficient to accurately predict promoter activity without such homology. Including biophysical or structural information for individual nucleotides or motifs could potentially have increased the predictive power of the sequence-function models. One study, for example modelled promoter activity as a function of the contribution of nucleotide 3-mers to the free energy barrier for RNA polymerase binding to the promoter sequence (Li & Zhang, 2014). Another study used Principle Components Analysis to derive nucleotide descriptors from 1,209 structural descriptors, which were then used by a Support Vector Machine (SVM) model to derive predictions of promoter strength (Liang & Li, 2007).

Different statistical learning approaches may also have potentially yielded sequence-function models with greater predictive accuracy. For example, SVMs have been applied in *E. coli* to derive sequence-activity models from microarray data (Kiryu *et al.*, 2005). SVMs were also used to derive predictions of activity from a library of 100 promoter sequences characterised upstream of GFP (Meng *et al.*, 2017), although the training library was the same homologous, mutagenesis-derived library to which ANNs were previously applied by Meng *et al*. (Meng *et al.*, 2013). Gaussian Process models have also been shown to model peptide structure-activity relationships more accurately than ANN and PLS models (Zhou *et al.*, 2009), and could potentially be applied to *Geobacillus* promoter sequences in future.

4.4 Summary

Three progressively larger sets of bioinformatically identified putative promoters were characterised in *G. thermoglucosidans*. The resulting characterisation data were used to derive ANN, PLS and Random Forest statistical models. In total, 105 *promoter::GFP* fusions and 82 *promoter::mOrange* fusions were characterised. ANN and PLS models were obtained that returned accurate fits of training, validation and primary test sets, but predictive accuracy was low when the models were applied to predicting activity levels for secondary test sets of bioinformatically-identified putative promoters or *de novo* designed synthetic sequences. This lack of predictive power was hypothesised to be the result of a lack of significant sequence homology in the training data and the relatively small size of the training data set as compared to the dimensionality of the promoter design space. Random Forest partitioning provided useful descriptive models, which suggested that regions upstream of the -35 and -10 consensus regions play a key role in transcription regulation in *Geobacillus.* This conclusion was supported by the presence of a conserved AT-rich region in active *Geobacillus* promoters that was comparable to previously characterised UP-elements that have been reported in *B. subtilis* and *E. coli.*

Although 2-log ranges of expression levels were observed for both reporter proteins, promoter activity was generally poorly conserved between reporters. This result highlighted the necessity of thoroughly characterising promoter activity in multiple contexts. If promoters do not function consistently when used in genetic or environmental contexts other than those used in their initial characterisation, significant “re-tuning” may be required. As the complexity of synthetic pathways increases, such re-tuning quickly becomes prohibitive and truly modular, context-independent promoters become invaluable. To assess the modularity of the bioinformatically identified promoter sequences, four genetic and environmental factors with the potential to alter promoter activity were investigated concurrently with the characterisation experiments that were discussed in this chapter. The results of these characterisation experiments are discussed in Chapter 5.

5 Analysing the effect of environmental and genetic context on promoter activity

Summary

The synthetic biology approach to metabolic engineering requires libraries of thoroughly characterised regulatory parts that function in a predictable, consistent manner. However, genetic and environmental context can significantly alter promoter output. This context-specificity can hinder the application of previously characterised promoters in novel scenarios, as the strength of a promoter in the conditions under which it was initially characterised may not be reflective of promoter performance in more industrially relevant scenarios. This discrepancy can necessitate time- and resource-consuming pathway re-tuning. As the complexity of synthetic pathways increases, such re-tuning quickly becomes prohibitive and truly modular, context-independent promoters become increasingly desirable. Therefore, to assess the modularity of the bioinformatically identified *Geobacillus* promoters, four genetic and environmental factors with the potential to impact promoter activity were investigated. These factors were the method by which *promoter::reporter* fusions were cloned, the effect of fluctuations in plasmid copy number on culture fluorescence, the reporter protein used to characterise promoter activity and the effect of culture oxygenation. The experiments discussed in the following chapter were performed concurrently with the sequence-function modelling described in Chapter 4.

5.1 Introduction

The promoter sequence-function models that were discussed in Chapters 3 & 4 assumed that the observed differences in fluorescence between disparate *promoter::GFP* fusions were the result of promoter effects. However, the environmental and genetic context in which regulatory parts are characterised can also impact upon gene expression (Cardinale & Arkin, 2012).

Ideally, the activity of a characterised regulatory sequence should be independent of context (Gilman & Love, 2016); given the effort that is expended in optimising a synthetic pathway at laboratory scale, said pathway should require minimal re-tuning when cultured at industrial scale (Segall-Shapiro *et al.*, 2018). The context-independence of regulatory elements becomes particularly important as the complexity of synthetic pathways increases; as the number of genetic parts in a system increases, the number of potential designs grows exponentially (Davidsohn *et al.*, 2014), precluding the use of trial-and-error optimisation of individual components (Rudge *et al.*, 2016). Truly modular, context-dependent genetic regulatory parts can facilitate such optimisation by reducing the number of candidate designs that must be analysed *in vivo*, and can aid the systematic, scalable, bottom-up design of genetic circuits that synthetic biology strives to achieve (Del Vecchio, 2015, Nielsen *et al.*, 2016).

To analyse the context-independence of the characterised *Geobacillus* promoter library, four genetic and environmental factors with the potential to effect promoter activity were investigated. These four factors were the method by which *promoter::reporter* fusions were cloned, the effect of fluctuations in plasmid copy number on culture fluorescence, the reporter protein used to characterise promoter activity and the effect of culture oxygenation on promoter activity.

5.1.1 Application of a type IIS restriction cloning strategy for promoter::reporter *fusion*

The *promoter::GFP* fusions that were characterised in *G. thermodenitrificans* and in *G. thermoglucosidans* data set A were synthesised as a single part and cloned directly into the pS797 vector by ATUM (previously DNA 2.0, California, United States of America). Whilst this approach expedited the characterisation process, it did not allow for routine switching of promoter elements between genetic contexts.

To facilitate routine application of promoter sequences for the control of CDS other than the *GFP* used in their characterisation, a type IIS restriction

cloning strategy was implemented (Engler *et al.*, 2008, Kirchmaier *et al.*, 2013). The requisite DNA parts (CDS, promoter, RBS, terminator and vector backbone) were flanked with unique 5 bp sequences that, when cut with the restriction enzyme BsaI, resulted in specific overhanging DNA sequences that ensured digested fragments could only ligate in the defined order.

Although the 100 bp *Geobacillus cis*-regulatory elements that are discussed in this study are referred to as promoters, they also contained Ribosome Binding Sites (RBS). The sequence-function models discussed in Chapters 3 & 4 treated promoter and RBS as a single regulatory unit. However, future applications of the regulatory sequences might require disparate promoter and RBS elements; the construction of multi-gene operons, for example, requires RBS sequences to be placed between individual CDS. The putative location of the RBS sequence was therefore identified, and regulatory elements were split *in silico* into 85 bp promoter and 15 bp RBS sequences.

The *in vitro* ligation of promoter, RBS and CDS parts by the type IIS cloning strategy resulted in the insertion of two scar DNA sequences. The 4 bp scar "ACCT" was inserted between the promoter and RBS sequence, and another 4 bp scar, "AATG" was inserted between the 3' terminus of the RBS and the start codon of the adjacent CDS.

Previous studies have shown that novel cryptic functionality can potentially arise through the fusion of previously characterised genetic parts (Lou *et al.*, 2012, Yao *et al.*, 2013, Zong *et al.*, 2017). In particular, any alterations to the mRNA secondary structure arising from scar sequences located between the RBS and CDS can potentially negatively impact the efficiency of translation initiation (Mirzadeh *et al.*, 2015), thereby altering protein production. The regulatory sequences that were characterised without scar sequences in *G. thermoglucosidans* were therefore cloned upstream of *GFP* using the type IIS restriction cloning system and re-characterised so that the impact of cloning scar sequences could be assessed.

5.1.2 The effect of plasmid copy number on fluorescence activity

The use of plasmid-based gene expression to quantify promoter activity is ubiquitous (Urtecho *et al.*, 2018). In this study, all *promoter::reporter* fusions were characterised in the pS797 vector, which used the repBST1 origin of replication for propagation in *Geobacillus* (Liao *et al.*, 1986, Taylor *et al.*, 2008). As compared to genome integration of heterologous pathways, multiple-copy plasmids offer increased signal (Jahn *et al.*, 2014) and facilitate easy manipulation of heterologous pathways (Jones *et al.*, 2000).

However, Plasmid Copy Number (PCN), and therefore transgene copy number, can fluctuate (Wong Ng *et al.*, 2010, Segall-Shapiro *et al.*, 2018). Given that the copy number of heterologously expressed genes is known to impact the level of output from gene networks (Jones *et al.*, 2000), such fluctuations in copy number can potentially result in inaccurate quantifications of promoter activity.

To ensure that the observed differences in fluorescence output between *G. thermoglucosidans* cultures expressing GFP under the control of different promoter sequences were not the result of fluctuations in PCN, quantitative PCR (qPCR) was employed to determine PCN. The qPCR methodology used a fluorescent dye that intercalated to dsDNA during the amplification step of the PCR, resulting in a fluorescence signal that was proportional to the number of amplicons in a sample (Thornton & Basu, 2011). Sample fluorescence was then compared to a standard curve consisting of serial 10-fold dilutions of purified DNA, allowing the concentration of the amplicon of interest to be calculated (Lee *et al.*, 2006b).

Two sets of PCR primers were designed, one set which amplified a unique region of the ampicillin resistance gene in the pS797 vector and one set which amplified a unique region of the *G. thermoglucosidans* genome. By calculating the ratio of plasmid- to genomic-amplicons, a per-genome estimate of PCN was calculated (Lee *et al.*, 2006a, Skulj *et al.*, 2008).

5.1.3 The effect of reporter sequence on promoter activity

The functional composability of regulatory sequences is a vital requirement for synthetic biology (Davis *et al.*, 2011). However, the context-specificity of regulatory sequences is well understood (Cardinale & Arkin, 2012, Kosuri *et al.*, 2013, Mutalik *et al.*, 2013b, Nielsen *et al.*, 2016). If promoter sequences do not display common functionality when placed upstream of CDS other than those used in their initial characterisation, the results of characterisation experiments can only ever have local validity. Therefore, to assess the contextual robustness of the putative *Geobacillus* regulatory elements, putative promoter sequences were characterised upstream of a second reporter protein CDS, the *RFP* derivate *mOrange*.

5.1.4 The effect of oxygen concentration on promoter activity

The *ldhA* promoter, variants of which have been previously employed for genetic engineering projects in *Geobacillus*, has been shown to be oxygen dependent (Cripps *et al.*, 2009, Bartosiak-Jentys *et al.*, 2012, Lin *et al.*, 2014, Kananavičiūtė & Čitavičius, 2015, Sheng *et al.*, 2017). Whilst the ability to induce expression under oxygen limitation may be advantageous in certain scenarios (Kananavičiūtė & Čitavičius, 2015), consistent, predictable output is often required for complex metabolic engineering projects and industrial-scale bio-production.

To assess whether the characterised promoter sequences functioned independently of culture aeration, *G. thermoglucosidans* transformants expressing *promoter::reporter* constructs were cultured in different growth formats. Specifically, culture fluorescence in 250 ml baffled and non-baffled Erlenmeyer flasks was compared, as baffles have previously been shown to increase culture aeration (Gupta & Rao, 2003, Running & Bansal, 2016).

5.1.5 Data set composition

The experiments outlined in this chapter were performed concurrently with the Artificial Neural Network (ANN) and Partial Least Squares (PLS) sequence-function modelling that is described in Chapter 4. The sequence-function modelling was performed using three progressively larger sets of putative promoter sequences, termed A, B and C. The experiments that are described in this chapter were therefore performed using one of these three data sets. The promoter set on which a given experiment was performed is specified in the relevant results section. A summary of data set composition is shown in Table 5-1.

Data Set	Characterised constructs	Contains cloning scars?	Notes
A	47 x *promoter::GFP*	No	The 31 *promoter::GFP* fusions that were characterised in *G. thermodenitrificans* (Chapter 3) were combined with 11 further putative promoter sequences that had not been previously characterised *in vivo*. 21 out of the resulting 42 promoter sequences were used to train and Cross-Validate (CV) PLS sequence-function models. Five further sequences were characterised during model testing.
B	45 x *promoter::GFP* 42 x *promoter::mOrange*	Yes	The 26 promoter sequences that were used in PLS model training, CV and testing in data set A were added to 10 previously uncharacterised putative promoters. Two of these 36 sequences contained restriction enzyme recognition sites that rendered them incompatible with the type IIS cloning strategy, and were therefore discarded. 14 further putative regulatory sequences were randomly selected to test ANN and PLS sequence-function models, of which 11 were successfully cloned upstream of *GFP* via the type IIS cloning strategy and characterised in *G. thermoglucosidans*. Six of the 48 data set B *promoter::mOrange* constructs could not be expressed in *G. thermoglucosidans*.
C	95 x *promoter::GFP* 82 x *promoter::mOrange*	Yes	Constructs from data set B were added to 52 previously uncharacterised putative promoter sequences that were selected at random from across the *Geobacillus* promoter phylogeny. Five *promoter::GFP* fusions could not be synthesised by ATUM (previously DNA 2.0, California, United States of America). 12 *promoter::mOrange* fusions could either not be synthesised or could not be expressed in *G. thermoglucosidans*.

Table 5-1: Summary of characterised *promoter::reporter* constructs from each of the three *G. thermoglucosidans* data sets.

5.2 Results and Discussion

5.2.1 Application of a type IIS restriction cloning strategy for promoter::reporter *fusion*

Alignment of 17 sequences with empirically quantified *in vivo* promoter activity revealed a heavily conserved region of purine-rich sequence, located between 15 and 7 bp upstream of the start codon of the adjacent CDS (Figure 5.1). Given the similarities in both location and sequence composition of this motif to the canonical Shine-Dalgrano sequence (Shine & Dalgarno, 1974), this motif was assumed to represent the location of the RBS.

In order to expedite *in vivo* characterisation, the 4 bp `ACCT` scar sequence that would have resulted from *in vitro* ligation of the promoter and RBS parts was inserted into each of the bioinformatically identified putative regulatory elements *in silico*, and the resulting 104 bp regulatory elements were synthesised as single parts. The regulatory sequences were then cloned upstream of *GFP* and inserted into the pS797 vector.

To assess the effect of the cloning scar sequences on promoter activity, the fluorescence activity of the 24 *promoter::GFP* constructs that were common to data sets A & B were compared. For the majority of *cis*-regulatory sequences, the insertion of scar sequences by the type IIS cloning strategy had no statistically significant impact on GFP fluorescence (Figure 5.2). Significance was determined by multiple t-tests, using the Holm-Šidak method to correct for multiple comparisons and a significance level of 0.05. Scar sequences significantly changed the activity levels of four promoter sequences. Activity was significantly decreased in three promoters (GKAU_00722, adjusted P value = 0.000; GKAU_00878, adjusted P value = 0.022; GSTEA_02162, adjusted P value = 0.024) and increased in one sequence (GSTEA_00548, adjusted P value = 0.009) when scar sequences were inserted.

The observed differences in fluorescence between scarred and un-scarred regulatory sequences were hypothesised to be the result of alterations

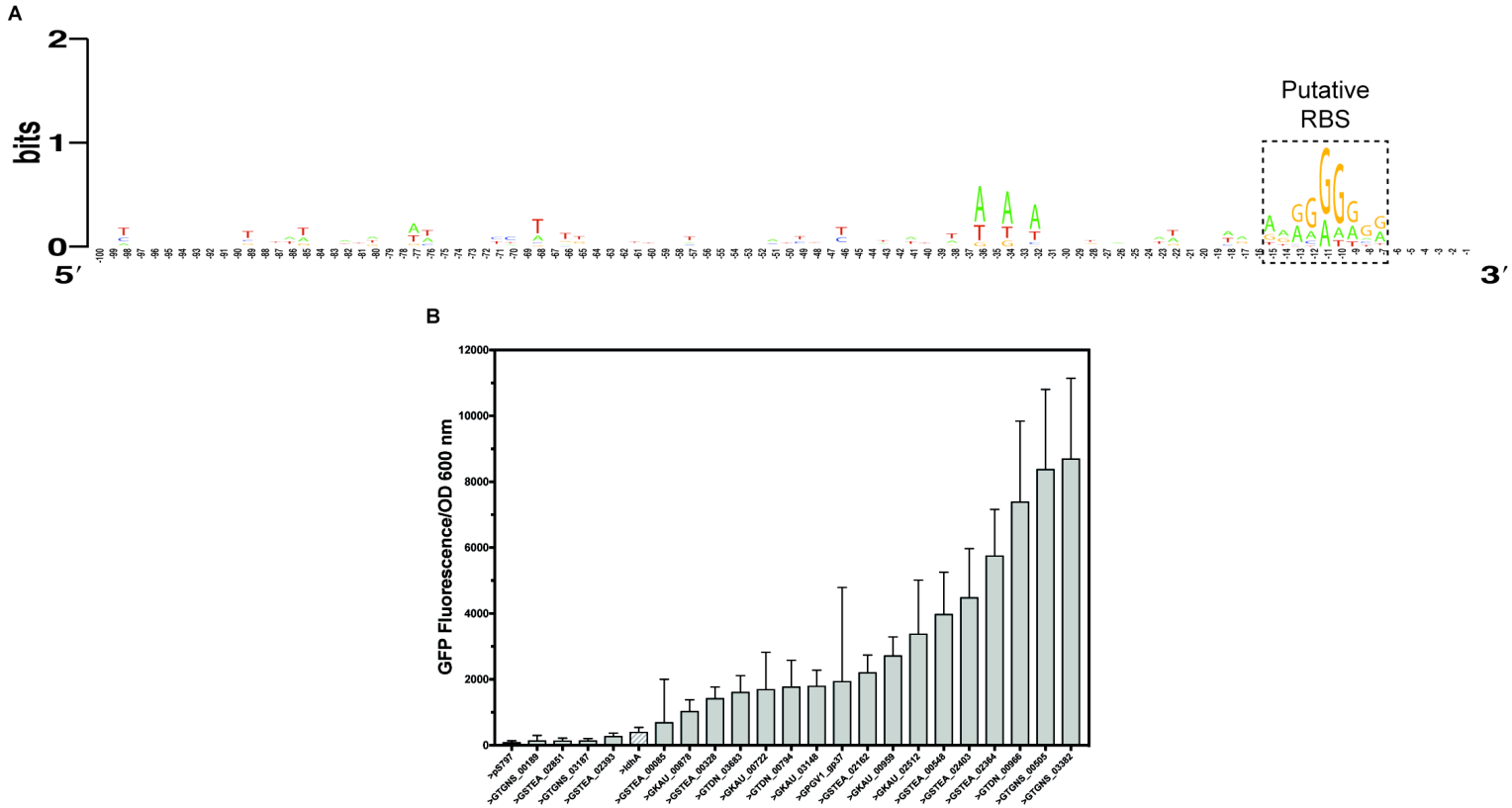


Figure 5.1: Visualisation of a sequence alignment used to identify the putative Ribosome Binding Site (RBS) and the empirically measured activity of the aligned sequences.

A) The location of the putative RBS sequence is highlighted by the dashed box. The overall height of individual stacks indicates the degree of sequence conservation at a given position, and the height of nucleotide symbols indicates the conservation of each nucleic acid at that position. Position numbering is relative to the start codon of the upstream CDS. Sequences were aligned and visualised using WebLogo version 2.8.2 (Crooks *et al.,* 2004).

B) Empirical measurements of fluorescence and absorbance taken after 24 h incubation at 60 °C in 96-well plate format. Bars represent the mean of 9 ≤ n ≤ 20 starter cultures, arising from independent transformation events. Standard deviation error bars are shown, unless hidden by the bar. The positive control, *ldhA::GFP* is represented by the hatched bar. The negative control, *G. thermoglucosidans* transformed to contain empty vector pS797, is represented by the black bar. The two controls were not included in the sequence alignment, but are included for reference.

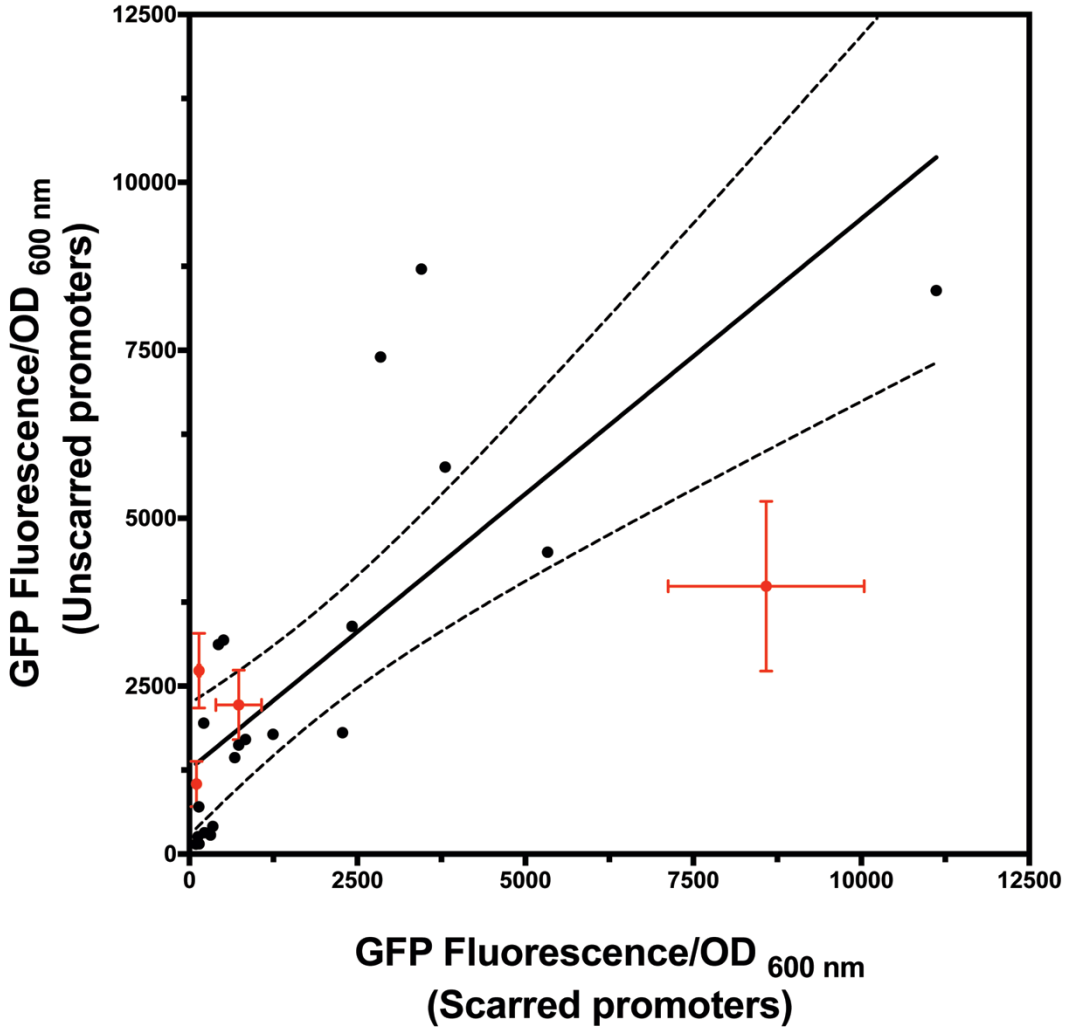


Figure 5.2: Fluorescence output of GFP under the control of scarred and un-scarred putative promoter sequences.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. Points represent the mean GFP output of each promoter, from $3 \leq n \leq 9$ independent starter cultures. The points coloured red represent promoters for which GFP fluorescence was statistically significantly changed when the cloning scar was introduced to the DNA sequence. Significance was calculated using multiple t-tests, using the Holm-Šidák method to correct for multiple comparisons and a significance limit of 0.05. For ease of visualisation, standard deviation error bars are shown only on the statistically significant points. The solid line represents a linear regression of the data, with 95% confidence limits represented by the dashed lines. The linear regression had an R^2 value of 0.5216.

to mRNA secondary structure. To test this hypothesis, the free-energy associated with mRNA folding was calculated using the mFold zipfold server (Zuker, 2003). The default settings were used, with RNA 2.3 energy rules. Folding energies were returned in kcal/mol. The temperature at which folding was simulated was set to 60 °C, to match the temperature at which *Geobacillus* cultures were incubated. The sequence window in which folding was analysed stretched from -20 to +20, relative to the adenine residue of the GFP start codon.

Three of the promoter sequences for which fluorescence output was significantly altered by the inclusion of scar sequences also showed the greatest changes in mRNA secondary structure free energy (Figure 5.3). *GSTEA_00548::GFP*, for which fluorescence activity was statistically significantly increased by the inclusion of scar sequences, showed the greatest relaxation of mRNA secondary structure out of the 24 analysed sequences. *GKAU_00878::GFP* also showed an increase in fluorescence output after scar insertion and relaxation of the mRNA secondary structure. Conversely, *GKAU_00959::GFP* and *GSTEA_02162::GFP* showed a statistically significant decrease in fluorescence and an increase in mRNA secondary structure.

The correlation between relaxed mRNA secondary structure at the RBS-CDS junction and increased protein production was corroborated by the literature (Kudla *et al.*, 2009, Bentele *et al.*, 2013, Mortimer *et al.*, 2014, Mirzadeh *et al.*, 2015). The presence of significant secondary structure surrounding the RBS is hypothesised to negatively impact on the ability of a mRNA transcript to sequester ribosomes, thereby reducing the rate at which transcripts are translated, although the strength of this correlation may be sensitive to genetic context and cellular concentrations of amino acids and tRNAs (Welch *et al.*, 2009, Tuller & Zur, 2014).

These results showed that novel functionality could result from the insertion of scar DNA sequences. Altered regulatory element activity was observed in 8% of the 24 characterised sequences. In future, *in silico* screening of any potentially unfavourable alterations to mRNA secondary structure

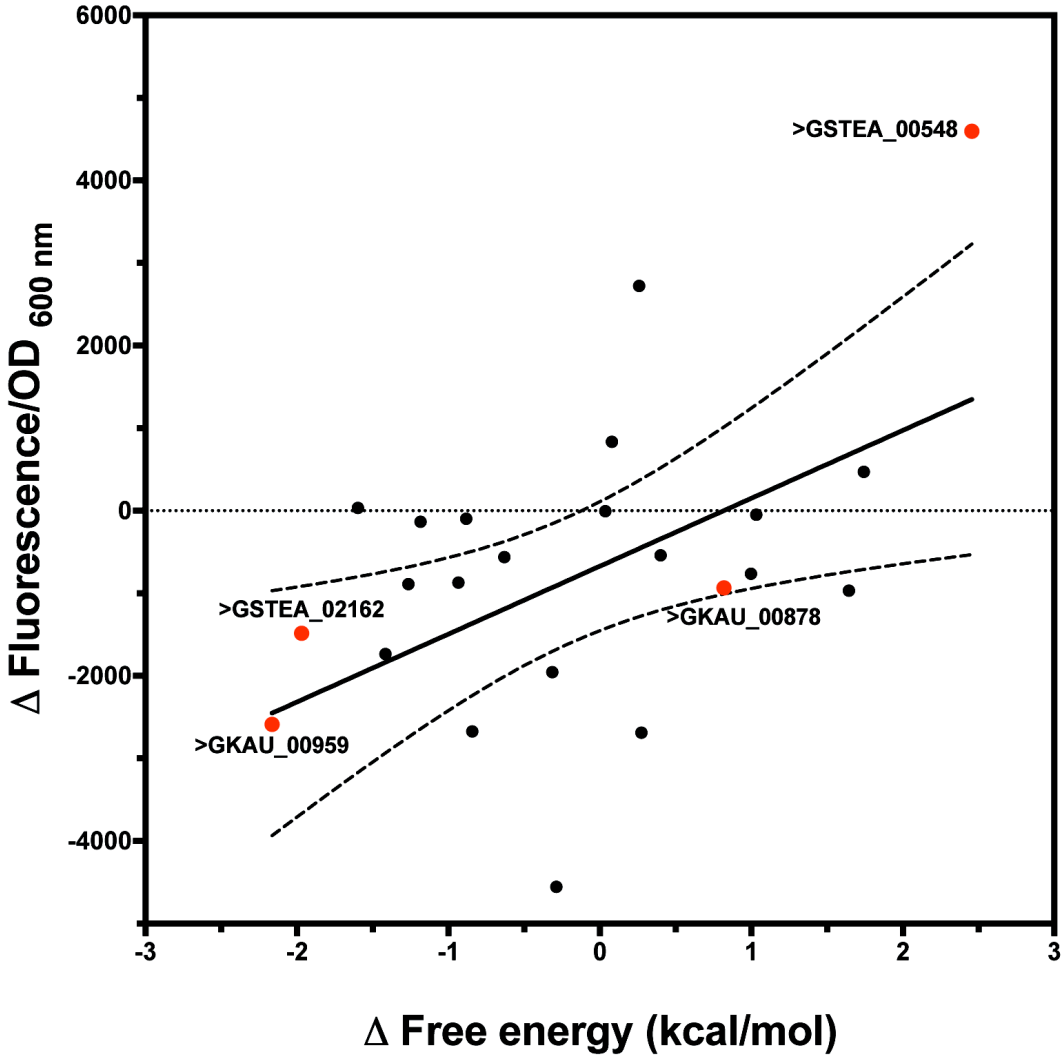


Figure 5.3: Comparing the change in GFP fluorescence and the change in free energy of the mRNA secondary structure of *promoter::GFP* fusions once cloning scar sequences were inserted.

Points represent individual *promoter::GFP* fusions. Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format, from $3 \leq n \leq 9$ starter cultures arising from independent transformants. The points that are coloured red represent promoters for which GFP fluorescence was statistically significantly different when the cloning scar was introduced to the DNA sequence. Significance was calculated using multiple t-tests, using the Holm-Šidák method to correct for multiple comparisons and a significance level of 0.05. The solid line represents a linear regression of the data, with 95% confidence limits represented by the dashed lines. The linear regression had an R^2 value of 0.2407

Free energies were calculated using the mFold zipfold server (Zuker, 2003), using default settings and RNA 2.3 energy rules. The sequence window for which secondary structure was calculated stretched from -20 to +20, relative to the adenine residue of the GFP start codon. The temperature at which folding was simulated was set to 60 °C, to reflect the temperature at which *G. thermoglucosidans* cultures were incubated.

resulting from the method of *in vitro* pathway construction could be employed to mitigate the need for the time-consuming *in vivo* optimisation.

5.2.2 The effect of Plasmid Copy Number on fluorescence activity

PCN varied by 88-fold across different cultures in Data Set B (Figure 5.4). However, no correlation was observed between PCN and population fluorescence. A linear regression of the data returned an R^2 value of 0.023. Additionally, a F-test of the null hypothesis that the slope of the linear regression curve was 0 returned a P value of 0.2342, showing that the slope of the curve did not deviate significantly from 0.

Further analysis of the effect of PCN on culture fluorescence was provided by a PLS model. The promoter that was used to control GFP expression in a given *G. thermoglucosidans* culture and the PCN of that culture were used as *x* variables, with culture fluorescence used as the *y* variable. The PLS model extracted eight Latent Variables (LVs) from the data, and was capable of explaining 98.089% of the empirical variation in GFP fluorescence. In total, nine *x* variables returned VIP values that were greater than the threshold value of 0.8, and were therefore judged to have statistically significant impact on determining GFP fluorescence (Figure 5.5A) (Eriksson *et al.*, 2006). Of the nine statistically significant *x* variables, eight were promoter sequences and one was PCN.

The promoter sequences that returned a VIP of greater than 0.8 were the six strongest promoters in the data set and two sequences (GKAU_01463 & GTDN_02731) with no promoter activity. Sequences were judged to have promoter activity if their mean fluorescence activity was statistically significantly different from the negative control, *G. thermoglucosidans* transformed to contain the empty pS797 vector. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, at a significance level of 0.05.

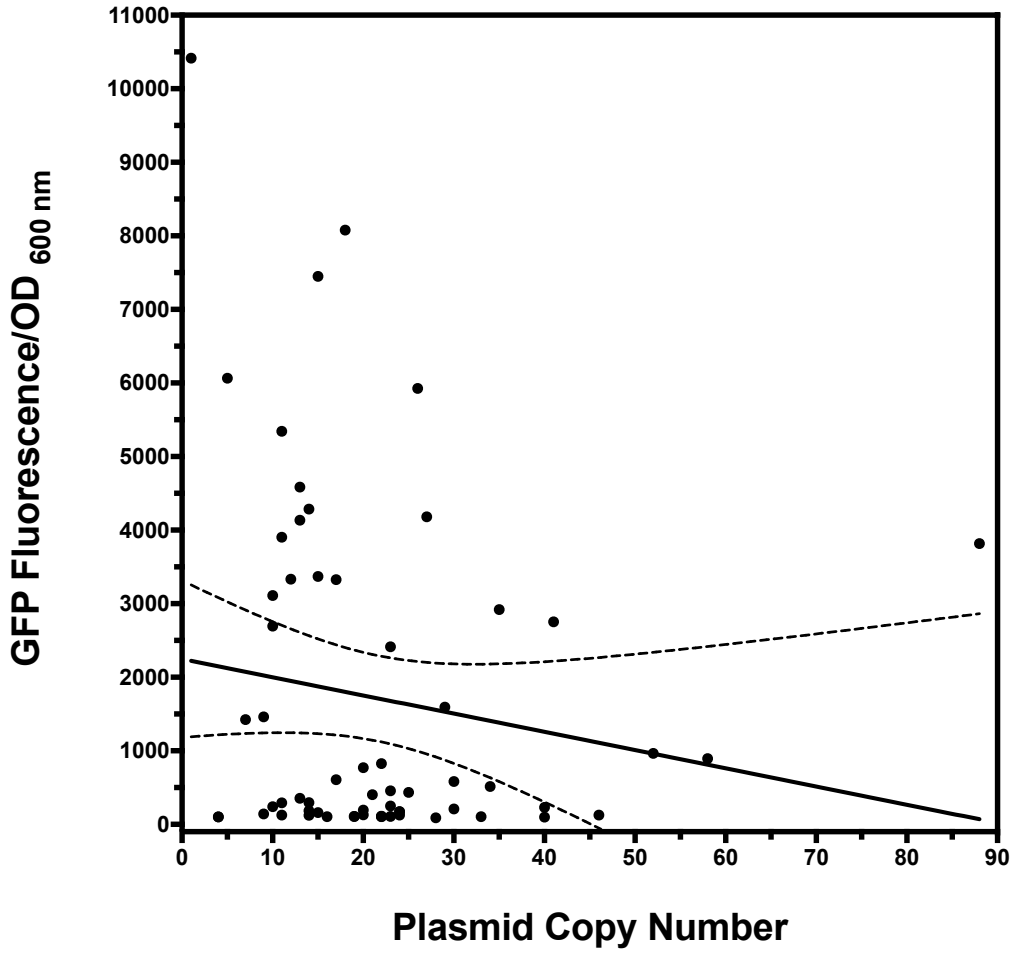


Figure 5.4: The effect of Plasmid Copy Number (PCN) on *G. thermoglucosidans* culture fluorescence.

Points represent individual populations of *G. thermoglucosidans*, expressing GFP under the control of one of 34 *Geobacillus* promoter sequences. The solid line represents a linear regression of the data, with 95% confidence limits represented by the dashed lines. The R^2 value of the linear regression was 0.023.

The six strong promoter sequences returned positive model coefficients, which suggested that the presence of these six sequences was having a statistically significant positive impact on culture fluorescence. The magnitude of the model coefficient was positively correlated with mean *promoter:GFP* fluorescence. The two sequences with no promoter activity returned negative coefficients.

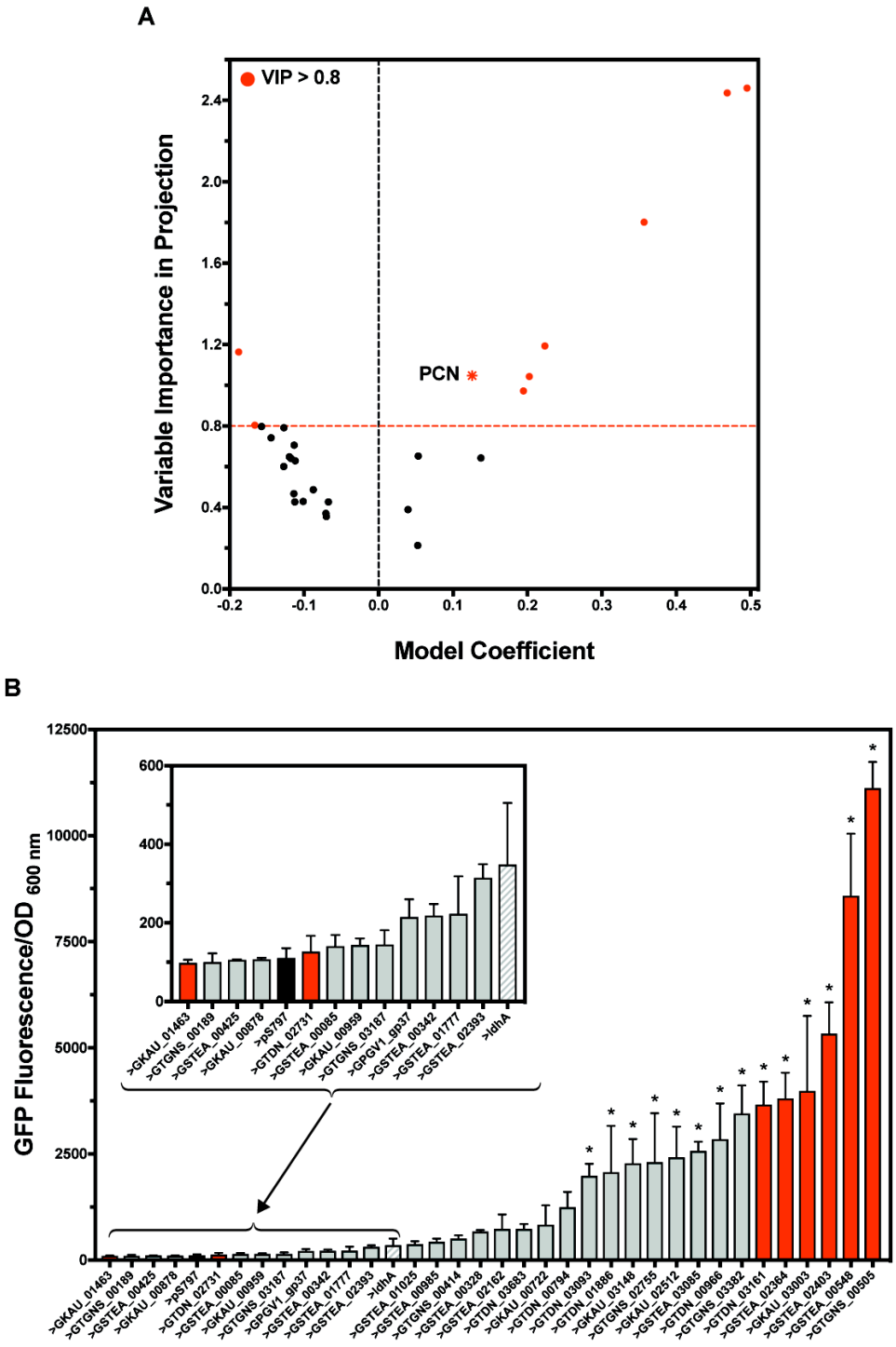


Figure 5.5: Model coefficients and Variable Importance in Projection (VIP) scores returned by a Partial Least Squares (PLS) model examining the relationship between Plasmid Copy Number (PCN) and culture fluorescence.

A) PLS VIP scores plotted against model coefficients. Circular points represent individual promoter sequences. The asterisk represents PCN. Points are coloured in red when their VIP value exceeded the threshold value of 0.8

B) Empirically measured fluorescence of *promoter::GFP* fusions. Bars are coloured red when the given promoter returned a VIP value greater than the threshold value of 0.8.

PCN returned a positive model coefficient, which suggested that variations in copy number were having a statistically significant positive impact on culture fluorescence. However, the model coefficient that was returned by PCN was of lesser magnitude than the coefficients of each of the six active promoters. PCN returned a model coefficient of 0.1258, whilst the six active promoter sequences returned coefficients in the range 0.1948 to 0.4952. This result suggested that although PCN did have a statistically significant impact on *G. thermoglucosidans* culture fluorescence, the magnitude of this effect was between 1.5- and 3.9-fold smaller than the effect of the strongest promoters.

The data were also analysed using a standard least squares model with effect screening emphasis. This model suggested that both promoter selection and PCN had a statistically significant impact on culture fluorescence, but that the effect of promoter sequence was greater than that of PCN (Promoter: FDR LogWorth = 21.865, FDR PVaule = 0.000, PCN: FDR LogWorth = 2.068, FDR PValue = 0.0085).

Taken together, these results suggested that fluctuations in PCN were having a statistically significant impact on *G. thermoglucosidans* GFP expression levels. However, both the PLS and standard least squares model attributed greater statistical significance to the effect of the strongest promoter sequences than PCN. This suggested that pathway tuning using strong promoters might be sufficient for simple overexpression of a heterologously expressed gene, but that sophisticated pathway tuning would require a more nuanced approach.

Integrating heterologous genes into the host genome offers a potential method by which copy number could be more tightly regulated than plasmid-based expression. However, studies have shown that the position of the integrated gene relative to the origin of replication can impact copy number (Chandler & Pritchard, 1975, Block *et al.*, 2012). The location within the genome at which synthetic pathways are integrated must therefore be carefully considered. Relatively simple synthetic pathways could be expressed from the same plasmid or be integrated into the same region of the genome such that

fluctuations in copy number effect all of the pathway components equally (Brophy & Voigt, 2014), although this approach would not necessary be applicable in the case of genetic circuits with large numbers of constituent parts.

For complex synthetic pathways, additional regulatory mechanisms, such as feedback loops that respond to fluctuations in copy number and regulate promoter activity accordingly, may be necessary to fully decouple regulatory sequence activity from the effects of copy number variation. For example, Segall-Shapiro *et al.* posited the use of transcription-activator-like effectors that bind to operator sequences within a promoter and repress expression of a gene of interest. By expressing the repressor from the same plasmid as the gene of interest, repressor concentration and copy number are positively correlated, such that any increase in copy number results in an increase in repressor concentration, thereby reducing expression of the gene of interest, offsetting the effect of the increased copy number (Segall-Shapiro *et al.*, 2018).

5.2.3 The effect of reporter sequence on promoter activity

In data set C, 80 regulatory sequences were characterised upstream of both GFP and mOrange (Figure 5.6). Moderate correlation was observed between promoter activity for the two reporter proteins; a linear regression of the data returned an R^2 value of 0.447. 18 regulatory sequences, including the *G. thermodenitrificans ldhA* promoter, fell within the 95% confidence limits of the linear regression, suggesting that promoter activity for these sequences was well conserved between the two reporters. Of these 18 sequences, eight, including *ldhA,* displayed mean fluorescence output that was statistically significantly greater than the negative control for both reporters at the 0.05 significance level. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test.

To better identify promoters that functioned independently of genetic context, and to account for the difference in magnitude between GFP and mOrange fluorescence signals, the mean fluorescence output of each *promoter::reporter* fusion was normalised to the fluorescence of the relevant

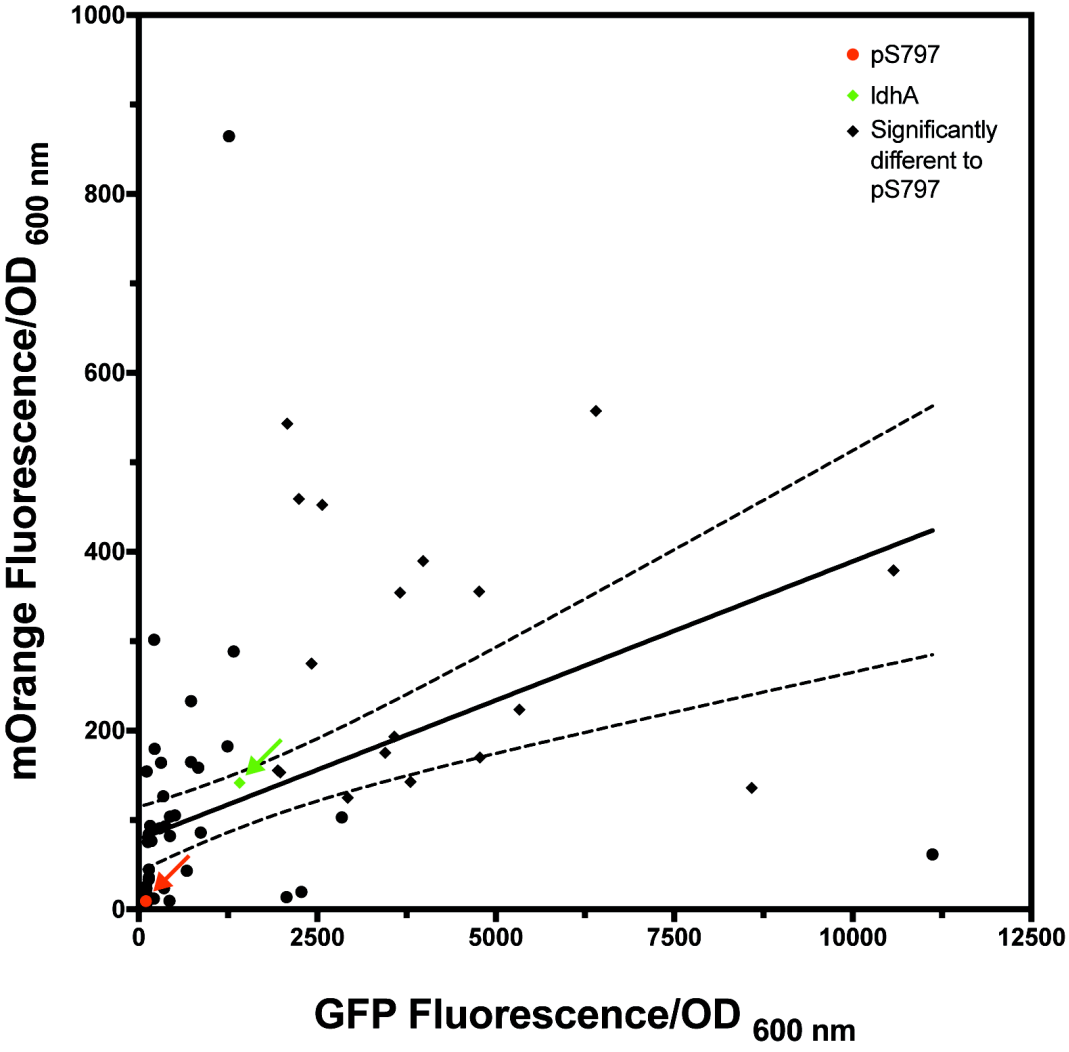


Figure 5.6: Fluorescence output of GFP & mOrange under the control of putative promoter sequences.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. Points represent the mean fluorescence output of individual promoter sequences from $3 \leq n \leq 9$ independent starter cultures. The negative control, *G. thermoglucosidans* transformed to contain the empty vector pS797, is shown in red. The positive control, the *G. thermodenitrificans ldhA* promoter, is shown in green. Those promoter sequences for which mean fluorescence output was statistically significantly greater than pS797 for both reporter proteins are represented by diamonds. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, at a significance level of 0.05. The solid line represents a linear regression of the data, with 95% confidence limits represented by the dashed lines. The linear regression had an R^2 value of 0.447.

negative control. Regulatory sequences were grouped into five clusters by *K*-means clustering, based on their Euclidean distance from the line $y = x$ (*i.e.* the point at which normalised GFP and mOrange fluorescence was equal) (Figure 5.7A). The clustering algorithm identified 58 regulatory sequences that fell close to the line $y = x$ (Cluster 1 in Figure 5.7A). Of these 58 sequences, seven displayed mean fluorescence output that was statistically significantly greater than the negative control. This result suggested that the *Geobacillus* promoter library contained seven active, context-independent regulatory sequences (Figure 5.7B). These seven regulatory sequences covered a range of activity of four-fold.

In addition to the seven context-independent active promoter sequences, the characterised promoter library also contained 20 sequences that showed no regulatory activity when placed upstream of either *GFP* or *mOrange*. The mean fluorescence activity of these 20 context-independent inactive sequences fell within two standard deviations of the mean of the negative control for both reporter proteins. Sequences that are known to never show regulatory activity, regardless of genetic context, could be of use in providing robust negative controls for future work (Mutalik *et al.*, 2013b).

The number of characterised context-independent, active regulatory sequences in the characterised *Geobacillus* promoter library was relatively small: only 9% of the characterised sequences showed activity that was independent of the downstream CDS. Additionally, the range of expression strengths provided by the seven context-independent sequences (four-fold) compared poorly with the range of expression strengths shown by the active context-dependent regulatory sequences (30-fold when GFP was the reporter, 8.4 fold when the reporter was mOrange). Future work could therefore focus on reducing the context-dependence of the characterised *Geobacillus* promoter library.

One approach by which the modularity of the *Geobacillus* promoter library could potentially be increased is through the insertion of insulating spacer DNA sequences that physically separate genetic parts (Davis *et al.*,

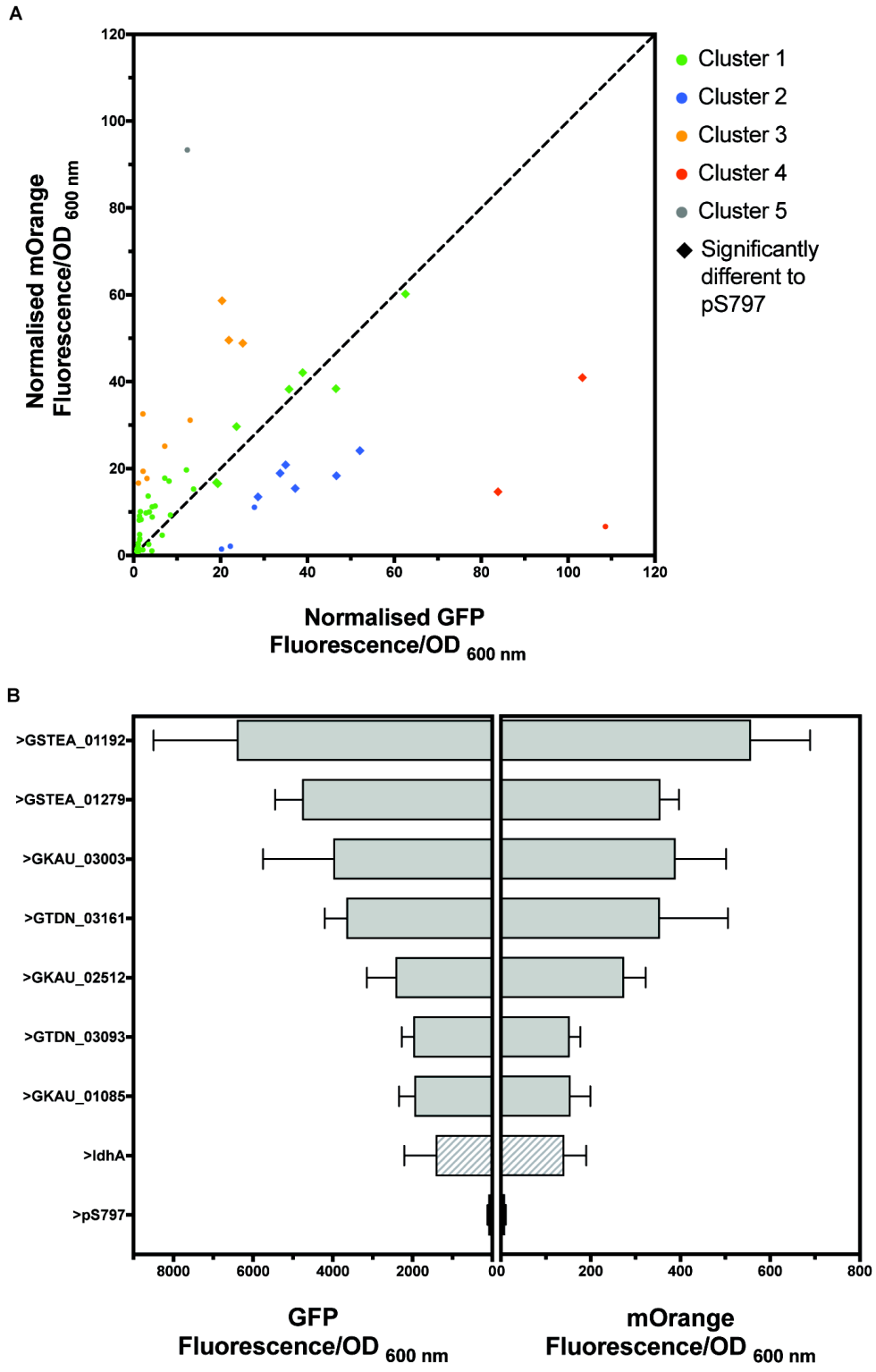


Figure 5.7: Identifying promoter sequences that functioned independently of genetic context.

A) GFP and mOrange fluorescence values are normalised to the negative control, pS797. Points represent individual promoter sequences. Colours represent groups of promoter sequences, determined by K-means clustering based on the Euclidean distance of the points from the line $y = x$. Diamond-shaped points represent active promoter sequences, *i.e.* those promoters for which both GFP and mOrange fluorescence was statistically significantly different from the negative control, *G. thermoglucosidans* transformed with pS797.

B) GFP & mOrange fluorescence activity of the seven most active promoters from Cluster 1.

2011, Carr *et al.*, 2017). For example, insulator sequences inserted between the promoter and RBS sequences aim to increase part modularity by defining the 5' leader sequence of the mRNA transcript. The secondary structure of the 5' untranslated region (UTR) and downstream CDS, as well as interactions between the RBS and 5' UTR sequences have been shown to play a role in determining translation efficiency (Kudla *et al.*, 2009, Salis *et al.*, 2009, Kosuri *et al.*, 2013). By standardising 5' UTR structure, insulator sequences therefore aim to normalise the rate at which a given regulatory element initiates translation, regardless of the genetic context.

Promoter sequences could also potentially be decoupled from the 5' UTR through the use of a second translation initiation element, placed downstream of the primary RBS, to disrupt secondary structure across the 5' UTR (Mutalik *et al.*, 2013a). In a pioneering study, a core *cis*-regulatory element, consisting of -35 and -10 consensus regions and an RBS were placed upstream of a leader sequence that contained a secondary Shine-Dalgrano sequence and that overlapped with the adenine residue of start codon of a downstream CDS of interest. The leader sequence was therefore translated by ribosomes that bound to the core RBS sequence, with the CDS of interest being translated from the secondary Shine-Dalgrano sequence. The core promoter sequence did therefore not contribute to the 5' UTR of the CDS, resulting in decoupling of promoter and CDS. This method was shown to increase the modularity of regulatory sequences in *E. coli* (Mutalik *et al.*, 2013a), and could potentially also be applied in *Geobacillus*.

Upstream of the promoter 5' terminus, insulator sequences have also previously been applied to mitigate any interactions between sequence regions upstream of the promoter and the RNA polymerase alpha subunit (Davis *et al.*, 2011, Carr *et al.*, 2017). However, given that the 100 bp *Geobacillus* promoter sequences are already of sufficient length to contain the promoter sequence region in which interactions with the RNA polymerase alpha subunit are most common (Ross *et al.*, 1993, Aiyar *et al.*, 1998, Estrem *et al.*, 1998, Meijer & Salas, 2004, Phan *et al.*, 2012), the utility of upstream insulator elements in this context was unclear.

If insulator sequences are to be applied, care must be taken to ensure that insulators themselves do not contain cryptic regulatory sequences. Additionally, the insulating activity of a given spacer can itself be context-specific (Carr *et al.*, 2017), potentially necessitating intensive screening to identify a significantly robust insulator.

In lieu of inserting DNA sequences to physically separate genetic regulatory parts, RNA processing could be applied to increase the modularity of regulatory sequences. For example, riboszymes have previously been used to cleave the 5' UTR from mRNA transcripts (Lou *et al.*, 2012). By cleaving the mRNA transcript at a defined location between 5' UTR and CDS, ribozymes remove any 5' leader sequences that may differentially arise from different upstream sequences and cause context-dependent expression (Bashor & Collins, 2012). The output of individual regulatory sequences therefore remains constant regardless of the surrounding sequence (Lou *et al.*, 2012, Nielsen *et al.*, 2016). CRISPR-mediated RNA cleavage could also be applied to decouple RBS activity from the contextual sequence (Qi *et al.*, 2012). Both ribozyme and CRISPR mediated insulation have proven capable of increasing the modularity of transcription and translation regulators in *E. coli*, and could also potentially be applicable in *Geobacillus*.

5.2.4 The effect of oxygen concentration on promoter activity

The seven promoter sequences that were shown to function independently of genetic context (Figure 5.7) were characterised upstream of both *GFP* and *mOrange* in 250 ml baffled and non-baffled flasks (Figure 5.8).

Congruent with literature concerning the *G. stearothermophilus ldhA* promoter (Bartosiak-Jentys *et al.*, 2012), the *G. thermodenitrificans ldhA* promoter resulted in differential expression between the two growth formats for both reporter proteins. A statistically significant decrease in GFP and mOrange expression ($P = 0.0012$ and $P = 0.0257$, respectively) was apparent when *G. thermoglucosidans* transformants were cultured in baffled flasks as compared to non-baffled flasks, indicating that expression from the *ldhA* promoter was

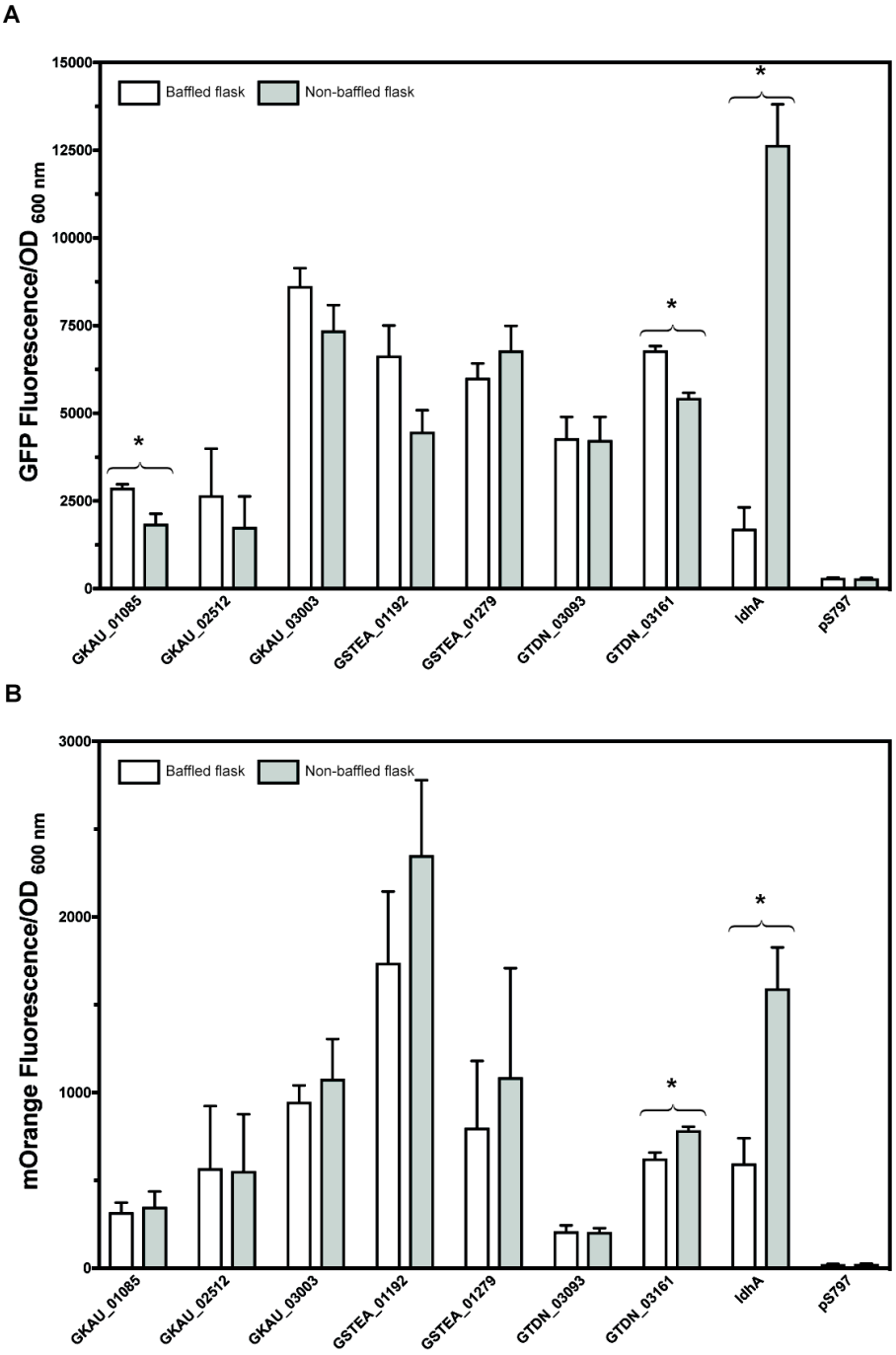


Figure 5.8: Fluorescence output of *G. thermoglucosidans* transformants expressing A) GFP and B) mOrange, cultured in baffled and non-baffled 250 ml flasks.

Fluorescence and absorbance measurements after 24h. Bars represent the mean of n = 3 starter cultures, arising from independent transformants. Each starter culture was used to inoculate one baffled and one non-baffled flask. Standard deviation error bars shown, unless hidden by the bar. Those promoter sequences for which mean fluorescence output was statistically significantly different between the two growth formats are indicated by an asterisk. Significance was determined by multiple t-tests at a significance level of 0.05, corrected for multiple comparisons using the Holm-Šidák method.

down-regulated in environments with increased oxygenation. Significance was determined by multiple t-tests at a significance level of 0.05, corrected for multiple comparisons using the Holm-Šidák method.

One of the seven characterised promoters showed a statistically significant difference in activity levels for both reporter proteins between the two growth formats. As compared to growth in non-baffled flasks, the promoter sequence GTDN_03161 caused a statistically significant increase in GFP fluorescence ($P = 0.0021$) and a statistically significant decrease in mOrange fluorescence ($P = 0.0172$) when cultured in a baffled flask. Additionally, the promoter sequence GKAU_01085 caused a statistically significant increase in GFP fluorescence ($P = 0.0274$) when cultured in non-baffled flasks, although the observed decrease in mOrange expression was not significant ($P = 0.9534$). The other five *cis*-regulatory sequences that were characterised in both flask growth formats showed no significant change in expression, suggesting that the activity of these five promoter sequences was independent of culture aeration.

5.3 Summary

The results of this chapter highlighted the importance of characterising candidate parts for synthetic biology applications in a variety of genomic and environmental contexts. If synthetic biology parts, such as promoters, are to be truly modular and scalable, they must function consistently regardless of genetic context and the growth conditions under which they are employed. Without considering the effect of context on the performance of synthetic biology parts, any resulting characterisation data cannot be generalised, necessitating time consuming empirical testing and optimisation when designing synthetic pathways.

Further work is required to decouple the majority of the characterised *Geobacillus* promoter sequences from the effects of environmental and genomic context. The presence of cloning scar sequences, the reporter protein used for part characterisation and culture oxygenation were all shown to impact,

to a greater or lesser extent, on the performance of the bioinformatically identified *Geobacillus cis*-regulatory sequences. In particular, only 9% of the *cis*-regulatory sequences that were characterised upstream of both *GFP* and *mOrange* showed promoter activity that was independent of the downstream coding sequence. The 73 characterised sequences for which promoter activity was dependent on CDS therefore require additional modularisation, potentially through the use of insulator sequences or ribozyme- or CRISPR-mediated RNA processing, if they are to become more generally applicable.

Furthermore, although PCN was not positively correlated with culture fluorescence, partial and standard least squares statistical models showed that PCN was having a statistically significant positive impact on culture fluorescence. However, the magnitude of this effect was less than the effect of six of the strongest and two of the weakest promoters characterised. This result had implications for potential future applications of the *Geobacillus cis*-regulatory sequences, as it suggested that some manner of copy-number regulation would be required to achieve nuanced control of gene expression in *Geobacillus*.

The promoter sequences that were characterised in Chapters 3, 4, and 5 were bioinformatically identified from the core genome of four *Geobacillus* species. The results of these characterisation experiments showed that bioinformatic prospecting of *cis*-regulatory elements, followed by deep empirical characterisation of *in vivo* activity, was an efficient method by which promoter toolkits can be expanded in non-model organisms. However, the bioinformatic approach to promoter identification is not the only method available for the identification of prokaryotic promoters. The mutagenesis of pre-existing, well-understood promoters, for example, is a widely used technique and it is unclear which approach to promoter discovery and design is the most applicable in non-model organisms such as *Geobacillus*. To provide a direct comparison between bioinformatic and mutagenic approaches to promoter discovery, two commonly employed methods, error-prone PCR and Saturation Mutagenesis of Flanking Regions, were investigated. The results of this investigation are discussed in the following chapter.

6 Comparing mutagenesis approaches for synthetic promoter production

Summary

The results that were discussed in Chapters 3, 4 and 5 showed that bioinformatic prospecting of *cis*-regulatory elements, followed by deep characterisation of *in vivo* activity, represents an efficient, practical method by which the promoter toolkit can be expanded in non-model organisms. However, other methods are available for promoter discovery, including mutagenesis-based approaches, and it is unclear which technique is most applicable in an industrial context. To provide a direct comparison between bioinformatic and mutagenic approaches to promoter discovery, two commonly employed mutagenic methods, Saturation Mutagenesis of Flanking Regions and error-prone PCR, were investigated. In both instances, the *G. thermodenitrificans ldhA* promoter, which has previously been used for metabolic engineering in *Geobacillus*, was used as the template sequence.

6.1 Introduction

The random mutation of pre-existing, well-understood promoter sequences or motifs has become a widely employed technique for the production of synthetic promoter libraries (SPLs) (Blazeck & Alper, 2013, Gilman & Love, 2016). The methods by which mutagenesis-derived SPLs are generated can be broadly split into two categories. In the first approach, Saturation Mutagenesis of Flanking Regions (SMFR), promoter consensus regions such as the -35, -10 and RBS motifs are held constant, while the flanking regions surrounding the core motifs are mutagenised (Jensen & Hammer, 1998a, 1998b). Alternatively, error-prone PCR (epPCR) may be used to introduce mutations throughout an entire promoter sequence (Alper *et al.*, 2005).

Both of the approaches to promoter mutagenesis have specific advantages and disadvantages. SMFR, for example has been shown to yield SPLs with broad expression profiles in a diverse range of species and genera, including *Clostridium acetobutylicum* (Yang *et al.*, 2017a)*, Corynebacterium glutamicum* (Rytter *et al.*, 2014), *Francisella novicida,* (McWhinnie & Nano, 2013)*, Geobacillus thermoglucosidans* (Jensen *et al.*, 2017, Pogrebnyakov *et al.*, 2017)*, Lactoccoccus lactis* (Jensen & Hammer, 1998a, 1998b) *Saccharomyces cerevisiae* (Ellis *et al.*, 2009, Blount *et al.*, 2012, Yang *et al.*, 2017c) and *Steptomyces coelicolor* (Sohoni *et al.*, 2014), and is therefore potentially applicable in non-model organisms of industrial relevance.

However, SMFR requires *a priori* knowledge of the optimal location and nucleotide composition of functional motifs, which may not be readily available in non-model organisms. Additionally, given that studies that apply SMFR to the development of synthetic promoters often use composite promoter scaffolds as starting points, establishing a definitive wild-type reference expression baseline is impossible. Definitively stating whether SMFR will improve wild-type expression capability *pre hoc* is therefore problematic (Blazeck & Alper, 2013, Gilman & Love, 2016).

The second mutagenesis-based approach to SPL production, epPCR, has also previously been employed in diverse species, including *Escherichia coli* (Alper *et al.*, 2005) *Geobacillus thermoglucosidans* (Reeve *et al.*, 2016), *Mycobacterium bovis* (Kanno *et al.*, 2016), *Saccharomyces cerevisiae* (Nevoigt *et al.*, 2006) and *Synechococcus* sp. strain PCC 7002 (Markley *et al.*, 2015). By introducing mutations across an entire promoter sequence, epPCR obviates any *a priori* knowledge of functional motif location (Gilman & Love, 2016). The epPCR approach to promoter derivation is therefore potentially more immediately applicable than SMFR in species in which the number of previously characterised promoter sequences is limiting; a single promoter sequence is required as a starting point, rather than multiple characterised sequences to permit the derivation of consensus motifs. However, if the rate at which epPCR introduces mutations is low, the resulting promoters can be highly homologous, potentially decreasing the stability of engineered pathways. For example, highly

homologous promoters could potentially recombine with the wild-type promoter in the host genome (Pogrebnyakov *et al.*, 2017). Alternatively, if an engineered pathway were to require multiple promoters from the same SPL, significant homology could lead to recombination between mutated sequences (Hammer *et al.*, 2006).

The selection of one method for SPL production over the other will depend, in part, on the aims of specific projects. Whichever method is selected, from an industrial perspective it is vital for the resulting SPL to contain robust promoters covering a wide range of expression levels. To provide a direct comparison between the two methods and to therefore assess which, if either, of the two approaches was best suited for use in an industrial context, libraries of synthetic *Geobacillus* promoters were generated using both SMFR and epPCR. In both cases, the *G. thermodenitrificans ldhA* promoter was used as a template sequence.

In the case of the SMFR-derived *Geobacillus* SPL, degenerate oligonucleotides were designed that maintained the putative -35, -10 and RBS motifs from the *G. thermodenitrificans ldhA* promoter, whilst randomising the remainder of the 150 bp sequence. At each position where degeneracy was specified, all four nucleobases had an equal probability of occurring. In the case of the epPCR-derived SPL, mutations were incorporated at random across the entire 150 bp sequence.

Previous studies have highlighted the relative inefficiency of random mutation as a strategy for deriving sequences with *in vivo* promoter activity. The screening of large numbers of mutated sequences is often necessary to identify a comparatively small number of sufficiently robust, active promoter elements. The number of screened mutants can, in some extreme cases, be three to four orders of magnitude greater than the number of fully characterised sequences in the final SPL (Alper *et al.*, 2005, Fischer *et al.*, 2006, Qin *et al.*, 2011, Yim *et al.*, 2013, Wei *et al.*, 2018).

Low transformation efficiencies precluded such large-scale screening of mutated sequences in *G. thermoglucosidans*. Therefore, to increase the number of mutated promoter sequences that could be analysed, *E. coli* was used as an intermediate host. Mutated promoter sequences were cloned upstream of the *GFP* CDS, and transformed into *E. coli* NEB 5-alpha *en masse*. FACS analysis was subsequently used to isolate mutated *promoter::GFP* fusions with *in vivo* fluorescence activity. Mutant promoter sequences were only selected for characterisation in *G. thermoglucosidans* if they resulted in GFP fluorescence in *E. coli* that was greater than the negative control, *E. coli* transformed to contain the empty vector pS797. A lack of fluorescence was assumed to indicate either a failed cloning reaction, or that the mutated within the analysed culture had no promoter activity. This pre-screening approach assumed that promoter activity in *E. coli* was representative of activity in *G. thermoglucosidans*. However, *in vivo* characterisation of promoter activity showed that this assumption was erroneous.

6.2 Results

6.2.1 Saturation Mutagenesis of Flanking Regions

Approximately 4,000 *E. coli* transformants, expressing GFP under the control of SMFR-derived putative promoter sequences, were screened for fluorescence activity. 28 of these 4,000 transformants, displaying a range of fluorescence levels, were isolated as described in Chapter 2, Section 2.4.3. The putative promoters were subsequently extracted, sequenced and characterised in *G. thermoglucosidans*.

The final SMFR-derived SPL covered a total GFP expression range of 19.4-fold when characterised in *G. thermoglucosidans*, and 53.7-fold when characterised in *E. coli* (Figure 6.1). However, the majority of the mutated sequences showed no promoter activity in *E. coli*. When characterised in *G. thermoglucosidans*, only two *promoter::GFP* fusions showed mean fluorescence that was statistically significantly greater than the negative control, *G. thermoglucosidans* transformed with the empty vector pS797. When characterised in *E. coli,* six promoters showed mean activity levels that were statistically significantly greater than that of the negative control. In all cases, significance was determined by ordinary one-way ANOVA, with Dunnett's multiple comparisons test at a significance level of 0.05.

The two sequences that showed statistically significant promoter activity in *G. thermoglucosidans*, GSYN_SMFR_020 and GSYN_SMFR_030, were also active in *E. coli.* Overall, promoter activity levels between *E. coli* and *G. thermoglucosidans* showed moderate conservation; a linear regression of the two data sets returned an R^2 value of 0.604.

Analysis of the DNA sequences of the mutated promoters showed that only five out of the 28 mutants had the correct sequence at all three of the conserved motifs (Figure 6.2). The particular dearth of sequences containing the specified -35 region may have been a result of the way in which the promoter sequences were cloned.

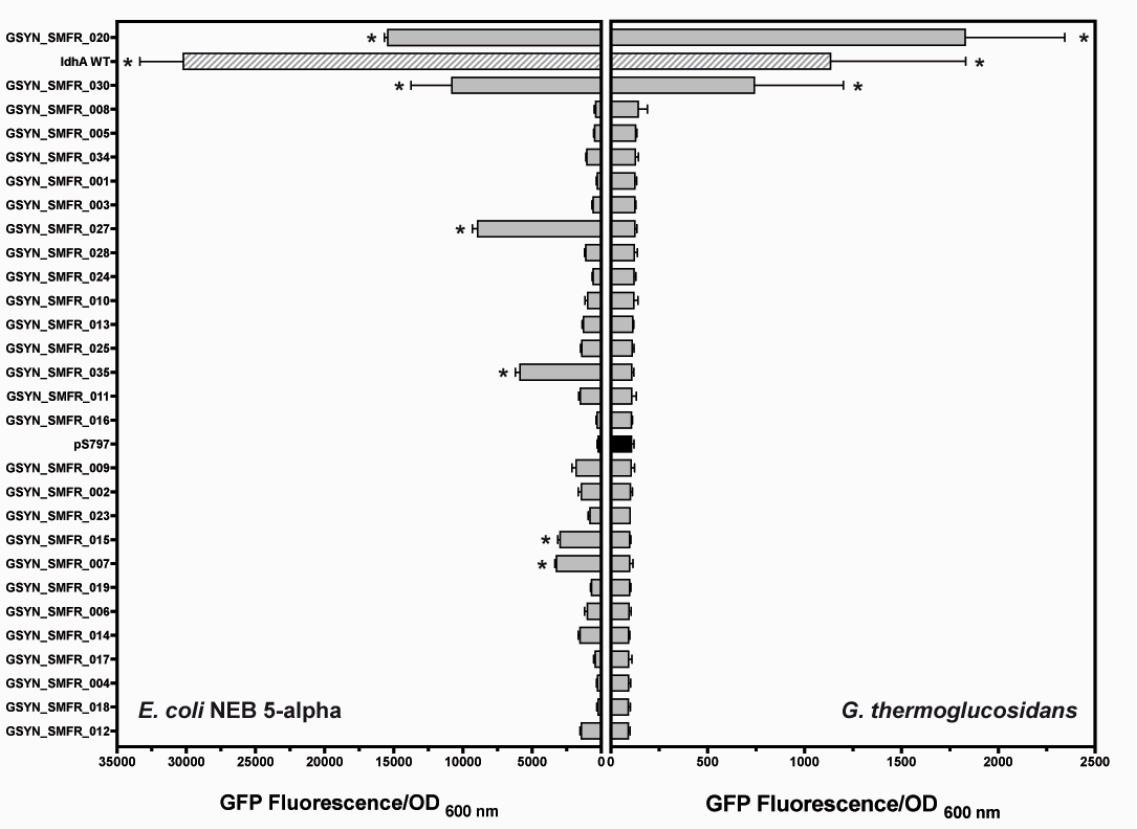


Figure 6.1: *In vivo* characterisation of a Synthetic Promoter Library derived by Saturation Mutagenesis of Flanking Regions.

Fluorescence and absorbance after 24 h incubation in 96-well plate format. Bars represent the mean of 3 ≤ n ≤ 16 starter cultures arising from independent transformants. Standard deviation error bars are shown, unless hidden by the bar. The hatched bars represent the positive control, the *G. thermodenitrificans ldhA* promoter. The black bars represent the negative control, *G. thermoglucosidans* or *E. coli* transformed to contain the empty vector pS797. Asterisks indicate promoter sequences that resulted in mean fluorescence output that was significantly different to the negative control. Significance was determined by ordinary one-way ANOVA with with Dunnett's multiple comparisons test, using a significance level of 0.05.

The length of the final promoter sequence and the requisite cloning affixes prohibited the synthesis of promoters as a single degenerate oligonucleotide. Two overlapping oligonucleotides were therefore annealed *in vitro* to form complete promoter sequences. The first oligonucleotide encoded the antisense strand of the final promoter sequence, up to and including the -35 motif. The second oligonucleotide encoded the sense strand of the remaining promoter sequence, including the -35 motif so that the two oligonucleotides

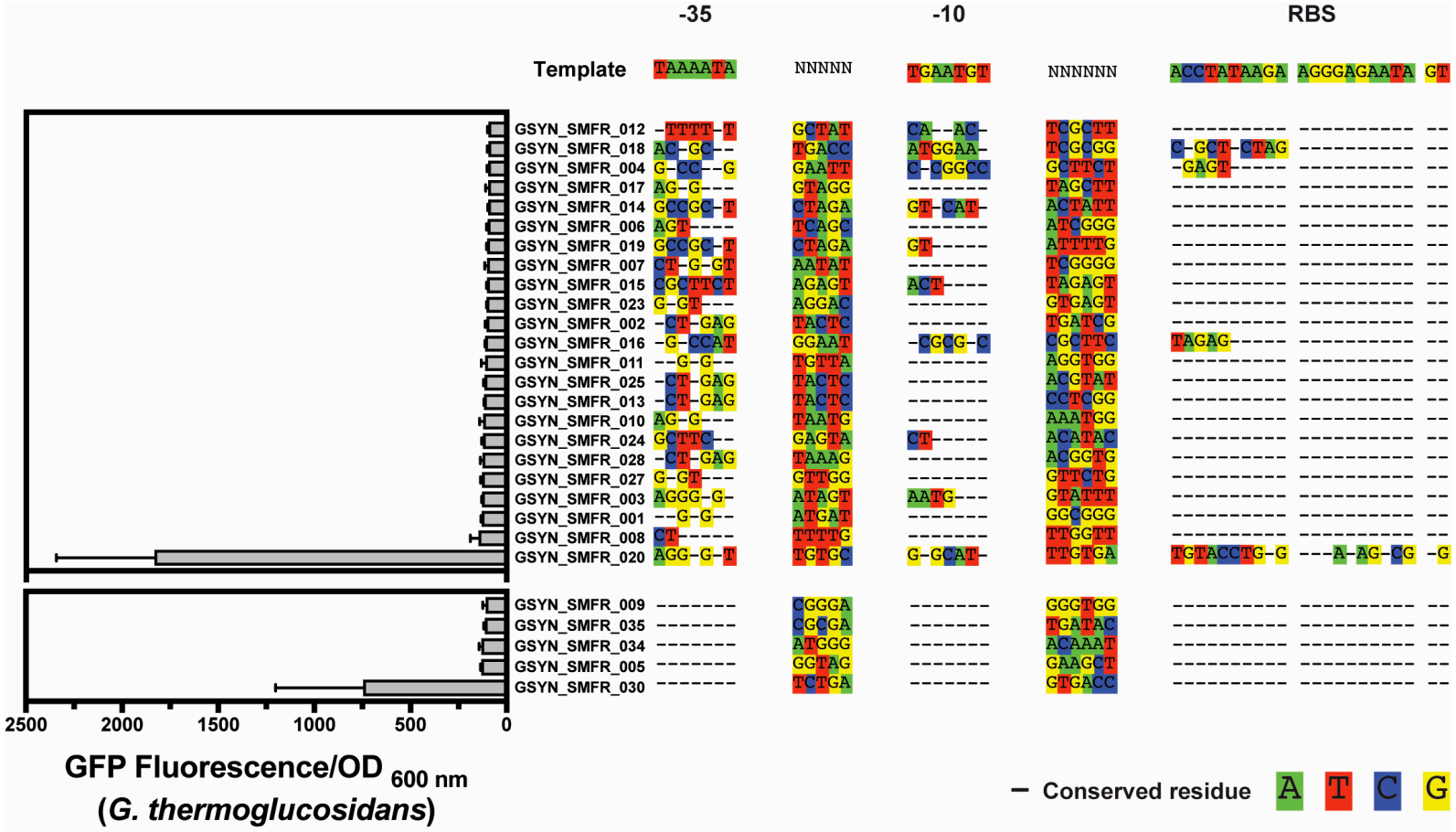


Figure 6.2: Sequence alignment showing the conserved motifs in the Synthetic Promoter Library derived by SMFR

Dashes represent sequence positions at which the mutated sequence was identical to the template. Sequences were aligned and visualised using SeaView version 4.6 (Gouy *et al.*, 2010). Fluorescence and absorbance after 24 h incubation in 96-well plate format. Bars represent the mean of 3 ≤ n ≤ 16 starter cultures arising from independent transformants, with standard deviation error bars shown, unless hidden by the bar.

could be annealed. However, the -35 motif was short, 7 bp, and was comprised solely of adenine and thymine residues. The -35 motif may therefore have been too short and AT-rich to allow for accurate ligation.

Interestingly, the mutated sequence that showed the strongest *in vivo* promoter activity in both *E. coli* and *G. thermoglucosidans*, GSYN_SMFR_020, showed significant deviation from the template in all three motifs (Figure 6.2). This result suggested that consensus sequences other than the putative motifs from the *ldhA* promoter were active in the *Geobacillus* promoter design space, and highlighted the limitations of applying SMFR in organisms for which regulatory motifs are not well understood; the sequences that were used for the production of the SMFR-derived SPL were likely inadequate, either with regards to nucleotide sequence or location, for efficient transcription initiation in *G. thermoglucosidans*.

6.2.2 Error-prone PCR

Approximately 7,000 *E. coli* transformants, expressing GFP under the control of epPCR-derived putative promoter sequences, were screened for fluorescence activity. 39 of these 7,000 transformants, displaying a range of fluorescence levels, were isolated as described in Chapter 2, Section 2.4.3. When characterised in *G. thermoglucosidans*, the 39 *ldhA* promoter mutants covered a GFP expression range of 7.5-fold (Figure 6.3). However, as with the SMFR-derived library, the promoter activity of the majority of sequences was, at best, minimal; only two of the mutated sequences caused GFP expression that was statistically significantly greater than the negative control. Significance was determined by ordinary one-way ANOVA with Dunnett's multiple comparisons test, at a significance level of 0.05. None of the mutated sequences resulted in GFP fluorescence that was stronger than the wild-type *ldhA* promoter.

The strongest mutant promoter, GSYN_EP_027, resulted in mean GFP expression that was approximately two-fold lower than that caused by the wild-type *ldhA* promoter. No correlation was apparent between promoter activity levels as characterised in *E. coli* and *G. thermoglucosidans.* A linear regression

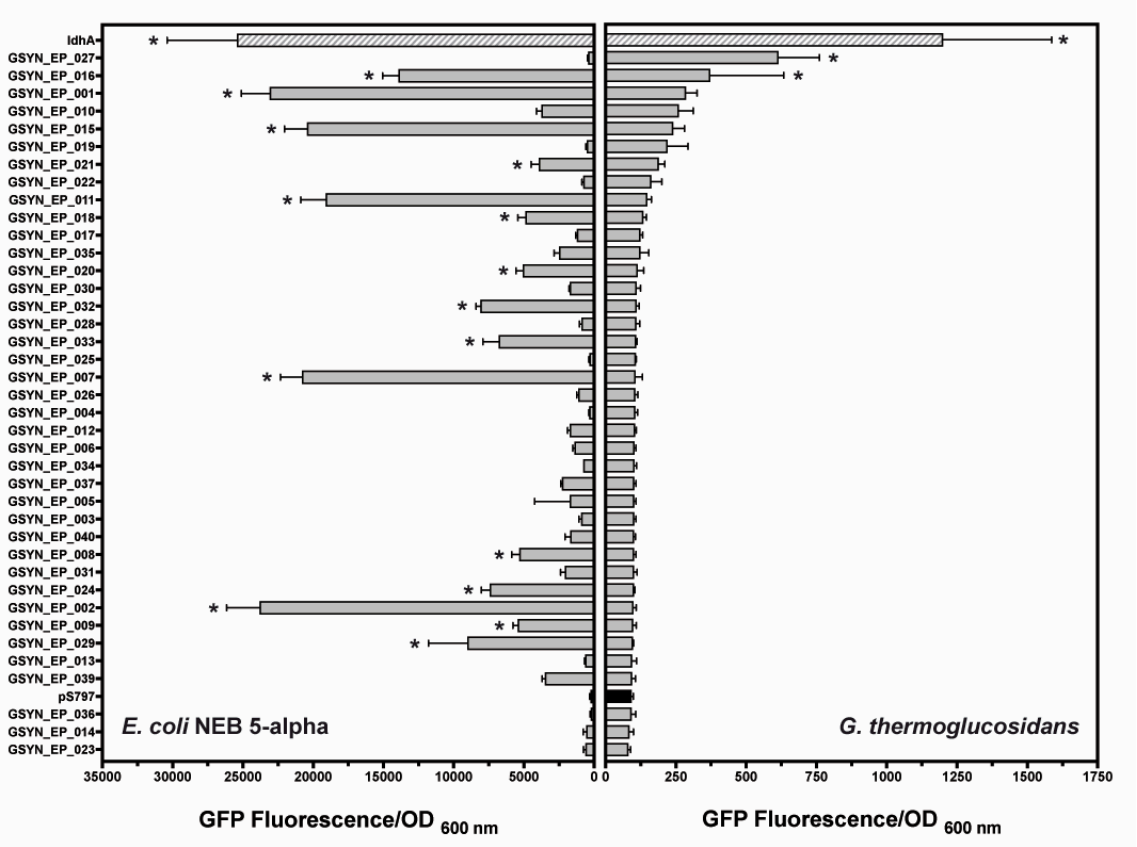


Figure 6.3: *In vivo* characterisation of a Synthetic Promoter Library derived by error-prone PCR.

Fluorescence and absorbance after 24 h incubation in 96-well plate format. Bars represent the mean of 3 ≤ n ≤ 8 starter cultures arising from independent transformants. Standard deviation error bars are shown, unless hidden by the bar. The hatched bars represent the positive control, the *G. thermodenitrificans ldhA* promoter. The black bars represent the negative control, *G. thermoglucosidans* or *E. coli* transformed to contain the empty vector pS797. Asterisks indicate promoter sequences that resulted in mean fluorescence output that was significantly different to the negative control. Significance was determined by ordinary one-way ANOVA with with Dunnett's multiple comparisons test, using a significance level of 0.05.

of the two data sets returned an R^2 value of 0.03. When characterised in *E. coli,* the 39 mutant sequences covered an expression range of 91.6-fold. 14 of the sequences resulted in GFP expression in *E. coli* that was statistically significantly greater than the negative control.

The mutation rate of the 39 *ldhA* promoter variants ranged from 8% to 76%. The average mutation rate was 21.5%. The mutation rates of three sequences (GSYN_EP_014, 72%; GSYN_EP_015, 68.7%; GSYN_EP_037, 76%) were outlying with regards to the other 36 sequences, as determined by Huber M-Estimation with a K value of 4. The 36 non-outlying sequences had an average error rate of 17.3%. In all of the characterised mutated promoter sequences, mutations were distributed throughout the complete 150 bp sequences, including within the putative consensus regions (Figure 6.4). No correlation was apparent between promoter mutation rate and activity level. A linear regression of the two data sets returned an R^2 value of 0.014.

6.3 Discussion

The sequence space explored by the two mutagenesis-derived SPLs did contain active *Geobacillus* promoters aside from the *ldhA* wild-type sequence, but at a low frequency; only 10% of the sequences characterised in the SMFR-derived library and 5% of the sequences in the epPCR-derived library showed activity in *G. thermoglucosidans* that was statistically significantly different to the negative control (Figure 6.1 and Figure 6.3, respectively). This result suggested that functional *Geobacillus* promoter sequences were extremely stringent, with a relatively small deviation from the wild-type sequence sufficient to destroy promoter activity (Mordaka & Heap, 2018). The epPCR-derived sequence GSYN_EP_029, for example, differed from the wild-type *ldhA* promoter sequence at only 8% of the 150 sequence positions. However, GSYN_EP_029 showed no *in vivo* promoter activity; mean GFP fluorescence under GSYN_EP_029 was within one standard deviation of the mean fluorescence of the negative control, *G. thermoglucosidans* transformed to contain the empty vector pS797.

The relatively high average error rates of both the SMFR- and epPCR-derived SPLs may have contributed to the observed lack of *in vivo* promoter activity. Only five of the sequences in the final SMFR-dervied SPL, for example, contained all three of the putative motifs that were identified in the wild-type

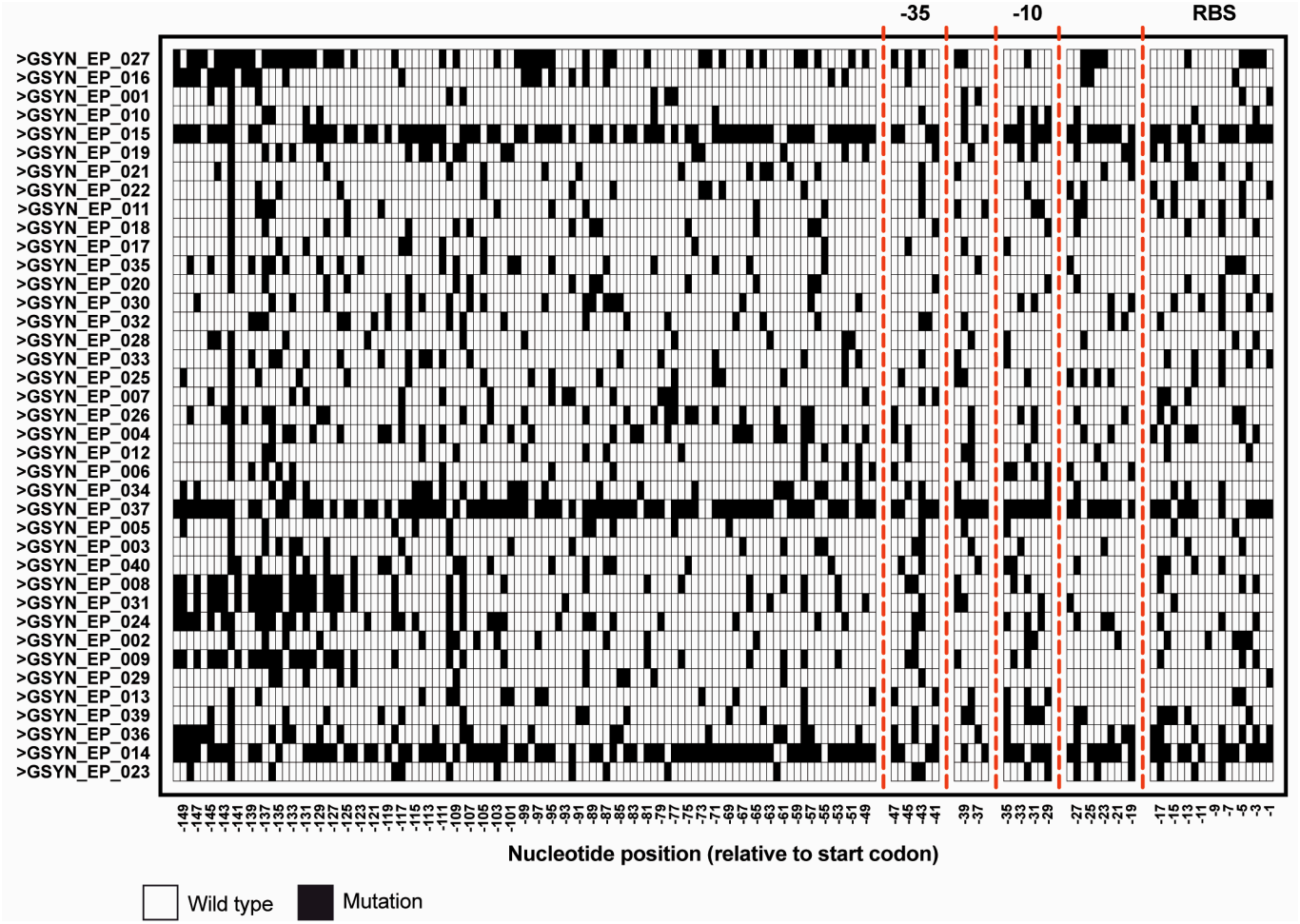


Figure 6.4: Heat map showing the location of mutated nucleobases within sequences in the synthetic promoter library derived by error-prone PCR.

Positions that matched the wild-type *G. thermodenitrificans ldhA* promoter are shown in white. Sequence positions that were mutated relative to the wild-type *ldhA* promoter are shown in black. Sequences are ranked in descending order of *in vivo* GFP expression strength, as measured in *G. thermoglucosidans*. The red dashed lines bound the locations of the putative -35, -10 and RBS motifs.

ldhA promoter, and the 29 mutated sequences shared, on average, only 40% of their sequence with the wild-type *ldhA* sequence. Reducing the mutation rate of the SMFR template so that the resulting promoters were closer to the wild-type sequence may have resulted in a SPL in which promoter activity was better conserved.

Reducing the mutation rate of an SMFR-derived SPL to better conserve promoter activity has shown to be an applicable strategy in *Clostridium*, and may have been useful in *Geobacillus*. Instead of using an SMFR template in which mutated positions were fully degenerate, Mordaka & Heap applied an SMFR in which the probability of a given sequence position in the SMFR template matching the wild-type sequence was 79%, rather than 25% as is the case in fully degenerate SMFR. The resulting “tuned” SPL contained 10 promoters with *in vivo* activity, as compared to a fully degenerate SMFR-derived SPL that contained no active promoters (Mordaka & Heap, 2018).

In addition to tuning the mutation rate of the SMFR-derived SPL to increase the number of active promoters, alternative -35, -10 and RBS motifs may have improved the *in vivo* activity of the *G. thermoglucosidans* SPL that was obtained in this investigation. Given that the promoter sequence from the SMFR-derived SPL with the strongest *in vivo* promoter activity matched the template sequence at only 34% of the conserved positions (Figure 6.2), the defined location and sequence composition of the putative motif sequences may itself have been sub-optimal.

In lieu of the sequence alignment that was used identify the location of putative *Geobacillus* -35, -10 and RBS consensus motifs, a DNA motif-locating software package, such as DMINDA 2.0 (Yang *et al.*, 2017b) or MEME (Bailey *et al.*, 2009) may have more accurately identified regulatory motifs. These motifs could then have been used to design the oligonucleotides that were used to generate the SMFR SPL, potentially increasing the characterised range of expression afforded by the SPL.

In the case of the epPCR-derived library, the average error rate of 21.5% was double that reported by Reeve *et al.*, whose SPL covered a reported GFP expression range of 100-fold in *G. thermoglucosidans* (Reeve *et al.*, 2016). By reducing the rate at which mutations were incorporated into the wild type *ldhA* promoter sequence may therefore have better maintained promoter activity and resulted in broader expression range from the final SPL. It is worth noting, however, that the SPL published by Reeve *et al.* only contained one promoter sequence with a greater activity level than the wild-type sequence that was mutagenised, and the published error bars suggested that this difference was not statistically significant. Furthermore, the published figure lacked a negative control, and the weakest eight members of the library showed, at best, minimal *in vivo* activity. It was therefore unclear how much of the stated 100-fold range represented statistically significant promoter activity.

A greater proportion of the mutated sequences that were derived in this investigation showed promoter activity in *E. coli* than in *G. thermoglucosidans;* approximately 20% of the sequences in the SMFR-derived SPL and approximately 40% of the sequences in the epPCR-derived SPL showed activity that was statistically significantly different to the negative control when characterised in *E. coli.* The relative lack of conservation of promoter activity levels between *E. coli* and *G. thermoglucosidans*, especially in the case of the epPCR-derived SPL, was congruent with the lack of inter-species correlation shown by the bioinformatically identified, natural *Geobacillus* promoter sequences discussed in Chapter 3.

The lack of inter-species conservation of promoter activity shown by both the bioinformatically identified and mutagenised promoter sequences underscored the difficulty of transferring genetic parts between phylogenetically distant host organisms (Cardinale & Arkin, 2012). The lack of conservation in the DNA binding affinity of homologous cellular machinery such as RNA polymerases and transcription factors complicates the transfer of parts (Adams, 2016), as it cannot be assumed that the characterised activity of a part in one host will be representative of activity in another. This inter-species variation in the activity of synthetic biology parts has been observed even when the two

host organisms being compared are closely related (Adams, 2016). Therefore, although collections of thoroughly characterised regulatory mechanisms of increasing sophistication are becoming commonplace in the literature, these collections cannot necessarily be easily used to facilitate metabolic engineering in species other than those in which they were initially developed. The lack of conservation in promoter activity between *E. coli* and *Geobacillus* therefore highlighted the necessity of the development of species-specific synthetic biology toolkits for non-model, industrially relevant organisms.

In addition to reducing the potential for inter-species transfer of synthetic biology tools, the lack of conservation in promoter activity levels between *E. coli* and *G. thermoglucosidans* may have inadvertently reduced the range of GFP expression levels shown by the SPLs when they were characterised in *G. thermoglucosidans*. By initially selecting for mutant sequences that showed promoter activity in *E. coli*, it is possible that promoters that were inactive in *E. coli* but active in *G. thermoglucosidans* were overlooked. Characterisation in *G. thermoglucosidans* of putative promoter sequences that showed no activity in *E. coli* may therefore have potentially increased the range of expression levels shown in the final SPLs.

Alternatively, bypassing the intermediate screening of mutated sequences in *E. coli* and assessing promoter activity directly in *G. thermoglucosidans* may have led to the identification of a greater number of active sequences. However, given the apparent stringency of *Geobacillus* promoters, the probability of identifying a randomly mutated sequence with *in vivo* promoter functionality in *G. thermoglucosidans* without screening large numbers of mutants was likely small.

Even if large numbers of mutated promoter sequences were screened in *G. thermoglucosidans*, a review of published mutagenesis-derived promoter libraries suggested that the resulting SPL would likely contain few promoter sequences of greater activity than the *ldhA* promoter starting point (Table 6-1).

A:

	Organism	Reference promoter (ref.)	Number of promoters: in SPL	>ref.	Reference
SMFR	*C. acetobutylicum*	*p_THL*	35	10	(Yang *et al.*, 2017a)
SMFR	*C. acetobutylicum*	*p_thl*	22	3	(Mordaka & Heap, 2018)
SMFR	*Corynebacterium glutamicum*	*p_tac*	69	4	(Zhang *et al.*, 2018)
SMFR	*C. glutamicum*	*p_Trc*	35	31	(Wei *et al.*, 2018)
SMFR	*E. coli*	*p_LacI*	71	56	(De Mey *et al.*, 2007)
SMFR	*E. coli*	*p_pgi*	20	12	(Braatsch *et al.*, 2008)
SFMR	*F. novicida*	*p_bfr*	15	0	(McWhinnie & Nano, 2013)
SMFR	*G. thermoglucosidans*	*p_groES*	17	2	(Pogrebnyakov *et al.*, 2017)
SMFR	*Homo sapiens* ARPE-19 & *Rattus rattus* HiB5 cell lines	*p_JeT*	27	0	(Tornøe *et al.*, 2002)
SMFR	*Lactobacillus plantarum*	*p_r-RNA3-a*	33	0	(Rud *et al.*, 2006)
SMFR*	*Rhodococcus opacus*	*p_ermEP1*	25	1	(DeLorenzo *et al.*, 2018)
SFMR	*S. cerevisiae*	*p_GAL1*	20	0	(Ellis *et al.*, 2009)
SMFR	*S. cerivisiae*	*PFY1*	36	0	(Blount *et al.*, 2012)
SMFR	*Streptomyces. coelicolor*	*p_act11 orf4*	11	6	(Sohoni *et al.*, 2014)
SMFR	*S. lividans*	*p_ermEP1*	56	8	(Siegl *et al.*, 2013)

B:

	Organism	Reference promoter (ref.)	Number of promoters: in SPL	>ref.	Reference
epPCR	*E. coli*	*p_Trc*	99	20	(Meng *et al.*, 2013)
epPCR	*E. coli*	*p_L-Y*	22	13	(Alper *et al.*, 2005)
epPCR	*G. thermoglucosidans*	*p_Rpls*	20	1	(Reeve *et al.*, 2016)
epPCR	*Pichia pastoris*	*p_GAP*	33	16	(Qin *et al.*, 2011)
epPCR	*S. cerevisiae*	*p_TEF*	11	5	(Alper *et al.*, 2005, Nevoigt *et al.*, 2006)
epPCR	*Synechococcus* sp. strain PCC 7002	*p_cpt*	29	2	(Markley *et al.*, 2015)

Table 6-1: Improvement of promoter strengths reported by studies that applied A) Saturation Mutagenesis of Flanking Regions or B) error-prone PCR to the production of Synthetic Promoter Libraries.

* Performed saturation mutagenesis of the -35 and -10 consensus regions, whilst holding the flanks constant.

Given the widespread use of composite scaffolds for SMFR, defining a wild-type promoter strength baseline was not always possible (Blazeck & Alper, 2013). However, out of 21 published SPLs for which a wild-type baseline could be established, 16 were predominately comprised of promoters of reduced strength compared to the baseline (Table 6-1). Five of these SPLs contained no promoters of greater strength than the wild-type reference. Furthermore, in instances when promoters of greater strength than the wild type were reported, the magnitude of the increase in strength was typically relatively small.

For example, Yang *et al.* reported a SMFR-derived *Clostridium acetobutylicum* SPL, in which 29% of the characterised mutants showed stronger promoter activity than the wild-type reference promoter, *pTHL.* However, the strongest of these mutated sequences afforded only a 0.4-fold upregulation in promoter activity as compared to *pTHL* (Yang *et al.*, 2017a).

Likewise, Sohoni *et al*. reported a SMFR-derived SPL for *Streptomyces coelicolor* in which 55% of the characterised sequences were stronger than the reference, but a strongest promoter that was only 1.75-fold stronger than said reference (Sohoni *et al.*, 2014). Approximate one- to three- fold increases in promoter strength were also reported in SPLs for *C. acetobutylicum* (Mordaka & Heap, 2018), *E. coli* (Meng *et al.*, 2013)*, S. cerevisiae* (Alper *et al.*, 2005, Nevoigt *et al.*, 2006) and *S. lividans* (Siegl *et al.*, 2013).

A minority of SPLs did report large upregulation in promoter activity. In *Pichia pastoris* for example, Qin *et al.* screened 30,000 epPCR-derived mutants of the *GAP* promoter and isolated a SPL of 33 characterised promoters. 16 members of the final SPL showed greater promoter activity than the wild-type *GAP* promoter. Of these 16 sequences, the strongest showed 74-fold stronger expression of *lacZ* at the transcript level than *pGAP* (Qin *et al.*, 2011). Wei *et al.* also employed high-throughput screening in *Corynebacterium glutamicum* to ultimately isolate 31 promoters that were stronger than the wild-type reference at the transcript level from an initial library of approximately 5 x 10^6 clones (Wei *et al.*, 2018).

Finally, De Mey *et al*. reported an apparently large increase in promoter strength in *E. coli.* Approximately 80% of the mutant sequences in the SMFR-derived SPL showed greater activity than the wild type reference, the *LacI* promoter. The strongest of these mutants showed an approximately 30-fold increase in GFP expression as compared to *pLacI.* However, the consensus regions that were used in the synthetic promoter sequences were not present in *pLacI* (De Mey *et al.*, 2007). The extent to which the wild type *pLacI* served as a truly comparable baseline was therefore debatable.

6.4 Summary

The SMFR- and epPCR- derived SPLs that were generated in this study contained few active promoters when characterised in *G. thermoglucosidans*. Although alterations to the experimental design (*i.e.* different conserved motifs in the SMFR-derived SPL, a lower average mutation rate in the epPCR-derived SPL) may have increased the number of sequences with *in vivo* promoter activity, a review of the literature highlighted the propensity of mutagenesis-based approaches to promoter engineering to predominately decrease promoter strength.

Of course, maximising the expression strength of a heterologous gene is not always desirable. Indeed, promoters of moderate and low strengths are equally valuable, and permit the nuanced tuning of synthetic pathways (Goldbeck *et al.*, 2012). However, given the proclivity of both SMFR and epPCR to reduce promoter activity, the final expression range of a mutagenesis-derived SPL is often highly dependent on the strength of the initial wild-type promoter or composite scaffold. In instances where the number of *a priori* characterised promoter elements in a species or genus of interest is low, or if the understanding of the location and functional composition of consensus motifs is limited, the use of *a posteriori*, mutagenesis-based approaches to promoter design is likely to be an inadequate method by which to explore the promoter design space.

In contrast to the lack of promoter activity shown by the mutagenesis-derived *Geobacillus* promoter libraries, the bioinformatic approach to promoter discovery that was discussed in Chapter 3 yielded libraries of endogenous promoter elements that covered two-log ranges of fluorescence, increasing in steady increments (Chapter 4). Bioinformatic promoter prospecting, followed by deep empirical characterisation of promoter activity, therefore represents a comparatively more efficient, practical method by which the promoter toolkit can be rapidly expanded in non-model industrially relevant organisms for which *a priori* knowledge of *cis*-regulatory elements is lacking.

7 General Discussion

Summary

This investigation aimed to ascertain which method for promoter discovery and design was most applicable in an industrial context, and in particular to assess the applicability of statistical learning approaches to promoter identification and characterisation. To that end, a novel bioinformatic approach to promoter identification was developed, and promoter sequences were rationally selected for *in vivo* characterisation. Collectively, the characterised sequences resulted in a 2-log range of expression strengths when placed upstream of two fluorescent reporter proteins, and promoter sequences were shown to individually yield homogenous control of gene expression. Minimal conservation of activity between reporter proteins was observed for the majority of the characterised promoters, indicating that further work is required to reduce the context-dependency of the characterised promoters if they are to be used to facilitate bottom-up pathway design in industrial contexts. Although ANN and PLS models derived from the characterisation data showed inadequate generality, Random Forest partition models were shown to be a useful tool for increasing understanding of transcription regulation in non-model organisms.

The results of this investigation can potentially provide a foundation for a number of future studies. Future work could aim to increase the context-independence of the characterised promoters, or make use of the increased understanding of the *Geobacillus* promoter design space provided by the Random Forest partition models to facilitate rational promoter optimisation. Alternatively, the novel bioinformatic approach to promoter identification that was developed, coupled with sequence-function statistical modelling, is potentially broadly applicable to expand promoter toolkits in other non-model synthetic biology host organisms.

7.1 Identification and characterisation of putative promoters

The aim of this investigation was to ascertain which method for promoter discovery and design was most applicable in an industrial context. In particular, we hypothesised that statistical learning approaches to promoter discovery were applicable to non-model organisms, and that such statistical models would accelerate the discovery and characterisation of promoter libraries. Industrial applicability was defined as the capability of a given method to produce libraries of promoter elements that collectively covered broad ranges of recombinant gene expression levels (*i.e.* "promoter strength") and that individually yielded homogeneous gene expression that was orthogonal to endogenous metabolic regulatory pathways. To that end, bioinformatic, mathematic and mutagenic approaches to promoter discovery and design were applied in the industrially relevant chassis, *Geobacillus*.

Bioinformatic screening of *G. kaustophilus* DSM7263, *G. stearothermophilus* DSM22, *G. thermodenitrificans* K1041 and *G. thermoglucosidans* DSM2542 resulted in the identification of 636 putative constitutive promoters from the *Geobacillus* core genome (Chapters 3 & 4). By isolating the 100 bp immediately upstream of coding sequences in the *Geobacillus* core genome as putative promoter sequences, the requirement for any *a priori* understanding of the sequence composition or statistical properties of *Geobacillus cis*-regulatory elements was obviated. Such understanding is requisite for more sophisticated *in silico* approaches to prokaryotic promoter identification (Song, 2011, Umarov & Solovyev, 2017). The approach taken in this investigation is therefore potentially more broadly applicable than other *in silico* approaches in non-model organisms in which the *a priori* understanding of *cis*-regulatory elements is minimal.

To maximise the sequence diversity of *in vivo* characterised promoters, and thereby maximise the proportion of the sequence design space that was empirically explored, promoters were selected for characterisation from across the *Geobacillus* promoter phylogeny. In total, 105 putative promoters were characterised upstream of *GFP* in *G. thermoglucosidans*. To assess the

functional composability of the endogenous *Geobacillus* promoters, 82 of these sequences were also characterised upstream of *mOrange* (Chapters 4 & 5). A subset of seven promoters whose activity was shown to be independent of the downstream coding sequence (CDS) were also characterised with regards to their oxygen-dependence (Chapter 5).

Flow cytometry analysis showed that the characterised *Geobacillus* promoters afforded tight control of protein expression across a three-log range of expression strengths (Chapter 4). Crucially, promoter sequences were identified which showed significantly reduced expression variation as compared to the *ldhA* promoter, which has previously been applied in *Geobacillus* genetic engineering (Cripps *et al.*, 2009, Lin *et al.*, 2014). 98% of the characterised *promoter::GFP* fusions and 73% of the characterised *promoter::mOrange* fusions returned a lower coefficient of variance than the *ldhA* promoter, indicating that these sequences afforded tighter control of expression than *ldhA.* Members of the characterised promoter library were therefore potentially broadly applicable to *Geobacillus* synthetic biology and metabolic engineering projects where tight control of expression is required, such as the optimisation of the complex heterologous pathways that would enable the large-scale bioproduction of fourth generation biofuels (Howard *et al*., 2013).

In contrast to the broad range of activity levels shown by the bioinformatically identified endogenous *Geobacillus* promoters, the two mutagenesis derived promoter libraries contained few active promoters when characterised in *G. thermoglucosidans*; only 10% of the sequences characterised in the library derived through Saturation Mutagenesis of Flanking Regions and 5% of the sequences in the error-prone PCR-derived library showed activity in *G. thermoglucosidans* that was statistically significantly different to the negative control. This result was congruent with a review of the literature, which highlighted the proclivity of mutagenesis to reduce, rather than enhance, promoter activity (Blazeck & Alper, 2013) (Chapter 6).

7.2 Mitigating the effect of genetic and environmental context on gene expression

Standards such as SBOL (Galdzicki *et al.*, 2014), in which engineered biological circuits are visualised in a manner analogous to an electrical wiring diagram, facilitate the view of synthetic biology as "building with legos" (Carlson, 2010). The resulting abstraction hierarchy permits difficulties that are encountered at a given level of the engineering process to be ignored by those working at another level. For example, those working on whole-system optimisation can abstract beyond the difficulties inherent in the design and optimisation of individual genetic parts (Zong *et al.*, 2017). Without this abstraction, the engineering of complex biological systems is, at best, challenging (Mutalik *et al.*, 2013a). However, this hierarchical approach to biological engineering relies upon the composability of the constituent genetic parts.

If the characterisation of synthetic biology parts does not consider the potentially synergistic, antagonistic or neutral effects of environmental and genomic context (Cardinale & Arkin, 2012, Del Vecchio, 2015), the performance of individual parts cannot be generalised, necessitating time consuming empirical testing and optimisation when designing synthetic pathways. As the complexity of engineered pathways increases, such trial-and-error optimisation of individual parts becomes prohibitive (Davidsohn *et al.*, 2014, Rudge *et al.*, 2016). Promoter sequences that display true modularity and functional composability *in vivo* would obviate, or at least reduce the scale of, large-scale *in vivo* tuning, thereby aiding the systematic, scalable, bottom-up design of genetic pathways that synthetic biology strives to achieve (Del Vecchio, 2015, Nielsen *et al.*, 2016).

In model organisms such as *E. coli* and *S. cerevisiae,* studies have considered the effect of environmental (Keren *et al.*, 2013, Johns *et al.*, 2018) and genetic context on the activity of *cis*-regulatory elements (Davis *et al.*, 2011, Kosuri *et al.*, 2013, Mutalik *et al.*, 2013a, 2013b, Zong *et al.*, 2017). However, the drive for composable, modular promoters in non-model organisms

is hindered by the fact that many studies still characterise the function of *cis*-regulatory elements in a restricted number of genomic contexts or under a single growth condition. The general applicability of the promoter sequences characterised in these studies is therefore limited. For example, two libraries of *Geobacillus* promoter sequences that have been published to date used only GFP to characterise promoter performance (Reeve *et al.*, 2016, Pogrebnyakov *et al.*, 2017). Such characterisation is unlikely to be fully representative of the genetic context in which the regulatory elements may ultimately be employed (Moser *et al.*, 2012).

In this investigation, two reporter proteins, GFP and mOrange, were used to assess the functional composability of the bioinformatically identified promoter sequences. Of the 80 promoters that were used to express both reporter proteins, only seven sequences (*i.e.* approximately 9% of the characterised *cis*-regulatory elements), covering an activity range of four-fold, showed promoter activity that was independent of the downstream CDS (Chapter 5). This result suggested that bioinformatic screening, followed by deep characterisation of *in vivo* promoter activity, represented a method by which endogenous *cis*-regulatory elements that function independently of the downstream CDS can be identified for synthetic biology or metabolic engineering applications.

Endogenous context-independent promoters occurred in the *Geobacillus* core genome at a low frequency; only 9% of characterised sequences showed consistent activity across the two reporters. Furthermore, the expression range of four-fold that was returned by the context-independent sequences compared poorly with the total expression range afforded by all the characterised sequences. When used to express GFP, for example, the 47 “active” promoters showed an expression range of 30-fold. When mOrange was the reporter, 31 “active” promoters covered an expression range of 8.4-fold. Future work could therefore use approaches such as RNA sequencing to characterise the performance of promoters in the genetic context of the engineered circuit or pathway that they are intended to control. This *in situ* approach to characterisation could obviate the need for reporter-protein based

characterisation of candidate promoters, which may not always be fully representative of the performance of promoters when they are used in more industrially relevant genomic contexts (Gorochowski *et al.*, 2017). Alternatively, work could focus on reducing the context-dependency of the *Geobacillus* promoter elements that were identified in this investigation, thereby increasing the range of predictable gene expression available for *Geobacillus* synthetic biology projects.

Mechanisms such as ribozyme or CRISPR mediated RNA processing of the 5' untranslated region of mRNA transcripts could potentially be used to decouple the activity of regulatory elements from their genomic context (Lou *et al.*, 2012, Qi *et al.*, 2012). However, the activity of ribozyme- or CRISPR-based regulatory systems is more likely than not to be uncharacterised in non-model, industrially relevant organisms. From an industrial perspective, consideration should also be given to the effect complicated *cis*-regulatory mechanisms may have on the host organism. Heterologous pathways are known to place a quantifiable burden on host metabolism, potentially resulting in reduced growth and lower yields of the product of interest (Borkowski *et al.*, 2016, Wu *et al.*, 2016, Borkowski *et al.*, 2018). If the expression of heterologous regulatory mechanisms is also required, the resulting increase in metabolic burden on the host could potentially lead to loss-of-function mutations within the heterologous pathways (Sleight & Sauro, 2013), with a corresponding reduction in product yield.

Insulator DNA sequences that physically separate genetic regulatory parts, thereby disrupting context-specific mRNA secondary structures or preventing unintended DNA-protein interactions (Davis *et al.*, 2011, Mutalik *et al.*, 2013a, Carr *et al.*, 2017), may offer a mechanism by which genomic context-dependency could be minimised without increasing burden on the host. However, care should be taken to ensure that such insulators are not themselves context-specific (Carr *et al.*, 2017), and that they do not encode cryptic regulatory sequences (Yao *et al.*, 2013, Zong *et al.*, 2017).

With regards to assessing the effect of environmental context on the performance of cis-regulatory sequences, further work is required. In this investigation, high-throughput screening of candidate *cis*-regulatory sequences was performed using growth in 96-well microplates. Although invaluable at the initial prospecting stage, microplate growth conditions are not necessarily representative of the conditions experienced by cultures in industrial-scale bioreactors (Schmidt, 2005, Moser *et al.*, 2012). Therefore, once bioinformatically identified and initially screened *in vivo*, the performance of candidate promoters should be characterised in as close a facsimile as possible to the environmental conditions in which they will ultimately be used.

Of the seven *Geobacillus* promoters that were shown to function independently of the downstream CDS, five were also shown to function independently of culture aeration (Chapter 5). This result was in direct contrast to the *G. thermodenitrificans ldhA* promoter, which showed statistically significant variation in expression levels depending on culture aeration, congruent with previous studies (Bartosiak-Jentys *et al.*, 2012, Kananavičiūtė & Čitavičius, 2015). Given that oxygen concentration can display spatial variation in large-scale bioreactors (Enfors *et al.*, 2001), constitutive *cis*-regulatory elements that function consistently and predictably under varying oxygen concentrations are potentially valuable.

However, oxygen concentration is only one of many environmental conditions that are known to vary in a spatiotemporal manner in industrial bioreactors (Moser *et al.*, 2012). Furthermore, the fed-batch fermentation process that can be used at larger scales can cause cultures to display different physiological and metabolic states than at laboratory scale (Chubukov *et al.*, 2016). Given that this investigation only examined the performance of *cis*-regulatory elements under three different oxygenation levels (*i.e.* 96-well microplates, baffled and non-baffled 250 ml Erlenmeyer flasks), future work should focus on characterising the performance of the *Geobacillus* promoter library under environmental conditions that are more representative of industrial-scale culture.

If screening of promoter activity in a range of industrially-relevant growth formats was combined with the suggested work on reducing the sensitivity of the 105 characterised *Geobacillus* promoters to any downstream CDS, the resulting library of *cis*-regulatory elements could potentially display consistent, predictable output under varying genetic and environmental contexts, thereby facilitating bottom-up engineering in this industrially relevant, non-model organism.

7.3 Promoter sequence-function modelling

Mathematical models with the *pre hoc* capability to determine promoter activity would reduce the need for *in* vivo characterisation of individual *cis*-regulatory elements. Once a training set of sufficient robustness was established, putative promoters of the desired strength could hypothetically be designed *de novo*. However, this investigation showed that sequence data alone was not sufficient to derive Artificial Neural Network (ANN) or Partial Least Squares (PLS) models with adequate predictive accuracy (Chapters 3 & 4). Despite providing accurate fits of the training data, the ANN and PLS models showed limited predictive power when applied to secondary test sets. Furthermore, synthetic promoter sequences that were derived using the sequence-function models showed no *in vivo* promoter activity. Given that the synthetic promoter sequences contained the same conserved motifs as active promoter sequences in the training data set, the lack of correlation between predicted and empirically measured activity levels was hypothesised to be a result of the inability of the sequence-function models to accurately infer the contribution to promoter strength of given nucleotides at given sequence positions.

Although sequence-function models of sufficient predictive accuracy to determine *pre hoc in vivo* promoter activity could not be derived from the *Geobacillus* promoter data set, the potential for statistical modelling to enhance our fundamental knowledge of genetic regulation in complex systems should not be overlooked. Partition modelling of the relationship between promoter

sequence and the fluorescence activity of the reporter proteins GFP and mOrange yielded potentially useful insights into the structure of *cis*-regulatory elements in *Geobacillus*; regions upstream of the canonical -35, -10 and RBS consensus motifs were predicted to be important for determining promoter activity. Furthermore, a sequence alignment of 21 active promoter sequences revealed conserved regions of AT-rich sequence towards the 5' terminus of the promoter sequence that were not as heavily conserved in sequences which did not show promoter activity. These results suggested that UP-elements similar to those described in *E. coli* (Ross *et al.*, 1993, Aiyar *et al.*, 1998, Estrem *et al.*, 1998) and *B. subtilis* (Meijer & Salas, 2004, Phan *et al.*, 2012) may also play a role in determining transcription activation in *Geobacillus*.

7.4 Potential future application of statistical learning approaches to promoter optimisation

The lack of generality that was shown by the ANN and PLS promoter sequence-function models that were derived in this investigation should not preclude the use of statistical learning approaches in future promoter optimisation studies. Statistical learning approaches to the design and optimisation of regulatory DNA have previously been most successful when the training data sets have been rationally designed. Design of Experiments (DoE) approaches to training set design, typically using full- or fractional-factorial libraries of sequence variants, have been used to characterise the activity and context-dependence of regulatory sequences in *E. coli* (Kosuri *et* al., 2013, Mutalik *et al.*, 2013b), and have been successfully used to optimise translation efficiency, also in *E. coli* (Cambray *et al.*, 2018). A DoE-based approach to promoter optimisation based on the rational design of variant libraries could therefore potentially be applied by future studies in *Geobacillus*.

Given the 100 bp length of the *Geobacillus* promoters that were characterised in this investigation, the number of sequence variants required for rational approaches to promoter optimisation could quickly become prohibitive; a full-factorial library of 100 bp promoters would contain 4^{100} sequences.

Although the proliferation of affordable DNA synthesis has increased the practicality of producing large sequence-variant libraries (Allert *et al.*, 2010), the synthesis and subsequent characterisation of such vast part collections remains a non-trivial concern. Although analysing promoters of shortened length could mitigate this issue, removing regions of promoter sequence could also potentially remove useful promoter activity. Removing the 5'-most 50 bp from the characterised *Geobacillus* promoters, for example, would have maintained the putative RBS, -35 and -10 motifs whilst halving the size of the design space, but would have removed the UP-elements that were shown by Random Forest partition modelling to be key in determining *Geobacillus* promoter activity (Figure 4.24).

In lieu of shortening the characterised promoter sequences to facilitate future DoE-guided sequence optimisation, the novel bioinformatic approach to promoter identification that was developed in this investigation, coupled with Random Forest partition modelling, could provide an initial screen of the promoter design space. The *in vivo* characterisation of a comparatively limited number of putative promoters from all clades of a sequence phylogeny maximised the portion of the endogenous *Geobacillus* promoter design space that was empirically explored, whilst maintaining experimental feasibility. Random forest partition models derived from the resulting characterisation data subsequently allowed key promoter sequence positions or motifs to be identified. Future studies could make use of this increased understanding of the relationship between sequence and function to guide the rational design of sequence libraries with variation only at the identified key positions. These rationally designed variant libraries could subsequently be synthesised, characterised and modelled.

Any future applications of DoE to promoter optimisation in *Geobacillus* would benefit from an increased application of automated liquid handling and laboratory automation to facilitate the construction and characterisation of large-scale sequence variant libraries (Rao, 2015). Although automated liquid handling was employed in this investigation, the scope of this automation was limited to loading *Geobacillus* culture aliquots onto 96-well microplates and

preparing qPCR reactions. In future, part characterisation could be expedited through the use of large-scale automated synthetic biology foundries (Clarke & Kitney, 2016), which use robotic arms to transfer microplates between automated liquid handling platforms, incubators and plate readers (Rao, 2015).

Future studies would also benefit from the development of a highly efficient protocol for the transformation of *Geobacillus*. In this investigation, the low efficiency of conjugal transformation proved to be a significant bottleneck to high-throughput part characterisation. The removal of this bottleneck could therefore expedite characterisation of large, rationally designed sequence libraries. The application of high-throughput characterisation techniques such as the combination of flow cytometry and multiplexed DNA or RNA sequencing (Kosuri *et al.,* 2013, Johns *et al.,* 2018) requires the acquisition of large numbers of transformants, as approximately 50-fold library coverage is required to achieve accurate characterisation of individual promoters (Kosuri *et al.,* 2013). A future study that rationally optimised either a conjugal transformation protocol, or an alternative method to transform *Geobacillus* such as electroporation (Kananavičiūtė & Čitavičius, 2015, Reeve *et al.*, 2016), would therefore be of considerable value.

Finally, the combination of bioinformatics, *in vivo* characterisation and statistical modelling that was implemented in this investigation is theoretically broadly applicable beyond *Geobacillus*. As the approach requires no *a priori* understanding of the structure and function of species-specific promoters, it can potentially be applied to promoter optimisation studies in a broad range of synthetic biology host organisms, opening up synthetic biology solutions in non-model systems.

7.5 Planned and published manuscripts

A manuscript is in preparation that discusses the principle results of this investigation. The novel bioinformatic approach to promoter discovery will be covered, and the results of the *in vivo* characterisation of the bioinformatically-

identified promoters will be outlined. The performance of the characterised promoters will be contrasted to the relative lack of activity that was observed in the mutagenesis-derived sequences. The results of the sequence-function statistical modelling will also be covered, and the general applicability of machine learning approaches to promoter optimisation will be discussed.

In lieu of a single paper, multiple publications discussing individual aspects of the investigation were also considered. For example, two separate manuscripts that covered the bioinformatic approach to promoter identification and the statistical sequence-function modelling approach respectively could have been prepared. However, it was felt that a single manuscript that discussed and directly compared each of the three approaches to promoter design and discovery (i.e. bioinformatic, mutagenic and mathematic) would be more compelling, and potentially more useful in guiding future studies aiming to develop promoter libraries in non-model organisms. Submission of the completed manuscript to Nature Communications is planned.

A review paper that summarised different strategies for promoter discovery was published in Biochemical Society Transactions in 2016 (Gilman & Love, 2016). In particular, the use of mutagenesis-based methods for promoter production was discussed, and the potential of *in silico* approaches to expedite promoter discovery and design was outlined. A copy of the review can be found in the appendix of this thesis.

8 Conclusion

This investigation hypothesised that statistical learning approaches to promoter discovery would be applicable in non-model organisms, and would accelerate the discovery and characterisation of promoter libraries in the industrially relevant chassis, *Geobacillus*.

A novel approach to bioinformatic promoter identification was employed, based upon the analysis of the *Geobacillus* core genome coupled with rational promoter selection based on phylogenetic analysis of putative promoter sequences that maximised the proportion of the promoter design space that was empirically explored. This approach to promoter discovery yielded libraries of endogenous promoter elements that covered two-log ranges of fluorescence, increasing in steady increments. Flow cytometry analysis of *in vivo* promoter activity showed that the characterised library afforded homogeneous expression of recombinant genes across a wide range of expression strengths.

Seven of the characterised promoters were also shown to function consistently regardless of the downstream coding sequence. Furthermore, five sequences were shown to function independently of culture oxygenation. Prior to this investigation, the constitutive promoters that were available for use in *Geobacillus* were limited to three endogenous regulatory elements (one of which, the *ldh* promoter, had been shown to be oxygen dependant) and three libraries of mutagenesis-derived synthetic promoters, all of which were only characterised in single genetic and environmental contexts. The promoters that were characterised in this investigation therefore represented an expansion of the synthetic biology toolkit in an industrially relevant host.

This investigation also demonstrated the utility of statistical modelling approaches to the design of synthetic promoter libraries in non-model organisms. Artificial Neural Network and Partial Least Squares models showed inadequate generality, possibly as a result of a lack of sequence homology in training data sets that were relatively small compared to the dimensionality of

the design space. The promoter DNA sequence alone was therefore not sufficient to derive statistical sequence-function models of adequate predictive accuracy to determine *pre hoc in vivo* promoter activity. However, the application of Random Forest partition models was shown to be a useful approach for increasing understanding of transcription regulation in non-model organisms. In particular, UP-elements, which have previously been reported in *B. subtilis* and *E. coli*, were suggested to play an important role in determining *Geobacillus* promoter activity. The increased understanding of the *Geobacillus* promoter design space that resulted from these models could potentially be used by future investigations to facilitate a rational, Design of Experiments approach to promoter optimisation.

In addition, the use of two *a posteriori,* mutagenesis-based approaches to promoter design, Saturation Mutagenesis of Flanking Regions and error-prone PCR, highlighted the inefficiency of random mutation as a technique for deriving promoter libraries, especially in organisms for which *a priori* understanding of the structure of *cis*-regulatory elements is limited. The propensity for mutagenesis to reduce, rather than increase, promoter activity that was shown in this investigation and corroborated by a review of the literature inherently limits the capability of mutagenesis to yield libraries of promoter elements that collectively cover broad ranges of recombinant gene expression levels.

To conclude, this investigation has shown that bioinformatic prospecting of *cis*-regulatory elements, followed by deep empirical characterisation of *in vivo* activity, represents an efficient, practical method by which the promoter toolkit can be rapidly expanded in non-model, industrially relevant organisms. The combination of bioinformatics, *in vivo* characterisation and statistical modelling that was implemented in this investigation is theoretically broadly applicable. As the approach requires no *a priori* understanding of the structure and function of species-specific promoters, it can potentially be applied to promoter optimisation studies in a broad range of synthetic biology host organisms, opening up synthetic biology solutions in non-model systems.

Bibliography

Adams, B.L. (2016) The Next Generation of Synthetic Biology Chassis: Moving Synthetic Biology from the Laboratory to the Field. *ACS Synthetic Biology*, **5**; 1328–1330.

Adhya, S., Gottesman, M., Garges, S. & Oppenheim, A. (1993) Promoter resurrection by activators - a minireview. *Gene*, **132**; 1–6.

Aiyar, S.E., Gourse, R.L. & Ross, W. (1998) Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase. *Proceedings of the National Academy of Sciences*, **95**; 14652–14657.

Ajikumar, P.K., Xiao, W.H., Tyo, K.E.J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T.H., Pfeifer, B. & Stephanopoulos, G. (2010) Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in *Escherichia coli*. *Science*, **330**; 70–74.

Allert, M., Cox, J.C. & Hellinga, H.W. (2010) Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames. *Journal of Molecular Biology*, **402**; 905–918.

Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences*, **102**; 12678–12683.

Anderson, J.C., Clarke, E.J., Arkin, A.P. & Voigt, C.A. (2006) Environmentally Controlled Invasion of Cancer Cells by Engineered Bacteria. *Journal of Molecular Biology*, **355**; 619–627.

Anderson, J.C., Voigt, C.A. & Arkin, A.P. (2007) Environmental signal integration by a modular AND gate. *Molecular Systems Biology*, **3**; 133.

Ang, J., Harris, E., Hussey, B.J., Kil, R. & McMillen, D.R. (2013) Tuning Response Curves for Synthetic Biology. *ACS Synthetic Biology*, **2**; 547–567.

Angermayr, M., Oechsner, U. & Bandlow, W. (2003) Reb1p-dependent DNA Bending Effects Nucleosome Positioning and Constitutive Transcription at the Yeast Profilin Promoter. *Journal of Biological Chemistry*, **278**; 17918–17926.

Angov, E. (2011) Codon usage: Nature's roadmap to expression and folding of proteins. *Biotechnology Journal*, **6**; 650–659.

Aro, E.M. (2015) From first generation biofuels to advanced solar biofuels. *Ambio*, **45**; 24–31.

Assareh, R., Zahiri, H.S., Noghabi, K.A., Aminzadeh, S. & Khaniki, G.B. (2012) Characterization of the newly isolated *Geobacillus* sp. T1, the efficient cellulase-producer on untreated barley and wheat straws. *Bioresource Technology*, **120**; 99–105.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. & Noble, W.S. (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, **37**; W202–W208.

Baltagi, Y. & Kussener, F. (2014) *Advantages of Bootstrap Forest for Yield Analysis*. SAS Institute Inc, North Carolina.

Bartosiak-Jentys, J., Eley, K. & Leak, D.J. (2012) Application of *pheB* as a Reporter Gene for *Geobacillus* spp., Enabling Qualitative Colony Screening and Quantitative Analysis of Promoter Strength. *Applied and Environmental Microbiology*, **78**; 5945–5947.

Bartosiak-Jentys, J., Hussein, A.H., Lewis, C.J. & Leak, D.J. (2013) Modular system for assessment of glycosyl hydrolase secretion in *Geobacillus thermoglucosidasius*. *Microbiology*, **159**; 1267–1275.

Basehoar, A.D., Zanton, S.J. & Pugh, B.F. (2004) Identification and Distinct Regulation of Yeast TATA Box-Containing Genes. *Cell*, **116**; 699–709.

Bashor, C.J. & Collins, J.J. (2012) Insulating gene circuits from context by RNA processing. *Nature Biotechnology*, **30**; 1061–1062.

Bataineh, M. & Marler, T. (2017) Neural network for regression problems with reduced training sets. *Neural Networks*, **95**; 1–9.

Beal, J., Weiss, R., Yaman, F., Davidsohn, N. & Adler, A. (2012) *A Method for Fast, High-Precision Characterization of Synthetic Biology Devices*. MIT CSAIL Tech Report 2012-008.

Bedbrook, C.N., Yang, K.K., Rice, A.J., Gradinaru, V. & Arnold, F.H. (2017) Machine learning to design integral membrane channel rhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Computational Biology*, **13**; e1005786.

Beleites, C. & Salzer, R. (2008) Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry*, **390**; 1261–1271.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & thgen, N.B.U. (2013) Efficient translation initiation dictates codon usage at gene start. *Molecular Systems Biology*, **9**; DOI: 10.1038/msb.2013.32.

Bezuidt, O.K., Pierneef, R., Gomri, A.M., Adesioye, F., Makhalanyane, T.P., Kharroub, K. & Cowan, D.A. (2015) Genomic analysis of six new *Geobacillus* strains reveals highly conserved carbohydrate degradation architectures and strategies. *Frontiers in Microbiology*, **6**; 41993.

Bhalla, A., Bischoff, K.M., Uppugundla, N., Balan, V. & Sani, R.K. (2014) Novel thermostable endo-xylanase cloned and expressed from bacterium *Geobacillus* sp. WSUCF1. *Bioresource Technology*, **165**; 314–318.

Blanchard, K., Robic, S. & Matsumura, I. (2014) Transformable facultative thermophile *Geobacillus stearothermophilus* NUB3621 as a host strain for metabolic engineering. *Applied Microbiology and Biotechnology*, **98**; 6715–6723.

Blazeck, J. & Alper, H.S. (2013) Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnology Journal*, **8**; 46–58.

Blazeck, J., Garg, R., Reed, B. & Alper, H.S. (2012) Controlling Promoter Strength and Regulation in *Saccharomyces cerevisiae* Using Synthetic Hybrid Promoters. *Biotechnology and Bioengineering*, **109**; 2884–2895.

Block, D.H.S., Hussein, R., Liang, L.W. & Lim, H.N. (2012) Regulatory consequences of gene translocation in bacteria. *Nucleic Acids Research*, **40**; 8979–8992.

Bloom, J.D., Meyer, M.M., Meinhold, P., Otey, C.R., MacMillan, D. & Arnold, F.H. (2005) Evolving strategies for enzyme engineering. *Current Opinion in Structural Biology*, **15**; 447–452.

Blount, B.A., Weenink, T., Vasylechko, S. & Ellis, T. (2012) Rational Diversification of a Promoter Providing Fine-Tuned Expression and Orthogonal Regulation for Synthetic Biology. *PLoS ONE*, **7**; e33279.

Bokinsky, G., Peralta-Yahya, P.P., George, A., Holmes, B.M., Steen, E.J., Dietrich, J., Soon Lee, T., Tullman-Ercek, D., Voigt, C.A., Simmons, B.A. & Keasling, J.D. (2011) Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered *Escherichia coli*. *Proceedings of the National Academy of Sciences*, **108**; 19949–19954.

Borkowski, O., Bricio, C., Murgiano, M., Rothschild-Mancinelli, B., Stan, G.B. & Ellis, T. (2018) Cell-free prediction of protein expression costs for growing cells. *Nature Communications*, **9**; 1457.

Borkowski, O., Ceroni, F., Stan, G.B. & Ellis, T. (2016) Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. *Current Opinion in Microbiology*, **33**; 123–130.

Boulesteix, A.L. & Strimmer, K. (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**; 32–44.

Box, G.E.P. & Draper, N.R. (1986) *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, Inc., New York.

Boyle, P.M. & Silver, P.A. (2012) Parts plus pipes: Synthetic biology approaches to metabolic engineering. *Metabolic Engineering*, **14**; 223–232.

Braatsch, S., Helmark, S., Kranz, H., Koebmann, B. & Ruhdal Jensen, P. (2008) *Escherichia coli* strains with promoter libraries constructed by Red/ET recombination pave the way for transcriptional fine-tuning. *BioTechniques*, **45**; 335–337.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**; 5–32.

Brophy, J.A.N. & Voigt, C.A. (2014) Principles of genetic circuit design. *Nature Methods*, **11**; 508–520.

Brown, J.A., Barne, K.A., Minchin, S.D. & Busby, S.J.W. (1997) Extended -10 promoters. In *Nucleic Acids and Molecular Biology. Mechanisms of Transcription* (ed. by Eckstein, F. & Lilley, D.M.J.). Springer, New York, pp. 41–52.

Brown, S., Loh, J., Aves, S.J. & Howard, T.P. (2018) Alkane Biosynthesis in Bacteria. In *Biogenesis of Hydrocarbons* (ed. by Stams, A.J.M. & Sousa, D.). Springer International Publishing, Cham.

Brown, S.R., Staff, M., Lee, R., Love, J., Parker, D.A., Aves, S.J. & Howard, T.P. (2018b) Design of Experiments Methodology to Build a Multifactorial Statistical Model Describing the Metabolic Interactions of Alcohol Dehydrogenase Isozymes in the Ethanol Biosynthetic Pathway of the Yeast *Saccharomyces cerevisiae*. *ACS Synthetic Biology*, **7**; 1676–1684.

Browning, D.F. & Busby, S.J.W. (2004) The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, **2**; 57–65.

Browning, D.F. & Busby, S.J.W. (2016) Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, **14**; 638–650.

Bryant, J.A. & Hughes, S. (2017) Biofuels and Bioenergy - Ethical Aspects. In *Biofuels and Bioenergy* (ed. by Love, J. & Bryant, J.A.). John Wiley & Sons, Ltd.

Buscema, P.M., Massini, G. & Maurelli, G. (2014) Artificial Neural Networks: An Overview and their Use in the Analysis of the AMPHORA-3 Dataset. *Substance Use & Misuse*, **49**; 1555–1568.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J. & Munafò, M.R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**; 365–376.

Cagnon, C., Valverde, V. & Masson, J.M. (1991) A new family of sugar-inducible expression vectors for *Escherichia coli*. *Protein Engineering*, **4**; 843–847.

Cambray, G., Guimaraes, J.C. & Arkin, A.P. (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nature Biotechnology*. DOI: 10.1038/nbt.4238.

Campbell, E.A., Muzzin, O., Chlenov, M., Sun, J.L., Olson, C.A., Weinman, O., Trester-Zedlitz, M.L. & Darst, S.A. (2002) Structure of the Bacterial RNA Polymerase Promoter Specificity. *Molecular Cell*, **9**; 527–539.

Canton, B., Labno, A. & Endy, D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nature Biotechnology*, **26**; 787–793.

Cardinale, S. & Arkin, A.P. (2012) Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems. *Biotechnology Journal*, **7**; 856–866.

Carlson, R. (2010) *Biology is technology: the promise, peril, and new business of engineering life*. Harvard University Press, Cambridge, MA.

Carr, S.B., Beal, J. & Densmore, D.M. (2017) Reducing DNA context dependence in bacterial promoters. *PLoS ONE*, **12**; e0176013.

Carriquiry, M.A., Du, X. & Timilsina, G.R. (2011) Second generation biofuels: Economics and policies. *Energy Policy*, **39**; 4222–4234.

Caschera, F., Gazzola, G., Bedau, M.A., Bosch Moreno, C., Buchanan, A., Cawse, J., Packard, N. & Hanczyc, M.M. (2010) Automated Discovery of Novel Drug Formulations Using Predictive Iterated High Throughput Experimentation. *PLoS ONE*, **5**; e8546.

Chandler, M.G. & Pritchard, R.H. (1975) The Effect of Gene Concentration and Relative Gene Dosage on Gene Output in *Escherichia coli*. *Molecular Genetics and Genomics*, **138**; 127–141.

Chavez, M., Ho, J. & Tan, C. (2016) Reproducibility of high-throughput plate-reader experiments in synthetic biology. *ACS Synthetic Biology*, **6**; 375-380.

Chen, J., Zhang, Z., Zhang, C. & Yu, B. (2015) Genome sequence of *Geobacillus thermoglucosidasius* DSM2542, a platform hosts for biotechnological applications with industrial potential. *Journal of Biotechnology*, **216**; 98–99.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuz'min, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A. & Tropsha, A. (2014) QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, **57**; 4977–5010.

Chubukov, V., Mukhopadhyay, A., Petzold, C.J., Keasling, J.D. & Martín, H.G. (2016) Synthetic and systems biology for microbial production of commodity chemicals. *npj Systems Biology and Applications*, **2**; 16009.

Clarke, B., Fokoue, E. & Zhang, H.H. (2009) *Principles and Theory for Data Mining and Machine Learning*. 1st edn. Springer-Verlag, New York.

Clarke, L.J. & Kitney, R.I. (2016) Synthetic and Systems Biotechnology. *Synthetic and Systems Biotechnology*, **1**; 243–257.

Clyde, M. (2002) Model Averaging. In *Subjective and Objective Bayesian Statistics* (ed. by Press, S.J.). John Wiley & Sons, Inc.

Collado-Vides, J., Magasanik, B. & Gralla, J.D. (1991) Control Site Location and Transcriptional Regulation in *Escherichia coli*. *Microbiological Reviews*, **55**; 371–394.

Contreras-Moreira, B. & Vinuesa, P. (2013) GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and Environmental Microbiology*, **79**; 7696–7701.

Contreras-Moreira, B. & Vinuesa, P. (2015) *GET_HOMOLOGUES Manual*. Estación Experimental de Aula Dei/CSIC, Fundación ARAID & Universidad Nacional Autónoma de México.

Cook, C., Dayananda, C., Tennant, R.K. & Love, J. (2017) Third-Generation Biofuels from the Microalga, *Botryococcus braunii*. In *Biofuels and Bioenergy* (ed. by Love, J. & Bryant, J.A.). John Wiley & Sons, Ltd.

Cook, V.M. & deHaseth, P.L. (2007) Strand Opening-deficient *Escherichia coli* RNA Polymerase Facilitates Investigation of Closed Complexes with Promoter DNA. *Journal of Biological Chemistry*, **282**; 21319–21326.

Cox, I. & Gaudard, M. (2013) *Discovering Partial Least Squares with JMP*. SAS Institute Inc., North Carolina.

Criminisi, A., Shotton, J. & Konukoglu, E. (2012) Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Foundations and Trends in Computer Graphics and Vision*, **7**; 81–227.

Cripps, R.E., Eley, K., Leak, D.J., Rudd, B., Taylor, M., Todd, M., Boakes, S., Martin, S. & Atkinson, T. (2009) Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. *Metabolic Engineering*, **11**; 398–408.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R. & Mockler, T. (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**; e122.

Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. (2004) WebLogo: A Sequence Logo Generator. *Genome Research*, **14**; 1188–1190.

Curran, K.A., Karim, A.S., Gupta, A. & Alper, H.S. (2013) Use of expression-enhancing terminators in *Saccharomyces cerevisiae* to increase mRNA half-life and improve gene expression control for metabolic engineering applications. *Metabolic Engineering*, **19**; 88–97.

Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., Xie, Z. & Weiss, R. (2014) Accurate Predictions of Genetic Circuit Behavior from Part Characterization and Modular Composition. *ACS Synthetic Biology*, **4**; 673–681.

Davis, J.H., Rubin, A.J. & Sauer, R.T. (2011) Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Research*, **39**; 1131–1141.

de Boer, H., Comstock, L.J. & Vasser, M. (1983) The *tac* promoter: A functional hybrid derived from the *trp* and *lac* promoters. *Proceedings of the National Academy of Sciences*, **80**; 21–25.

de Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**; 251–263.

De Mey, M., Maertens, J., Boogmans, S., Soetaert, W.K., Vandamme, E.J., Cunin, R. & Foulquié-Moreno, M.R. (2010) Promoter knock-in: a novel rational method for the fine tuning of genes. *BMC Biotechnology*, **10**; 26.

De Mey, M., Maertens, J., Lequeux, G.J., Soetaert, W.K. & Vandamme, E.J. (2007) Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering. *BMC Biotechnology*, **7**; 34.

DeLorenzo, D.M., Rottinghaus, A.G., Henson, W.R. & Moon, T.S. (2018) Molecular Toolkit for Gene Expression Control and Genome Modification in *Rhodococcus opacus* PD630. *ACS Synthetic Biology*, **7**; 727–738.

Del Vecchio, D. (2015) Modularity, context-dependence, and insulation in engineered biological circuits. *Trends in Biotechnology*, **33**; 111–119.

Demirbas, A. (2009) Political, economic and environmental impacts of biofuels: A review. *Applied Energy*, **86**; S108–S117.

Deng, B.C., Yun, Y.H., Liang, Y.Z., Cao, D.S., Xu, Q.S., Yi, L.Z. & Huang, X. (2015) A new strategy to prevent over-fitting in partial least squares models based on model population analysis. *Analytica Chimica Acta*, **880**; 32–41.

Deng, C., Li, J., Shin, H.D., Du, G., Chen, J. & Liu, L. (2018) Efficient expression of cyclodextrin glycosyltransferase from *Geobacillus stearothermophilus* in *Escherichia coli* by promoter engineering and downstream box evolution. *Journal of Biotechnology*, **266**; 77–83.

Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J. & Keasling, J.D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nature Biotechnology*, **27**; 753–759.

Eck, D.J. (2018) Bootstrapping for multivariate linear regression models. *Statistics and Probability Letters*, **134**; 141–149.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**; 1792–1797.

Ellis, T., Wang, X. & Collins, J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotechnology*, **27**; 465–471.

Elmore, J.R., Furches, A., Wolff, G.N., Gorday, K. & Guss, A.M. (2017) Development of a high efficiency integration system and promoter library for rapid modification of *Pseudomonas putida* KT2440. *Metabolic Engineering Communications*, **5**; 1–8.

Elowitz, M.B. & Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**; 335–338.

Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. (2002) Stochastic Gene Expression in a Single Cell. *Science*, **297**; 1183–1186.

Elvin, C.M., Thompson, P.R., Argall, M.E., Hendry, P., Stamford, N.P.J., Lilley, P.E. & Dixon, N.E. (1990) Modified bacteriophage lambda promoter vectors for overproduction of proteins in *Escherichia coli*. *Gene*, **87**; 123–126.

Endy, D. (2005) Foundations for engineering biology. *Nature*, **438**; 449–453.

Enfors, S.O., Jahic, M., Rozkov, A., Xu, B., Hecker, M., Jürgen, B., Krüger, E., Schweder, T., Hamer, G., O'Beirne, D., Noisommit-Rizzi, N., Reuss, M., Boone, L., Hewitt, C., McFarlane, C., Nienow, A., Kovacs, T., Trägårdh, C., Fuchs, L., Revstedt, J., Friberg, P.C., Hjertager, B., Blomsten, G., Skogman, H., Hjort, S., Hoeks, F., Lin, H.Y., Neubauer, P., van der Lans, R., Luyben, K., Vrabel, P. & Manelius, Å. (2001) Physiological responses to mixing in large scale bioreactors. *Journal of Biotechnology*, **85**; 175–185.

Engler, C., Kandzia, R. & Marillonnet, S. (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE*, **3**; e3647.

Eriksson, L., Andersson, P.L., Johansson, E. & Tysklind, M. (2006) Megavariate analysis of environmental QSAR data. Part I – A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Molecular Diversity*, **10**; 169–186.

Estrem, S.T., Gaal, T., Ross, W. & Gourse, R.L. (1998) Identification of an UP element consensus sequence for bacterial promoters. *Proceedings of the National Academy of Sciences*, **95**; 9761–9766.

Falcón, R.M. (2015) Enumeration and classification of self-orthogonal partial Latin rectangles by using the polynomial method. *European Journal of Combinatorics*, 1–9.

Farrés, M., Platikanov, S., Tsakovski, S. & Tauler, R. (2015) Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, **29**; 528–536.

Fischer, C.R., Alper, H., Nevoigt, E., Jensen, K.L. & Stephanopoulos, G. (2006) Response to Hammer *et al*.: Tuning genetic control - importance of thorough promoter characterization versus generating promoter diversity. *Trends in Biotechnology*, **24**; 55–56.

Forman, G. & Scholz, M. (2010) Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *ACM SIGKDD Explorations Newsletter*, **12**; 49-57.

Galdzicki, M., Clancy, K.P., Oberortner, E., Pocock, M., Quinn, J.Y., Rodriguez, C.A., Roehner, N., Wilson, M.L., Adam, L., Anderson, J.C., Bartley, B.A., Beal, J., Chandran, D., Chen, J., Densmore, D., Endy, D., nberg, R.G.U., Hallinan, J., Hillson, N.J., Johnson, J.D., Kuchinsky, A., Lux, M., Misirli, G., Peccoud, J., Plahar, H.A., Sirin, E., Stan, G.-B., Villalobos, A., Wipat, A., Gennari, J.H., Myers, C.J. & Sauro, H.M. (2014) The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology*, **32**; 545–550.

Gardner, T.S., Cantor, C.R. & Collins, J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**; 339–342.

Gilman, J. & Love, J. (2016) Synthetic promoter design for new microbial chassis. *Biochemical Society Transactions*, **44**; 731–737.

Gasser, B., Steiger, M.G. & Mattanovich, D. (2015) Methanol regulated yeast promoters: production vehicles and toolbox for synthetic biology. *Microbial Cell Factories*, **14**; 196.

Golbraikh, A. & Tropsha, A. (2002) Beware of q^2! *Journal of Molecular Graphics and Modelling*, **20**; 269–276.

Goldbeck, C.P., Jensen, H.M., TerAvest, M.A., Beedle, N., Appling, Y., Hepler, M., Cambray, G., Mutalik, V., Angenent, L.T. & Ajo-Franklin, C.M. (2012) Tuning Promoter Strengths for Improved Synthesis and Function of Electron Conduits in *Escherichia coli*. *ACS Synthetic Biology*, **2**; 150–159.

Gorochowski, T.E., Espah Borujeni, A., Park, Y., Nielsen, A.A., Zhang, J., Der, B.S., Gordon, D.B. & Voigt, C.A. (2017) Genetic circuit characterization and debugging using RNA-seq. *Molecular Systems Biology*, **13**; 952.

Gotwalt, C.M. (2011) *JMP 9 Neural Platform Numerics*. SAS Institute Inc., North Carolina.

Gouy, M., Guindon, S. & Gascuel, O. (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, **27**; 221–224.

Gowen, A.A., Downey, G., Esquerre, C. & O'Donnell, C.P. (2010) Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics*, **25**; 375–381.

Granitto, P.M., Verdes, P.F. & Ceccatto, H.A. (2005) Neural network ensembles: evaluation of aggregation algorithms. *Artificial Intelligence*, **163**; 139–162.
Green, A.A., Silver, P.A., Collins, J.J. & Yin, P. (2014) Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell*, **159**; 925–939.

Gries, T.J., Kontur, W.S., Capp, M.W., Saecker, R.M. & Record, M.T., Jr. (2010) One-step DNA melting in the RNA polymerase cleft opens the initiation bubble to form an unstable open complex. *Proceedings of the National Academy of Sciences*, **107**; 10418–10423.

Gruber, T.M. & Gross, C.A. (2003) Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. *Annual Review of Microbiology*, **57**; 441–466.

Gualerzi, C.O. & Pon, C.L. (1990) Initiation of mRNA translation in prokaryotes. *Biochemistry*, **29**; 5881–5889.

Guido, N.J., Wang, X., Adalsteinsson, D., McMillen, D., Hasty, J., Cantor, C.R., Elston, T.C. & Collins, J.J. (2006) A bottom-up approach to gene regulation. *Nature*, **439**; 856–860.

Gupta, A. & Rao, G. (2003) A study of oxygen transfer in shake flasks using a non-invasive oxygen sensor. *Biotechnology and Bioengineering*, **84**; 351–358.

Guzman, L.-M., Belin, D., Carson, M.J. & Beckwith, J. (1995) Tight Regulation, Modulation and High-Level Expression by Vectors containing the Arabinose P_{BAD} Promoter. *Journal of Bacteriology*, **177**; 4121–4130.

Haahr, M. & Haahr, S. (1998) *RANDOM.ORG* [Online]. Available: http://www.random.org [accessed on 30 March 2018].

Hahn, S. & Young, E.T. (2011) Transcriptional Regulation in *Saccharomyces cerevisiae*: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators. *Genetics*, **189**; 705–736.

Hammer, K., Mijakovic, I. & Jensen, P.R. (2006) Synthetic promoter libraries – tuning of gene expression. *Trends in Biotechnology*, **24**; 53–55.

Hanahan, D. (1985) Techniques for transformation of *E. coli.* In *DNA Cloning: A Practical Approach* (ed. by Glover, D.M.). IRL Press, Oxford, pp. 109-135.

Hansen, L.K. & Salamon, P. (1990) Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**; 993–1001.

Hampsey, M. (1998) Molecular Genetics of the RNA Polymerase II General Transcriptional Machinery. *Microbiology and Molecular Biology Reviews*, **62**; 465–503.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd edn. Springer, New York.

Hess, J.R., Wright, C.T. & Kenney, K.L. (2007) Cellulosic biomass feedstocks and logistics for ethanol production. *Biofuels, Bioproducts and Biorefining*, **1**; 181–190.

Heyduk, E. & Heyduk, T. (2014) Next Generation Sequencing-Based Parallel Analysis of Melting Kinetics of 4096 Variants of a Bacterial Promoter. *Biochemistry*, **53**; 282–292.

Ho, T.K. (1995) Random Decision Forests. *Presented at the International Conference on Document Analysis and Recognition, Montreal, Que*., pp. 278–282.

Hoekman, S.K., Broch, A., Robbins, C., Ceniceros, E. & Natarajan, M. (2012) Review of biodiesel composition, properties and specifications. *Renewable and Sustainable Energy Reviews*, **16**; 143–169.

Hornik, K. (1989) Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, **2**; 359–366.

Howard, T.P. (2017) Engineering Microbial Metabolism for Biofuel Production. In *Biofuels and Bioenergy* (ed. by Love, J. & Bryant, J.A.). John Wiley & Sons, Ltd.

Howard, T.P., Middelhaufe, S., Moore, K., Edner, C., Kolak, D.M., Taylor, G.N., Parker, D.A., Lee, R., Smirnoff, N., Aves, S.J. & Love, J. (2013) Synthesis of customised petroleum-replica fuel molecules by targeted modification of free fatty acid pools in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, **110**; 7636–7641.

Huang, H.H., Camsund, D., Lindblad, P. & Heidorn, T. (2010) Design and characterization of molecular tools for a Synthetic Biology approach towards developing cyanobacterial biotechnology. *Nucleic Acids Research*, **38**; 2577–2593.

Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. (2000) Computational identification of Cis-regulatory elements associwated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, **296**; 1205–1214.

Hussein, A., Lisowska, B. & Leak, D.J. (2015) The Genus *Geobacillus* and Their Biotechnological Potential. In *Advances in Applied Microbiology* (ed. by Sariaslani, S. & Gadd, G.M.). Academic Press, pp. 1–48.

iGEM (2018) *Promoters/Catalog/Anderson* [Online]. Available: http://parts.igem.org/Promoters/Catalog/Anderson [accessed on 26 September 2018].

Integrated DNA Technologies. (2016) *Calculations: Converting from nanograms to copy number* [Online] Available: http://eu.idtdna.com/pages/decoded/decoded-articles/pipet-tips/decoded/2013/10/21/calculations-converting-from-nanograms-to-copy-number [accessed on 30 March 2018].

International Energy Agency. (2017) *World Energy Statistics*. OECD Publishing, Paris.

Jahn, M., Vorpahl, C., Türkowsky, D., Lindmeyer, M., Bühler, B., Harms, H. & Müller, S. (2014) Accurate Determination of Plasmid Copy Number of Flow-Sorted Cells using Droplet Digital PCR. *Analytical Chemistry*, **86**; 5969–5976.

Jensen, K., Alper, H., Fischer, C. & Stephanopoulos, G. (2006) Identifying Functionally Important Mutations from Phenotypically Diverse Sequence Data. *Applied and Environmental Microbiology*, **72**; 3696–3701.

Jensen, P.R. & Hammer, K. (1998a) The Sequence of Spacers between the Consensus Sequences Modulates the Strength of Prokaryotic Promoters. *Applied and Environmental Microbiology*, **64**; 82–87.

Jensen, P.R. & Hammer, K. (1998b) Artificial Promoters for Metabolic Optimization. *Biotechnology and Bioengineering*, **58**; 191–195.

Jensen, P.R., Westerhoff, H.V. & Michelsen, O. (1993) The use of *lac*-type promoters in control analysis. *European Journal of Biochemistry*, **211**; 181–191.

Jensen, T.Ø., Pogrebnyakov, I., Falkenberg, K.B., Redl, S. & Nielsen, A.T. (2017) Application of the thermostable β-galactosidase, *BgaB* from *Geobacillus stearothermophilus* as a versatile reporter under anaerobic and aerobic conditions. *AMB Express*, **7**; 169.

Johansson, U., Löfström, T. & Niklasson, L. (2007) The Importance of Diversity in Neural Network Ensembles - An Empirical Investigation. *Presented at the International Joint Conference on Neural Networks*, pp. 661–666.

Johns, N.I., Gomes, A.L.C., Yim, S.S., Yang, A., Blazejewski, T., Smillie, C.S., Smith, M.B., Alm, E.J., Kosuri, S. & Wang, H.H. (2018) Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nature Methods*. DOI: 10.1038/nmeth.4633.

Jones, J.A., Toparlak, Ö.D. & Koffas, M.A. (2015) Metabolic pathway balancing and its role in the production of biofuels and chemicals. *Current Opinion in Biotechnology*, **33**; 52–59.

Jones, K.L., Kim, S.-W. & Keasling, J.D. (2000) Low-Copy Plasmids can Perform as Well as or Better Than High-Copy Plasmids for Metabolic Engineering of Bacteria. *Metabolic Engineering*, **2**; 328–338.

Jonsson, J., Eriksson, L., Helberg, S., Lindgren, F., Sjöstrom, M. & Wold, S. (1991) A Multivariate Representation and Analysis of DNA Sequence Data. *Acta Chemica Scandinavica*, **45**; 186–192.

Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C. & Wold, S. (1993) Quantitative sequence-activity models (QSAM)-tools for sequence design. *Nucleic Acids Research*, **12**; 733–739.

Jung, Y. & Hu, J. (2015) A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, **27**; 167–179.

Kananavičiūtė, R. & Čitavičius, D. (2015) Genetic engineering of *Geobacillus* spp. *Journal of Microbiological Methods*, **111**; 31–39.

Kanhere, A. & Bansal, M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research*, **33**; 3165–3175.

Kanno, A.I., Goulart, C., Rofatto, H.K., Oliveira, S.C., Leite, L.C.C. & McFadden, J. (2016) New Recombinant *Mycobacterium bovis* BCG Expression Vectors: Improving Genetic Control over Mycobacterial Promoters. *Applied and Environmental Microbiology*, **82**; 2240–2246.

Keasling, J.D. (1999) Gene-expression tools for the metabolic engineering of bacteria. *Trends in Biotechnology*, **17**; 452–460.

Keren, L., Zackay, O., Lotan-Pompan, M., Barenholz, U., Dekel, E., Sasson, V., Aidelberg, G., Bren, A., Zeevi, D., Weinberger, A., Alon, U., Milo, R. & Segal, E. (2013) Promoters maintain their relative activity levels under different growth conditions. *Molecular Systems Biology*, **9**; 701.

Khlebnikov, A., Datsenko, K.A., Skaug, T., Wanner, B.L. & Keasling, J.D. (2001) Homogeneous expression of the P_{BAD} promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology*, **147**; 3241–3247.

Kibria, A., Akhundjanov, S.B. & Oladi, R. (2018) Fossil fuel share in the energy mix and economic growth. *International Review of Economics and Finance*. DOI: 10.1016/j.iref.2018.09.002.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**; R36.

Kirchmaier, S., Lust, K. & Wittbrodt, J. (2013) Golden GATEway Cloning - A Combinatorial Approach to Generate Fusion and Recombination Constructs. *PLoS ONE*, **8**; e76117.

Kiryu, H., Oshima, T. & Asai, K. (2005) Extracting relations between promoter sequences and their strengths from microarray data. *Bioinformatics*, **21**; 1062–1068.

Kosuri, S., Goodman, D., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Edny, D. & Church, G.M. (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, **110**; 140242–14029.

Kudla, G., Murray, A.W., Tollervey, D. & Plotkin, J.B. (2009) Coding-sequence Determinants of Gene Expression in *Escherichia coli*. *Science*, **324**; 255–258.

Kumar, V., Bhalla, A. & Rathore, A.S. (2013) Design of experiments applications in bioprocessing: Concepts and approach. *Biotechnology Progress*, **30**; 86–99.

Langmead, B. & Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**; 357–359.

Laudani, A., Lozito, G.M., Riganti Fulginei, F. & Salvini, A. (2015) On Training Efficiency and Computational Costs of a Feed Forward Neural Network: A review. *Computational Intelligence and Neuroscience*, **818243**; DOI: 10.1155/2015/818243.

Laursen, B.S., Sorensen, H.P., Mortensen, K.K. & Sperling-Petersen, H.U. (2005) Initiation of Protein Synthesis in Bacteria. *Microbiology and Molecular Biology Reviews*, **69**; 101–123.

Le Boedec, K. (2016) Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size: a computer-simulation study. *Veterinary Clinical Pathology*, **45**; 648–656.

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

Lee, C.L., Ow, D.S.W. & Oh, S.K.W. (2006b) Quantitative real-time polymerase chain reaction for determination of plasmid copy number in bacteria. *Journal of Microbiological Methods*, **65**; 258–267.

Lee, S.Y. & Kim, H.U. (2015) Systems strategies for developing industrial microbial strains. *Nature Biotechnology*, **33**; 1061–1072.

Lendrem, D.W., Lendrem, B.C., Rowland-Jones, R., D'Agostino, F., Linsley, M., Owen, M.R. & Isaacs, J.D. (2015a) Teaching examples for the design of experiments: geographical sensitivity and the self-fulfilling prophecy. *Pharmaceutical Statistics*, **15**; 90-92.

Lendrem, D.W., Lendrem, B.C., Woods, D., Rowland-Jones, R., Burke, M., Chatfield, M., Isaacs, J.D. & Owen, M.R. (2015b) Lost in space: design of experiments and scientific exploration in a Hogarth Universe. *Drug Discovery Today*, **20**; 1365–1371.

Leong, W.H., Lim, J.W., Lam, M.K., Uemura, Y. & Ho, Y.C. (2018) Third generation biofuels: A nutritional perspective in enhancing microbial lipid production. *Renewable and Sustainable Energy Reviews*, **91**; 950–961.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**; 2078–2079.

Li, J. & Zhang, Y. (2014) Relationship between promoter sequence and its strength in gene expression. *The European Physical Journal E*, **37**; 86.

Li, S., Wang, J., Li, X., Yin, S., Wang, W. & Yang, K. (2015) Genome-wide identification and evaluation of constitutive promoters in streptomycetes. *Microbial Cell Factories*, **14**; 172.

Liang, G. & Li, Z. (2007) Scores of generalized base properties for quantitative sequence-activity modeling for *E. coli* promoters based on support vector machine. *Journal of Molecular Graphics and Modelling*, **26**; 269–281.

Liang, Y., Woodle, S.A., Shibeko, A.M., Lee, T.K. & Ovanesov, M.V. (2013) Correction of microplate location effects improves performance of the thrombin generation test. *Thrombosis Journal*, **11**; 12.

Liao, H., McKenzie, T. & Hageman, R. (1986) Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *Proceedings of the National Academy of Sciences*, **83**; 576–580.

Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C. & Minshull, J. (2007) Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnology*, **7**; 16.

Lin, P.P., Rabe, K.S., Takasumi, J.L., Kadisch, M., Arnold, F.H. & Liao, J.C. (2014) Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*. *Metabolic Engineering*, **24**; 1–8.

Lin, Z., Xu, Z., Li, Y., Wang, Z., Chen, T. & Zhao, X. (2014b) Metabolic engineering of *Escherichia coli* for the production of riboflavin. *Microbial Cell Factories*, **13**; 104.

Liu, B., Zhou, F., Wu, S., Xu, Y. & Zhang, X. (2009) Genomic and proteomic characterization of a thermophilic *Geobacillus* bacteriophage GBSV1. *Research in Microbiologoy*, **160**; 166–171.

Liu, B., Wei, Y., Zhang, Y. & Yang, Q. (2017) Deep Neural Networks for High Dimension, Low Sample Size Data. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence IJCAI-17*. pp. 2287–2293.

Lisser, S. & Margalit, H. (1993) Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Research*, **21**; 1507–1516.

Lou, C., Stanton, B., Chen, Y.J., Munsky, B. & Voigt, C.A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature Biotechnology*, **30**; 1137–1142.

Lu, T.K., Khalil, A.S. & Collins, J.J. (2009) Next-generation synthetic gene networks. *Nature Biotechnology*, **27**; 1139–1150.

Luo, Y., Zhang, L., Barton, K.W. & Zhao, H. (2015) Systematic Identification of a Panel of Strong Constitutive Promoters from *Streptomyces albus*. *ACS Synthetic Biology*, **4**; 1001–1010.

Makoff, A.J. & Oxer, M.D. (1991) High level heterologous expression in *E. coli* using mutant forms of the *lac* promoter. *Nucleic Acids Research*, **19**; 2417–2421.

Manilich, E.A., Ozsoyoglu, Z.M., Trubachev, V. & Radivoyevitch, T. (2011) Classification of large microarray datasets using fast random forest construction. *Journal of Bioinformatics and Computational Biology*, **9**; 251–267.

Mann, S., Li, J. & Chen, Y.P.P. (2006) A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts. *Nucleic Acids Research*, **35**; e12.

Markley, A.L., Begemann, M.B., Clarke, R.E., Gordon, G.C. & Pfleger, B.F. (2015) Synthetic Biology Toolbox for Controlling Gene Expression in the Cyanobacterium *Synechococcus* sp. strain PCC 7002. *ACS Synthetic Biology*, **4**; 595–603.

Mathews, J.A. (2008) Carbon-negative biofuels. *Energy Policy*, **36**; 940–945.

McWhinnie, R.L. & Nano, F.E. (2013) Synthetic Promoters Functional in *Francisella novicida* and *Escherichia coli*. *Applied and Environmental Microbiology*, **80**; 226–234.

Meijer, W.J.J. & Salas, M. (2004) Relevance of UP elements for three strong *Bacillus subtilis* phage 29 promoters. *Nucleic Acids Research*, **32**; 1166–1176.

Mellado, R.P. & Salas, M. (1982) High level synthesis in *Escherichia coli* of the *Bacillus subtillis* phage phi 29 proteins p3 and p4 under the control of phage lambda P_L promoter. *Nucleic Acids Research*, **10**; 5773–5784.

Mellin, J.R. & Cossart, P. (2015) Unexpected versatility in bacterial riboswitches. *Trends in Genetics*, **31**; 150–156.

Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B., Jimenez-Jacinto, V., Salgado, H., Juárez, K., Contreras-Moreira, B., Huerta, A.M., Collado-Vides, J. & Morett, E. (2009) Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*. *PLoS ONE*, **4**; e7526.

Meng, H. & Wang, Y. (2015) Cis-acting regulatory elements: from random screening to quantitative design. *Quantitative Biology*, **3**; 107–114.

Meng, H., Ma, Y., Mai, G., Wang, Y. & Liu, C. (2017) Construction of precise support vector machine based models for predicting promoter strength. *Quantitative Biology*, **5**; 90–98.

Meng, H., Wang, J., Xiong, Z., Xu, F., Zhao, G. & Wang, Y. (2013) Quantitative Design of Regulatory Elements Based on High-Precision Strength Prediction Using Artificial Neural Network. *PLoS ONE*, **8**; e60288.

Meng, W., Belyaeva, T., Savery, N.J., Busby, S.J.W., Ross, W.E., Gaal, T., Gourse, R.L. & Thomas, M.S. (2001) UP element-dependent transcription at the *Escherichia coli rrnB* P1 promoter: positional requirements and role of the RNA polymerase α subunit linker. *Nucleic Acids Research*, **29**; 4166–4178.

Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M.J., Elofsson, A., Nørholm, M.H.H. & Daley, D.O. (2015) Enhanced Protein Production in *Escherichia coli* by Optimization of Cloning Scars at the Vector–Coding Sequence Junction. *ACS Synthetic Biology*, **4**; 959–965.

Mordaka, P.M. & Heap, J.T. (2018) Stringency of Synthetic Promoter Sequences in *Clostridium* Revealed and Circumvented by Tuning Promoter Library Mutation Rates. *ACS Synthetic Biology*, **7**; 672–681.

Morgan-Kiss, R.M., Wadler, C. & Cronan, J.E., Jr. (2002) Long-term and homogeneous regulation of theEscherichia coli araBAD promoter by use of alactose transporter of relaxed specificity. *Proceedings of the National Academy of Sciences*, **99**; 7373–7377.

Mortimer, S.A., Kidwell, M.A. & Doudna, J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, **15**; 469–479.

Moser, F., Broers, N.J., Hartmans, S., Tamsir, A., Kerkman, R., Roubos, J.A., Bovenberg, R. & Voigt, C.A. (2012) Genetic Circuit Performance under Conditions Relevant for Industrial Bioreactors. *ACS Synthetic Biology*, **1**; 555–564.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P. & Endy, D. (2013a) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature Methods*, **10**; 354–360.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Mai, Q.A., Christoffersen, M.J., Martin, L., Yu, A., Lam, C., Rodriguez, C., Bennett, G., Keasling, J.D., Endy, D. & Arkin, A.P. (2013b) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nature Methods*, **10**; 347–353.

Nair, T.M. & Kulkarni, B.D. (1994) On the consensus structure within the *E. coli* promoters. *Biophysical Chemistry*, **48**; 383–393.

Nielsen, A.A.K., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., Ross, D., Densmore, D. & Voigt, C.A. (2016) Genetic circuit design automation. *Science*, **352**; aac7341-1–aac7341-11.

Nevoigt, E., Kohnke, J., Fischer, C.R., Alper, H., Stahl, U. & Stephanopoulos, G. (2006) Engineering of Promoter Replacement Cassettes for Fine-Tuning of Gene Expression in *Saccharomyces cerevisiae*. *Applied and Environmental Microbiology*, **72**; 5266–5273.

Niu, H., Leak, D., Shah, N. & Kontoravdi, C. (2015) Metabolic characterization and modeling of fermentation process of an engineered *Geobacillus thermoglucosidasius* strain for bioethanol production with gas stripping. *Chemical Engineering Science*, **122**; 138–149.

Novotny, R., Berger, H., Schinko, T., Messner, P., Schäffer, C. & Strauss, J. (2008) A temperature-sensitive expression system based on the *Geobacillus stearothermophilus* NRS 2004/3a sgsE surface-layer gene promoter. *Biotechnology and Applied Biochemistry*, **49**; 35.

Oh, Y.K., Hwang, K.R., Kim, C., Kim, J.R. & Lee, J.S. (2018) Recent developments and key barriers to advanced biofuels: A short review. *Bioresource Technology*, **257**; 320–333.

Parsons, J.B., Frank, S., Bhella, D., Liang, M., Prentice, M.B., Mulvihill, D.P. & Warren, M.J. (2010) Synthesis of Empty Bacterial Microcompartments, Directed Organelle Protein Incorporation, and Evidence of Filament-Associated Organelle Movement. *Molecular Cell*, **38**; 305–315.

Pasini, A. (2015) Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease*, **7**; 953–960.

Palermo, G., Piraino, P. & Zucht, H.D. (2009) Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and Applications in Bioinformatics and Chemistry*, **2**; 57–70.

Paul, N., Adelfinskaya, O., Hidalgo, A.E., Le, T. & Shore, S. (2013) Modified dNTPs: A Toolbox for Use in PCR. In *PCR Technology: Current Innovations* (ed. by Nolan, T. & Bustin, S.A.). CRC Press, New York, pp. 103–122.

Phan, T.T.P., Nguyen, H.D. & Schumann, W. (2012) Development of a strong intracellular expression system for *Bacillus subtilis* by optimizing promoter elements. *Journal of Biotechnology*, **157**; 167–172.

Pogrebnyakov, I., Jendresen, C.B. & Nielsen, A.T. (2017) Genetic toolbox for controlled expression of functional proteins in *Geobacillus* spp. *PLoS ONE*, **12**; e0171313.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, **26**; 1641–1650.

Prieto, A., Prieto, B., Ortigosa, E.M., Ros, E., Pelayo, F., Ortega, J. & Rojas, I. (2016) Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing*, **214**; 242–268.

Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A. & Arkin, A.P. (2012) RNA processing enables predictable programming of gene expression. *Nature Biotechnology*, **30**; 1002–1006.

Qin, X., Qian, J., Yao, G., Zhuang, Y., Zhang, S. & Chu, J. (2011) GAP Promoter Library for Fine-Tuning of Gene Expression in *Pichia pastoris*. *Applied and Environmental Microbiology*, **77**; 3600–3608.

Quax, T.E.F., Claassens, N.J., Söll, D. & van der Oost, J. (2015) Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, **59**; 149–161.

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**; 841–842.

Raccuglia, P., Elbert, K.C., Adler, P.D.F., Falk, C., Wenny, M.B., Mollo, A., Zeller, M., Friedler, S.A., Schrier, J. & Norquist, A.J. (2016) Machine-learning-assisted materials discovery using failed experiments. *Nature*, **533**; 73–76.

Raita, M., Ibenegbu, C., Champreda, V. & Leak, D.J. (2016) Production of ethanol by thermophilic oligosaccharide utilising *Geobacillus thermoglucosidasius* TM242 using palm kernel cake as a renewable feedstock. *Biomass and Bioenergy*, **95**; 45–54.

Rambaut, A. (Ed.). (2017) *FigTree*. Institute of Evolutionary Biology, University of Edinburgh [Online]. Available: http://tree.bio.ed.ac.uk/software/figtree/ [accessed on 30 March 2018].

Rao, C.V. (2015) Control Challenges in Synthetic Biology. *IFAC-PapersOnLine*, **48**; 996–1001.

Rao, L., Ross, W., Appleman, J.A., Gaal, T., Leirmo, S., Schlax, P.J., Record, M.T., Jr & Gourse, R.L. (1994) Factor Independent Activation of *rrnB* P1: An “Extended” Promoter with an Upstream Element that Dramatically Increases Promoter Strength. *Journal of Molecular Biology*, **235**; 1421–1435.

Ravasi, P., Peiru, S., Gramajo, H. & Menzella, H.G. (2012) Design and testing of a synthetic biology framework for genetic engineering of *Corynebacterium glutamicum*. *Microbial Cell Factories*, **11**; 147.

Razali, N.M. & Wah, Y.B. (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, **2**; 21–33.

Reeve, B., Martinez-Klimova, E., De Jonghe, J., Leak, D.J. & Ellis, T. (2016) The *Geobacillus* Plasmid Set: A Modular Toolkit for Thermophile Engineering. *ACS Synthetic Biology*, **5**; 1342–1347.

Reeve, B., Hargest, T., Gilbert, C. & Ellis, T. (2014) Predicting translation initiation rates for designing synthetic biology. *Frontiers in Bioengineering and Biotechnology*, **2**; 1.

Rhodius, V.A. & Mutalik, V.K. (2010) Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, sigmaE. *Proceedings of the National Academy of Sciences*, **107**; 2854–2859.

Rhodius, V.A., Mutalik, V.K. & Gross, C.A. (2012) Predicting the strength of UP-elements and full-length E. coli sigmaE promoters. *Nucleic Acids Research*, **40**; 2907–2924.

Rosipal, R. & Krämer, N. (2006) Overview and Recent Advances in Partial Least Squares. In *Subspace, Latent Structure and Feature Selection. Lecture Notes in Computer Science.* (ed. by Saunders, C., Grobelnik, M., Gunn, S. & Shawe-Taylor, J.). Springer, Berlin, pp. 34–51.

Ross, W., Aiyar, S.E., Salomon, J. & Gourse, R.L. (1998) *Escherichia coli* Promoters with UP Elements of Different Stregths: Modular Structure of Bacterial Promoters. *Journal of Bacteriology*, **180**; 5375–5383.

Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K. & Gourse, R.L. (1993) A Third Recognition Element in Bacterial Promoters: DNA Binding by the α Subunit of RNA Polymerase. *Science*, **262**; 1407–1413.

Rud, I., Jensen, P.R., Naterstad, K. & Axelsson, L. (2006) A synthetic promoter library for constitutive gene expression in *Lactobacillus plantarum*. *Microbiology*, **152**; 1011–1019.

Rudge, T.J., Brown, J.R., Federici, F., Dalchau, N., Phillips, A., Ajioka, J.W. & Haseloff, J. (2016) Characterization of Intrinsic Properties of Promoters. *ACS Synthetic Biology*, **5**; 89–98.

Ruff, E., Record, M., Jr. & Artsimovitch, I. (2015) Initial Events in Bacterial Transcription Initiation. *Biomolecules*, **5**; 1035–1062.

Running, J.A. & Bansal, K. (2016) Oxygen Transfer Rates in Shaken Culture Vessels From Fernbach Flasks to Microtiter Plates. *Biotechnology and Bioengineering*, **113**; 1729–1735.

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

Rytter, J.V., Helmark, S., Chen, J., Lezyk, M.J., Solem, C. & Jensen, P.R. (2014) Synthetic promoter libraries for *Corynebacterium glutamicum*. *Applied Microbiology and Biotechnology*, **98**; 2617–2623.

Saecker, R.M., Record, M.T., Jr & deHaseth, P.L. (2011) Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis. *Journal of Molecular Biology*, **412**; 754–771.

Salis, H.M., Mirsky, E.A. & Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, **27**; 946–950.

Sanderson, H., Dyer, S.D. & Nabholz, J.V. (2008) (Q)SAR and Extrapolation. In *Extrapolation Practice for Ecotoxilogical Effect Characterization of Chemicals* (ed. by Solomon, K.R., Brock, T.C.M., De Zwart, D., Dyer, S.D., Posthuma, L., Richards, S., Sanderson, H., Sibley, P. & van den Brink, P.J.). CRC Press, Florida.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**; 5463–5467.

Sargeant, L.A., Jenkins, R.W. & Chuck, C.J. (2017) Lipid-based Biofuels from Oleaginous Microbes. In *Biofuels and Bioenergy* (ed. by Love, J. & Bryant, J.A.). John Wiley & Sons, Ltd.

SAS Institute Inc. (2016a) *JMP (R) 13 Multivariate Methods*. SAS Institute Inc., North Carolina.

SAS Institute Inc. (2016b) *JMP(R) 13 Predictive and Specialized Modeling*. SAS Institute Inc., North Carolina.

Schmidt, A.F. & Finan, C. (2018) Linear regression and the normality assumption. *Journal of Clinical Epidemiology*. DOI: https://doi.org/10.1016/j.jclinepi.2017.12.006.

Schmidt, F.R. (2005) Optimization and scale up of industrial fermentation processes. *Applied Microbiology and Biotechnology*, **68**; 425–435.

Segall-Shapiro, T.H., Sontag, E.D. & Voigt, C.A. (2018) Engineered promoters enable constant gene expression at any copy number in bacteria. *Nature Biotechnology*, **36**; 352-358.

Setiono, R. (1997) A penalty-function approach for pruning feedforward neural networks. *Neural Computation*, **9**; 185–204.

Sharkey, A.J. (1996) On Combining Artificial Neural Nets. *Connection Science*, **8**; 299–313.

Shell International BV. (2016) *A better life with a healthy planet*. Shell International.

Sheng, L., Kovács, K., Winzer, K., Zhang, Y. & Minton, N.P. (2016) Development and implementation of rapid metabolic engineering tools for chemical and fuel production in *Geobacillus thermoglucosidasius* NCIMB 11955. *Biotechnology for Biofuels*, **10**; 5.

Sheridan, R.P. (2013) Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *Journal of Chemical Information and Modeling*, **53**; 783–790.

Shine, J. & Dalgarno, L. (1974) The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Proceedings of the National Academy of Sciences*, **71**; 1342–1346.

Siegl, T., Tokovenko, B., Myronovskyi, M. & Luzhetskyy, A. (2013) Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. *Metabolic Engineering*, **19**; 98–106.

Siegele, D.A. & Hu, J.C. (1997) Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proceedings of the National Academy of Sciences*, **94**; 8168–8172.

Simon, R., Priefer, U. & Pühler, A. (1983) A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Nature Biotechnology*, **1**; 784–791.

Simons, G., Remaut, E., Allet, B., Devos, R. & Fiers, W. (1984) High-level expression of human interferon gamma in *Escherichia coli* under control of the p_L promoter of bacteriophage lambda. *Gene*, **28**; 55–64.

Singh, V. (2014) Recent advancements in synthetic biology: Current status and challenges. *Gene*, **535**; 1–11.

Sinha, S. (2006) On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, **22**; e454–e463.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.Y., Hjalmarsson, H. & Juditsky, A. (1995) Nonlinear Black-box Modeling in System Identification: a Unified Overview*. *Automatica*, **31**; 1691–1724.

Skulj, M., Okrslar, V., Jalen, S., Jevsevar, S., Slanc, P., Strukelj, B. & Menart, V. (2008) Improved determination of plasmid copy number using quantitative real-time PCR for monitoring fermentation processes. *Microbial Cell Factories*, **7**; 6.

Sleight, S.C. & Sauro, H.M. (2013) Visualization of Evolutionary Stability Dynamics and Competitive Fitness of Escherichia coli Engineered with Randomized Multigene Circuits. *ACS Synthetic Biology*, **2**; 519–528.

Sohoni, S.V., Fazio, A., Workman, C.T., Mijakovic, I. & Lantz, A.E. (2014) Synthetic Promoter Library for Modulation of Actinorhodin Production in *Streptomyces coelicolor* A3(2). *PLoS ONE*, **9**; e99701.

Solem, C. & Jensen, P.R. (2002) Modulation of Gene Expression Made Easy. *Applied and Environmental Microbiology*, **68**; 2397–2403.

Solovyev, V. & Salamov, A. (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies* (ed. by Li, R.W.). Nova Science Publishers, New York, pp. 61–78.

Song, K. (2011) Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. *Nucleic Acids Research*, **40**; 963–971.

Song, Y., Nikoloff, J.M., Fu, G., Chen, J., Li, Q., Xie, N., Zheng, P., Sun, J. & Zhang, D. (2016) Promoter Screening from *Bacillus subtilis* in Various Conditions Hunting for Synthetic Biology and Industrial Applications. *PLoS ONE*, **11**; e0158447.

Strainic, M.G., Sullivan, J.J., Velevis, A. & deHaseth, P.L. (1998) Promoter Recognition by *Escherichia coli* RNA Polymerase: Effects of the UP Element on Open Complex Formation and Promoter Clearance. *Biochemistry*, **37**; 18074–18080.

Stromo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**; 16–23.

Struchtemeyer, C.G., Davis, J.P. & Elshahed, M.S. (2011) Influence of the Drilling Mud Formulation Process on the Bacterial Communities in Thermogenic Natural Gas Wells of the Barnett Shale. *Applied and Environmental Microbiology*, **77**; 4744–4753.

Studier, F.W. & Moffatt, B.A. (1986) Use of Bacteriophage T7 RNA Polymerase to Direct Selective High-level Expression of Cloned Genes. *Journal of Molecular Biology*, **189**; 113–130.

Suzuki, H. (2012) Genetic Transformation of *Geobacillus kaustophilus* HTA426 by Conjugative Transfer of Host-Mimicking Plasmids. *Journal of Microbiology and Biotechnology*, **22**; 1279–1287.

Suzuki, H., Murakami, A. & Yoshida, K.I. (2012) Counterselection System for *Geobacillus kaustophilus* HTA426 through Disruption of *pyrF* and *pyrR*. *Applied and Environmental Microbiology*, **78**; 7376–7383.

Suzuki, H., Yoshida, K.I. & Ohshima, T. (2013) Polysaccharide-Degrading Thermophiles Generated by Heterologous Gene Expression in *Geobacillus kaustophilus* HTA426. *Applied and Environmental Microbiology*, **79**; 5151–5158.

Sztiller-Sikorska, M., Heyduk, E. & Heyduk, T. (2011) Promoter spacer DNA plays an active role in integrating the functional consequences of RNA polymerase contacts with −10 and −35 promoter elements. *Biophysical Chemistry*, **159**; 73–81.

Taylor, M.P., Esteban, C.D. & Leak, D.J. (2008) Development of a versatile shuttle vector for gene expression in *Geobacillus* spp. *Plasmid*, **60**; 45–52.

Tenenbaum, D.J. (2008) Food vs. Fuel: Diversion of Crops Could Cause More Hunger. *Environmental Health Perspectives*, **116**; A254–A257.

Terpe, K. (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, **72**; 211–222.

Thornton, B. & Basu, C. (2011) Real-time PCR (qPCR) primer design using free online software. *Biochemistry and Molecular Biology Education*, **39**; 145–154.

Tobias, R.D. (1995) An Introduction to Partial Least Squares Regression. In *Proceedings of the Twentieth Annual SAS Users Group International Conference*. SAS Institute Inc, North Carolina, pp. 1250-1257.

Tornøe, J., Kusk, P., Johansen, T.E. & Jensen, P.R. (2002) Generation of a synthetic mammalian promoter library by modification of sequences spacing transcription factor binding sites. *Gene*, **297**; 21–32.

Tuller, T. & Zur, H. (2014) Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Research*, **43**; 13–28.

Tye, H. (2004) Application of statistical “design of experiments” methods indrug discovery. *Drug Discovery Today*, **9**; 485–491.

Tyner, W.E. (2015) Biofuel economics and policy: the renewable fuel standard, the blend wall, and future uncertainties. In *Bioenergy* (ed. by Dahiya, A.). Academic Press, pp. 511–521.

Tyo, K.E.J., Ajikumar, P.K. & Stephanopoulos, G. (2009) Stabilized gene duplication enables long-term selection-free heterologous pathway expression. *Nature Biotechnology*, **27**; 760–765.

Umarov, R.K. & Solovyev, V.V. (2017) Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE*, **12**; e0171410.

Urtecho, G., Tripp, A.D., Insigne, K., Kim, H. & Kosuri, S. (2018) Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically-Encoded Multiplexed Reporter Assay in E. coli. *Biochemistry*. DOi: 10.1021/acs.biochem.7b01069

US Energy Information Administration. (2016) *International energy outlook*. 484 edn. US Energy Information Administration, Washington DC.

van der Voet, H. (1994) Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*, **25**; 313–323.

Valdez-Cruz, N.A., Caspeta, L., Pérez, N.O., Ramírez, O.T. & Trujillo-Roldán, M.A. (2010) Production of recombinant proteins in *E. coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. *Microbial Cell Factories*, **9**; 18.

Wei, L., Xu, N., Wang, Y., Zhou, W., Han, G., Ma, Y. & Liu, J. (2018) Promoter library-based module combination (PLMC) technology for optimization of threonine biosynthesis in *Corynebacterium glutamicum*. *Applied Microbiology and Biotechnology*, **102**; 4117-4130.

Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J. & Gustafsson, C. (2009) Design Parameters to Control Synthetic Gene Expression in *Escherichia coli*. *PLoS ONE*, **4**; e7002.

Wojcik, E.Z., Singleton, C., Chapman, L.N., Parker, D.A. & Love, J. (2017) Plant Biomass as Biofuels. In eLS (ed. by John Wiley & Sons Ltd). DOI: 10.1002/9780470015902.a0023716.

Wold, S., Høy, M., Martens, H., Trygg, J., Westad, F., MacGregor, J. & Wise, B.M. (2009) The PLS model space revisited. *Journal of Chemometrics*, **23**; 67–68.

Wold, S., Sjöstrom, M. & Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 109–130.

Wong Ng, J. Chatenay, D., Robert, J. & Poirier, M.G. (2010) Plasmid copy number noise in monoclonal populations of bacteria. *Physical Review E*, **81**; 011909.

Wu, G., Yan, Q., Jones, J.A., Tang, Y.J., Fong, S.S. & Koffas, M.A.G. (2016) Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. *Trends in Biotechnology*, **34**; 652–664.

Wycuff, D.R. & Matthews, K.S. (2000) Generation of an AraC-araBAD Promoter-Regulated T7 Expression System. *Analytical Biochemistry*, **277**; 67–73.

Yan, Q. & Fong, S.S. (2017) Study of *in vitro* transcriptional binding effects and noise using constitutive promoters combined with UP element sequences in *Escherichia coli*. *Journal of Biological Engineering*, **11**; 33.

Yang, G., Jia, D., Jin, L., Jiang, Y., Wang, Y., Jiang, W. & Gu, Y. (2017a) Rapid Generation of Universal Synthetic Promoters for Controlled Gene Expression in Both Gas-Fermenting and Saccharolytic *Clostridium* Species. *ACS Synthetic Biology*, **6**; 1672–1678.

Yang, J., Chen, X., McDermaid, A. & Ma, Q. (2017b) DMINDA 2.0: Integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics*, **33**; 2586–2588.

Yang, J., Zeng, X., Zhong, S. & Shengli, W. (2013) Effective Neural Network Ensemble Approach for Improving Generalization Performance. *IEEE Transactions on Neural Networks and Learning Systems*, **24**; 878–887.

Yang, S., Liu, Q., Zhang, Y., Du, G., Chen, J. & Kang, Z. (2017c) Construction and Characterization of Broad-spectrum Promoters for Synthetic Biology. *ACS Synthetic Biology*, **7**; 287-291.

Yao, A.I., Fenton, T.A., Owsley, K., Seitzer, P., Larsen, D.J., Sit, H., Lau, J., Nair, A., Tantiongloc, J., Tagkopoulos, I. & Facciotti, M.T. (2013) Promoter Element Arising from the Fusion of Standard BioBrick Parts. *ACS Synthetic Biology*, **2**; 111–120.

Yim, S.S., An, S.J., Kang, M., Lee, J. & Jeong, K.J. (2013) Isolation of Fully Synthetic Promoters for High-Level Gene Expression in *Corynebacterium glutamicum. Biotechnology and Bioengineering*, **110**; 2959–2969.

Yona, A.H., Alm, E.J. & Gore, J. (2017) Random Sequences Rapidly Evolve Into *de novo* Promoters. *bioRxiv*. DOI: https://doi.org/10.1101/111880

Zambare, V.P., Bhalla, A., Muthukumarappan, K., Sani, R.K. & Christopher, L.P. (2011) Bioprocessing of agricultural residues to ethanol utilizing a cellulolytic extremophile. *Extremophiles*, **15**; 611–618.

Zaccolo, M., Williams, D.M., Brown, D.M. & Gherardi, E. (1996) An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. *Journal of Molecular Biology*, **255**; 589–603.

Zeigler, D.R. (2014) The *Geobacillus* paradox: why is a thermophilic bacterial genus so prevalent on a mesophilic planet? *Microbiology*, **160**; 1–11.

Zeigler, D.R. (2001) *The Genus Geobacillus - Introduction and Strain Catalog*. Bacillus Genetic Stock Center.

Zhang, S., Liu, D., Mao, Z., Mao, Y., Ma, H., Chen, T., Zhao, X. & Wang, Z. (2018) Model-based reconstruction of synthetic promoter library in *Corynebacterium glutamicum*. *Biotechnology Letters*. DOI: https://doi.org/10.1007/s10529-018-2539-y.

Zhang, Z.G., Yi, Z.L., Pei, X.Q. & Wu, Z.L. (2010) Improving the thermostability of *Geobacilus stearothermophilus* xylanase XT6 by directed evolution and site-directed mutagenesis. *Bioresource Technology*, **101**; 9272–9278.

Zhou, J., Wu, K. & Rao, C. (2016) Evolutionary Engineering of *Geobacillus thermoglucosidasius* for Improved Ethanol Production. *Biotechnology and Bioengineering*, **113**; 2156–2167.

Zhou, P., Chen, X., Wu, Y. & Shang, Z. (2009) Gaussian process: an alternative approach for QSAM modeling of peptides. *Amino Acids*, **38**; 199–212.

Zhou, S., Du, G., Kang, Z., Li, J., Chen, J., Li, H. & Zhou, J. (2017) The application of powerful promoters to enhance gene expression in industrial microorganisms. *World Journal of Microbiology and Biotechnology*, **33**; 23.

Zong, Y., Zhang, H.M., Lyu, C., Ji, X., Hou, J., Guo, X., Ouyang, Q. & Lou, C. (2017) Insulated transcriptional elements enable precise design of genetic circuits. *Nature Communications*, **8**; DOI: 10.1038/s41467-017-00063-z.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**; 3406–3415.

Appendix

Synthetic promoter design for new microbial chassis

James Gilman* **and John Love*[1]**

*Biosciences, College of Life and Environmental Sciences, The University of Exeter, Stocker Road, Exeter EX4 4QD, U.K.

Abstract

The judicious choice of promoter to drive gene expression remains one of the most important considerations for synthetic biology applications. Constitutive promoter sequences isolated from nature are often used in laboratory settings or small-scale commercial production streams, but unconventional microbial chassis for new synthetic biology applications require well-characterized, robust and orthogonal promoters. This review provides an overview of the opportunities and challenges for synthetic promoter discovery and design, including molecular methodologies, such as saturation mutagenesis of flanking regions and mutagenesis by error-prone PCR, as well as the less familiar use of computational and statistical analyses for *de novo* promoter design.

Introduction

Predictable output is a defining aspiration of synthetic biology. A number of factors affect the output from synthetic gene networks to a greater or lesser extent, including transgene copy number [1], integration into the genome or expression from plasmids [2], promoter activity [3], ribosome-binding sites [4–6], codon bias of the host [7], transcription rate and tRNA abundance [8], half-life of mRNA [9], substrate and cofactor availability [10], adjustment of enzyme kinetics [11], protein scaffolding [12] and sub-cellular localization through the use of microcompartments [13,14]. The use of RNA as a control mechanism, either through the application of riboswitches [15] or toehold switches [16] has also emerged as a powerful tool for pathway control. Each of these aspects can be investigated and improved individually, and then integrated by a model, a suite of experiments or ideally, a combination of modelling and empiricism.

Several investigations, including the now archetypal 'repressilator' [17] and the genetic toggle switch [18] have modelled promoters and generated bacteria that display patterns of gene expression consistent with mathematical predictions. However, despite these successes, when individual bacteria are investigated, strong variations in transgene expression levels become apparent, even within clonal populations [19].

Controlling transcription is often the simplest way to balance expression of a transgene or synthetic pathway, and constitutive promoters with different and predictable activation characteristics are a desirable feature of any synthetic biology toolkit. However, in practise, promoter availability tends to be restricted to relatively few sequences [20], which do not always perform as required and may not necessarily be transferrable to new microbial chassis. The fact that many promoters are characterized as merely 'weak' or 'strong' [21] highlights this issue – such definitions are hardly sufficient to allow adequate promoter selection.

A number of inducible promoter systems are available for which the concentration of inducer can, in theory, be modulated in order to achieve the desired level of protein production [22]. Although the use of inducible promoter systems has been successful in some instances, in others it can prove inadequate. Promoter hypersensitivity to the inducer [23], the cost of adding large quantities of inducer to an industrial-scale fermenter [24] or heterogeneous expression levels across a population [25] all complicate the use of inducible promoters in industrial-scale cultures. Consequently, for large-scale production applications, constitutive promoters with 'hard-wired', predictable properties are often preferred and are the focus of this review.

In this article, we review the potential and methodologies for designing and characterizing new constitutive promoter sequences with predictable outputs, including conventional PCR-based techniques, hybrid promoter engineering and the expanding use of computational analysis for *de novo* promoter design.

Characteristics of promoters for synthetic biology applications

A promoter can be broadly defined as a cis-regulatory element containing a somewhat modular suite of key motifs that control the transcription of individual ORFs or operons. In prokaryotes, the structure and organization of natural promoter motifs is relatively well understood (Figure 1A). Eukaryotic promoters are somewhat more complex than their prokaryotic counterparts (Figure 1B),

Key words: artificial neural networks, partial least squares modelling, promoter, synthetic biology, systems biology.

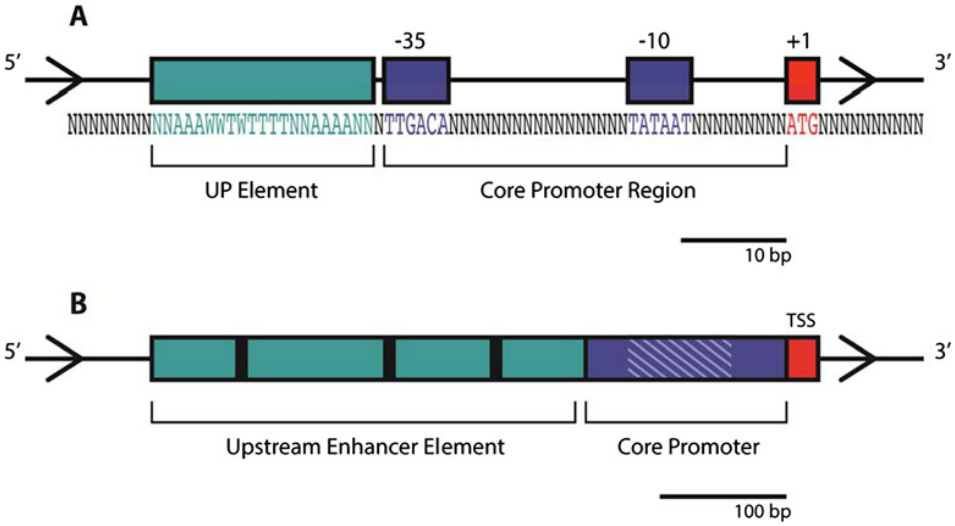
Abbreviations: ANN, artificial neural network; epPCR, error-prone PCR; PLS, partial least squares; PWM, position weight matrix; SMFR, saturation mutagenesis of flanking regions; SPL, synthetic promoter library; TSS, transcription start site; UAS, upstream activation sequence.

[1] **To whom correspondence should be addressed (email J.Love@Exeter.ac.uk).**

Biochem. Soc. Trans. (2016) 44, 731–737; doi:10.1042/BST20160042

Figure 1 | Schematic representations of typical promoter sequences

(**A**) Schematic representation of a typical prokaryotic promoter sequence. The transcription start site (TSS) is shown in red. Two conserved hexamers, at approximately 10 and 35 bp upstream of the TSS [68], highlighted here in blue, serve as key binding regions for RNA polymerase [69]. No such conserved motifs have been found in the region of sequence separating the two hexamers, although a consensus length of 17 bp has been observed in some species [70]. In addition to these core promoter elements, an upstream region (highlighted here in turquoise) is present in some promoters. Typically adenine/thymine rich, these UP elements boost transcription rate through interactions with the C-terminal domain on the RNA polymerase α-subunit [71]: Estrem, S.T., Gaal, T., Ross, W. and Gourse, R.L. (1998) Identification of an UP element consensus sequence for bacterial promoters. Proc. Natl. Acad. Sci. U.S.A. **95**, 9761–9766. The UP element consensus sequence is as derived by [71]. −10 and −35 consensus sequences are from *E. coli* and are reproduced from [3]: Blazeck, J. and Alper, H.S. (2013) Promoter engineering: recent advances in controlling transcription at the most fundamental level. Biotechnol. J. **8**, 46–58 and [72]: Ross, W., Aiyar, S.E., Salomon, J. and Gourse, R.L. (1998) *Escherichia coli* promoters with UP elements of different strengths: modular structure of bacterial promoters. J. Bacteriol. **180**, 5375–5383. N represents any deoxyribonucleotide. W represents adenine (A) or thymine (T). G and C represent guanine and cytosine respectively. (**B**) Schematic representation of a *S. cerevisiae* promoter sequence. The TSS is highlighted in red. Eukaryotic promoters can be broadly split into two regions, a core promoter element (shown in blue) and an upstream enhancer [3] (shown in turquoise), both of which can be modified in order to modulate expression levels. The core region provides the minimal sequence necessary for initiation of basal transcription and may contain key motifs, the most widely studied of which is the TATA box, which typically occurs 40–120 bp upstream of the TSS [73]. However, such motifs are by no means requisite for transcription initiation, as TATA boxes appear in only 20% of *S. cerevisiae* promoter elements [74]. Diagonal lines represent the region in which TATA boxes are most common. Upstream of the core promoter, the enhancer element serves to localize transcription factors, with interactions between bound transcription factors and the transcriptional machinery serving as a determinant of promoter strength and control [56]. Transcription factor binding sites do not display uniform distribution across the enhancer element, and are represented here as solid vertical lines in arbitrary positions. The highest concentration of such binding motifs has been reported between 50–150 bp prior to the TSS [75], although they may be present as much as 500 bases upstream of the TSS [76].



with localization of the transcriptional apparatus resulting from interactions between highly specific transcription factors, the promoter elements and co-activators [26]. The activity of promoters is typically quantified through measures of cellular mRNA or reporter proteins [27], linking the levels of promoter activity (or 'strength') to both transcription and translation. In reality, the promoter regulates only transcription but in practise, experimental constraints use protein quantification as a useful proxy for promoter activity.

From an industrial perspective, it is preferable to have a system that displays little variation, even if the overall output of that system is, on average, slightly less than that of an alternative that displays irregularities; synthetic biology aims to be boringly predictable rather than wonderfully complex. Candidate promoters for synthetic biology must therefore be well-characterized and yield consistent results, and also be insulated from the background metabolisms and molecular control systems. However, consistency is often confounded by the inherently stochastic nature of gene expression, which

subjects both promoters and any downstream proteins used in their characterization to large degrees of noise [28], as well as the all-or-nothing phenomenon [29] in inducible systems, wherein expression is typically fully induced in a subset of the population whereas the remaining cells display no expression [22,30,31].

Natural promoter sequences

The promoters available for use in synthetic systems have generally been limited to those endogenous elements isolated from model organisms, for instance, the *Escherichia coli lac* promoter and derivatives thereof [32–35] and the arabinose-inducible P_{BAD} [36–38] promoter.

Phage genomes can also be used to generate novel promoters. For example the p_L promoter, isolated from bacteriophage lambda, provides medium to high expression levels, and is tightly thermally-regulated by the cI repressor [34,39]. p_L has been successfully employed to increase yield of various proteins in *E. coli* expression systems [40–42]. Similarly, the T7 RNA polymerase-based promoter system, also initially isolated from bacteriophage, has been widely adopted [34,43].

Although natural promoters are widely used in relatively simple, laboratory applications, the relative paucity of sufficiently characterized elements makes their use in control in industrial contexts problematic. Additionally, natural promoter activity is often context-specific [3] and subject to interaction with a multitude of regulatory proteins, rendering prediction of activity levels under varying conditions non-trivial [44]. As a result of these inherent limitations, researchers have increasingly turned to libraries of synthetic promoter elements to meet their needs.

Molecular approaches for the production of synthetic promoter libraries

Saturation mutagenesis of flanking regions

A key method of forming synthetic promoter libraries (SPLs) is based on the observation that the flanking regions surrounding consensus motifs within the promoter sequence have a role in determining activity [45]. Degenerate oligonucleotides allow known consensus motifs to be maintained whereas the flanking regions are mutagenized, leading to altered promoter activity. For example saturation mutagenesis of flanking regions (SMFR) was successfully used to produce a SPL with a 400-fold activity range in *Lactococcus lactis*, with greater range being reported as a result of synthesis errors in the consensus sequences and alteration to flank length [24,45]. However, the initial approach taken to saturation mutagenesis by Jensen and Hammer [24,45] does not take into account the context-dependant nature of promoter activity. Consequently, current SPL generation uses a single PCR stage, with degenerate oligonucleotides coupled to either a full-length or truncated version of the gene that the promoter is intended to drive. This improvement allows for ectopic analysis or replacement of a wild-type promoter with a synthetic alternative, although maintaining the 5' mRNA of the target gene [23,46]. Promoter function is maintained due to the preservation of the key consensus regions within the sequence, with altered expression levels likely being the result of minor changes in DNA confirmation within the randomized flanks [45].

SMFR has been successfully applied in a variety of prokaryotes and eukaryotes, including *Corynebacterium glutamicum* [47] and *Streptomyces coelicolor* [48], yielding robust libraries with broad expression profiles. The methodology has also shown applicability in *Saccharomyces cerevisiae,* wherein screening of an initial large library of colonies ultimately yielded 20 characterized promoters, displaying expression levels of yeast-EGFP that varied by approximately 22-fold [21]. In a separate study, a selection of constitutive promoters was initially isolated from the *S. cerevisiae* genome, and expression levels were subsequently characterized using expression profiles available from public databases. The promoter of the gene *PFϒ1* was chosen as a starting point for its robust expression profile [49]. Knowledge of *PFϒ1* structure enabled identification of a rDNA enhancer-binding protein and a poly-dT that were important for transcription initiation [50]. These regions were therefore held constant whereas a 48 bp section of the promoter core was randomized, providing a library of 36 promoter elements with a broad range of expression levels. It must be noted that none of the new sequences provided higher expression levels than the original *PFϒ1* promoter [49]. This inability to produce a synthetic promoter with higher expression levels than a natural alternative was also reported by McWhinnie and Nano [51].

Although SMFR has successfully provided many new promoters, the technique requires labour intensive cloning and an *a priori* knowledge of promoter structure in the organism of interest, something that may not be immediately available in industrially relevant microbes. Furthermore, as many libraries use composite promoter scaffolds as a starting point, establishing a definitive wild-type reference expression baseline is impossible. Definitively stating whether SMFR will improve wild-type expression capability *pre hoc*, is therefore problematic [3]. Additionally, by restricting mutagenesis to only the flanking regions, SMFR fails to take into account alterations to consensus sequences, which are known to play a significant role in modulating expression strength.

Error-prone PCR

Generating a SPL by applying error-prone PCR (epPCR) to an entire promoter sequence obviates any *a priori* knowledge of functional motif location and can potentially result in promoters with entirely new characteristics [3]. This methodology was successfully used to mutagenize a bacteriophage P_L-λ promoter that was subsequently placed upstream of a green fluorescent protein (GFP) coding sequence and transformed into *E. coli*, resulting in a library containing approximately 9000–12000 functional clones [52,53]. Visual screening of the colonies resulted in a subset of 200 promoters, of which 27, representing 22 discrete promoter sequences, were found to give homogeneous expression levels. Subsequently,

thorough characterization of this promoter subset resulted in a promoter library which was successfully employed to modulate levels of phosphoenolpyruvate carboxylase and lycopene production in *E. coli* [53]. epPCR for promoter production has also been employed in *C. glutamicum*, where iterative rounds of high-throughput sorting and analysis at the single-cell level ultimately yielded a library of 20 well-characterized sequences from an initial library of 10^5 mutagenized cells [54]. The technique has also been successfully applied in yeast [55].

Despite these successes, the epPCR approach to SPL production has certain limitations: a reliance on a selection of a small subset of colonies for further analysis [53,54] renders discovery of a true optimum problematic. Moreover, the extensive screening required to isolate said subset should not be underestimated; it is typical for initial libraries of hundreds or thousands of bacterial colonies to ultimately yield relatively few fully characterized promoters. Both these problems become less of an issue if visual selection of colonies is replaced by high-throughput analytical techniques such as fluorescence-activated cell sorting and/or imaging cytometry.

Hybrid promoter engineering

In addition to the two mutagenic techniques discussed above, the generation of synthetic promoters through hybridization of existing promoter elements provides an alternative strategy for promoter genesis. By combining minimal core promoter elements with various combinations of modular upstream activation sequences (UAS), Blazeck et al. [56] demonstrated that expression levels could be increased compared with a wild-type baseline in *S. cerevisiae*. A roughly linear relationship was observed between the number of UAS modules added and promoter strength, with the addition of four such elements boosting expression of a weak constitutive promoter to levels comparable with the strongest endogenous promoter [56]. Transcriptional increase was shown to depend both on the core element and UAS, but all core promoters were amenable to improvement [56].

Computational methods for synthetic promoter discovery

Although the above molecular methodologies have certainly provided new promoters of varying activities, these approaches do not represent a systematic, theoretical examination of the promoter design space. If, for arguments sake, a promoter sequence is 100 bp in length, there are 4^{100} potential promoter sequences. Therefore, although the best sequence discovered by molecular-based SPL may be sufficient for some experimental purposes, it is possible that other optima are present. *In silico* methods that are capable of deciphering the effect of individual DNA bases and motifs, or predicting promoter activity level in advance of *in vivo* characterization have, in this context, considerable potential [57]. Conventionally, the use of computational techniques in pathway design and optimization has been limited to *post hoc* data analytics [21]. However, computational modelling in biological systems design and optimization is becoming more widespread, and a number of computational methodologies are available to facilitate the *de novo* design of synthetic promoter sequences.

Position weight matrix models

Position weight matrix (PWM) models have been widely applied for the detection of transcription factor binding sites [58,59], and have also shown some promise in predicting promoter strength. By breaking promoter sequences into constitutive motifs, PWM models were able to predict the strength of *E. coli* core promoter sequences recognized by sigma factor σ^E to a relatively high degree of accuracy [60]. The core promoter PWM was subsequently combined with a score describing the activity of upstream elements to provide a model capable of predicting the strength of entire promoter sequences [61]. In addition to this predictive power, PWM models provide increased understanding of promoter structure, something that is often limited in novel microbial chassis.

Although PWM models certainly have the potential to be applied to *de novo* sequence design, they are not without limitations. PWMs may prove inadequate for modelling in promoter families with a less conserved nature than those which interact with σ^E, as poorly conserved sequences required greater complexity within the model [60]. Application of PWMs in novel microbial chassis, where understanding of interactions between proteins and promoter sequences can be limited, may therefore be challenging.

Additionally, by assuming that the contribution of individual nucleotides to DNA-protein binding is independent and additive [61], PWMs fail to account for the effect of interactions between positions. Despite these limitations, the application of PWMs for the *pre hoc* determination of strength in certain promoter families carries great potential.

Partial least squares regression

The use of statistical modelling to quantitatively link DNA sequence to function is not a new concept [62], although as a method for the generation of synthetic promoters it remains underutilized. In a pioneering study, 25 *E. coli* promoters were analysed using a partial least squares (PLS) methodology, resulting in a statistical model that analysed the contribution of each individual nucleotide at any given position in the DNA sequence. In order to validate the model, two synthetic sequences with predicted high activity levels were synthesized. The -35, -10 and $+1$ sites were determined using the consensus sequence of the training set of 25 promoters, whereas the remainder of the synthetic sequences were determined using regression coefficients provided by the modelling process [62]. *In vivo* characterization of the synthetic promoters revealed activity levels within approximately 8% of the strength predicted by the model. Furthermore, the synthetic sequences were shown to provide higher expression levels than any of those sequences found within the training set [62].

Similar statistical methods were later applied to quantitatively link promoter structure with function for a library

of synthetic *E. coli* promoters that were generated through the randomization of flanking regions [29]. The generated model was able to predict, with reasonable accuracy, the strength of promoter sequences that had not been used in the construction of the model [29]. In further validation of this computational technique, the promoter strength predictive model was subsequently utilized to predict the strength of an endogenous *E. coli* promoter, that of the *ppc* gene [63]. Based on this information, stronger promoters were selected from the previously characterized promoter library [29] in order to fine-tune *ppc* expression levels. This knock-in approach resulted in an increase in expression levels roughly in line with the model's predictions, with a 3–4-fold increase in mRNA levels seen at flask scale [63]. Although the PLS regression doubtlessly aided in the optimization process, it was not applied, in this instance to the *de novo* design of synthetic promoter sequences.

Artificial neural networks

The linear nature of PLS modelling is a drawback when applied to the analysis of promoter sequences, confounding the effects of any interactions between bases with the main effects for each individual nucleotide position [62]. PLS models therefore may not accurately account for the complexity inherent in promoter structure, thereby increasing the probability of prediction errors and inadequate generality [64]. Indeed, many such models lack robust prediction accuracy [65], rendering their use in *de novo* sequence design challenging.

Artificial neural networks (ANNs) may provide a solution to these issues. Based upon a network of interconnected nodes designed to act as a rudimentary mimic of the brain, ANNs permit machine learning, as the order and force of connections may be altered [66]. By systematically altering node structure during the analysis of a training data set, ANN models can potentially better represent the complex, non-linear interactions occurring within a promoter sequence [64]. ANN modelling has proven successful for *de novo* promoter design [64]; using a set of synthetic promoters derived from the random mutagenesis of a wild-type *E. coli* promoter as a training set for an ANN model, strength predictions of sequences generated by *in silico* mutagenesis were used to select 16 synthetic sequences for *in vivo* verification [64]. The predicted expression levels displayed good correlation with empirical testing, suggesting that such models are indeed applicable to synthetic promoter design. Indeed, the fact that approximately 30% of *de novo* designed sequences displayed greater expression levels that the wild-type control [64] compares extremely favourably to the more traditional mutagenesis-based techniques discussed above, where much lower success rates are not uncommon.

The importance of insulation

Whichever method is applied to the generation of SPLs, promoter elements must be sufficiently insulated if they are to be efficiently used in synthetic regulatory systems. Empirical or predictive data regarding promoter strength from characterization using a reporter protein must be comparable to promoter performance when coupled to a protein of interest within a synthetic pathway; context-dependent effects should be minimal. However, achieving context dependency is non-trivial, as fluctuation in promoter activity levels may be the result of a wide array of experimental and/or genetic factors [27,53].

A possible solution to this problem is to separate core elements from their genetic context through the use of insulator sequences, such as a defined 5' mRNA sequence [67]. By using such insulators, promoter elements from a SPL can produce constant relative levels of various reporter proteins when used for both plasmid and chromosomal expression [67].

Conclusion

The ability to select a reliable promoter of known activity is of paramount importance for synthetic biology. Indeed, promoters with different and, most importantly, predictable effects on transcription may be used to regulate complex gene circuits, balance engineered metabolic pathways and exploit new chassis for industrial-scale applications. As reviewed here, a number of molecular and computational methodologies are available for the discovery and design of new constitutive promoters. Each technique has advantages and weaknesses, and a selection of one over the other will depend on the aims of specific projects. However, to date, computational approaches to promoter design remain underutilized aside from proof of principle studies in model organisms. As the applications of synthetic biology become more entrenched in the future bio-economy, which may require the development of different chassis, the application of computational modelling to promoter design can enhance and accelerate the design process and ultimately enhance our fundamental knowledge of genetic regulation in complex systems.

References

1 Ajikumar, P.K., Xiao, W.H., Tyo, K.E.J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Heng Phon, T., Pfeifer, B. and Stephanopoulos, G. (2010) Isoprenoid pathway optimization for taxol precursor overproduction in *Escherichia coli*. Science **330**, 70–74 CrossRef PubMed

2 Tyo, K.E.J, Ajikumar, P.K. and Stephanopoulos, G. (2009) Stabilized gene duplication enables long-term selection-free heterologous pathway expression. Nat. Biotechnol. **27**, 760–765 CrossRef PubMed

3 Blazeck, J. and Alper, H.S. (2013) Promoter engineering: recent advances in controlling transcription at the most fundamental level. Biotechnol. J. **8**, 46–58 CrossRef PubMed

4 Lin, Z., Xu, Z., Li, Y., Wang, Z., Chen, T. and Zhao, X. (2014) Metabolic engineering of *Escherichia coli* for the production of riboflavin. Microb. Cell Fact. **13**, 1–12 CrossRef PubMed

5 Ravasi, P., Peiru, S., Gramajo, H. and Menzella, H.G. (2012) Design and testing of a synthetic biology framework for genetic engineering of *Corynebacterium glutamicum*. Microb. Cell Fact. **11**, 147–158 CrossRef PubMed

6 Markley, A.L., Begemann, M.B., Clarke, R.E., Gordon, G.C. and Pfleger, B.F. (2015) Synthetic Biology toolbox for controlling gene expression in the Cyanobacterium *Synechococcus* sp. strain PCC 7002. ACS Synth. Biol. **4**, 595–603 CrossRef PubMed

7 Quax, T.E.F., Claassens, N.J., Söll, D. and van der Oost, J. (2015) Codon bias as a means to fine-tune gene expression. Mol. Cell **59**, 149–161 CrossRef PubMed
8 Angov, E. (2011) Codon usage: nature's roadmap to expression and folding of proteins. Biotechnol. J. **6**, 650–659 CrossRef PubMed
9 Curran, K.A., Karim, A.S., Gupta, A. and Alper, H.S. (2013) Use of expression-enhancing terminators in *Saccharomyces cerevisiae* to increase mRNA half-life and improve gene expression control for metabolic engineering applications. Metab. Eng. **19**, 88–97 CrossRef PubMed
10 Jones, J.A., Toparlak, Ö.D. and Koffas, M.A. (2015) Metabolic pathway balancing and its role in the production of biofuels and chemicals. Curr. Opin. Biotech. **33**, 52–59 CrossRef
11 Bloom, J.D., Meyer, M.M., Meinhold, P., Otey, C.R., MacMillan, D. and Arnold, F.H. (2005) Evolving strategies for enzyme engineering. Curr. Opin. Struct. Biol. **15**, 447–452 CrossRef PubMed
12 Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L.J. and Keasling, J.D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. Nat. Biotechnol. **27**, 753–759 CrossRef PubMed
13 Boyle, P.M. and Silver, P.A. (2012) Parts plus pipes: synthetic biology approaches to metabolic engineering. Metab. Eng. **14**, 223–232 CrossRef PubMed
14 Parsons, J.B., Frank, S., Bhella, D., Liang, M., Prentice, M.B., Mulvihill, D.P. and Warren, M.J. (2010) Synthesis of empty bacterial microcompartments, directed organelle protein incorporation, and evidence of filament-associated organelle movement. Mol. Cell **38**, 305–315 CrossRef PubMed
15 Mellin, J.R. and Cossart, P. (2015) Unexpected versatility in bacterial riboswitches. Trends Genet. **31**, 150–156 CrossRef PubMed
16 Green, A.A., Silver, P.A., Collins, J.J. and Yin, P. (2014) Toehold switches: de-novo-designed regulators of gene expression. Cell **159**, 925–939 CrossRef PubMed
17 Elowitz, M.B. and Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. Nature **403**, 335–338 CrossRef PubMed
18 Gardner, T.S., Cantor, C.R. and Collins, J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature **403**, 339–342 CrossRef PubMed
19 Guido, N.J., Wang, X., Adalsteinsson, D., McMillen, D., Hasty, J., Cantor, C.R., Elston, T.C. and Collins, J.J. (2006) A bottom-up approach to gene regulation. Nature **439**, 856–860 CrossRef PubMed
20 Lu, T.K., Khalil, A.S. and Collins, J.J. (2009) Next-generation synthetic gene networks. Nat. Biotechnol. **27**, 1139–1150 CrossRef PubMed
21 Ellis, T., Wang, X. and Collins, J.J. (2009) Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat. Biotechnol. **27**, 465–471 CrossRef PubMed
22 Siegele, D.A. and Hu, J.C. (1997) Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. Proc. Natl. Acad. Sci. U.S.A. **94**, 8168–8172 CrossRef PubMed
23 Hammer, K., Mijakovic, I. and Jensen, P.R. (2006) Synthetic promoter libraries – tuning of gene expression. Trends Biotechnol. **24**, 53–55 CrossRef PubMed
24 Jensen, P.R. and Hammer, K. (1998) Artificial promoters for metabolic optimization. Biotechnol. Bioeng. **58**, 191–195 CrossRef PubMed
25 Khlebnikov, A., Datsenko, K.A., Skaug, T., Wanner, B.L. and Keasling, J.D. (2001) Homogeneous expression of the PBAD promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. Microbiol. **147**, 3241–3247 CrossRef
26 Hahn, S. and Young, E.T. (2011) Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics **189**, 705–736 CrossRef PubMed
27 Kelly, J.R., Rubin, A.J., Davis, J.H., Ajo-Franklin, C.M., Cumbers, J., Czar, M.J., de Mora, K., Glieberman, A.L., Monie, D.D. and Endy, D. (2009) Measuring the activity of BioBrick promoters using an *in vivo* reference standard. J. Biol. Eng. **3**, 4 CrossRef PubMed
28 Rudge, T.J., Brown, J.R., Federici, F., Dalchau, N., Phillips, A., Ajioka, J.W. and Haseloff, J. (2016) Characterization of intrinsic properties of promoters. ACS Synth. Biol. **5**, 89–98 CrossRef PubMed
29 De Mey, M., Maertens, J., Lequeux, G.J., Soetaert, W.K. and Vandamme, E.J. (2007) Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering. BMC Biotechnol. **7**, 34 CrossRef PubMed
30 Morgan-Kiss, R.M., Wadler, C. and Cronan, Jr, J.E. (2002) Long-term and homogeneous regulation of the *Escherichia coli araBAD* promoter by use of a lactose transporter of relaxed specificity. Proc. Natl. Acad. Sci. U.S.A. **99**, 7373–7377 CrossRef PubMed
31 Keasling, J.D. (1999) Gene-expression tools for the metabolic engineering of bacteria. Trends Biotechnol. **17**, 452–460 CrossRef PubMed
32 Makoff, A.J. and Oxer, M.D. (1991) High level heterologous expression in *E. coli* using mutant forms of the *lac* promoter. Nucleic Acids Res. **19**, 2417–2421 CrossRef PubMed
33 de Boer, H., Comstock, L.J. and Vasser, M. (1983) The *tac* promoter: a functional hybrid derived from the *trp* and *lac* promoters. Proc. Natl. Acad. Sci. U.S.A. **80**, 21–25 CrossRef PubMed
34 Terpe, K. (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. Appl. Microbiol. Biotechnol. **72**, 211–222 CrossRef PubMed
35 Jensen, P.R., Westerhoff, H.V. and Michelsen, O. (1993) The use of *lac*-type promoters in control analysis. Eur. J. Biochem. **211**, 181–191 CrossRef PubMed
36 Wycuff, D.R. and Matthews, K.S. (2000) Generation of an AraC-*ara*BAD promoter-regulated T7 expression system. Anal. Biochem. **277**, 67–73 CrossRef PubMed
37 Guzman, L.M., Belin, D., Carson, M.J. and Beckwith, J. (1995) Tight regulation, modulation and high-level expression by vectors containing the Arabinose PBAD promoter. J. Bacteriol. **177**, 4121–4130 PubMed
38 Cagnon, C., Valverde, V. and Masson, J.M. (1991) A new family of sugar-inducible expression vectors for *Escherichia coli*. Protein Eng. **4**, 843–847 CrossRef PubMed
39 Valdez-Cruz, N.A., Caspeta, L., Pérez, N.O., Ramírez, O.T. and Trujillo-Roldán, M.A. (2010) Production of recombinant proteins in *E. coli* by the heat inducible expression system based on the phage lambda pL and/or pR promoters. Microb. Cell Fact. **9**, 18 CrossRef PubMed
40 Elvin, C.M., Thompson, P.R., Argall, M.E., Hendry, P., Stamford, N.P.J., Lilley, P.E. and Dixon, N.E. (1990) Modified bacteriophage lambda promoter vectors for overproduction of proteins in *Escherichia coli*. Gene **87**, 123–126 CrossRef PubMed
41 Mellado, R.P. and Salas, M. (1982) High level synthesis in *Escherichia coli* of the *Bacillus subtillis* phage phi 29 proteins p3 and p4 under the control of phage lambda PL promoter. Nucleic Acids Res. **10**, 5773–5784 CrossRef PubMed
42 Simons, G., Remaut, E., Allet, B., Devos, R. and Fiers, W. (1984) High-level expression of human interferon gamma in *Escherichia coli* under control of the *pL* promoter of bacteriophage lambda. Gene **28**, 55–64 CrossRef PubMed
43 Studier, F.W. and Moffatt, B.A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol. **189**, 113–130 CrossRef PubMed
44 Collado-Vides, J., Magasanik, B. and Gralla, J.D. (1991) Control site location and transcriptional regulation in *Escherichia coli*. Microbiol. Rev. **55**, 371–394 PubMed
45 Jensen, P.R. and Hammer, K. (1998) The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. Appl. Environ. Microb. **64**, 82–87
46 Solem, C. and Jensen, P.R. (2002) Modulation of gene expression made easy. Appl. Environ. Microb. **68**, 2397–2403 CrossRef
47 Rytter, J.V., Helmark, S., Chen, J., Lezyk, M.J., Solem, C. and Jensen, P.R. (2014) Synthetic promoter libraries for *Corynebacterium glutamicum*. Appl. Microbiol. Biotechnol. **98**, 2617–2623 CrossRef PubMed
48 Sohoni, S.V., Fazio, A., Workman, C.T., Mijakovic, I. and Lantz, A.E. (2014) Synthetic promoter library for modulation of Actinorhodin production in *Streptomyces coelicolor* A3(2). PLoS One **9**, e99701 CrossRef PubMed
49 Blount, B.A., Weenink, T., Vasylechko, S. and Ellis, T. (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. PLoS One **7**, e33279 CrossRef PubMed
50 Angermayr, M., Oechsner, U. and Bandlow, W. (2003) Reb1p-dependent DNA bending effects nucleosome positioning and constitutive transcription at the yeast profilin promoter. J. Biol. Chem. **278**, 17918–17926 CrossRef PubMed
51 McWhinnie, R.L. and Nano, F.E. (2013) Synthetic promoters functional in *Francisella novicida* and *Escherichia coli*. Appl. Environ. Microb. **80**, 226–234 CrossRef

52 Fischer, C.R., Alper, H., Nevoigt, E., Jensen, K.L. and Stephanopoulos, G. (2006) Response to Hammer et al.: tuning genetic control – importance of thorough promoter characterization versus generating promoter diversity. Trends Biotechnol. **24**, 55–56 CrossRef PubMed
53 Alper, H., Fischer, C., Nevoigt, E. and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. Proc. Natl. Acad. Sci. U.S.A. **102**, 12678–12683 CrossRef PubMed
54 Yim, S.S., An, S.J., Kang, M., Lee, J. and Jeong, K.J. (2013) Isolation of fully synthetic promoters for high-level gene expression in *Corynebacterium glutamicum*. Biotechnol. Bioeng. **110**, 2959–2969 CrossRef PubMed
55 Nevoigt, E., Kohnke, J., Fischer, C.R., Alper, H., Stahl, U. and Stephanopoulos, G. (2006) Engineering of promoter replacement cassettes for fine-tuning of gene expression in *Saccharomyces cerevisiae*. Appl. Environ. Microbiol. **72**, 5266–5273 CrossRef PubMed
56 Blazeck, J., Garg, R., Reed, B. and Alper, H.S. (2012) Controlling promoter strength and regulation in *Saccharomyces cerevisiae* using synthetic hybrid promoters. Biotechnol. Bioeng. **109**, 2884–2895 CrossRef PubMed
57 Jensen, K., Alper, H., Fischer, C. and Stephanopoulos, G. (2006) Identifying functionally important mutations from phenotypically diverse sequence data. Appl. Environ. Microb. **72**, 3696–3701 CrossRef
58 Sinha, S. (2006) On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics **22**, e454–e463 CrossRef PubMed
59 Stromo, G.D. (2000) DNA binding sites: representation and discovery. Bioinformatics **16**, 16–23 CrossRef PubMed
60 Rhodius, V.A. and Mutalik, V.K. (2010) Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, σE. Proc. Natl. Acad. Sci. U.S.A. **107**, 2854–2859 CrossRef PubMed
61 Rhodius, V.A., Mutalik, V.K. and Gross, C.A. (2012) Predicting the strength of UP-elements and full-length *E. coli* σE promoters. Nucleic Acids Res. **40**, 2907–2924 CrossRef PubMed
62 Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C. and Wold, S. (1993) Quantitative sequence-activity models (QSAM)-tools for sequence design. Nucleic Acids Res. **12**, 733–739 CrossRef
63 De Mey, M., Maertens, J., Boogmans, S., Soetaert, W.K., Vandamme, E.J., Cunin, R. and Foulquié-Moreno, M.R. (2010) Promoter knock-in: a novel rational method for the fine tuning of genes. BMC Biotechnol. **10**, 26 CrossRef PubMed
64 Meng, H., Wang, J., Xiong, Z., Xu, F., Zhao, G. and Wang, Y. (2013) Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network. PLoS One **8**, e60288 CrossRef PubMed
65 Meng, H. and Wang, Y. (2015) *Cis*-acting regulatory elements: from random screening to quantitative design. Quant. Biol. **3**, 107–114 CrossRef
66 Buscema, P.M., Massini, G. and Maurelli, G. (2014) Artificial neural networks: an overview and their use in the analysis of the AMPHORA-3 dataset. Subst. Use Misuse **49**, 1555–1568 CrossRef PubMed
67 Davis, J.H., Rubin, A.J. and Sauer, R.T. (2011) Design, construction and characterization of a set of insulated bacterial promoters. Nucleic Acids Res. **39**, 1131–1141 CrossRef PubMed
68 Harley, C.B. and Reynolds, R.P. (1987) Analysis of *E. coli* promoter sequences. Nucleic Acids Res. **15**, 2343–2361 CrossRef PubMed
69 Kanhere, A. and Bansal, M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. Nucleic Acids Res. **33**, 3165–3175 CrossRef PubMed
70 Nair, T.M. and Kulkarni, B.D. (1994) On the consensus structure within the *E. coli* promoters. Biophys. Chem. **48**, 383–393 CrossRef PubMed
71 Estrem, S.T., Gaal, T., Ross, W. and Gourse, R.L. (1998) Identification of an UP element consensus sequence for bacterial promoters. Proc. Natl. Acad. Sci. U.S.A. **95**, 9761–9766 CrossRef PubMed
72 Ross, W., Aiyar, S.E., Salomon, J. and Gourse, R.L. (1998) *Escherichia coli* promoters with UP elements of different strengths: modular structure of bacterial promoters. J. Bacteriol. **180**, 5375–5383 PubMed
73 Hampsey, M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol. Mol. Biol. Rev. **62**, 465–503 PubMed
74 Basehoar, A.D., Zanton, S.J. and Pugh, B.F. (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell **116**, 699–709 CrossRef PubMed
75 Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J. Mol. Biol. **296**, 1205–1214 CrossRef PubMed
76 Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J. et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature **431**, 99–104 CrossRef PubMed

Received 1 February 2016
doi:10.1042/BST20160042