Supporting Information

Analysing the effect of type IIs restriction cloning scars on promoter sequence activity.

The use of type IIs restriction cloning to join DNA parts resulted in the insertion of a 4 bp scar sequence (ACCT) between the putative Distal Regulatory Sequence (DRS) and RBS regions of the bioinformatically identified 100 bp *Geobacillus* promoters. Previous studies have shown that novel cryptic functionality can potentially arise through the fusion of previously characterised genetic parts[45,47,84]. 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[85], thereby altering protein production. To assess the impact of the type IIS scar sequence on the activity of *Geobacillus* promoters, a subset of 24 bioinformatically identified sequences were characterised upstream of GFP both with and without scars (Supporting Figure 2).

For the majority of the analysed *promoter::GFP* constructs, the insertion of scar sequences between DRS and RBS had no statistically significant impact on GFP fluorescence (Supporting Figure 2A). Scar sequences significantly changed the regulatory activity of 4 sequences. Activity was significantly decreased in 3 promoters (P_34, adjusted P value = 0.000; P_18, adjusted P value = 0.022; P_56, adjusted P value = 0.024) and increased in 1 sequence (P_79, adjusted P value = 0.009) when scar sequences were inserted.

The observed differences in fluorescence between scarred and un-scarred *promoter::GFP* constructs were hypothesised to be the result of alterations to mRNA secondary structure. To test this hypothesis, the free-energy associated with mRNA folding was calculated using the mFold zipfold server[86]. 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 reflect 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.

3 of the promoters for which fluorescence output was significantly altered by the inclusion of scar sequences also showed the greatest changes in mRNA

secondary structure free energy (Supporting Figure 2B). *P_79::GFP*, for which GFP fluorescence was statistically significantly increased by the inclusion of scar sequences, showed the greatest relaxation of mRNA secondary structure out of the 24 analysed sequences. *P_18::GFP* also showed an increase in fluorescence output after scar insertion and relaxation of the mRNA secondary structure. Conversely, *P_34::GFP* and *P_56::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 was corroborated by the literature[85,87-89]. The presence of significant secondary structure surrounding the RBS is hypothesised to negatively impact on the ability of an 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[90,91].

Artificial Neural Network sequence-function modelling

ANNs were fit using the multilayer perceptron algorithm of JMP software. All ANNs used a single hidden layer, as a single hidden layer had previously been shown to be sufficiently complex to describe 224 bp *E. coli* promoters[57]. Furthermore, the increase in model complexity arising from additional hidden layers carried the risk of increasing model variance and overfitting the model to the training data[92].

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 Supporting Table 4. The specified parameters were combined in a full-factorial manner, which resulted in 81 ANN architectures being specified. As the result to which an individual ANN converges is dependent on the random seed used to generate the starting network weights[92], each of the 81 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 value when applied to the test set was identified.

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 sequence positions included in the ANN and the personality of the activation function were having a significant effect on model performance at the 0.05 significance level (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 for cross validation did not have a significant effect (P = 0.582 and P = 0.671, respectively).

The results of the screening design were also analysed using a PLS model, using the NIPALS algorithm and KFold cross validation where K = 7. The R^2 value returned when the models were applied to the test set was used as the *y* variable, and the parameters summarised in Supporting Table 4 were used as *x* variables. The resulting model extracted a single latent variable and was capable of explaining 12.5% of the cumulative variation in *X* and 52.525% of the cumulative variation in *Y.*

In PLS models, the Variable Importance in Projection (VIP) statistic can be used to determine the importance of a given variable in determining model output[54]. Variables with a VIP score above the threshold value of 0.8 are commonly accepted to have a significance impact on model output[93]. In this instance, 4 variables exceeded the threshold VIP value (Supporting Figure 10). The Linear and TanH activation functions were shown to have a significant effect on PLS model output, as did the effect of including 20 or 100 sequence positions in the ANNs. Analysis of the model coefficients showed that the TanH activation function was predicted to positively contribute to ANN predictive power, whereas the linear activation function was predicted to negatively impact predictive power, suggesting that linear modelling provided an inadequately sophisticated mathematical abstraction to adequately describe the promoter design space.

Taken together, the results of the screening design suggested that the TanH activation function and the number of promoter sequence positions included in the ANN were likely to contribute most to predictive power. A second iteration of ANN design was therefore undertaken using only TanH activation functions. The number of promoter nucleotide sequence positions that were included in the models ranged from 10 to 100, increasing in increments of 10 positions, and the number of nodes in the hidden layer was between 3 and 15, increasing in increments of 2. The parameters were combined in a full factorial manner, and each of the 70 resulting architectures was fit 1,000 times, with each run using a unique, defined random

seed. 500 runs of each ANN architecture used K = 4 KFold cross validation, and the remaining 500 runs used K = 5. All ANNs used the squared penalty method.

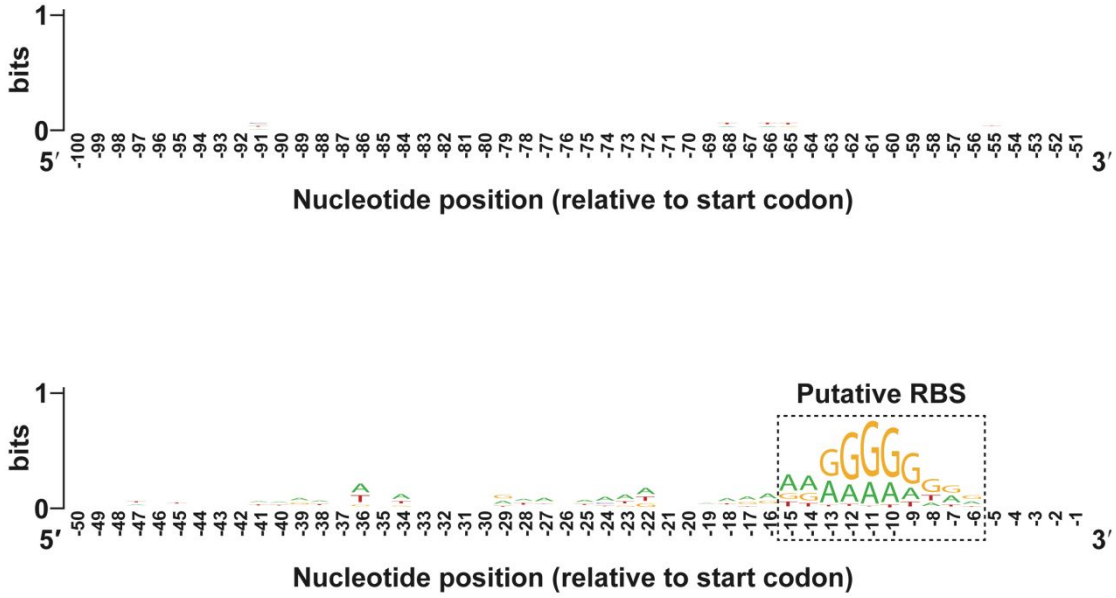
The individual ANN that returned the highest R^2 value when applied to the test set was identified for each of the 70 architectures. The best performing ANN that was obtained returned an R^2 value of 0.9304 when applied to the test set, and modelled GFP fluorescence as a function of 20 promoter nucleotide sequence positions.

In the case of ANNs that modelled GFP fluorescence as a function of complete 104 bp 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, additional single layer ANNs were trained using 17, 19 or 21 nodes. 1,000 ANNs were trained for each of the defined architectures, but none performed better than the 9-node model when applied to the test set, either in terms of R^2 or RASE (Supporting Figure 11).

Although single “best” performing models are most often presented in support of conclusions, this approach falsely assumes that only 1 model explains the response surface of interest[94]. Model ensembling aims to avoid this issue by combining the outputs of multiple individual models trained on the same task [95]. Theoretically, individual constituents of the ensemble can counteract deficiencies in other members, thereby improving the generalisation performance of the ensemble as compared to the individual base models[95]. The simplest method for ensembling multiple quantitative models is to calculate the arithmetic mean of the predicted values for a given set of *x* variables (*e.g.* the average predicted activity for a given promoter nucleotide sequence) and subsequently asses the accuracy of the resulting predictions.

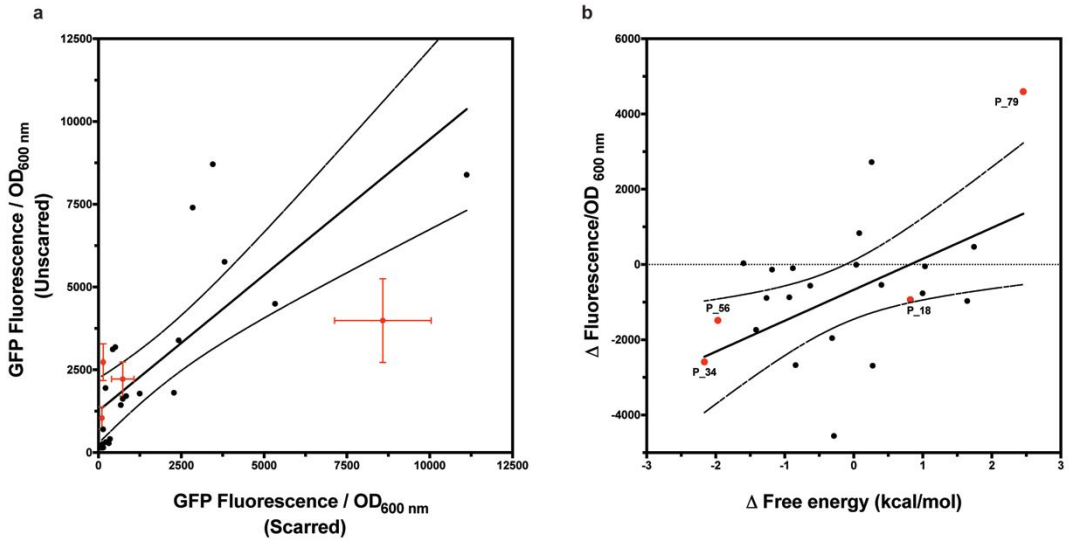
Model ensembling was therefore applied in attempt to increase predictive power. For each number of promoter nucleotide sequence positions that were modelled, the network architecture that returned the single ANN with the highest test set R^2 value was identified. The top 10 highest performing ANNs with the chosen architecture (as judged by the test set R^2 value) were used to create 9 progressively larger ensembles, with ANNs being included in descending order of their test set R^2 value. In each ensemble, the mean predicted value for each of the 10 promoter

sequences in the test set was calculated and the accuracy of the resulting predictions was assessed. The ensemble that was obtained with the lowest predictive error contained 2 constituent ANNs, each of which modelled promoter activity as a function of 20 sequence positions using 5 nodes in the hidden layer and TanH activation functions. The optimal model returned an R^2 value of 0.9746 when applied to the training and validation sets and an R^2 value of 0.9691 when applied to the test set.



Supporting Figure 1: Visualisation of a sequence alignment of 100 putative promoters used to identify the putative location of the Ribosome Binding Site (RBS).

The dashed box highlights the location of the putative RBS. The overall height of individual stacks indicates the degree of sequence conservation at a given position, and the height of the nucleotide symbols indicates the conservation of each nucleic acid at that position. Position numbering is relative to the start codon of the downstream coding sequence. Sequences were aligned and visualised using WebLogo version 2.8.2[96].



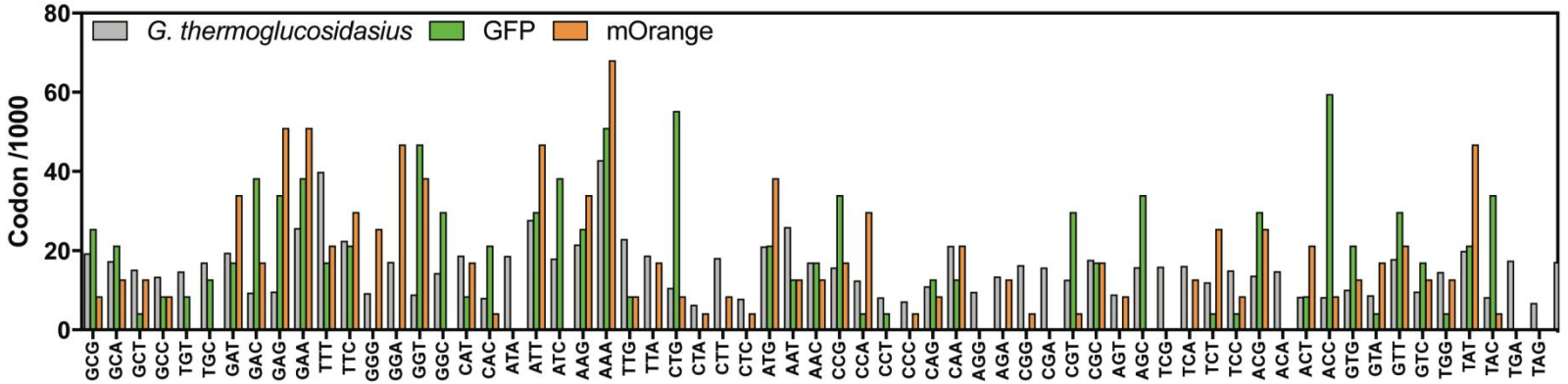
Supporting Figure 2: The effect of cloning scar sequences on promoter activity.

Fluorescence and absorbance measurements after 24 h incubation in 96-well plate format. In both instances, points that are coloured red represent promoters for which GFP fluorescence was statistically significantly changed when the cloning scar was inserted in the DNA sequence. Significance was determined using multiple t-tests, using the Holm-Šidak method to correct for multiple comparisons and a significance level of 0.05.

A) Fluorescence activity levels of scarred and unscarred promoters. Points represent the mean GFP output of each promoter, from $3 \leq n \leq 9$ starter cultures arising from independent transformation events. 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 dashed lines. The linear regression had an R^2 value of 0.5216.

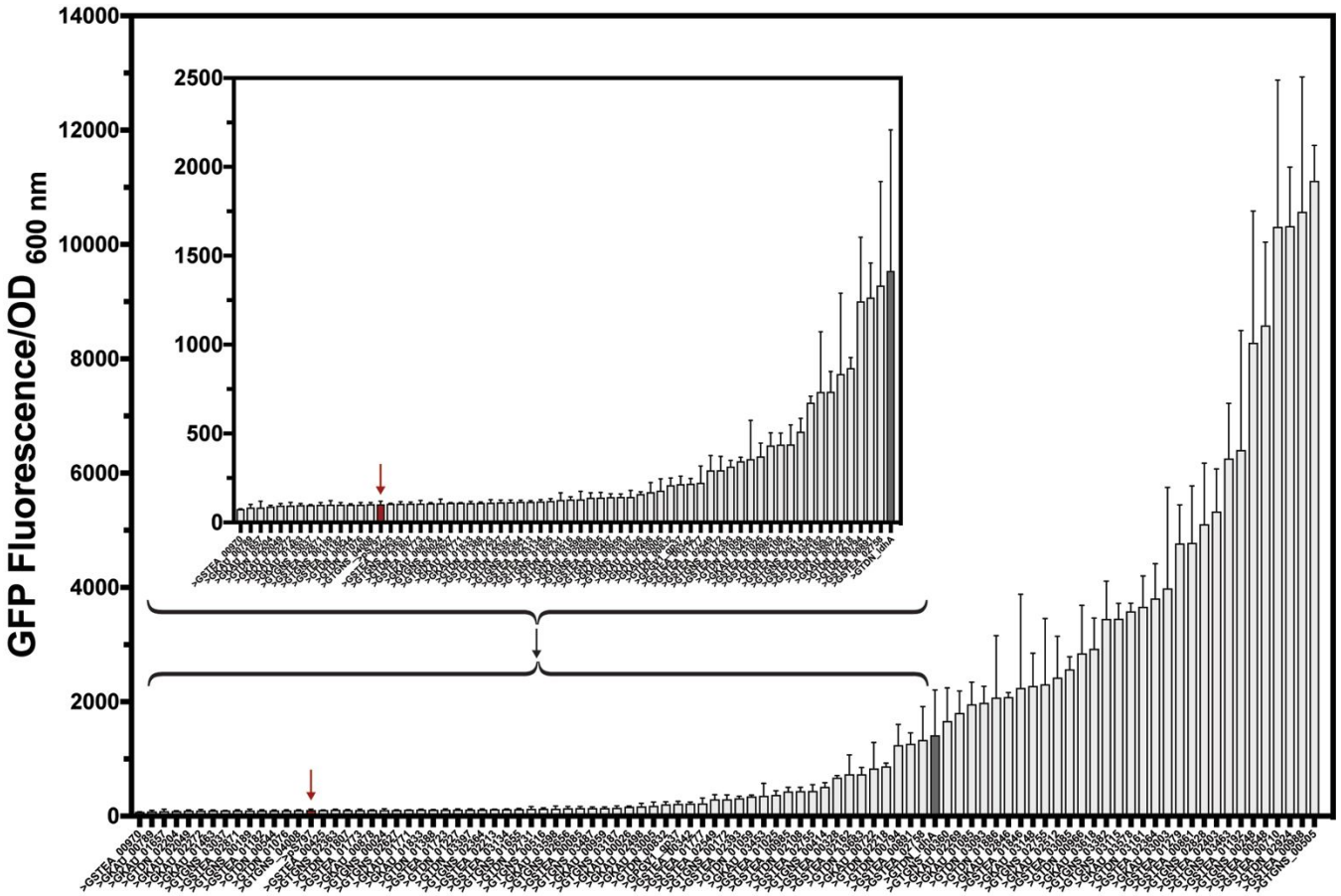
B) Comparing the change in GFP fluorescence and the change in free energy of the mRNA secondary structure of the *promoter::GFP* fusions when cloning scar sequences were inserted. Free energies were calculated using the mFold zipfold server[86], using default settings and RNA 2.3 energy rules. The sequence 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. thermoglucosidasius* cultures were incubated. The solid line represents a linear

194 regression of the data, with 95% confidence limits represented by dashed lines. The
195 linear regression had an R^2 value of 0.2407.

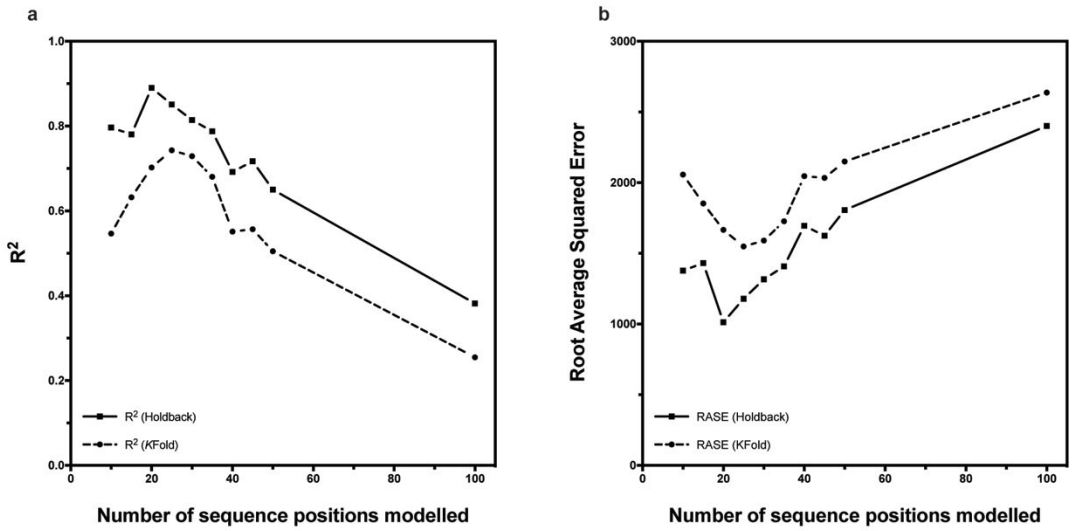


197 **Supplementary Figure 3: A comparison of codon usage in the GFP and mOrange reporter sequences.**

198

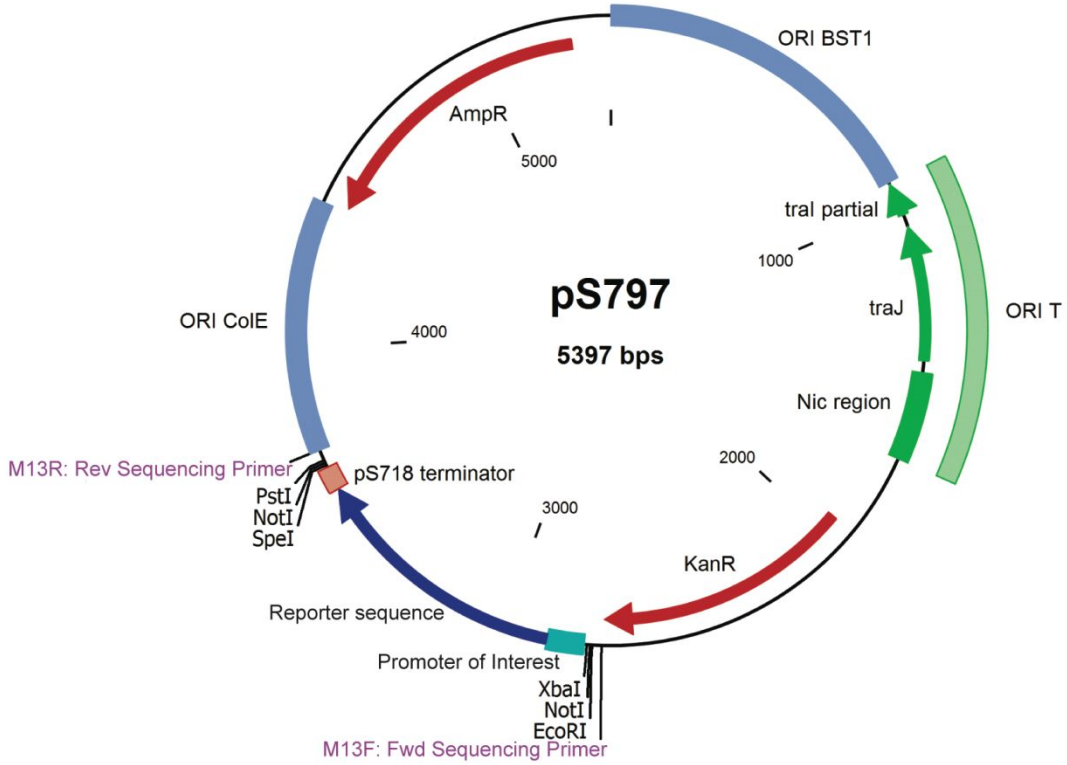


Supporting Figure 4: Activity levels of putative promoter sequences characterised upstream of GFP in *G. thermoglucosidasius*. Fluorescence and absorbance measurements after 24 h incubation. The positive control, the *G. thermodenitrificans ldhA* promoter, is represented by the dark grey bar. The negative control, *G. thermoglucosidasius* transformed with the empty pS797 vector, is shown in red. Standard deviation error bars shown, unless hidden by the bar. Bars represent the mean of n = 3 starter cultures, arising from independent transformation events, with the following exceptions: n = 4 (GSTEA_02871, GTGNS_02828); n = 5 (GSTEA_02755); n = 6 (GTDN_00832, GTDN_01227, GSTEA_02755); n = 11 (pos. control); n = 14 (neg. control) .



Supporting Figure 5: R^2 (a) and Root Average Squared Error (RASE, b) values returned by PLS sequence-function models when applied to a test data set.

PLS models were trained using the NIPALS algorithm, with a maximum of 10 latent variables permitted per model. For each number of sequence positions modelled, 4 PLS regressions were fit using KFold cross validation with K values of 4, 5, 7 and 10. 2 models were also fit at each position using holdback cross validation, with 20% or 33% of the training data set withheld to act as a validation set. Points represent the individual model that returned the highest R^2 value for the given number of promoter sequence positions. The circular points and dashed lines represent models trained using KFold cross validation. The square points and solid lines represent models trained using holdback cross validation.

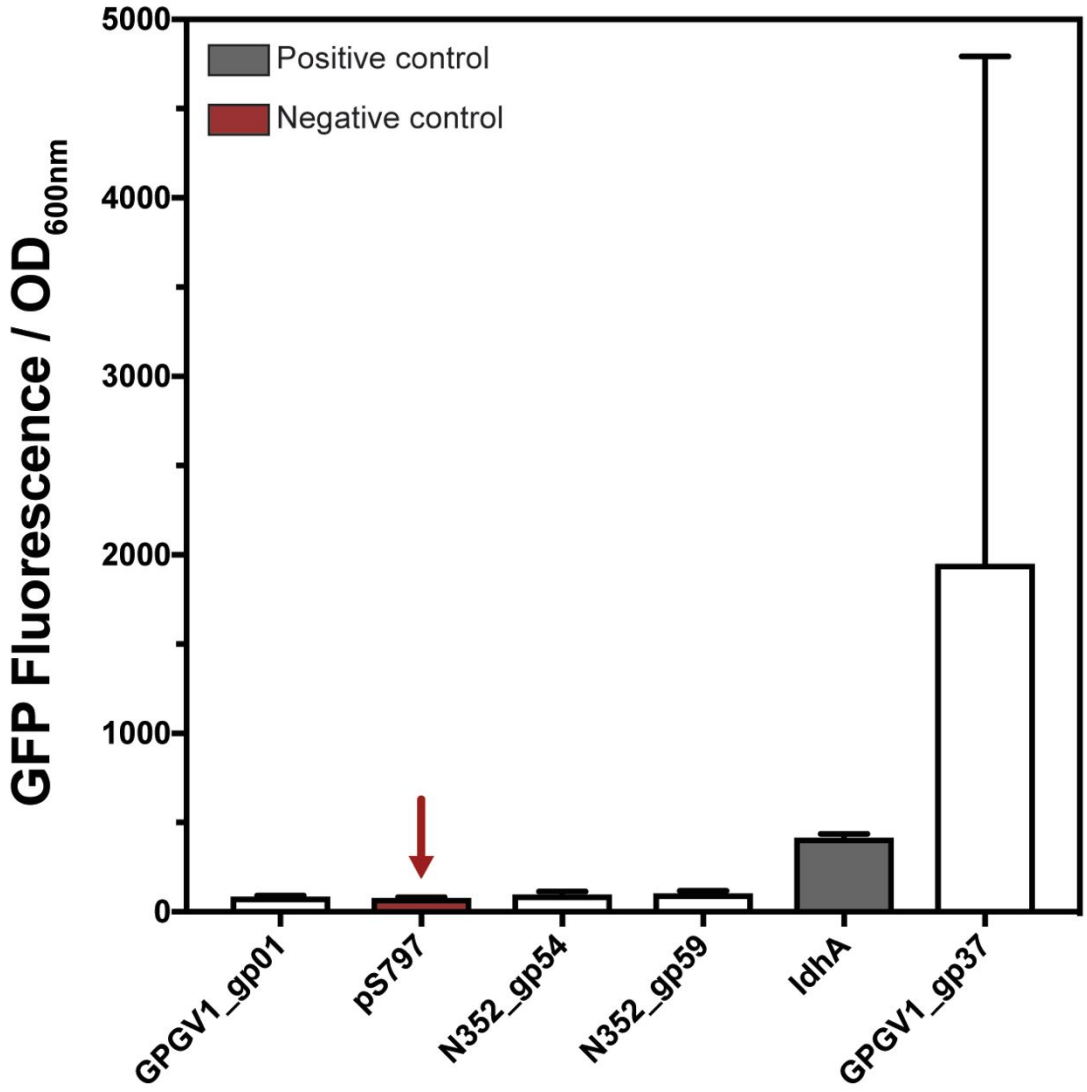


Supporting Figure 6: Plasmid map of the pS797 expression vector used for *in vivo* characterisation of putative promoter sequences.

The origin of transfer, ORI T, which contains the machinery necessary for mobilisation of the vector during conjugal transformation, is shown in green. Antibiotic resistance genes, conferring resistance to ampicillin (AmpR) and kanamycin (KanR) are shown in red. 2 origins of replication are present: ORI ColE for replication in *E. coli*, and ORI BST1 for replication in *Geobacillus*. Both origins are shown in blue. The binding sites of primers used for sequence verification of the promoter and reporter sequences are shown in purple.

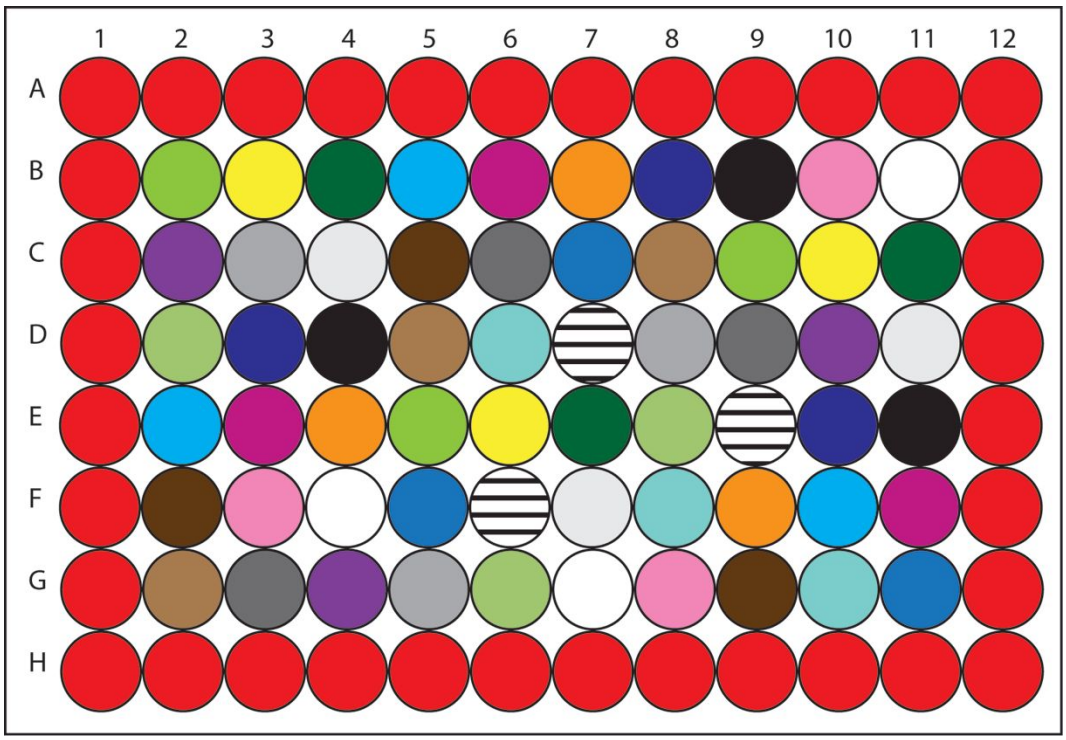
The promoter of interest, shown in cyan, is located between multiple cloning sites (MCS) containing the listed restriction sites. The reporter protein used for promoter characterisation is also located between the MCS and is followed by the S718 terminator sequence. The reporter and terminator sequences are shown in dark blue and red, respectively.

The conjugal plasmid pS797 was developed by ZuvaSyntha Ltd. (Hertfordshire, UK) from the electroporation plasmid pNC2, which in turn was developed from plasmid pUCG18 (GenBank EU547236)[92].



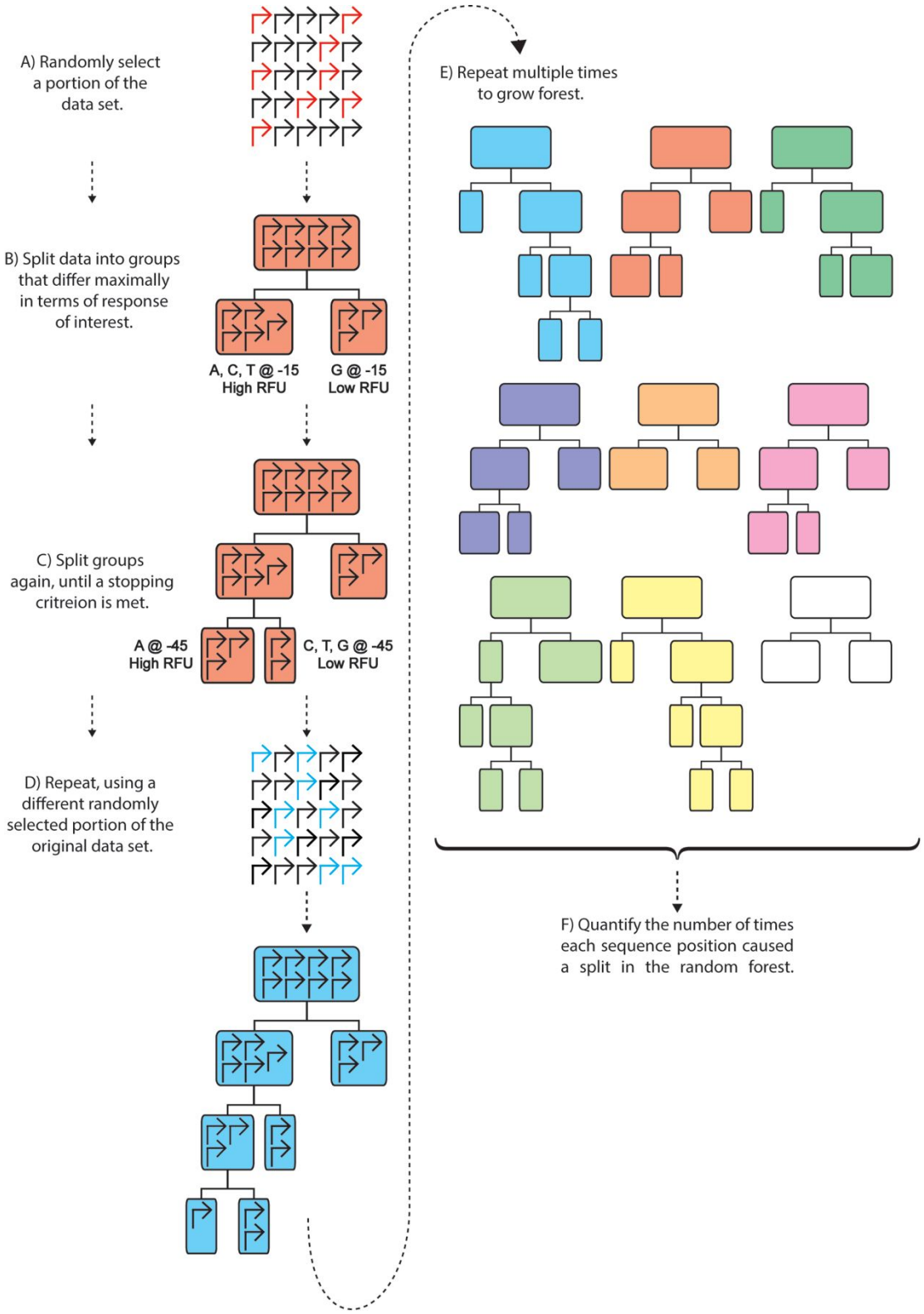
Supporting Figure 7: Initial characterisation of putative promoter sequences isolated from bacteriophage genomes.

Promoter activity in *G. thermoglucosidasius* was characterised after 24 h growth. 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. Bars represent the mean of n = 3 independent starter cultures arising from independent transformation events, except in the case of GPGV1_gp37, where n = 9. Standard deviation error bars shown, unless hidden by the bar.

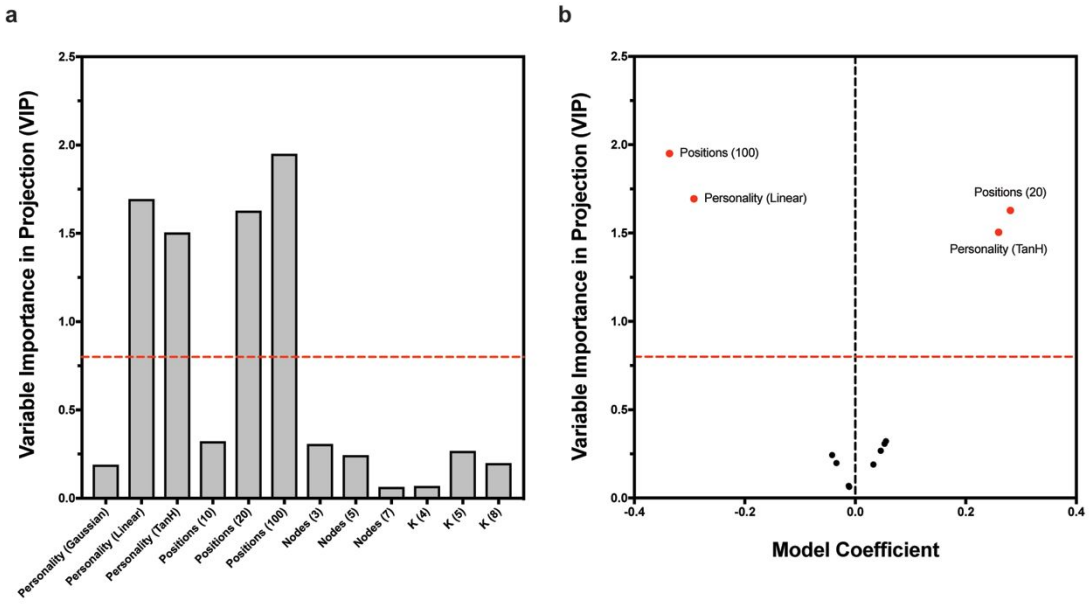


Supporting Figure 8: Schematic representation of Latin rectangle 96-well plate layout used in promoter characterisation.

96-well plates contained 3x 200 µl aliquots from each of 20 starter cultures. Disparate colours and patterns represent aliquots taken from the same starter culture. The outermost wells, highlighted in red, contained sterile growth media to reduce edge effects.



Supporting Figure 9: Schematic representation of a random forest partition model, as applied to promoter sequences.



285

286 **Supporting Figure 10: Assessing the contribution of ANN model parameters to**
287 **determining predictive power using a PLS model**.

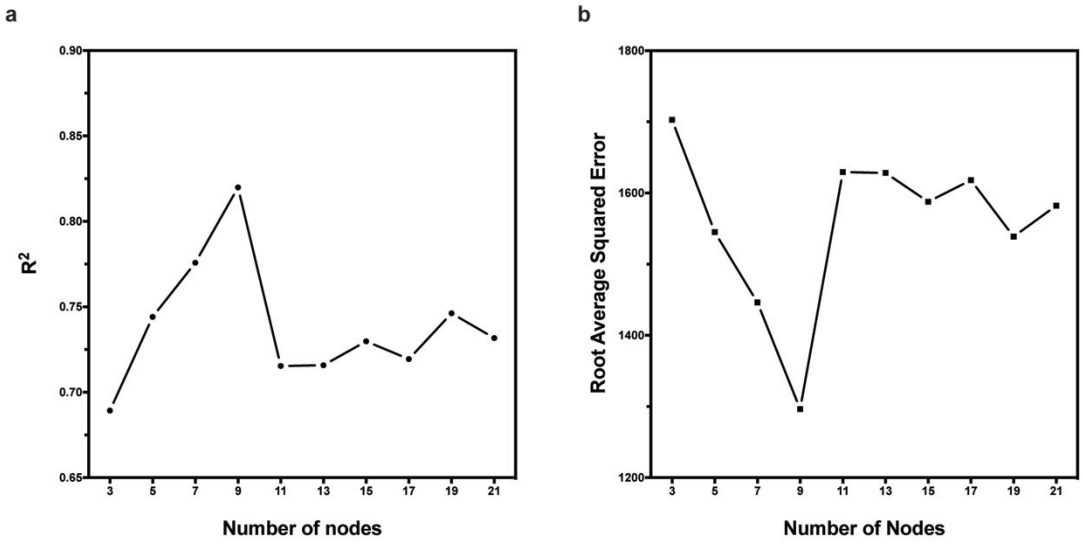
288

289 a) Variable Importance in Projection (VIP) plot and b) VIP v. model coefficient plot.

290

291 In both panels, the dashed red line represents the VIP threshold value of 0.8, above
292 which variables are predicted to have a significant impact on model output. VIP and
293 coefficient values were returned by a PLS regression that modelled the R^2 value
294 returned by ANNs when applied to a test data set as a function of activation function
295 personality (Personality), the number of promoter nucleotide sequence positions
296 included in the ANN (Positions), the number of nodes included in the hidden layer
297 (Nodes) and the value of K used in ANN KFold cross validation (K).

298



Supporting Figure 11: Model performance statistics for ANNs modelling GFP fluorescence as a function of complete 104 bp promoters.

R^2 (A) and Root Average Squared Error (RASE, B) values were returned when ANNs were applied to an independent test 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 KFold cross validation with K = 4, and 500 models were fit with K = 5. Points represent the R^2 and RASE values returned by the single highest performing ANN for each number of nodes.

	Sequence		Sequence		Sequence		Sequence
ada	ATAAGAAT	cpxR	AGTGTAAA	cysB	TCTTGCAT	fruR	TGAATCGA
araC	AAGATTAG	cpxR	GGAGTAAC	cysB	TGTATATA	fur	ATAATTGT
araC	TAACAAAA	crp	TGTCGTAA	cytR	TTCTGTAA	fur	ATTTGCAT
araC	TCCGACCT	crp	TTTTGTTA	cytR	ACTTGTAA	fur	TAATGCTT
araC	GACATAAG	crp	AAGTGTGA	cytR	ATAGAACT	fur	ATGATAAT
araC	TCCTTGTT	crp	TAGACACT	deoR	AATTCTAA	fur	ATAATGAT
araC	ACAGAGGG	crp	AGATCACA	deoR	TTAGAATA	fur	ATCATTTT
arcA	ATTTGTAA	crp	TAATGTGA	deoR	AATTTTAT	fur	CATTATAG
arcA	TTAATTAA	crp	TTAAATTG	dnaA	ATTCACAA	galR	TGTAAGCG
arcA	TAATTAAA	crp	ATGCGAGG	dnaA	CCACAAGT	galR	ATGTAACC
arcA	TTAACTAA	crp	TCACACTT	dnaA	AGTTATCC	gcvA	AACTAATT
arcA	TAACTCAA	crp	AGATGTGA	dnaA	TTGTTATC	gcvA	TTATATTT
arcA	TAATTATA	crp	TTAGATTA	dnaA	TATACACA	gcvA	ATAAGCTA
arcA	TCATGTTT	crp	GATGTGTA	dnaA	TTTGGATA	gcvA	ACTAATAG
arcA	AATAAAAA	crp	AAATGTAA	fadR	CCGACCTA	glpR	TTCAAAAT
arcA	AAAAGGGA	crp	AATAAATT	fadR	GGACTTGT	glpR	TCGAATTA
arcA	TAACAATT	crp	TCACATTT	fadR	AACTCATC	glpR	CACACATT
arcA	AATTGTAA	crp	TATACATA	farR	TGTATTAT	glpR	ACGATAAG
arcA	TCACAAAA	crp	ACACACAT	fhlA	TCATTTTC	glpR	AATTTGAG
argR	ATAAAAAT	crp	CTCGTTTT	fis	TCTTAACT	hipB	CCCTTAAG
argR	ATAATCAT	crp	AAATAACA	fis	TCAGAGGA	hipB	AAAGGATA
argR	TTTTTTAT	crp	AACCCCTC	fis	CTCATTTT	hipB	TAAGATGT
argR	TTATAATT	crp	GTTAGCTC	fis	TGTAAATT	hipB	CCTTTTTA
argR	ATATTCAT	crp	ATAACAAT	fis	TTATCTAA	hns	TAGGCTGA
argR	AATAATTA	crp	TGTGATCT	fis	ACAGTTGT	hns	ACAGAGTA
argR	AATTCAAT	crp	CGCACATA	fis	TATACTTA	hns	AAAGGAAT
argR	CATAAAAA	crp	TTTTGCAA	fis	TATTCTAT	hns	TAATTTAA
argR	CATCCATA	crp	TTAGTTCA	fis	AATTATTT	hns	ATTAAATT
argR	TAATTCAT	crp	CACAGTGC	fis	ACAATTAT	ihf	TTTTATTT
argR	AATTCATC	crp	GTGCTCCC	fis	AAATGTGA	ihf	GCTTAGAG
argR	CACAATAA	crp	ATCACAAA	fis	CCCTCCGT	ihf	TTGGAGTC
argR	AATTAATA	crp	TTTGTGAG	fis	AAAAATAA	ihf	ATCATACA
argR2	ATACACTA	crp	TCACAATT	fis	TCTTTAAT	ihf	TAGGATAA
argR2	CATATTTT	crp	AGGTAACA	flhCD	TGAGGGGT	ihf	AGATATAT
argR2	TTTTTATT	crp	ATGCCTGA	flhCD	GGGCTTTT	ihf	CCGGCCTA
argR2	ACATATAA	crp	TCATATCA	fnr	ATCAATTT	ihf	CCATGTAG
argR2	TTTATTTT	crp	TCAAACTT	fnr	TTTTTTGA	ihf	AATAAAAT
argR2	ATATAAAT	crp	GTTCAATT	fnr	ATCAATAA	ihf	TGTAGGCC
argR2	GCAAATAA	crp	ATTTGTGA	fnr	TTATACAA	ihf	AAAGTATA
carP	CACTTTTT	crp	GTGTGTAA	fnr	ACAATTTA	ihf	TAGATTAA
carP	CTGTAAAA	crp	GAAACATA	fnr	TAAATTGT	ihf	TATACTTT
cpxR	GTAAATTA	crp	TAACCAAT	fnr	ATTTGTTT	ihf	ACAGACAA
cpxR	CTTGTAAG	cspA	GCAAGACA	fnr	ATAAATAT	ihf	TGATATGA
cpxR	TAAAATTA	cynR	TAAGGTAA	fnr	TGTAAATC	ihf	AGTATACA
cpxR	TAAAACAA	cynR	ATAAGTAA	fnr	TCAAGAGT	ihf	CTCCTTGA
cpxR	TAAAAAGA	cysB	TATAGTTA	fnr	CCTACCTC	ihf	TAATCTAT

	Sequence		Sequence		Sequence		Sequence
ihf	TTTCAAAA	melR	TCCCATAA	oxyR	ATTAGTGT	rpoD15	TAATGTTT
ihf	AAATAAAA	metJ	ATAAGCGT	oxyR	CGGAGTAA	rpoD15	CTAAATAA
ihf	TTAATTTA	metJ	ATACATCT	oxyR	GATTAATT	rpoD15	TCCTCTGT
ihf	GCTAATAG	metJ	ACATCTAA	oxyR	TAGAATAG	rpoD15	AGAGTCAA
ihf	CTTCGGGA	metJ	AGACATCC	pdhR	TTGTTAAA	rpoD15	TTTGTGTA
ihf	CTAAGGGC	metJ	TGTAAACA	phoB	TTTAATTA	rpoD15	TTAACACA
ihf	TGTAAGAA	metJ	GCTAAAAT	phoB	CTGTCATA	rpoD15	CCTCTCCC
ihf	ACAAAAAA	metJ	CTTTACAT	phoB	TTTTCATA	rpoD15	CCCTTAAA
ilvY	ATATATCA	metJ	TAAACGGA	phoB	ACGCATAA	rpoD15	TATAAGTC
lacI	AGTAACAA	metJ	TAGATGTG	phoB	TAATATAT	rpoD15	GGTATACT
lexA	TATATAAA	metR	ATTTTTCC	phoB	TAATCTGT	rpoD15	AGGATTAA
lexA	TATACAGT	metR	TTTTTTCA	phoB	AATAAAAG	rpoD15	TGTCTCAC
lexA	TCCATACA	metR	TTCTTTTC	phoB	TTTATTAA	rpoD15	TTTTAACA
lexA	ATAAATAA	metR	CAAATTTT	phoB	TGTCATCA	rpoD15	GTGAATAG
lexA	TATATACT	modE	TATACAAG	phoB	TCATAAAA	rpoD15	AAGGTAGA
lexA	TATATACA	modE	CTATATAC	phoB	CTTACATA	rpoD15	TTTTCGCA
lexA	AAACCACA	modE	CCTACATA	phoB	CATCTTTC	rpoD16	CACATTTC
lexA	TACTGTAT	nagC	CATTTCAC	phoB	ATGTAACA	rpoD16	GCTTTTTA
lexA	ATATAAAA	nagC	GAAATAAG	phoB	TAATTCGA	rpoD16	TTGCAAAT
lexA	TGTACACA	nagC	ATATTTTA	phoB3	TCCTTACA	rpoD16	TCATAACA
lexA	ATATATAC	nagC	ATTTTAGA	purR	ATAAAAAG	rpoD16	CGTATAAT
lexA	ATAGATAA	nagC	TTTAATTT	purR	ATTTCAAG	rpoD16	ACCTAAAT
lexA	TAAAAACA	nagC	CTTATTTT	purR	TAATCGTT	rpoD16	ACGGTATA
lexA	TATATTCA	narL	TGCTCCTT	purR	CGCGAGGT	rpoD16	TCGCCCCT
lexA	TTTTTTTA	narL	TAACTCTT	purR	TGAGGAAA	rpoD16	TTTCAAAT
lexA	TATATATA	narL	GTAAGGGT	purR	TTTTTAAG	rpoD16	TAGCGTAA
lrp	ATCTTTTA	narL	TCCCCATG	purR	CAAGGAAA	rpoD16	CAGATATA
lrp	TATTTTTT	narL	TCCTAAAG	purR	CTTTTTCT	rpoD16	ATTTCATA
lrp	TTTAGTGT	narL	AAGTAGTA	purR	TTTTCGTT	rpoD16	TTTTTATA
lrp	ATTTTTTT	narL	TCGGGGTA	purR	CTTTCCCT	rpoD16	TGTCAAGA
lrp	TTGTCTTA	narL	ACTCCTTA	purR	GAAACGAG	rpoD16	CCCCATAA
lrp	TGCATTTT	narL	TAACTCTA	purR	TTTCGTTT	rpoD16	TTCAATCT
lrp	TGTAGAAT	narP	TTGAGGTA	purR	CGTTTTTT	rpoD16	TAGTTCGG
lrp	ATTTATTA	narP	AAGGAGTA	rhaS	TCCTGTCA	rpoD16	TCTTTAGG
lrp	TTTCTTTT	narP	TTTAGAGT	rpoD15	AGCCTAAA	rpoD16	TAGAATGC
lrp	TATTCTTA	ntrC	TGCACTAA	rpoD15	AACGCATA	rpoD16	CTCCTGAT
lrp	TTGACAAT	ompR	AAATCACA	rpoD15	ACATACTA	rpoD16	CCTTATAA
lrp	TATTTATT	ompR	TCATATTT	rpoD15	CATGATAG	rpoD16	AAATAATT
lrp	TAAGATAA	ompR	GTAACATA	rpoD15	TAGCTTAT	rpoD16	TGTAGACT
lrp	AATACATA	ompR	TGTTTACA	rpoD15	TTTTGTAA	rpoD16	GTTAGGGT
lrp	ATAACTAA	ompR	TTTACATT	rpoD15	TTTTGTTT	rpoD16	GAGTATAA
malT	AGGGAGGA	ompR	GTTACATA	rpoD15	TAAGGTTA	rpoD16	ACAGTATA
malT	GGGGAGGA	ompR	TTCTTTTT	rpoD15	TTTTCACG	rpoD16	TTATAAAA
marR	TATACTTG	ompR	TTGTAGCA	rpoD15	GCTCAAGA	rpoD16	TAATTAGA
marR	CAACTAAT	ompR	ATATTACA	rpoD15	TGTGTAGG	rpoD16	TAGAGCAC

	Sequence		Sequence		Sequence		Sequence
rpoD16	TGTATAAT	rpoD17	CCCGTAGG	rpoD17	GTGTCATA	rpoH2	TTTTTTTT
rpoD16	AGTCTCAA	rpoD17	AATTTTCT	rpoD17	ATTAGAGT	rpoH2	TCTCCCTT
rpoD16	AGTGTAAT	rpoD17	CTCCTTTT	rpoD17	TGCAATAA	rpoH2	CTCTCCCA
rpoD16	CCTACAAT	rpoD17	CCAAATAG	rpoD17	TTGTCTGA	rpoH2	AAATAATG
rpoD16	AACTAGTT	rpoD17	TTTTACTT	rpoD18	GATAGAAT	rpoH2	ACAAAAGA
rpoD16	CATAGTGT	rpoD17	CTAAACCA	rpoD18	ACTTGTTA	rpoH2	ATTCTACC
rpoD16	TGATATAA	rpoD17	TTTTATAG	rpoD18	ATCATTGT	rpoH2	CCCTTTAA
rpoD17	AACTAAAC	rpoD17	CTCTGTAG	rpoD18	AATTGAGG	rpoH3	CTCCCCCT
rpoD17	CGTTGTAA	rpoD17	CAAGAGGG	rpoD18	AGTGTAGT	rpoH3	CAAAAAAA
rpoD17	ACTTTTGT	rpoD17	AGTAGAAT	rpoD18	TTGCTTTT	rpoH3	CTTCCCTT
rpoD17	TTTGTTTT	rpoD17	ATACTTAA	rpoD18	AGGCCCTC	rpoH3	CTTGAATA
rpoD17	TATAGATT	rpoD17	AATCTTTA	rpoD18	ACGACCCC	rpoH3	CTGATAAG
rpoD17	ATTTTGTA	rpoD17	GTTATACT	rpoD18	AAATATAT	rpoH3	CCCCCCAT
rpoD17	CCAAATTG	rpoD17	GAGTAAAA	rpoD18	TTGAGATA	rpoH3	AATAACTC
rpoD17	CTTAAAAT	rpoD17	CTAACCCT	rpoD18	GTGAGGGA	rpoN	AAATTGTA
rpoD17	ACTTTTAG	rpoD17	TAGCAACA	rpoD18	TTACACTT	rpoS17	TTATGTTT
rpoD17	TGATAATT	rpoD17	TAGCCTTT	rpoD18	TTGATAGG	rpoS17	CCCCCTCC
rpoD17	CCCATAAC	rpoD17	GCATAATG	rpoD18	TGTATGAT	rpoS17	CTATTATA
rpoD17	GAGACACA	rpoD17	CTTACTTT	rpoD18	CTTATACT	rpoS17	GGAGGGTG
rpoD17	TAATGTAA	rpoD17	TTTTCCTT	rpoD18	AGTACGGG	rpoS17	CTCACAAA
rpoD17	CCTATAGT	rpoD17	TTCCCTGT	rpoD18	GGCATAAT	rpoS17	CAGACATA
rpoD17	ATTAGTTA	rpoD17	CCCCTTTT	rpoD18	TTAATTTT	rpoS17	CGGCTAGT
rpoD17	TTAGGCTA	rpoD17	TTGTGAAT	rpoD18	GTACAATC	rpoS17	AGAGGGAG
rpoD17	ATACTATA	rpoD17	ATAATAAT	rpoD18	TTTAAAAT	rpoS17	TTATATTA
rpoD17	ACAAGTGT	rpoD17	CAGCATAC	rpoD18	AGACTACT	rpoS18	GTAGTCTA
rpoD17	GACCCCAC	rpoD17	TCCCTTTT	rpoD18	GGTAGACT	rpoS18	TGTCATGA
rpoD17	AATAGTTA	rpoD17	TAGGAATA	rpoD19	ACGTGCTA	rpoS18	TAGTTTAG
rpoD17	CAATTCTA	rpoD17	GCTACAAT	rpoD19	ATTGTTTT	rpoS18	CAGCTAGT
rpoD17	AGTTAATA	rpoD17	CATGCTAA	rpoD19	TGTGTTAG	soxS	GATAAGCG
rpoD17	CTTATAAG	rpoD17	ACAATCTT	rpoD19	CGGAGTAG	soxS	TATCATTT
rpoD17	TACCTAAA	rpoD17	TTGGAATA	rpoD19	ACAGATTT	soxS	TTGCTCCT
rpoD17	CAATTCAT	rpoD17	GGAGTGTA	rpoD19	TTTCATAT	soxS	AACTGTAA
rpoD17	AATAAGTA	rpoD17	CGTATACT	rpoD19	GCTATAAT	soxS	ATTTGTTA
rpoD17	AATAAATA	rpoD17	ATAACAGG	rpoD19	TTACCTCA	soxS	AACCCCCG
rpoD17	TGCTTGTA	rpoD17	CCCCTGGA	rpoD19	GAATGTAG	soxS	ATAATTCA
rpoD17	CATATAGT	rpoD17	CCTATACT	rpoD19	ACTAGCGA	soxS	CTTTTCCC
rpoD17	AGATCGGA	rpoD17	ACAGGAAT	rpoD19	CACCTAAA	soxS	CAAGTATA
rpoD17	TTCCTCCT	rpoD17	TGTTATAA	rpoD19	GTGATCTA	soxS	AAACCATA
rpoD17	AAAAATAG	rpoD17	GTTCGAGG	rpoD19	TTAACATG	soxS	AATTTCTT
rpoD17	TTGATATA	rpoD17	CCCTAAAA	rpoD19	TACTTAAA	torR	GTTCATAT
rpoD17	TAACGGAT	rpoD17	TTATGATA	rpoD19	CAAGCTTG	trpR	TTAACTAG
rpoD17	AAATCAGA	rpoD17	TTTATAAT	rpoD19	TGTTGTGT	trpR	TACTCTTT
rpoD17	TTTTGTAT	rpoD17	CTCCATAG	rpoD19	TCCTGCTA	tus	TAGTATGT
rpoD17	CGCCTTTT	rpoD17	CTTAACCT	rpoE	ACTTTACA	tus	ATAAGTAT
rpoD17	TGTAAAAT	rpoD17	GTCCTACA	rpoE	GTCTGATA	tus	TAACTAAT

	Sequence
tus	CATTAGTA
tus	TAACTAAG
tus	TTGTAACG
tyrR	TAAATAAA
tyrR	TATACAGA
tyrR	GTGTAAAA
tyrR	TTTACAAT
tyrR	TGAGATTT
tyrR	TGTAATTT
tyrR	CTTACACA
tyrR	TTAATACA
tyrR	AATATATA
tyrR	TATGTAAC
tyrR	ATTGTACA

Supporting Table 1: List of Transcription Factor Binding Sites (TFBS) used in promoter identification.

Part	Prefix sequence	Suffix sequence
5'-Homologous region	CAGGAAACAGCTATGACCATGGA ATTCGCGGCCGCTTCTAGAGACT CTGTGGTCTCAGGAG	TACTTGAGACCACGAAGTTACTA GTAGCGGCCGCTGCAGGACTGGC CGTCGTTTTACA
Distal Regulatory Sequence (DRS)	CAGGAAACAGCTATGACCATGGA ATTCGCGGCCGCTTCTAGAGACT CTGTGGTCTCATACT	ACCTTGAGACCACGAAGTTACTA GTAGCGGCCGCTGCAGGACTGGC CGTCGTTTTACA
RBS	CAGGAAACAGCTATGACCATGGA ATTCGCGGCCGCTTCTAGAGACT CTGTCTGGAGTCACTGGTCTCAA CCT	AATGTGAGACCACGAAGTTACTA GTAGCGGCCGCTGCAGGACTGGC CGTCGTTTTACA
CDS	CAGGAAACAGCTATGACCATGGA ATTCGCGGCCGCTTCTAGAGACT CTGTGGTCTCAAATG	CTGCTGAGACCAGTGGCTCCAGA CGAAGTTACTAGTAGCGGCCGCT GCAGGACTGGCCGTCGTTTTACA
Terminator	CAGGAAACAGCTATGACCATGGA ATTCGCGGCCGCTTCTAGAGACT CTGTGGTCTCACTGC	TTCGTGAGACCACGAAGTTACTA GTAGCGGCCGCTGCAGGACTGGC CGTCGTTTTACA
3'-Homologous region	CAGGAAACAGCTATGACCATGGA ATTCGCGGCCGCTTCTAGAGACT CTGTGGTCTCATTCG	GCTTTGAGACCACGAAGTTACTA GTAGCGGCCGCTGCAGGACTGGC CGTCGTTTTACA
pS797 vector	GGTCTCTTTCGTACTAGTAGCGG CCGCTGCAGG	ATGGAATTCGCGGCCGCTTCTAG AGTACTAGAGACC

Supporting Table 2: DNA sequences of cloning affixes used in type IIs restriction cloning.

Prefixes were joined to the 5' end of the relevant parts and suffixes were joined to the 3' end.

Promoter	Source	GFP RFU	mOrange RFU	Putative Identity	Gene Ontology Biological process / Molecular function
ATPase/GTPase					
42	GSTEA	432.51	9.59	nirQ Denitrification regulatory protein	ATPase activity
44	GSTEA	674.39	43.08	typA GTP-binding protein	GTPAse activity
47	GTDN	2071.08	13.69	ychF Ribosome and GTP binding protein	ATPase activity
51	GSTEA	222.90	179.52	cmpC Bicarbonate transport ATP-binding protein	ATPase activity
55	GTGNS	510.41	105.22	hflX GTPase	GTPAse activity
60	GSTEA	1265.28	864.58	sugC Sugar transport ATP-binding protein	ATPase activity
64	GKAU	2083.20	543.27	ravA ATPase	ATPase activity
Biosynthetic processes-related					
4	GTDN	88.81	14.38	clsB Cardiolipin synthase	Cardiolipin biosynthetic process
9	GSTEA	99.76	12.53	ilvC Ketol-acid reductoisomerase	Isoleucine biosynthetic process
18	GKAU	107.35	9.76	hemN Oxygen-independent coproporphyrinogen III oxidase	Porphyrin-containing compound biosynthetic process
39	GSTEA	292.71	90.63	hemL Glutamate-1-semialdehyde 2,1-aminomutase	Protoporphyrinogen IX biosynthetic process
41	GSTEA	371.83	92.95	plsY Glycerol-3-phosphate acyltransferase	Phospholipid biosynthetic process
53	GTDN	344.88	126.45	modB Molybdenum transport system permease	Mo-molybdopterin cofactor biosynthetic process
65	GSTEA	2243.22	459.15	moaB Molybdenum cofactor biosynthesis protein B	Mo-molybdopterin cofactor biosynthetic process
69	GKAU	2928.95	124.90	dapH 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase	Diaminopimelate biosynthetic process
Catabolic processes-related					
6	GKAU	96.15	9.17	lon2 Lon protease 2	Protein catabolic process
8	GTGNS	98.59	8.76	deoB Phosphopentomutase	Deoxyribonucleotide catabolic process
43	GSTEA	439.98	82.08	lytC N-acetylmuramoyl-L-alanine amidase	Peptidoglycan catabolic process

Promoter	Source	GFP RFU	mOrange RFU	Putative Identity	Gene Ontology Biological process / Molecular function
Cell membrane-related					
15	GTGNS	106.07	17.17	Integral membrane protein TerC family protein	Integral membrane component
58	GKAU	835.97	158.54	mcpA Methyl-accepting chemotaxis protein	Integral membrane component
72	GTDN	3662.86	354.38	ydjM Inner membrane protein	Integral membrane component
77	GSTEA	5331.79	223.34	5-bromo-4-chloroindolyl phosphate hydrolysis protein	Integral membrane component
Miscellaneous					
7	GKAU	98.28	11.75	pepA Cytosol aminopeptidase	Aminopeptidase activity
12	GTDN	101.12	9.72	hmuU Hemin transport system permease	Transport
14	GSTEA	105.80	10.65	mtnK Methylthioribose kinase	L-methionine salvage
16	GTDN	106.67	25.62	Ferredoxin--NADP reductase	Cellular iron ion homeostasis
38	GPGV1	215.12	12.08	YqaJ viral recombinase	Recombinase activity
52	GSTEA	314.30	164.02	cobB NAD-dependent protein deacylase	Glutamine metabolic process
61	GSTEA	1332.51	288.34	amaA N-acyl-L-amino acid amidohydrolase	Aminoacylase activity
66	GKAU	2422.96	274.73	yhdN Aldo-keto reductase	Stress response
70	GTGNS	3451.39	175.18	zapA Cell division protein	Cell Division
73	GSTEA	3808.74	142.70	hcnC D-amino-acid-oxidase	Oxioreductase activity
74	GKAU	3983.82	389.54	Putative heme peroxidase	Oxidation-reduction process
79	GSTEA	8585.09	135.90	Enterobactin/ferric enterobactin esterase	Iron ion transport
80	GSTEA	10571.15	379.11	hmp Flavohemoprotein	Response to nitrosative stress
Nucleic acid-related					
2	GKAU	83.99	14.59	dnaG DNA primase	DNA replication
17	GSTEA	106.99	13.00	dps DNA protection during starvation protein	Cellular iron ion homeostasis
21	GKAU	109.52	16.29	dps DNA protection during starvation protein	Cellular iron ion homeostasis
46	GTGNS	11115.08	61.66	glcR HTH-type transcriptional repressor GlcR	Transcription Regulation
54	GTDN	438.67	103.75	yfhQ Adenine DNA glycosylase	Base-excision repair
63	GTDN	1982.95	153.12	sigW RNA polymerase sigma factor	Transcription
76	GSTEA	4778.75	169.79	gltX Glutamate-tRNA ligase	Glutamyl-tRNA aminoacylation

Promoter	Source	GFP RFU	mOrange RFU	Putative Identity	Gene Ontology Biological process / Molecular function
Protein folding and secretion-related					
24	GSTEA	112.45	16.50	groL Chaperonin	Protein refolding
48	GKAU	2279.05	19.69	SecG preprotein translocase subunit	Protein secretion
50	GSTEA	218.37	301.62	Putative bifunctional phosphatase/peptidyl-prolyl cis-trans isomerase	Protein folding
Sporulation-related					
3	GKAU	84.71	11.38	yabP Sporulation protein	Sporulation resulting in formation of a cellular spore
10	GTGNS	100.32	9.87	divIB Cell division protein	Sporulation resulting in formation of a cellular spore
11	GSTEA	100.94	9.26	Gamma-D-glutamyl-L-diamino acid endopeptidase	Sporulation resulting in formation of a cellular spore
19	GTGNS	108.24	12.04	sspD Small, acid-soluble spore protein D	Sporulation resulting in formation of a cellular spore
20	GTGNS	108.83	9.52	cwlJ Spore coat protein, cell wall hydrolase	Sporulation resulting in formation of a cellular spore
25	GTDN	114.89	8.04	YabP family Sporulation protein	Sporulation resulting in formation of a cellular spore
26	GTGNS	116.61	10.85	mecA Adapter protein	Negative regulation of sporulation
30	GTDN	126.93	11.70	ctpB Carboxy-terminal processing protease	Sporulation resulting in formation of a cellular spore
32	GTGNS	139.46	83.92	spo0F Sporulation initiation phosphotransferase F	Sporulation resulting in formation of a cellular spore
40	GKAU	356.39	23.72	cotJA Spore coat associated protein	Sporulation resulting in formation of a cellular spore
59	GTDN	1244.59	182.38	ecsA ABC-type transporter ATP-binding protein	Sporulation resulting in formation of a cellular spore
Transferases					
36	GKAU	160.56	93.60	Aminoalkylphosphonic acid N-acetyltransferase	Transferase activity
37	GKAU	179.38	76.61	lipL Octanoyl-[GcvH]:protein N-octanoyltransferase	Transferase activity
75	GSTEA	4768.29	355.45	yrrT Methyltransferase	Transferase activity

Promoter	Source	GFP RFU	mOrange RFU	Putative Identity	Gene Ontology Biological process / Molecular function
Unknown Function					
1	GSTEA	75.01	8.74	Hypothetical protein	n/a
5	GKAU	95.81	12.39	Hypothetical protein	n/a
13	GTGNS	101.81	9.00	Hypothetical protein	n/a
22	GKAU	109.63	12.29	Hypothetical protein	n/a
23	GTDN	110.48	23.59	Hypothetical protein	n/a
27	GSTEA	117.47	11.07	Hypothetical protein	n/a
28	GSTEA	118.11	14.48	YugN-like family protein (uncharacterised)	n/a
29	GTGNS	121.19	13.72	Hypothetical protein	n/a
31	GKAU	130.44	75.44	Hypothetical protein	n/a
33	GSTEA	140.90	33.40	Hypothetical protein	n/a
34	GKAU	143.87	44.62	Hypothetical protein	n/a
35	GTGNS	144.61	35.58	Hypothetical protein	n/a
45	GTDN	868.61	86.09	Hypothetical protein	n/a
49	GTDN	113.02	154.27	Hypothetical protein	n/a
56	GSTEA	734.10	233.05	Hypothetical protein	n/a
57	GTDN	735.03	164.92	Hypothetical protein	n/a
62	GKAU	1950.40	155.49	Hypothetical protein	n/a
67	GSTEA	2571.00	452.36	Hypothetical protein	n/a
68	GTDN	2846.23	102.78	Hypothetical protein	n/a
71	GKAU	3582.83	193.07	yugN-like family protein, unknown function	n/a
78	GSTEA	6405.40	557.41	Hypothetical protein	n/a

Supporting Table 3: Analysis of the native genes with which the characterised promoters were originally associated.

Promoters highlighted in yellow are those which were identified as functioning consistently independent of the downstream CDS. Source organisms: GKAU = *G. kaustophilus*; GTDN = *G. thermodenitrificans*; GTGNS = *G. thermoglucosidasius*; GSTEA = *G. stearothermophilus*. Gene Ontology information was obtained from the UniProt database.

Parameter	Levels specified
Activation function personality	Gaussian, Linear or TanH
Number of nodes in the hidden layer	3, 5 or 7
Cross validation methodology	KFold with K = 4, K = 5 or K = 8
Number of sequence positions modelled	10, 20 or 100

Supporting Table 4: Artificial Neural Network parameters included in the architecture optimisation screening design, and the values specified for each parameter.

Promoter	DNA Sequence
1	GTCATGCAGCCATCTTTTCTCCTTTTTTCTCCATTTATAATGCCGCTCGATCGAT TCCATAATGTAAATGGATAGAACGAACAGGACCTGGAAGGATGCCGCCA
2	TACGGGCAAGGGGGCGAAACGAACGTTTTGAAAAAATTTGTCGAAATGGAAAGGA AAATCGGCTTGCTGCCGAGAATAAATACGAACCTTACGGGGTTGTTCGT
3	AACCATCCCCCGCTGACATACATATGGTACAAACAACGGATGAGAAAAGGCGGTT GGCCTTGCCGGCTTTCTATACTAAATGGAAACCTATGCGGGGGATGAAC
4	TTCGGTAAGATGGGAGAGAATCGGCGGCTTCAGCTCTGCTATTTTTGTGCGCTGT AGGATACATTAATAAGTACTGAAGCGACGCACCTACAGGAGGGGAAGGC
5	CATCAAGAAATAACTGAAATCCGTGAGAGATGCCTATACTTCGCTTCGCCTCCTT GCATAGTCTAACAGTGCAACTAGGCGAAAGACCTGAAGGGGGGAAGGGC
6	ATGCCCTGGGCGCCGGTCGGCGCTTTTTTTGTTTATTCCCGCCAGTTCGCGGAGA TACTATCATACTACATACCATCTAAAGGCAACCTGGAGGATATCGGGGC
7	CCGTTTCAGCTGATTTTTTTGCATGAACGCAAGGAGAGAAAAGTTAGATAGTCGA AAGAATCGCTTTGCTTCGCTATAATGAAGGACCTTGGGAATGAATCAAC
8	GTCAGACTTCTGACGAGTGCAATGCTCAGGAGTTTCTATCAGTATTTCGGGAATG ATAAAACAGAAAGATCACGTTATGAAGACAACCTGGAGGAATTATGATT
9	CGATCGTTGTCAAAACATTGTCGATACGAATAACCGCGCCTGTGCGCAATGAGGG CGCATGGCGGCCCATTTACACTACACGACAACCTAAGGAGAGGGTAATC
10	AAAAACATGTTTGTCTGCGTTAGGAACGTAAAATATAACGGCATCAGATGTGCCG TTTGTTTTTATAAACATGCTTTCGAGTTTAACCTTTCGGAAAGCGGTGA
11	CCGCAACGGCTTGCTCCCTTGCTGGAGCAAGCCGTTGTTTTCATTTTTTGGCGCC TGCCGCATATGTTGAAGGTAGAAGAAAGGGACCTAAGGGGAGATGGCGC
12	GTTAGGTTGCGCTCGGTTGTCACTTGATTGAGTTTTTATGAGGCGGTTGAACGAG TGGGATGAGCGCAGATAAAAGAAGACAAACACCTGGTAGGATGAGGACA
13	TTCCTCCCTTGTTATTGTGTCATTGTTAGTTTGTACGGTTTCGGCGGATTCATTG GTGGAAAATTTTCGCATTGGATGTTAAAGGACCTGAAGGGGAAACGGTG
14	TCTGAAAGATAAGAGGCGCGAAGACGATGCATCTTCAAGCCTCTTTCCGTATGCG AAAGAGGCTTTTTCATTTGCCAAAACCTAAACCTAGGGGGAACGTGAAA
15	AAAGAAGGGAAATAAAAAAACCACGTTTTTAAAAATTGGTTATGAAATGAAACAT GCCCTACTATAATGGTGGTGAAAACAAAAGACCTAAATGAGGGGCACGC
16	AACCGTGGACGTTGAAAATCGCAACATTCCTATTTTATCATAGTTGGCAACGGAA CGAAAAACATTGTCTGCTTGTAGACAACGAACCTAAGGATGGAATCCAT
17	TCATATTTCCGGTGAAATCGGGCAAGCTATCCATCGACACGTTTTCCATGTTTCA TCAACAAGCCAGCTGGGTACAAGAACCATAACCTAGGAGGAGTGAGAGG
18	CCCCCCTCGCTTAAAGCCAAATGAAAGCTTCCAATTGCATCTATAAATTATAGAC AAAGGAGTTTTGCTGCTATATTATAGACAAACCTAGGGGTTTTGTTGCG
19	GCAGAAAACAGCATAATTTTCATCAACTTCCGTTATAGAGGATTCTCGATTGACA CATACTAACATTGCCAGACAACTAACTATGACCTTTAGGAGGAAATACA
20	ATCAGAAATTCACTTTCAAAATGGATTGTCCAATTGTCACCATTTTTTCCTTATA

	TCAATAGCCTAATATTGTTGAGAAAAAGACACCTAGTTGAGGTGACAAT
21	CTTCTGGGGAGTTGCAAGCCCCCACTTCAAGCGTTCGCGAAGTGGGGGTAGTTGA
	CCAACAAGCCAGCTGGGTACAAGAACCATAACCTAGGAGGAGTGAGAGG
22	AAGCACCCAAAGAAGTTTCATCGCCAACTGGACGGCGCGCATAGGACAATGGCCG
	TTTCGAATAGGCTATAGTAACAGGGCCATCACCTGAAAGGGGAATCGGG
23	ATAATAGGAAAATCGCGCAGGATATCCGGACGAAATCGAAGCATATTCGAGGATG
	TCTGGCATATGCTCGAACAAACAGAGGAAGACCTAAAGGGGAGAGAGGC
24	TAAACATTCCGGAAAACTGACATATCCTTAACTTTATGACACGAATACTTACCTT
	TTTCACGGCATGCAGTGAAACAAATTCATAACCTAGGAGGTAACGGGGT
25	GCAATGGGCCTTGCCGACAGGCGGGCCCGTTGTTTTTGAAAAGCGGTGCTAGAAC
	CCCCTTTTGTCTCATACATATGAGATGAAAACCTGGGGGTTCTTTTTAA
26	AATAAAAGTACAAGATGTTATCTTGTATTTAAATAGAATGTTTTGGGGGTAAAAA
	TCCCTTCAACTTTTATTCTTTATTTAGGAAACCTGGGAGAGTCTGCGCT
27	CCCCTTTTTTGGCAAAGTTCAGTCTATATGTATGATGGAGCCATGTCGTTCATGC
	CTGTTTGCCATGACGAATAATGAGGATCGGACCTGCTAAGGAGGAAACC
28	CTATATCGCCCCCTGCCGTTTTTCACCAGGCGTCTGAAGTCGCCGCGGTCAAATA
	ATCCCCATCAAACGGGAAACGATAAAAAAGACCTAAGGAGGGGTTGTCT
29	ACTTCAACCATCATGGAGGCGGAAAACCATTTCTTGTTTTCTCGCATACTTTATG
	AAAGGGATATGTTTCTATTCTCAGCCAAGAACCTAAGGAGGCAAAAACA
30	GAAGCATACTCAGTACAAGCCTGACATATGATGGAGATAAGGGCATGGCACGGCA
	TGATGCGGTGCTTTTCCTTTATGTATAGGGACCTTGAGGAGGAAACAAC
31	GGCAAACTGCTTGTCTATTATATCACAACCGATAGAAAATCGGCATGTTTTCCGG
	TACACTAGAAAAAGACGCCCACTTTGCCAGACCTAAAGGAGTCAACCCA
32	ATTTATCCAAATGCAAAAAAAAATGGCAGGAAAATAACTGAAGGATGCGAAACTTA
	TAACATGGAACAAAATAAATGAAAGAATTGACCTAGGTGTGTCTATAAG
33	CACCGTCGGTCTGCTAGTTCGTTCCAAAAAAATGATTGATTTTCCGCCGCTTTTG
	CCCCTACAATGATACAAGAACACTGCACGGACCTAAAGGAACAGCACGA
34	AAAAGAGGAATTCCTTTCTTTAAAACGTCGATTTTGTTGATTTCTTCTTCGCGAT
	GGGACCCCGTTTTATGGTATAGTGGTCTATACCTGTAGGAGGGACTGTG
35	CTTTCTTTGCCGAAATGGTTCTAGTTTTGTCGTGCCGGAAGGGAGTCAAGGTTTT
	ACCGCGAATTGATTATCATCATTGTGAAAGACCTATGGGGGAGATGACA
36	CTTTCGAAACGATGATCAACGTTGCGGAAAGGCGGAAGTTCCATTTTTGGACAAA
	CATGGATATCATGAATGTAGAAATAGAAGAACCTAGGGAGATGGGAGTC
37	GAGGGAGGCATGCGCCCCTTGTCCTTTTTGCGCCATTTGCTAATGGCGATTATGT
	TATAATGAATAGAGTTGGAATGATCTATGCACCTAAGGGGTTGCTCGCG
38	ATAACTCATCTTATAAATACTTCATCAAGTCAATTTCCTCCTTCCTTTCCAATGT
	AAATATATGTATAATTTTCTTGTATCAAAAACCTAGGGGGGGAATTTGT
39	TACATAAATCGTTGCTATTTTCACCTTTTTTGTTTTTCCAGACAAGTGTGCGCGC
	TATGGTAAAATAAAGAGGAGTCATCGGGAAACCTCAGGAGGATGAAGCG
40	TTCCTCCTTATTATTGCCATGATTCGCCATGACAAACAGGCATGAACGGTACTGC
	TCTATCATAATATAGACTGAACTTTGCCAAACCTAAAAGGAGGGGTTTC

41	CAAAACAACCGCTTTTCTTCATTTCCAGTCTACCGTTTTAACCGGCGCATGATAA AATAGAAACATGCACAAACTTGTTGAGTAAACCTAGGGTGTTTGATTCG
42	TTTACATAACATTTTTATGGTGAATTCTGTTTGTTTTGCATCCCACCGAACGATT TAGTAGAATGGGAGAAGGACACGACAACAGACCTAGAGGAAGACGTAGG
43	CGCCTGTTCTGTCGAGCGGTTTTCCTGGCCGACAGACAGGAAATGGCCTATTGTG CGCGAACCATATAGTAGAATGAGAGGGAACACCTTAGGAGGGAGCGCGC
44	ACGGCTCTTCACCAAACGTTATGTTGCATCTATATATGCTTGATTTTCTATGTTT TGTGATATAATAGACGAGCAGCTTGTCTAAACCTCAGGAGTGACCAACA
45	AGGATTGGCAAATTGCAAGAATGAAAGCCGCCCGCTTGCAATTTTCCTGCGAAAG CGGTAGAATACACAAAGCAAACGCTGTGGAACCTAAGGGATGACAAGTG
46	TTTTTTTGAGTGTTTTTGATTGAAAATGATTGATTTTTAAATCGTTTTCATTTAT GATAAGGGTACAAAGTAATCATTTCCATCCACCTAAGGAGGTGGGATCG
47	AAAAACGCTTTTTCCAGAACAGACTGGCGGTACGCTTGTCATTTTTCTTGCCAGC CCATATAATGAGGGAAGATGCGAACGGGAAACCTAGAGGAGTGGGACGA
48	TGTTCCTTTCGTCGTGTTGAGACAGGCCCGCCAGTTGCCGCCAACGGCGATCTAT GGTATAGTGAAATTATTGCAACGTCGGTCGACCTTTGGAGGTGTCAGGC
49	CCAGCTGCCACCCTGCCAGTTCGTTCAAAAAATGATTGATTTCCCGCCGTTTTTG CCCCTACAATGATACAGGAAAACAGCATGGACCTAAAGGAACAGCGAGA
50	TCATGGGGAAAAATGGAACGAACCGTGTATTTCGGTTGACATTCCAATTGTTTTT CAGTACGGTAAACAGTGAATACATTTTCGGACCTCAAGGAGAAGATAAC
51	ATGTCAGTCCACTCACTGCCAATAGTGACAAAAATCATTCCTCGCTTTGGACGAT TTCATTATAATAACCATATCAATGATTCATACCTTTAGGAGGGAAACCA
52	CTAGAAACGAGTGAAAAGGAAAAGTAAAATGATTGGTGGAAAGAGATGAAGCATC AGAAATGAGAGTCGCTATGGTTGCGACAGTACCTAAGGAGGGAAACGGA
53	TTTTCCCACTCCTATGATACAATCATTGTGAGTAGTGTCACAGCGTTTGTCTGCT TGTTTCCGTATAGTAGATGGGGATGTGGCAACCTAAGAGAGGTTGATTC
54	TGTAGTGGATCACCCAGCCGTTTGTCTTCCTGCATGCTGCAAGTCATGTTGGGGG TTTTTTGGTATAATGGGATGGACAATAGACACCTACCATGGGGGACATA
55	ATATAATAACAAACTGTATGCCCGCTATATAGCGATTTGCAGCAAGATTTGCTAT GATGGAGGCAGATTGAGTCATGCAGATGGAACCTAAGAAGGTGCTGCAT
56	GCCTTGTGCCAAATCATCGCGCTTCCGCGCCATCATTGCCTGGGCGGACAAGTTC CTCTATAATGAAGACGCCGACTATTTTTGCACCTGGACGGAGGGACGAT
57	GCCAGTCGTCAATTTTCCCTGCTTTCGACCATCGTTGCCTGCGCAGGCAAGTTCC TCTATAATTAAGGACGCCGACTATTTTTGTACCTGGACGGAGGGACGAT
58	ATTATGGCCAGGAAGGAAAATGAATGATCATTCACGAATCTATTTAAGCGGTGAA GATAAAGAAAAGACGATTGCATAATGGAAAACCTAGGAGGAGAACGACT
59	CATTTATACAGGACGACGAAAAGAACTTTTTTGGTAACGATAGACGTTCTTTGGT AAAATGGGGACAAACGTGATTCAAGATACGACCTGAAAGGAAGCGAACG
60	ACGGGCGACTTTTTTTGTGCACCTTGTCTATTCACGAATATTCACATTCATCTAT ATAATATGAGTGTGAGAAAACGATTTCAATACCTAAAGGGGGATTCGTG
61	AAACGTGCAGCGGGCTGCCTTTCTTCATGAATATAGAGAAATGAAAATTCAAAAA

	TTTCCGTTTATAATATTAGGAAAGCGGAAAACCTAGAGGAGGAAACGTG
62	AGCGTAAGCGAAACGCTCCTTTTTTTGCGCCCTCGGCGGTACAACAGCGGCGAAA
	ATCATGGTATGATGGAAGAAAAAGAATGACACCTTAGGGAGACGTGGCA
63	CCTCCCCTCTTTTTTCAGTTCTATGATACAATAAACATGTTGCATGAAACTTTTC
	TTCATCAAGCGCCGTAATATAAGATGAGCCACCTGTAAGGGCGGGGGTA
64	TTCCAAAAGTTGTGTATATTCGATATTATATTAATAGTATAAACTAGATATTGTC
	TACTGCTAAAAGTCCCATTGATCACCATAAACCTAAGGAGTTTGATGTA
65	GTCTTTTCCTTTCGTTTCGCCAGCGAGAAGCAGCCGGCAAGAGTTGGGTAATTTT
	TGCTATACTATAGCTAACAACACAATGACAACCTAAAGGAGATGAGGGA
66	ATTGATGGACTCTTTTCCATTGTACTACATGTTTGGCCATTTTTTCGCTTCCGTT
	GTACAATGAACATATCATCATCGATCGCGAACCTAAGGGGGCGCAAAAG
67	ATACATGGCTCGTCCTCCTTTTTTCTTCCCCGCCTTCCTCCCTTTTTCGCCGTTC
	GTCTGCTATGATGGACATAGGCGAAAAATAACCTGAGGGGGGAATGGAG
68	TTTTTACATAATCTTTTTTGGCCATCCTCCCCTCTGACATCTCTCCGCTTTCCAT
	TTTTCTTTCATTCATGGTATGATGAATAATACCTGGAGGTGCCAGATTT
69	AGATATTTACAGCATTTTGCTCGATGAATTTGTGCTTGTCACAATGATCTCTGTT
	TTGCTAAAGTAAAAAAGAATACAGGTGTAAACCTAGGAGGATATCGCCC
70	CATCGTTTTTTCTTAGGCATATCTAGTAGTCGTATCTTTTCATTATATCAGCATT
	CATGTTATGATAAACAATAGGATTTTGAAGACCTAATGGGGGAAATCAT
71	AACAATTCGTTGAAAAGCAAACCGTTTTCGTACTATTATAAATGAGGGGATGATT
	TCCTGTTTGGCGCTTTTACATACTGTGTTGACCTTTAGGAGGAAGAAAG
72	GTCACATTTGGTCGCATTGTCACAAATTCGTCAAAGAAGCGAACAAATATTCTTT
	GTTATGGTATAATGGGGAAAACAAGGAAGTACCTTTGAGGGGATGTTCG
73	CGCCTTTTTGACGCCTGCCGGCTTTCGGGCCGGCCCACACCGGTTCCCGATTGGC
	GCCGTCTATGGTATACTAGTAGAAAAAAGGACCTAAAAGGTGTTCATGA
74	ACAATTCAGTCTACTTCCACTTTGATACGCTGATAGCCGCTTTTTTGGGCAAACT
	GTGATATAGTGATACTGTACAATTTGTTGAACCTTAAGGAGTGACGATG
75	ATCATTGAACGGCGGAAGAAGTTATGGTACAGTGGGAGCAGGTTATCATCTAAAG
	AATACTATGTTTATAGGAATTATAAACTTTACCTATGGAGTGAGGCCAT
76	ACCGTTAGGCGCCGAATGGGGCGAGCCGTCCGTTTTGAACTTTGCCGCGATAAAC
	GGTACAATAAGGGTACAAGTTTGGATGAAAACCTGTGGAGGGTTGGACG
77	TGCGCGCCGTCTAGTTGGTGATGAAACTTCTTTGGGCGCTTGTACGTATCATGGC
	GTGGACATGTTATAATGGAAGAAAAGATGAACCTAGAAAGGGATGGACA
78	GATTTCCACCTCTAAATTTCATTTCCCTCCATTTTATCTTTTTCCACAAGACCGG
	CCACAAATAGAATGCGATATGATATAGTGAACCTAAAGAGGAGGGATTC
79	TGTTCGGCAAGTTCACGAAAAATTCGAGCCCTTCGTTAACATCTTTTTGCCTAAA
	TATGTTACAATAACAAGTGAACTTGCGCTTACCTTCATGGAGGATGATC
80	CAATTTATTAAAGATGTATTTGAAATGTTGCTTTTTCACTATGCGTATGATTGAA
	TAAAACATGTATTTAAAATACATGTTTTATACCTGAGAGGAGATGAAAC
ldhA	CTGCCTCGTCCATTTTTTTGCTTAATGGAGGTTGTCATGAAAATGACAAACAACG
	TCCAAACAATTGCCATAATCGTTTACGCATAGTTTCGATTTCATCGCGTAAAATA

ATTTGTGAATGTATTCACACCTATAAGAAGGGAGAATAGT

Supporting Table 5: DNA sequences of the characterised promoters.

Supplementary References

84. Yao, A. I., Fenton, T. A., Owsley, K., Seitzer, P., Larsen, D. J., Sit, H., Lau, J., Nair, A., Tantiongloc, J., Tagkopoulos, I., and Facciotti, M. T. (2013) Promoter Element Arising from the Fusion of Standard BioBrick Parts. *ACS Synth. Biol. 2*, 111–120.
85. Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M. J., Elofsson, A., Nørholm, M. H. H., and Daley, D. O. (2015) Enhanced Protein Production in *Escherichia coli* by Optimization of Cloning Scars at the Vector–Coding Sequence Junction. *ACS Synth. Biol. 4*, 959–965.
86. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res. 31*, 3406–3415.
87. Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009) Coding-sequence Determinants of Gene Expression in *Escherichia coli*. *Science 324*, 255–258.
88. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Blüthgen, N. (2013) Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol. 9*, 1–10.
89. Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet. 15*, 469–479.
90. Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009) Design Parameters to Control Synthetic Gene Expression in *Escherichia coli*. *PLoS ONE 4*, e7002.
91. Tuller, T., and Zur, H. (2014) Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Res. 43*, 13–28.
92. Hastie, T., Tibshirani, R., and Friedman, J. (2009) Neural Networks. In *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* 2nd ed., pp 389–416, Springer, New York.
93. Eriksson, L., Andersson, P. L., Johansson, E., and 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). *Mol. Divers. 10*, 169–186.
94. Clyde, M. (2002) Model Averaging. In *Subjective and Objective Bayesian Statistics* (Press, S. J., Ed.), John Wiley & Sons, Inc., New York.
95. Yang, J., Zeng, X., Zhong, S., and Wu, S. (2013) Effective Neural Network Ensemble Approach for Improving Generalization Performance. *IEEE Trans. Neural Netw. Learning Syst. 24*, 878–887.
96. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: A Sequence Logo Generator. *Genome Res. 14*, 1188–1190.