Rapid, Heuristic Discovery and Design of Promoter Collections in Non-Model Microbes for Industrial Applications.

James Gilman [1], Chloe Singleton [1], Richard K. Tennant [1], Paul James [1], Thomas P. Howard [2], Thomas Lux [3], David A. Parker [4] & John Love [1*]

[1] The BioEconomy Centre, Biosciences, College of Life and Environmental Sciences, Stocker Road, University of Exeter, Exeter, EX4 4QD, U.K.

[2] School of Natural and Environmental Sciences, Devonshire Building, Newcastle University, Newcastle-upon-Tyne NE1 7RU, U.K.

[3] Plant Genome and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Munich, Germany.

[4] Biodomain, Shell Technology Center Houston, 3333 Highway 6 South, Houston, Texas 77082-3101, U.S.A.

Running Title:

Promoter Design for Industrial Applications.

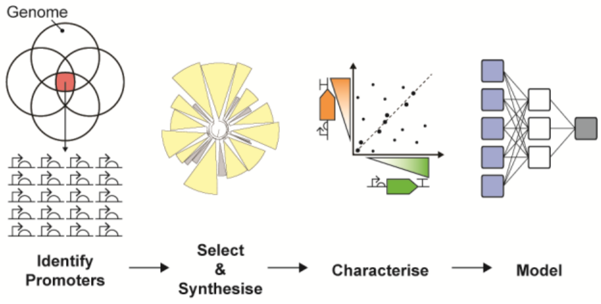
***Author for Correspondence:** J.Love@exeter.ac.uk

Keywords: Promoter Design ; Modelling ; *Geobacillus* ; Industrial Chassis ;

Abstract

Well-characterised promoter collections for synthetic biology applications are not always available in industrially relevant hosts. We developed a broadly applicable method for promoter identification in atypical microbial hosts that requires no *a priori* understanding of *cis*-regulatory element structure. This novel approach combines bioinformatic filtering with rapid empirical characterisation to expand the promoter toolkit, and uses machine learning to improve the understanding of the relationship between DNA sequence and function. Here, we apply the method in *Geobacillus thermoglucosidasius,* a thermophilic organism with high potential as a synthetic biology chassis for industrial applications. Bioinformatic screening of *G. kaustophilus*, *G. stearothermophilus*, *G. thermodenitrificans* and *G. thermoglucosidasius* resulted in the identification of 636 100 bp putative promoters, encompassing the genome-wide design space and lacking known transcription factor binding sites. 80 of these sequences were characterised *in vivo* and activities covered a 2-log range of predictable expression levels. 7 sequences were shown to function consistently regardless of the downstream coding sequence. Partition modelling identified sequence positions upstream of the canonical -35 and -10 consensus motifs that were predicted to strongly influence regulatory activity in *Geobacillus*, and Artificial Neural Network and Partial Least Squares regression models were derived to assess if there was a simple, forward, quantitative method for *in silico* prediction of promoter function. However, the models were insufficiently general to predict *pre hoc* promoter activity *in vivo*, most probably as a result of the relatively small size of the training data set as compared to the size of the modelled design space.

Visual Abstract



The predictable control of genetic modules or engineered metabolic pathways is a defining aspiration of synthetic biology[1] requiring thoroughly characterised, robust genetic parts. Although synthetic biology parts and tools of increasing sophistication are available[2-5], the majority have been designed for use in a small number of model organisms[6] and characterised only or mainly in these biological contexts[7]. Model organisms such as *Escherichia coli* or *Saccharomyces cerevisiae* are invaluable for laboratory-scale, proof-of-principle investigations and are used in some industrial applications[8] but there is a real, practical need to expand the range of microbial chassis available for industrial applications that present more extreme environments for the biocatalyst[9,6,10-13].

Different control points affect the output of gene networks, including levels of transcription, translation, protein half-life and enzyme kinetics[14]. On a practical level, the use of promoters with varied and predictable activation and output characteristics ("strengths") are an essential feature of any synthetic biology toolkit[3,15,14] and are particularly useful for balancing differential expression levels in "hard-wired", steady state genetic modules[16]. Promoter collections for synthetic biology applications should therefore cover a broad range of recombinant gene expression levels for nuanced tuning of synthetic pathways[17], with individual promoters providing homogeneous, consistent and predictable outputs independently of the associated downstream coding sequence[18].

Conventionally, promoters in atypical chassis may be isolated from upstream of genes or operons[15] that are homologous to well-understood regions in model organisms, or identified using genomic or transcriptomic analyses of the host[7] followed by in-depth characterisation in a range of genetic and environmental contexts. Alternatively, synthetic promoter libraries may be manufactured by mutagenesis of wild-type promoter sequences, again followed by deep analysis of novel activity[14,19,20], though this approach tends to reduce, rather than enhance, promoter strength[9,21-25]. Finally, recent advances in DNA synthesis have facilitated systematic approaches to promoter and regulatory sequence design by enabling the production and high-throughput screening of comprehensive sequence libraries[26,27]. Due to the scale of DNA synthesis required, however, this approach remains relatively expensive compared to mutagenesis and dependent on ready access to appropriate DNA synthesis facilities.

In this investigation, we used a bioinformatic approach to explore the promoter design space in *Geobacillus thermoglucosidasius*, a metabolically versatile[11,28-30], thermophilic microbe[31] with high potential as a synthetic biology chassis for industrial applications[6,32]. To date, engineering projects in *Geobacillus* have relied on 1 of 3 endogenous promoter sequences[11,33,34], the most widely used being the oxygen-dependent *ldh*A promoter[9,11,31,35,36]. Mutagenesis-derived, synthetic promoters have also been reported for the genus[9,37,38], though their characterisation is limited to single genetic contexts.

Here, we selected 100 putative promoter sequences from the *Geobacillus* core genome encompassing the genome-wide design space and lacking known transcription factor binding sites. The sequences were synthesised, cloned upstream of 2 different reporter CDS and their activities assessed *in vivo*. This process was relatively rapid and resulted in a collection of 7 characterised promoter sequences that displayed a range of activities with low internal variance and that functioned independently of the downstream reporter sequence. Additionally, to better understand the relationship between promoter sequence and activity, the data from the *in vivo* characterisation were used to train and validate a variety of *in silico* models, including Random Forest partition, Artificial Neural Network (ANN) and Partial Least Squares regression (PLS).

The method presented here is broadly applicable to any potential bacterial chassis and could be used to expand synthetic biology tools for other biocatalysts and ultimately enhance our fundamental knowledge of genetic regulation in synthetic and natural systems.

Results & Discussion

Bioinformatic identification of putative promoters from the core genome of 4 Geobacillus *species*

Different *Geobacillus* species have the potential to be used as host organisms for industrial bioproduction[6, 9, 33]. We therefore aimed to identify promoters that could potentially be used across the entire genus. To obtain a suite of promoters that were representative of the *Geobacillus* genus, we sequenced and assembled *de novo* the genomes of 4 *Geobacillus* species that were available when the project started; *G. kaustophilus* (DSM7263), *G. stearothermophilus* (DSM22), *G. thermodenitrificans* (K1041) and *G. thermoglucosidasius* (DSM2542). To identify genes that were common to all 4 *Geobacillus* species, single-copy coding sequences (CDS) were clustered into homologous gene families using the GET_HOMOLOGUES software package[39]. To increase calculation robustness, 3 separate clustering algorithms were used, and the resulting gene families compared. Bidirectional best-hit (BDBH), COG triangles (COG) and OrthoMCL (OMCL) algorithms returned 1,924, 1,914, and 1,902 CDS clusters respectively, with 1,886 homologous clusters being identified by all 3 algorithms (Figure 1A). The core genome of the selected *Geobacillus* species therefore contained 1,886 CDS; *i.e.* a total of 7,544 homologous core CDS.

In prokaryotes, the majority of motifs that affect the initiation of both transcription and translation occur in the 100 bp sequence window immediately upstream of the CDS start codon[40,41]. 100 bp sequences from immediately upstream of the start codon of the 7,544 core CDS were therefore identified as putative *Geobacillus* promoter sequences. BPROM software was subsequently used to classify the 100 bp sequences as putative promoters based on the presence and nucleotide composition of known conserved functional motifs[42]. To isolate sequences that were likely orthogonal to endogenous regulatory pathways, putative promoters were screened against BPROMs list of known Transcription Factor Binding Sites (TFBS, Supporting Table 1), and sequences that contained any known TFBS were discarded. A phylogeny of the 1,489 putative, generic sequences that remained after screening was constructed as a representation of the *Geobacillus* promoter design space (Figure 1B). Although BPROMs list of *E. coli* TFBS may not be exhaustively representative of binding sites that are functional in *Geobacillus*, the lack of extensive genus-specific TFBS characterisation in these non-model organisms renders a genus-specific approach impractical. Given previous successfully applications of

BPROM software for promoter identification[28], the utilised list of TFBS was judged likely to provide an adequately generic reference for binding site recognition in *Geobacillus*.

Multiple studies have used promoters isolated from the genomes of bacteriophage for the control of heterologous expression in *E. coli*[14]. Putative promoters were therefore also identified from the genomes of 2 bacteriophages, *Thermus* phage Phi OH2 and *Geobacillus* phage GBSV1, which were chosen due to their ready availability on the GenBank public database. Intergenic regions of at least 100 bp were identified in both genomes. 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 analysed using BPROM software to identify putative promoters, and any sequences that contained known TFBS were discarded. 9 putative promoters were identified from *Thermus* phage Phi OH2, and 7 putative promoters were identified from *Geobacillus* phage GBSV1.

In vivo *characterisation of putative promoters*

A number of studies have considered the effect of genetic context on promoter function in model organisms such as *E. coli* and *S. cerevisiae*[18,41,43-45]. However, the drive for composable, modular regulatory elements in non-model systems is hindered by the fact that many studies still characterise the function of promoter sequences in a single genetic context. 2 previously published *Geobacillus* synthetic promoter libraries, for example, used only GFP to characterise promoter performance[9,37]. Putative promoters were therefore characterised upstream of both Dasher GFP and mOrange fluorescent reporters.

A trade-off was required between the desire to empirically explore large portions of the *Geobacillus* promoter design space and the experimental feasibility of characterising large numbers of putative sequences in a host organism with low transformation efficiencies. The promoter phylogeny (Figure 1B) was therefore used to rationally select 100 putative promoters from across the *Geobacillus* promoter design space for *in vivo* characterisation using both reporters.

A sequence alignment of the 100 selected putative promoters revealed a heavily conserved purine-rich region located at the 3' terminus of the 100 bp sequence space (Supporting Figure 1). Given the similarities in both location and nucleotide composition of the motif to the canonical Shine-Dalgrano sequence[46], this region was identified as the RBS. We therefore changed the terminology, whereby "promoter" refers to the complete 100 bp sequence, RBS refers to the 15 bp of sequence at the 3' terminus of the sequence space and Distal Regulatory Sequence (DRS) refers to the sequence from -100 to -15 bp upstream of the start codon.

To facilitate potential future applications of the promoter sequences in which disparate DRS and RBS might be required, the 100 selected putative promoters were split *in silico* into DRS and RBS parts that were subsequently flanked with type IIs restriction cloning affixes (Supporting Table 2). *In vitro* cloning of the DRS and RBS parts resulted in the insertion of a 4 bp scar sequence at -19 to -16 bp upstream of the start codon, increasing the length of the promoters to 104 bp. The inclusion of the scar sequence was empirically shown to have no statistically significant effect on promoter activity for 20 out of a set of 24 characterised sequences, with significant alterations in regulatory activity hypothesised to be the result of extreme alterations to mRNA secondary structure (Supporting Information, Supporting Figure 2).

Of the 100 selected putative *Geobacillus* promoters, 5 *promoter::GFP* and 9 *promoter::mOrange* constructs could not be successfully synthesised. Furthermore, 11 *promoter::mOrange* constructs could not be transformed into *G. thermoglucosidasius*; 80 sequences were therefore characterised *in vivo* upstream of both reporters (Figure 2A). The characterised sequences covered a 148-fold range of activity when characterised upstream of GFP, and a 107-fold range of activity when characterised upstream of mOrange. 45 of the characterised promoters showed expression levels for both reporter proteins that were not statistically significantly greater than the negative control, *G. thermoglucosidasius* transformed with the empty pS797 vector. We therefore defined these 45 sequences as inactive. 19 out of the 100 screened promoters showed statistically significant activity with both reporters; 3 sequences were active with GFP only, and 13 sequences were active with mOrange only (Figure 2B). A comparison of the codon usage of the 2 reporter proteins showed them to be broadly comparable (Supporting Figure 3). The discrepancies in gene expression between the 2 reporters were therefore assumed to be a result of promoter activity, rather than differential codon utilisation.

To identify the promoters that functioned predictably and independently of the downstream CDS, K-means clustering was used to group the characterised sequences into 5 clusters based on their Euclidean distance from the line of equivalence between GFP and mOrange activity, $y = x$ (Figure 2C). No correlation in *in vivo* activity between the two reporter proteins was observed for the majority of the characterised sequences; clusters 2 and 4 contained promoters that resulted in stronger GFP expression than mOrange expression, whereas clusters 3 and 5 resulted in stronger mOrange than GFP expression. Clustering identified 13 promoters (cluster 1) with activity that fell close to the line of equivalence, of which 7 displayed mean expression levels that were significantly greater than the negative control. The characterised *Geobacillus* promoter library therefore contained 7 functionally composable, active sequences, covering activity levels that were between 1.1 and 4.5 times greater than the *G. thermodenitrificans ldhA* positive control.

Such functional composability of *cis*-regulatory sequences is crucial if information regarding promoter performance derived from laboratory-scale characterisation experiments is to be applied to the systematic, scalable, bottom-up engineering of increasingly complex synthetic biological systems[4,18]. The development of species-specific insulator mechanisms, that reduce the context-specificity of regulatory parts through either molecular transcript processing[47,48] or by physically separating genetic regulatory parts to disrupt context-specific mRNA secondary structures[18,41], is required if the majority of the identified promoters are to be used modularly in alternative contexts.

In addition to being functionally composable, promoter sequences for synthetic biology applications should ideally yield homogenous, predictable expression of the protein of interest at the single-cell level[49]. Flow cytometry was therefore used to analyse the intra-population variation in fluorescence activity of the characterised *promoter::reporter* fusions in transformed, clonal cultures. Compared to the positive control, the *G. thermodenitrificans ldhA* promoter, 98% of the characterised *promoter::GFP* fusions and 73% of the *promoter::mOrange* fusions returned lower coefficients of variance, indicating that the majority of the characterised sequences offered more predictable regulation of protein expression than the current benchmark *Geobacillus* promoter. Furthermore, the 7 promoters that functioned independently of coding sequence all returned lower coefficients of variation than the positive control *ldhA* promoter (Figure 2D). Although subpopulations of cells expressing the reporters were apparent for 4 of the characterised promoters, the performance of these promoters was less variable and therefore more predictable than that of the *ldhA* promoter which has been widely used in studies with potential industrial applications[9,11,31,35,36].

Analysis of the genes with which the 80 characterised promoters were natively associated in their source genomes showed that the majority of the sequences homogeneously regulate basic cellular functions, and were therefore likely to be constitutive (Supporting Table 3). Cellular functions with which the promoters were natively associated included biosynthesis, cell membrane formation, catabolism, transcription and protein folding. However, 11 of the characterised promoters were natively associated with proteins relating to sporulation, and may therefore result in altered expression levels under sporulation conditions. The failure of the bioinformatic screening to identify and exclude these sequences highlights the limitations of applying bioinformatic tools that were developed in *E. coli* in non-model organisms; as *E. coli* is non-sporulating, a list of *E. coli* TFBS will naturally not contain sporulation-specific TFBS.

Sequence-function modelling

Mathematical models with the *pre hoc* capability to determine promoter function could potentially reduce the need for *in vivo* characterisation of large numbers of individual *cis*-regulatory elements. Once a training set of sufficient robustness is established, regulatory elements of the desired strength for a given application could hypothetically be identified from the genome or designed *de novo*, in a manner analogous to tools such as the RBS calculator[3]. To better understand the basis of promoter function in *Geobacillus,* and to assess if there was a simple, forward method for *in silico* prediction of promoter function, statistical learning approaches were used to derive models of the design space.

We used a variety of techniques to mathematically describe the relationship between DNA sequence and function of the promoters characterised above. Partition modelling was used to identify positions within the sequence space that were having the greatest impact on promoter activity, and ANN and PLS models were subsequently used to make quantitative predictions of promoter activity.

Partition modelling

Recursive partition modelling is a powerful technique for determining the relationship between a response variable and a set of independent variables without the use of a mathematical model[50]. Partition models were fit to both the GFP and mOrange characterisation data sets. The number of times each promoter sequence position caused partitions in the data set across 100 random forests was quantified; the larger the number of partitions caused by a sequence position, the more important that position was predicted to be in determining promoter activity.

Sequence positions across the entirety of the sequence space were predicted to strongly influence regulatory activity for both reporters (Figure 3). In particular, sequence positions towards the 5' terminus of the sequence space were predicted to be important in determining promoter activity. This result suggested that UP elements, sequence motifs that are further upstream than the canonical RBS, -10 and -35 motifs and that boost transcription initiation through interactions with the C-terminal domain of the RNA polymerase alpha subunit[51,52], are active in *Geobacillus*.

Artificial Neural Network & Partial Least squares sequence-function modelling

Although the partition models provided useful insights to the relationship between promoter nucleotide sequence and function, they did not provide quantitative predictions of regulatory activity. We therefore applied 2 quantitative modelling approaches, linear Partial Least Squares (PLS) regression and non-linear Artificial Neural Networks (ANN).

To asses the predictive capability of PLS and ANNs when applied to *Geobacillus cis*-regulatory sequences, models were trained using data derived from the 95 characterised *promoter::GFP* fusions (Supporting Figure 4). In all instances, each of the 104 nucleotide positions within the promoter sequence was modelled as an individual *x* variable and GFP fluorescence was used as the response variable, *y*.

ANNs have previously been shown to return insufficiently accurate predictions when the response surface under investigation is complex and the number of observations in the training data set is small[53]. Furthermore, although the PLS algorithm was specifically designed to model data sets in which the number of predictor variables is greater than the number of observations in the training set [54], the extreme scale of the promoter design space (there are 4^{100} potential 100 bp nucleotide sequences) compared to the number of empirically characterised promoters was thought likely to result in models with limited predictive power. A reduction in the dimensionality of the modelled design space was therefore deemed necessary.

Characterising promoters of shorter length would have immediately reduced the dimensionality of the modelled design space. For example, 50 bp sequences would have been of sufficient length to contain the canonical location of the RBS, -10 and -35 consensus motifs. However, the partition results showed that sequence positions upstream of the -50 position were likely to be important in determining regulatory activity (Figure 3). Sequences of reduced length would therefore not have contained vital upstream regulatory motifs and may therefore have shown reduced activity as compared to the longer sequences.

The results of the partition modelling were therefore used to reduce the dimensionality of the modelled design space. PLS and ANN sequence-function models were derived that modelled GFP fluorescence as a function of varying number of nucleotide positions. Sequence positions were selected in descending order of the number of partitions caused in the 100 partition models (Figure 3). In all instances, model performance was quantified using an independent test set of 10 promoter sequences that were held-back from model training and validation.

The optimum PLS model that was obtained inferred promoter activity as a function of 20 nucleotide positions (Supporting Figure 5). The model returned an R^2 value of 0.6024 when applied to the training and validation data sets, and an R^2 value of 0.8901 when applied to the test set (Figure 4). These results suggested that the obtained PLS model provided a reasonable fit of the training data and had good predictive power when applied to previously unseen data.

A Design of Experiments (DoE) approach was used to optimise ANN architecture (Supporting Information). In total, over 113,500 single-layer ANNs were fit, varying in terms of the personality of the activation function used, the number of nodes in the hidden layer, the cross validation methodology and the number of promoter sequence positions modelled.

The optimal obtained ANN was an ensemble model that contained 2 constituent ANNs. Each of the constituent models used sigmoidal activation functions with 5 nodes in the hidden layer, and modelled promoter activity as a function of 20 nucleotide sequence positions. The optimal model returned an R^2 value of 0.9746 when applied to the training and validation data sets, and an R^2 value of 0.9691 when applied to the test set, suggesting a good fit of the training data and strong predictive power (Figure 4). For both ANN and PLS, models that inferred promoter activity as a function of complete 100 bp sequences showed lower predictive accuracy than models of reduced numbers of sequence positions (Supporting Information). This result validated the use of partition modelling to reduce the size of the modelled design space.

Predicting the function of previously uncharacterised promoters

To further test the predictive power of the putatively high-performing PLS and ANN models, a secondary test set of previously uncharacterised *Geobacillus* promoters was selected. 10 putative regulatory sequences were selected at random from across the promoter phylogeny (Figure 1A) and characterised in *G. thermoglucosidasius* upstream of GFP. However, despite the strong performance of the 2 models on the primary test set, neither model returned accurate predictions of promoter activity for the selected sequences (Figure 4); the PLS model returned an R^2 value of 0.3595 and the ANN returned an R^2 value of 0.2283. Consequently, the derived models were insufficiently general to permit accurate predictions of endogenous promoter activity or facilitate rational, forward promoter design.

Future applications of promoter sequence-function modelling

The lack of generality shown by the models derived in this investigation was probably the result of the limited number of characterised promoter sequences as compared to the scale of the design space, resulting in training set that does not adequately capture the complexity of the response surface. Although PLS and ANN promoter sequence-function models using comparatively small data sets have been described [55-57], the promoter libraries used in these studies contained considerable sequence homology, thereby restricting the complexity of the response surface under investigation. If accurate predictive models of more complex promoter design spaces are to be obtained, a training data set that contains several orders of magnitude more promoter sequences than the 80 sequences used here is likely necessary[7, 26, 43]. However, the scale of the required promoter libraries might be impractical in non-model organisms.

Although high-throughput characterisation of libraries containing thousands of genetic parts using techniques such as a combination of flow cytometry and multiplexed DNA or RNA sequencing has been previously described[7, 26, 43], such approaches require the acquisition of large numbers of transformants; approximately 50-fold library coverage is necessary to achieve accurate characterisation of individual promoters[43]. However, low transformation efficiencies in many non-model organisms, including *Geobacillus*, preclude the production of libraries of the required scale, potentially limiting the usefulness of statistical sequence-function modelling in these contexts.

In lieu of a massive increase in the number of characterised sequences, the novel bioinformatic approach to promoter identification that was developed in this investigation, coupled with partition modelling to identify those sequence positions that are key for determining promoter activity, could be used to provide an initial screen of the design space in organisms for which understanding of *cis*-regulatory sequences is limited. This information could subsequently be used for DoE inspired promoter optimisation in future studies by facilitating the rational design of limited sequence libraries that vary only at the identified key positions. *In vivo* characterisation and *in silico* modelling of the designed libraries could potentially yield models of greater predictive power than those derived here without the need for a large-scale increase in characterisation throughput.

The models that were derived in this study were based purely on the statistical likelihood of a given nucleotide occurring at a given position within the promoter sequence. Measures of biophysical promoter properties, such as mRNA secondary structures, AT content or the free energy barrier for promoter-RNA polymerase binding were not included on the basis that unsupervised ANN models could potentially learn the effect of biophysical promoter properties without specific terms being explicitly defined in the model. The inclusion of biophysical terms in future modelling attempts may facilitate the derivation of more accurate predictive models[26,43,58] by providing more information about promoter function than can be gleaned from sequence data alone. Alternatively, the use of distance metrics[59] as model terms to quantitatively define differences in nucleotide sequence between promoters might also allow for more accurate mapping of the promoter sequence-function design space[60].

Finally, although the quantitative sequence-function models derived in this investigation were insufficiently general to determine *pre hoc in vivo* promoter activity, the potential for statistical modelling to enhance our fundamental knowledge of genetic regulation in complex systems cannot be overlooked. For example, partition modelling of the relationship between nucleotide sequence and *in vivo* promoter function yielded potentially useful insights into the structure of cis-regulatory elements in *Geobacillus*; regions of sequence upstream of the likely position of canonical promoter motifs were predicted to be important in determining promoter activity (Figure 3).

Conclusion

We developed a generally applicable method for the identification of constitutive promoters that combines bioinformatic filtering, empirical characterisation and machine learning to expand promoter toolkits in atypical host organisms and increase the understanding of the relationship between DNA sequence and function. The method was used to identify 80 promoters, covering a 2-log range of predictable expression levels, in *G. thermoglucosidasius*, of which 7 were shown to function consistently regardless of downstream coding sequence. Although sufficiently general *in silico* models of promoter activity could not be obtained using ANN or PLS, partition modelling identified regions of sequence upstream of the canonical prokaryotic promoter consensus regions that strongly influenced regulatory activity in *Geobacillus*.

Materials & Methods

Bacterial strains & plasmids

Type strains of *Geobacillus kaustophilus* (DSM7263), *G. stearothermophilus* (DSM22) and *G. thermoglucosidasius* (DSM2542) were obtained from the DSMZ (Brunswick, Germany). Cultures were freeze-dried ampoules and rehydrated as required following the DSMZ standard protocol. *G. thermodenitrificans* (K1041) was obtained from ZuvaSyntha Ltd. (Hertfordshire, UK).

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 (genotype: *recA pro hsdRm RP4-Tc::Mu-Km::Tn7)* was used as the mobilisation host for the conjugal transformation of *Geobacillus* spp. 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[7,61].

All putative promoter sequences were characterised *in vivo* using the pS797 vector (Supporting Figure 6). To facilitate conjugal transformation of *Geobacillus* spp., pS797 contained an origin of transfer (ORI T), comprised of the Nic region and *traJ* gene from the conjugal plasmid RP4. pS797 also contained 2 origins of replication, ColE and BST1, to allow for propagation in *E. coli* and *Geobacillus* spp., respectively. 2 antibiotic selection markers were also present, allowing for selection by Ampicillin in *E. coli* and by Kanamycin in *Geobacillus*.

Both *E. coli* S17-1 and pS797 were obtained from ZuvaSyntha Ltd. (Hertfordshire, UK).

Growth media

All complex growth media were purchased from Becton Dickson UK (Berkshire, UK). *E. coli* cultures were propagated in Lysogeny Broth (LB; 10 g l^{-1} tryptone, 10 g l^{-1} NaCl, 5 g l^{-1} yeast extract). Lennox Lysogeny Broth (LLB; 10 g l^{-1} tryptone, 5 g l^{-1} NaCl, 5 g l^{-1} yeast extract) was used for co-culture of *E. coli* and *G. thermoglucosidasius* during conjugal transformation of *G. thermoglucosidasius*. All *Geobacillus* species were propagated in modified LB (mLB). mLB used a basal composition of LLB, supplemented with 1.05 mM $C_6H_9NO_6$, 0.91 mM $CaCl_2$, 0.59 mM $MgSO_4$ and 0.04 mM $FeSO_4$[62].

For all media types, agar was supplemented as required to 15 g l^{-1}. When required, *E. coli* growth media was supplemented with 100 µg ml^{-1} ampicillin. *G. thermoglucosidasius* growth media was supplemented with 12.5 µg ml^{-1} kanamycin.

Bioinformatic identification of putative promoters from the core genome of 4 Geobacillus *species*

The genomes of 4 *Geobacillus* species, *G. kaustophilus* (DSM7263), *G. stearothermophilus* (DSM22), *G. thermodenitrificans* (K1041) and *G. thermoglucosidasius* (DSM2542) were sequenced and *de novo* assembled. Genomes were sequenced using an Illumina MiSeq system, using reads with 300 bp paired end sequencing. The resulting raw sequencing reads were trimmed based on quality score using the fastq-mcf tool[63] and assembled using SPAdes software (Version 3.5[64]). Following assembly, the genome scaffolds were annotated using Prokka software (Version 1.9[65]).

The GET_HOMOLOGUES software package[39] was used to identify gene families with homologues in all 4 of the *Geobacillus* species of interest. To increase calculation robustness, 3 disparate algorithms were used to cluster homologous gene families: Bidirectional best-hit (BDBH), COGtriangles (COG) and OrthoMCL (OMCL). In all instances, the "-t" option was used to isolate only those clusters that contained single-copy proteins. All other software parameters were set as default. Only those clusters that were common to all 3 algorithms were selected for further analysis.

Once identified, the core coding sequences were extracted from the 4 genomes. Output files were parsed, reformatted to GenBank file format and imported into the Artemis genome browser[66]. For each entry, the 100 bp immediately upstream of the start codon was extracted. BPROM software[42] was subsequently used to screen the extracted 100 bp sequences for the presence and nucleotide composition of functional regulatory motifs. Additionally, putative promoters were screened against BPROM's list of known Transcription Factor Binding Sites (TFBS, Supporting Table 2). Any putative promoters containing TFBS were discarded.

The nucleotide sequences of the putative promoters were aligned using MUSCLE software[67] and the resultant alignments were used to construct a phylogenetic tree using FastTree software[68]. Putative promoters were subsequently manually clustered into 21 clades using FigTree software[69]. Putative regulatory sequences sequences were selected at random for *in vivo* characterisation from these 21 clades. True randomness was achieved by using a random number generator that converted atmospheric noise into numerical values[70]. Initially, those promoters that were selected for *in vivo* characterisation were manually checked using the Artemis genome browser to ensure that they did not overlap with any adjacent coding sequences. Later, to expedite this process, BEDTools intersect[71] was used to identify those putative promoters which were non-overlapping.

Putative promoters were aligned to transcripts of each of the 4 *Geobacillus* species using Bowtie 2 software[72]. Indexes of the genome files were prepared using the "build" command. Putative regulatory sequences were subsequently aligned to each *Geobacillus* genome using Bowtie 2, with the resultant alignments provided in .sam format. The alignment .sam files were converted to .bam format, sorted and indexed using SAMtools[73]. The resultant alignments were compared against the 4 selected *Geobacillus* genomes using BEDTools intersect. The "-v" command was used to report only those putative promoters that 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 software[74].

Bioinformatic identification of putative promoter sequences from bacteriophage

The genomes of 2 bacteriophages, *Thermus* phage Phi OH2 (NC_021784) and *Geobacillus* phage GBSV1 (NC_008376[75]) were selected for analysis based on their ready availability from the GenBank database. The retrieved GenBank files were loaded into the Artemis genome browser[66] and suitable intergenic regions of at least 100 bp length were manually identified. The 100 bp nucleotide sequences immediately upstream of the adjacent CDS were extracted and analysed using BPROM software[42] to identify putative promoters. Putative promoter sequences were screened against BPROMs list of known TFBS, and any sequences that contained known TFBS were discarded.

Selection, synthesis and cloning of putative promoters for in vivo *characterisation*

Following bioinformatic filtering, putative promoters were synthesised and independently cloned upstream of the coding sequences of 2 reporter proteins, Dasher GFP and mOrange[76] (Supporting Figure 6). The *Geobacillus* promoter phylogeny (Figure 1B) was used to rationally select putative regulatory sequences for *in vivo* characterisation in *G. thermoglucosidasius*. To maximise the portion of the design space that was empirically explored, at least 2 putative promoters were selected at random from each of the 13 clades of the phylogeny that contained more than 50 sequences. 2 putative promoters were also selected from each of the analysed phage genomes. Initial characterisation of the bacteriophage promoters showed that only 1 out of the 4 selected sequences was active in *G. thermoglucosidasius* (Supporting Figure 7). This 1 active bacteriophage promoter was added to 99 putative promoters from the *Geobacillus* phylogeny to create a set of 100 putative regulatory sequences.

The 100 selected putative promoters were synthesised and cloned into the pS797 vector (Supporting Figure 6). In all instances, the reporter CDS (GFP or mOrange) was followed by the S718 terminator from the *G. thermodenitrificans* NG80 2-oxoglutarate ferrodoxin oxioreductase subunit beta[77]. Putative regulatory sequences were either directly synthesised upstream of the relevant reporter CDS in pS797 by ATUM (Previously DNA 2.0, California, USA), or were synthesised as double stranded fragments by IDT (Illinois, USA) and cloned *in vitro* upstream of the relevant reporter CDS.

A type IIs restriction cloning methodology [78,79] was used to join DNA parts. Parts were flanked with unique cloning affixes (Supporting Table 3) containing BsaI restriction sites. Part-specific post-digestion overhangs ensured that digested fragments were only able to ligate in a defined manner. In instances where putative promoters were synthesised by ATUM, the scar sequences that would have resulted form *in vitro* cloning of DRS and RBS were inserted into the sequence *in silico* prior to synthesis.

For *in vitro* cloning, terminator and reporter sequences were synthesised by ATUM in the pJ201 cloning vector. Cloning reactions consisted of 20 fmol of each of the pS797 destination vector and the relevant cloning vectors, with 10 U Bsa1 restriction endonuclease and 1 U T4 DNA ligase in 2 µl ligation buffer (10x Thermo Scientific FastDigest buffer supplemented with 0.5 mM ATP). Final reactions were made up to 20 µl with ddH$_2$O. 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 below. Plasmid construction was verified by diagnostic digest, gel electrophoresis and Sanger sequencing.

Transformation of chemically competent E. coli

E. coli S17-1 were made chemically competent using a modified version of the protocol described by Hanahan[80]. 5 ml overnight cultures of *E. coli* S17-1 were used to inoculate 40 ml LB at a 1:1000 dilution. Inoculated 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 resuspended in 8 ml transformation buffer 1 (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 $MnCl_2.4H_2O$). Resuspended cells were subsequently incubated on ice for 15 min, and harvested as above. The resulting cell pellet was resuspended 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.

For 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 LB was added and the resulting samples were incubated at 37 °C, with shaking at 220 rpm, for 60 min. After incubation, samples were harvested by centrifugation at 4,300 *g* for 5 min, and 500 µl of the supernatant was removed. The cell pellet was resuspended in the remaining supernatant, 200 µl of which was subsequently plated out onto LB agar plates, with antibiotic selection as required. Plates were incubated at 37 °C for 16 h.

Conjugal transformation of G. thermoglucosidasius

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 and centrifuged at 4,300 *g* for 5 min. The supernatant was removed, and the resultant pellet re-suspended in a further 600 µl LLB. Approximately 10-15 µl wild-type *G. thermoglucosidasius* 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, and used to create dilutions of 1:10 and 1:5 biomass to sterile LLB. 200 µl aliquots of each dilution were spread onto separate mLB agar plates containing 12.5 µg ml^{-1} kanamycin. Plates were incubated at 55 °C for approximately 65 h.

In vivo *characterisation of promoter activity*

To prepare starter cultures of *G. thermoglucosidasius* for promoter characterisation, transformants 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. Bacterial suspensions were then used to inoculate mLB to an OD_{600} of 0.1, with antibiotic selection as required.

3 200 µl sample aliquots per transformant were loaded onto 96-well plates using either a Corbett Robotics CAS-1200 (Qiagen, Netherlands) or a Gilson Pipetmax 268 (Gilson Inc., Wisconsin, USA). To minimise the effect of position dependant bias, to which assays performed in a 96-well plate format can be susceptible[81], sample aliquots were loaded in a Latin rectangle design; no transformant was represented more than once on any given row or column of the microplate (Supporting Figure 8). 96-well plates with lid covers have been shown to suffer from significant loss of culture in the outermost wells through evaporation[82]. To account for such edge effects, wells at the plate periphery were filled with 200 µl aliquots of sterile growth media. Microplates were incubated using PHMP Thermoshakers (Grant Instruments, UK). Incubation was at 60 °C, with shaking at 800 rpm.

Population-level measurements of culture absorbance and fluorescence were taken using a Tecan Infinite 200 PRO microplate reader (Tecan, Switzerland). 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, the gain of the instrument was set at 56. Absorbance of all cultures was measured at 600 nm.

Single-cell measurements of fluorescence activity were obtained using a BD FACS Aria II Fluorescence Activated Cell Sorter (FACS), equipped with a 100 µm nozzle. A sheath fluid of Phosphate Buffered Saline was used. Culture fluorescence was excited at 488 nm and fluorescence intensity was recorded using a 530/30 nm detector in the case of GFP fluorescence, and a 585/42 detector in the case of mOrange fluorescence. 100,000 events were recorded per population.

Promoter sequence-function modelling

All sequence-function modelling was performed using JMP pro versions 12 & 13 (SAS Institute Inc., North Carolina, USA).

Partition modelling

100 random forest models were generated for each of the GFP and mOrange characterisation 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. Each tree was trained on a data set of 26 randomly selected promoter sequence positions, drawn with replacement.

To generate partition trees, the selected sequences were divided into groups that differed maximally in terms of the response of interest. For example, the maximum difference in expression activity between 2 groups of promoters might be obtained by splitting the training data into a group of sequences with guanine residues at the -15 position, and another group where adenine, cytosine or thymine residues are present at the -15 position (Supporting Figure 9). The resulting sub-groups were further divided, resulting in the formation of a tree like structure. By repeating the process multiple times on different, randomly selected portions of the training data, a “forest”[83] of decision trees was formed. Across the entire forest, the more times a given factor caused a split in the data set, the better that factor was predicted to be at explaining variation in the response of interest.

Selection of an independent test set for PLS & ANN modelling

To provide an independent test set on which to measure the predictive power of the derived models, 10 promoter sequences were selected and withheld from model training and validation. So that the test set contained promoters with a range of activity levels, the distribution of GFP expression levels of the 95 characterised sequences was analysed. 2 sequences were subsequently selected at random from the 1st distribution quartile, 5 promoters were selected from the interquartile range and 3 sequences were selected from the 4th quartile.

Partial Least squares sequence-function modelling

PLS models were trained that modelled GFP fluorescence as a function of varying numbers of sequence positions. The number of sequence positions modelled was systematically increased from 10 to 50 in increments of 5. Models that fit fluorescence as a function of the complete 104 bp promoters were also generated. For each of the 10 potential groups of *x* variables, multiple PLS models were fit using the non-iterative linear PLS (NIPALS) algorithm and using either KFold or holdback cross validation to optimise the number of latent variables that were extracted from the original data, with a maximum of 10 latent variables permitted per model. Once trained and validated, the models were used to make predictions of activity for the 10 promoters in the withheld test set (Supporting Figure 5). The optimum model was judged to be the one that returned the highest R^2 and lowest Root Average Squared Error (RASE) value when applied to the test set; i.e. the model that had the lowest prediction error.

Artificial Neural Network sequence-function modelling

ANNs were fit using the multilayer perceptron algorithm of JMP software with sigmoidal activation functions. Network architecture was optimised using a Design of Experiments approach (Supporting Information).

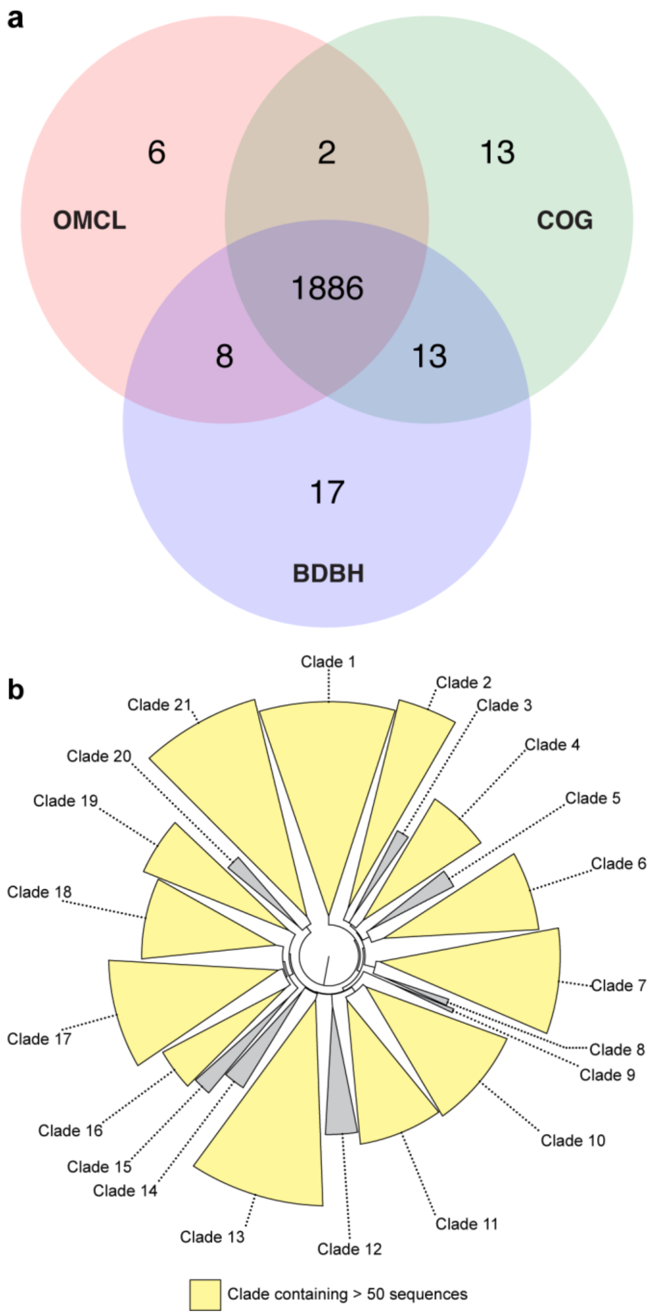


Figure 1: Bioinformatic identification of putative promoter sequences.

A) Venn diagram showing the number of homologous gene families identified in the genomes of the 4 selected *Geobacillus* species by Bidirectional best-hit (BDBH), COG triangles (COG) and OrthoMCL (OMCL) clustering algorithms.

B) Phylogeny of putative promoters, rooted at the midpoint. At least 2 putative promoters were selected at random for *in vivo* characterisation from each of the clades containing > 50 sequences (highlighted in yellow).

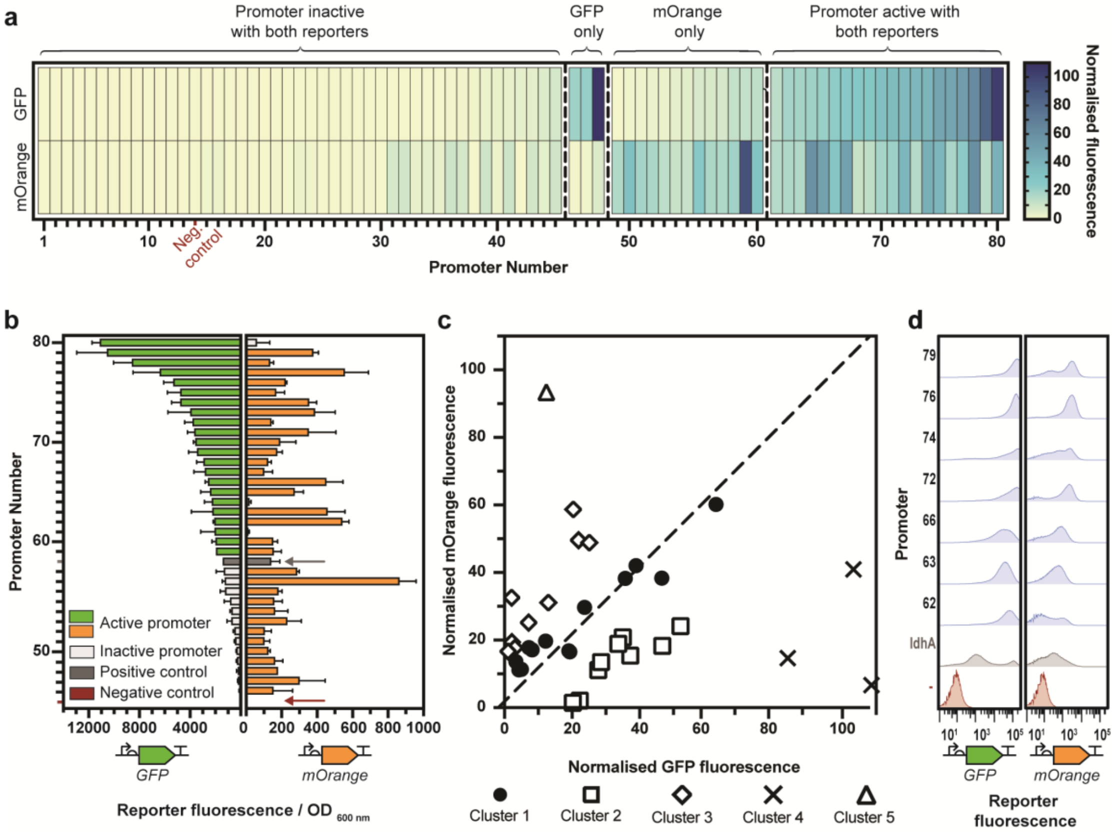


Figure 2: *In vivo* characterisation of bioinformatically identified promoter sequences.

Bioinformatically identified putative promoter sequences were synthesised upstream of *GFP* and *mOrange* reporter sequences, and promoter activity in *G. thermoglucosidasius* was characterised after 24 h growth. In all instances, the positive control, the *G. thermodenitrificans ldhA* promoter is shown in dark grey, and the negative control, *G. thermoglucosidasius* transformed with an empty pS797 vector, is shown in red.

A) Heat map of GFP and mOrange expression levels of the 80 characterised promoters. Each column represents a disparate promoter. To account for differences in intensity between GFP and mOrange fluorescence signals, the mean fluorescence output of each *promoter::reporter* fusion was normalised to the fluorescence output of the negative control, *G. thermoglucosidasius* transformed to express the empty pS797 vector, at the relevant excitation and emission wavelengths. Regulatory sequences were defined as active if reporter fluorescence was statistically significantly greater than the negative control at the relevant wavelengths.

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

B) Expression levels of the promoters for which fluorescence activity was statistically significant. Bars represent the mean of n = 3 independent starter cultures arising from independent transformation events, except in the case of the negative controls, where n = 14, and the positive controls, where n = 11. Error bars represent standard deviation.

C) GFP and mOrange expression levels are normalised to the negative control. Points represent individual promoter sequences. Promoter groupings were determined by K-means clustering based on the Euclidian distance of the points from the line of equivalence, $y = x$, which is represented by the dashed line.

D) Expression levels of the 7 promoters that functioned consistently regardless of CDS, as determined by flow cytometry. For each *promoter::reporter* fusion and the negative control, 100,000 events from each of 3 independent starter cultures arising from independent transformation events were combined to form a single "meta" population of 300,000 events. + = *ldhA* positive control; - = negative control.

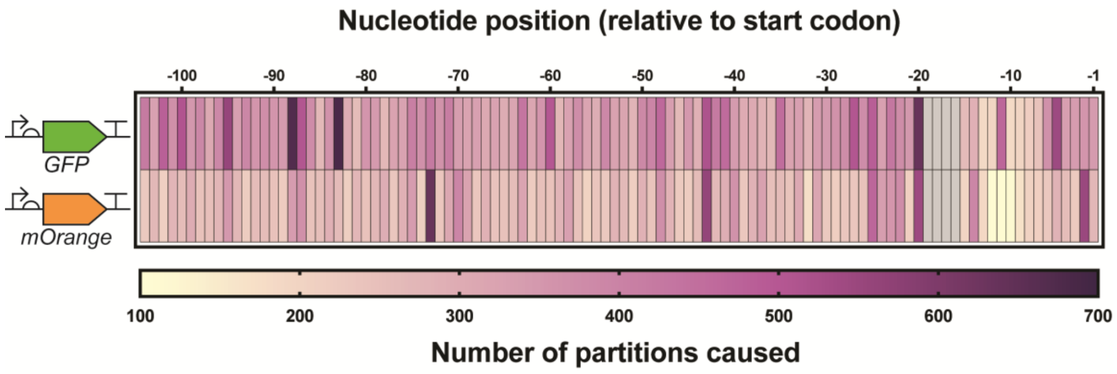


Figure 3: Heat map showing the number of data set partitions caused in 100 random forests by individual regulatory sequence nucleotide positions when either GFP or mOrange fluorescence was used as the response variable.

The grey region represents the ACCT cloning scar between the Distal Regulatory Sequence (DRS) and RBS regions. As all of the characterised promoters were identical in these locations, these 4 positions were not included in the partition modelling.

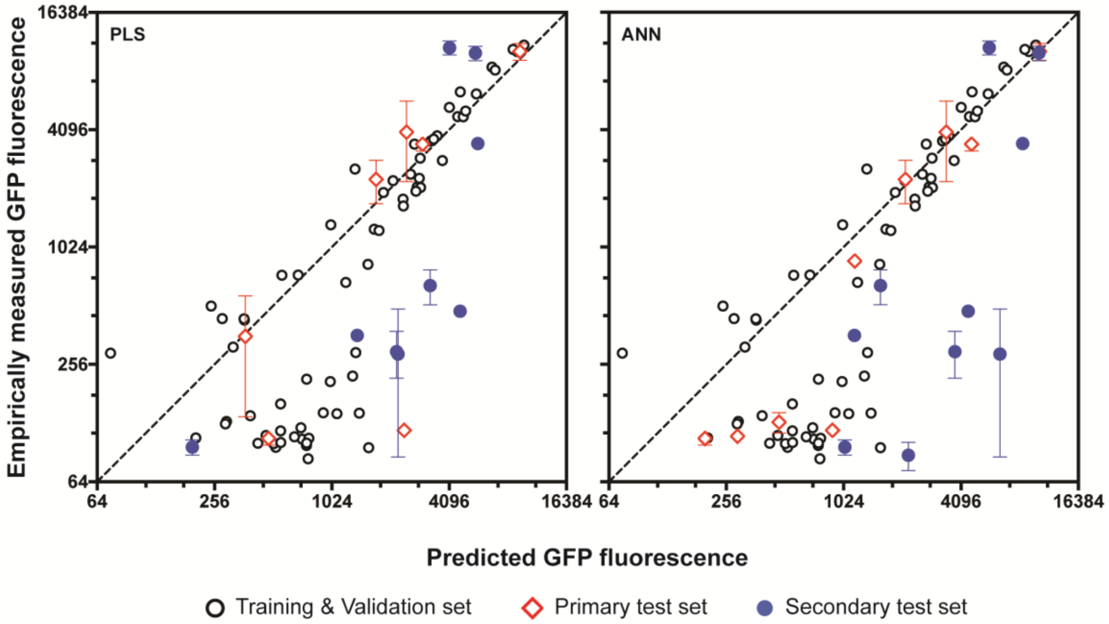


Figure 4: Empirically measured promoter activity levels plotted against activity levels as predicted by the optimum obtained PLS and ANN models

Points represent individual promoter sequences. Promoters that were used in model training and validation are shown in black, promoters that were part of the primary test set are shown in red, and sequences from the secondary test set are shown in blue. Empirical values are the mean of n = 3 starter cultures arising from independent transformation events. Standard deviation error bars are shown for both primary and secondary test sets, unless hidden by the points. The dashed lines represent the lines of equivalence, where empirically measured and predicted values are equal.

List of Abbreviations

ANN	Artificial Neural Network
BDBH	Bidirectional Best-Hit
bp	Base Pair
CDS	Coding Sequence
COG	COG triangles
DoE	Design of Experiments
DRS	Distal Regulatory Sequence
OMCL	OrthoMCL
PLS	Partial Least Squares
RBS	Ribosome Binding Site
TFBS	Transcription Factor Binding Site

Supporting Information

The Supporting Information file for this submission contains the following:

- Supporting Text:
    - Analysis of the effect of type IIs restriction cloning scars on the activity of promoter sequences
    - Extended methods for Artificial Neural Network sequence-function modelling
- Supporting Figures:
    - Supporting Figure 1: Visualisation of a sequence alignment of 100 putative promoters used to identify the putative location of the Ribosome Binding Site (RBS).
    - Supporting Figure 2: The effect of cloning scar sequences on promoter activity.
    - Supporting Figure 3: A comparison of codon usage in the GFP and mOrange reporter sequences.
    - Supporting Figure 4: Activity levels of putative promoter sequences characterised upstream of GFP in *G. thermoglucosidasius*.
    - Supporting Figure 5: R^2 and Root Average Squared Error (RASE) values returned by PLS sequence-function models when applied to a test data set.

  - Supporting Figure 6: Plasmid map of the pS797 expression vector used for *in vivo* characterisation of putative promoter sequences.
  - Supporting Figure 7: Initial characterisation of putative promoter sequences isolated from bacteriophage genomes.
  - Supporting Figure 8: Schematic representation of Latin rectangle 96-well plate layout used in promoter characterisation.
  - Supporting Figure 9: Schematic representation of a random forest partition model, as applied to promoter sequences.
  - Supporting Figure 10: Assessing the contribution of ANN model parameters to determining predictive power using a PLS model.
  - Supporting Figure 11: Model performance statistics for ANNs modelling GFP fluorescence as a function of complete 104 bp promoters.
- Supporting Tables:
  - Supporting Table 1: List of Transcription Factor Binding Sites (TFBS) used in promoter identification.
  - Supporting Table 2: DNA sequences of cloning affixes used in type IIs restriction cloning.
  - Supporting Table 3: Analysis of the native genes with which the characterised promoters were originally associated.
  - Supporting Table 4: Artificial Neural Network parameters included in the architecture optimisation screening design, and the values specified for each parameter.
  - Supporting Table 5: DNA sequences of the characterised promoters.

Author Information

Current Address: The BioEconomy Centre, Biosciences, College of Life and Environmental Sciences, Stocker Road, University of Exeter, Exeter, EX4 4QD, U.K.

Author Contributions

J.G., T.P.H., D.A.P. and J.L. designed the study. R.K.T. and T.L. assisted with Bioinformatic analyses. J.G. and C.S. performed the characterisation experiments. J.G. and R.K.T. performed flow cytometry experiments. J.G. analysed the data and performed the sequence-function modelling. J.G. and J.L. wrote the manuscript. All authors commented on and revised the manuscript.

Acknowledgements

This work was supported by a grant from Shell International Exploration and Production. The authors acknowledge the Exeter Sequencing Service for their assistance in sequencing the Illumina libraries.

Data Availability

The sequence data for the 4 *Geobacillus* spp. used in this study have been submitted to the NCBI Sequence Read Archive and are available under the accession number PRJNA521450.

References

1. Canton, B., Labno, A., and Endy, D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol. 26*, 787–793.
2. Segall-Shapiro, T. H., Sontag, E. D., and Voigt, C. A. (2018) Engineered promoters enable constant gene expression at any copy number in bacteria. *Nat. Biotechnol. 1*, 1399.
3. Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946–950.
4. Nielsen, A. A. K., Der, B. S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E. A., Ross, D., Densmore, D., and Voigt, C. A. (2016) Genetic circuit design automation. *Science 352*, aac7341.
5. Boyle, P. M., and Silver, P. A. (2012) Parts plus pipes: Synthetic biology approaches to metabolic engineering. *Metab. Eng. 14*, 223–232.
6. Adams, B. L. (2016) The Next Generation of Synthetic Biology Chassis: Moving Synthetic Biology from the Laboratory to the Field. *ACS Synth. Biol. 5*, 1328–1330.
7. 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., and Wang, H. H. (2018) Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nat. Methods 15,* 323–329.
8. Brown, S., Loh, J., Aves, S. J., and Howard, T. P. (2018) Alkane Biosynthesis in Bacteria. In *Biogenesis of Hydrocarbons* (Stams, A. J. M., and Sousa, D., Eds.), pp 1–20. Springer International Publishing, Cham.
9. Reeve, B., Martinez-Klimova, E., De Jonghe, J., Leak, D. J., and Ellis, T. (2016) The *Geobacillus* Plasmid Set: A Modular Toolkit for Thermophile Engineering. *ACS Synth. Biol. 5*, 1342–1347.
10. Yan, Q., and Fong, S. S. (2017) Challenges and Advances for Genetic Engineering of Non-model Bacteria and Uses in Consolidated Bioprocessing. *Front. Microbiol. 8*, 2060. doi:10.3389/fmicb.2017.02060
11. Cripps, R. E., Eley, K., Leak, D. J., Rudd, B., Taylor, M., Todd, M., Boakes, S., Martin, S., and Atkinson, T. (2009) Metabolic engineering of *Geobacillus thermoglucosidasius* for high yield ethanol production. *Metab. Eng. 11*, 398–408.
12. Jiang, Y., Xin, F., Lu, J., Dong, W., Zhang, W., Zhang, M., Wu, H., Ma, J., and Jiang, M. (2017) State of the art review of biofuels production from lignocellulose by thermophilic bacteria. *Bioresource Technol. 245*, 1498–1506.
13. Olson, D. G., McBride, J. E., Shaw, A. J., and Lynd, L. R. (2012) Recent progress in consolidated bioprocessing. *Curr. Opin. Biotech. 23*, 396–405.
14. Gilman, J., and Love, J. (2016) Synthetic promoter design for new microbial chassis. *Biochem. Soc. T. 44*, 731–737.
15. Blazeck, J., and Alper, H. S. (2013) Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnol. J. 8*, 46–58.
16. Brockman, I. M., and Prather, K. L. J. (2015) Dynamic metabolic engineering: New strategies for developing responsive cell factories. *Biotechnol. J. 10*, 1360–1369.
17. Goldbeck, C. P., Jensen, H. M., TerAvest, M. A., Beedle, N., Appling, Y., Hepler, M., Cambray, G., Mutalik, V., Angenent, L. T., and Ajo-Franklin, C. M. (2012) Tuning Promoter Strengths for Improved Synthesis and Function of Electron Conduits in *Escherichia coli*. *ACS Synth. Biol. 2*, 150–159.
18. 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., and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods 10*, 354–360.
19. Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *PNAS 102*, 12678–12683.
20. 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.
21. Mordaka, P. M., and Heap, J. T. (2018) Stringency of Synthetic Promoter Sequences in Clostridium Revealed and Circumvented by Tuning Promoter Library Mutation Rates. *ACS Synth. Biol. 7*, 672–681.
22. Zhang, S., Liu, D., Mao, Z., Mao, Y., Ma, H., Chen, T., Zhao, X., and Wang, Z. (2018)

Model-based reconstruction of synthetic promoter library in *Corynebacterium glutamicum*. *Biotechnol Lett. 40,* 819–827.

23. McWhinnie, R. L., and Nano, F. E. (2013) Synthetic Promoters Functional in *Francisella novicida* and *Escherichia coli*. *Appl. Environ. Microb. 80*, 226–234.
24. DeLorenzo, D. M., Rottinghaus, A. G., Henson, W. R., and Moon, T. S. (2018) Molecular Toolkit for Gene Expression Control and Genome Modification in *Rhodococcus opacus* PD630. *ACS Synth. Biol. 7*, 727–738.
25. 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.
26. Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat. Biotechnol. 36,* 1005–1015.
27. Kosuri, S., and Church, G. M. (2014) Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods 11*, 499–507.
28. Bartosiak-Jentys, J., Hussein, A. H., Lewis, C. J., and Leak, D. J. (2013) Modular system for assessment of glycosyl hydrolase secretion in *Geobacillus thermoglucosidasius*. *Microbiology 159*, 1267–1275.
29. Bezuidt, O. K., Pierneef, R., Gomri, A. M., Adesioye, F., Makhalanyane, T. P., Kharroub, K., and Cowan, D. A. (2015) Genomic analysis of six new *Geobacillus* strains reveals highly conserved carbohydrate degradation architectures and strategies. *Front. Microbiol. 6*, 430 doi:10.3389/fmicb.2015.00430.
30. Zhou, J., Wu, K., and Rao, C. (2016) Evolutionary Engineering of *Geobacillus thermoglucosidasius* for Improved Ethanol Production. *Biotechnol. Bioeng.113*, 2156–2167.
31. Kananavičiūtė, R., and Čitavičius, D. (2015) Genetic engineering of *Geobacillus* spp. *J. Microbiol. Meth. 111*, 31–39.
32. Studholme, D. J. (2015) Some (bacilli) like it hot: genomics of *Geobacillus* species. *Microb. Biotechnol. 8*, 40–48.
33. Blanchard, K., Robic, S., and Matsumura, I. (2014) Transformable facultative thermophile *Geobacillus stearothermophilus* NUB3621 as a host strain for metabolic engineering. *Appl. Microbiol. Biot. 98*, 6715–6723.
34. Suzuki, H., Murakami, A., and Yoshida, K. I. (2012) Counterselection System for *Geobacillus kaustophilus* HTA426 through Disruption of pyrF and pyrR. *Appl. Environ. Microb. 78*, 7376–7383.
35. Bartosiak-Jentys, J., Eley, K., and Leak, D. J. (2012) Application of pheB as a Reporter Gene for *Geobacillus* spp., Enabling Qualitative Colony Screening and Quantitative Analysis of Promoter Strength. *Appl. Environ. Microb. 78*, 5945–5947.
36. Lin, P. P., Rabe, K. S., Takasumi, J. L., Kadisch, M., Arnold, F. H., and Liao, J. C. (2014) Isobutanol production at elevated temperatures in thermophilic *Geobacillus thermoglucosidasius*. *Metab. Eng. 24*, 1–8.
37. Pogrebnyakov, I., Jendresen, C. B., and Nielsen, A. T. (2017) Genetic toolbox for controlled expression of functional proteins in Geobacillus spp. *PLoS ONE 12*, e0171313.
38. Jensen, T. Ø., Pogrebnyakov, I., Falkenberg, K. B., Redl, S., and 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. doi:10.1186/s13568-017-0469-z.
39. Contreras-Moreira, B., and Vinuesa, P. (2013) GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl. Environ. Microb. 79*, 7696–7701.
40. 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., and Morett, E. (2009) Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli. *PLoS ONE 4*, e7526.
41. 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.
42. Solovyev, V., and Salamov, A. (2011) Automatic Annotation of Microbial Genomes and

Metagenomic Sequences. In *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies* (Li, R. W., Ed.), pp 61–78, Nova Science Publishers, New York.

43. Kosuri, S., Goodman, D., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Edny, D., and Church, G. M. (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *PNAS 110,* 14024–14029.
44. 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., and Arkin, A. P. (2013) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods 10*, 347–353.
45. Zong, Y., Zhang, H. M., Lyu, C., Ji, X., Hou, J., Guo, X., Ouyang, Q., and Lou, C. (2017) Insulated transcriptional elements enable precise design of genetic circuits. *Nat. Commun. 8,* 52. doi:10.1038/s41467-017-00063-z.
46. Shine, J., and Dalgarno, L. (1974) The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *PNAS 71*, 1342–1346.
47. Lou, C., Stanton, B., Chen, Y.J., Munsky, B., and Voigt, C. A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol. 30*, 1137–1142.
48. Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A., and Arkin, A. P. (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol. 30*, 1002–1006.
49. Gasser, B., Steiger, M. G., and Mattanovich, D. (2015) Methanol regulated yeast promoters: production vehicles and toolbox for synthetic biology. *Microb. Cell Fact. 14*, 196. doi:10.1186/s12934-015-0387-1.
50. Baltagi, Y., and Kussener, F. (2014) Advantages of Bootstrap Forest for Yield Analysis. SAS Institute Inc, Cary.
51. Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., and Gourse, R. L. (1993) A Third Recognition Element in Bacterial Promoters: DNA Binding by the α Subunit of RNA Polymerase. *Science 262*, 1407–1413.
52. Estrem, S. T., Gaal, T., Ross, W., and Gourse, R. L. (1998) Identification of an UP element consensus sequence for bacterial promoters. *PNAS 95*, 9761–9766.
53. Bataineh, M., and Marler, T. (2017) Neural network for regression problems with reduced training sets. *Neural Networks 95*, 1–9.
54. Wold, S., Sjöstrom, M., and Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst. 58,* 109–130.
55. 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.
56. 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.
57. 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.
58. Li, J., and Zhang, Y. (2014) Relationship between promoter sequence and its strength in expression. Eur. Phys. J. E 37, 86.
59. Chen, B., and Yin, H. (2018) Learning category distance metric for data clustering. *Neurocomputing 306*, 160–170.
60. Li, D., and Tian, Y. (2018) Survey and experimental study on metric learning methods. *Neural Networks 105*, 447–462.
61. Simon, R., Priefer, U., and Pühler, A. (1983) A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Nat. Biotechnol. 1*, 784–791.
62. Zeigler, D. R. (2001) Media for growth of *Geobacillus* strains. In *The Genus Geobacillus - Introduction and Strain Catalog*, 7th ed., pp 20. Bacillus Genetic Stock Center.
63. Aronesty, E. (2013) Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal 7*, 1–8.
64. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012) SPAdes: A New Genome

Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol. 19*, 455–477.

65. Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics 30*, 2068–2069.

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

67. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res. 32*, 1792–1797.

68. Price, M. N., Dehal, P. S., and Arkin, A. P. (2009) FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol. 26*, 1641–1650.

69. Rambaut, A. (Ed.). FigTree. Institute of Evolutionary Biology, University of Edinburgh. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed Jan. 22, 2019).

70. Haahr, M., and Haahr, S. (Eds.) (1998) RANDOM.ORG. Available online: https://www.random.org/ (accessed Jan. 22, 2019).

71. Quinlan, A. R., and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics 26*, 841–842.

72. Langmead, B., and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods 9*, 357–359.

73. 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.

74. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol. 14*, R36.

75. Bin Liu, Zhou, F., Wu, S., Xu, Y., and Zhang, X. (2009) Genomic and proteomic characterization of a thermophilic *Geobacillus* bacteriophage GBSV1. *Res. Microbiol. 160*, 166–171.

76. Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., and Tsien, R. Y. (2004) Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat. Biotechnol. 22*, 1567–1572.

77. Feng, L., Wang, W., Cheng, J., Ren, Y., Zhao, G., Gao, C., Tang, Y., Liu, X., Han, W., Peng, X., Liu, R., and Weng, L. (2007) Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *PNAS 104*, 5602–5607.

78. Engler, C., Kandzia, R., and Marillonnet, S. (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE 3*, e3647.

79. Kirchmaier, S., Lust, K., and Wittbrodt, J. (2013) Golden GATEway Cloning - A Combinatorial Approach to Generate Fusion and Recombination Constructs. *PLoS ONE 8*, e76117.

80. Hanahan, D. (1985) DNA Cloning: A Practical Approach (Glover, D. M., Ed.). IRL Press, Oxford.

81. Liang, Y., Woodle, S. A., Shibeko, A. M., Lee, T. K., and Ovanesov, M. V. (2013) Correction of microplate location effects improves performance of the thrombin generation test. *Thromb. J. 11*, 12. doi:10.1186/1477-9560-11-12.

82. Chavez, M., Ho, J., and Tan, C. (2016) Reproducibility of high-throughput plate-reader experiments in synthetic biology. *ACS Synth. Biol. 6,* 375–380.

83. Ho, T. K. (1995) Random Decision Forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, pp 278–282, Montreal.