

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

3

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

6

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

9

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

12

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

15

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

18

19

20 **Running Title:**

21 Promoter Design for Industrial Applications.

22

23

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

25

26

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

28

29

30

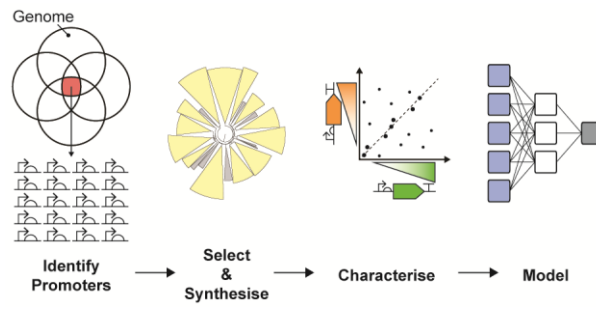
31 **Abstract**

32

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

56

57 **Visual Abstract**
58
59



60 The predictable control of genetic modules or engineered metabolic pathways
61 is a defining aspiration of synthetic biology¹ requiring thoroughly characterised,
62 robust genetic parts. Although synthetic biology parts and tools of increasing
63 sophistication are available²⁻⁵, the majority have been designed for use in a small
64 number of model organisms⁶ and characterised only or mainly in these biological
65 contexts⁷. Model organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*
66 are invaluable for laboratory-scale, proof-of-principle investigations and are used in
67 some industrial applications⁸ but there is a real, practical need to expand the range of
68 microbial chassis available for industrial applications that present more extreme
69 environments for the biocatalyst^{9,6,10-13}.

70

71 Different control points affect the output of gene networks, including levels of
72 transcription, translation, protein half-life and enzyme kinetics¹⁴. On a practical level,
73 the use of promoters with varied and predictable activation and output characteristics
74 (“strengths”) are an essential feature of any synthetic biology toolkit^{3,15,14} and are
75 particularly useful for balancing differential expression levels in “hard-wired”, steady
76 state genetic modules¹⁶. Promoter collections for synthetic biology applications
77 should therefore cover a broad range of recombinant gene expression levels for
78 nuanced tuning of synthetic pathways¹⁷, with individual promoters providing
79 homogeneous, consistent and predictable outputs independently of the associated
80 downstream coding sequence¹⁸.

81

82 Conventionally, promoters in atypical chassis may be isolated from upstream
83 of genes or operons¹⁵ that are homologous to well-understood regions in model
84 organisms, or identified using genomic or transcriptomic analyses of the host⁷
85 followed by in-depth characterisation in a range of genetic and environmental
86 contexts. Alternatively, synthetic promoter libraries may be manufactured by
87 mutagenesis of wild-type promoter sequences, again followed by deep analysis of
88 novel activity^{14,19,20}, though this approach tends to reduce, rather than enhance,
89 promoter strength^{9,21-25}. Finally, recent advances in DNA synthesis have facilitated
90 systematic approaches to promoter and regulatory sequence design by enabling the
91 production and high-throughput screening of comprehensive sequence libraries^{26,27}.
92 Due to the scale of DNA synthesis required, however, this approach remains
93 relatively expensive compared to mutagenesis and dependent on ready access to
94 appropriate DNA synthesis facilities.

95

96 In this investigation, we used a bioinformatic approach to explore the
97 promoter design space in *Geobacillus thermoglucosidasius*, a metabolically
98 versatile^{11,28-30}, thermophilic microbe³¹ with high potential as a synthetic biology
99 chassis for industrial applications^{6,32}. To date, engineering projects in *Geobacillus*
100 have relied on 1 of 3 endogenous promoter sequences^{11,33,34}, the most widely used
101 being the oxygen-dependent *ldhA* promoter^{9,11,31,35,36}. Mutagenesis-derived, synthetic
102 promoters have also been reported for the genus^{9,37,38}, though their characterisation
103 is limited to single genetic contexts.

104

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

116

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

121 **Results & Discussion**

122

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

125

126 Different *Geobacillus* species have the potential to be used as host organisms
127 for industrial bioproduction^{6,9,33}. We therefore aimed to identify promoters that could
128 potentially be used across the entire genus. To obtain a suite of promoters that were
129 representative of the *Geobacillus* genus, we sequenced and assembled *de novo* the
130 genomes of 4 *Geobacillus* species that were available when the project started; *G.*
131 *kaustophilus* (DSM7263), *G. stearothermophilus* (DSM22), *G. thermodenitrificans*
132 (K1041) and *G. thermoglucosidasius* (DSM2542). To identify genes that were
133 common to all 4 *Geobacillus* species, single-copy coding sequences (CDS) were
134 clustered into homologous gene families using the GET_HOMOLOGUES software
135 package³⁹. To increase calculation robustness, 3 separate clustering algorithms were
136 used, and the resulting gene families compared. Bidirectional best-hit (BDBH), COG
137 triangles (COG) and OrthoMCL (OMCL) algorithms returned 1,924, 1,914, and 1,902
138 CDS clusters respectively, with 1,886 homologous clusters being identified by all 3
139 algorithms (Figure 1A). The core genome of the selected *Geobacillus* species
140 therefore contained 1,886 CDS; *i.e.* a total of 7,544 homologous core CDS.

141

142 In prokaryotes, the majority of motifs that affect the initiation of both
143 transcription and translation occur in the 100 bp sequence window immediately
144 upstream of the CDS start codon^{40,41}. 100 bp sequences from immediately upstream
145 of the start codon of the 7,544 core CDS were therefore identified as putative
146 *Geobacillus* promoter sequences. BPROM software was subsequently used to
147 classify the 100 bp sequences as putative promoters based on the presence and
148 nucleotide composition of known conserved functional motifs⁴². To isolate sequences
149 that were likely orthogonal to endogenous regulatory pathways, putative promoters
150 were screened against BPROMs list of known Transcription Factor Binding Sites
151 (TFBS, Supporting Table 1), and sequences that contained any known TFBS were
152 discarded. A phylogeny of the 1,489 putative, generic sequences that remained after
153 screening was constructed as a representation of the *Geobacillus* promoter design
154 space (Figure 1B). Although BPROMs list of *E. coli* TFBS may not be exhaustively
155 representative of binding sites that are functional in *Geobacillus*, the lack of extensive
156 genus-specific TFBS characterisation in these non-model organisms renders a
157 genus-specific approach impractical. Given previous successfully applications of

158 BPROM software for promoter identification²⁸, the utilised list of TFBS was judged
159 likely to provide an adequately generic reference for binding site recognition in
160 *Geobacillus*.

161

162 Multiple studies have used promoters isolated from the genomes of
163 bacteriophage for the control of heterologous expression in *E. coli*¹⁴. Putative
164 promoters were therefore also identified from the genomes of 2 bacteriophages,
165 *Thermus* phage Phi OH2 and *Geobacillus* phage GBSV1, which were chosen due to
166 their ready availability on the GenBank public database. Intergenic regions of at least
167 100 bp were identified in both genomes. From these intergenic regions, the 100 bp
168 sequences immediately upstream of the start codon of the adjacent CDS were
169 extracted. The extracted sequences were subsequently analysed using BPROM
170 software to identify putative promoters, and any sequences that contained known
171 TFBS were discarded. 9 putative promoters were identified from *Thermus* phage Phi
172 OH2, and 7 putative promoters were identified from *Geobacillus* phage GBSV1.

173

174 *In vivo characterisation of putative promoters*

175

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

184

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

191

192 A sequence alignment of the 100 selected putative promoters revealed a
193 heavily conserved purine-rich region located at the 3' terminus of the 100 bp
194 sequence space (Supporting Figure 1). Given the similarities in both location and

195 nucleotide composition of the motif to the canonical Shine-Dalgrano sequence⁴⁶, this
196 region was identified as the RBS. We therefore changed the terminology, whereby
197 “promoter” refers to the complete 100 bp sequence, RBS refers to the 15 bp of
198 sequence at the 3’ terminus of the sequence space and Distal Regulatory Sequence
199 (DRS) refers to the sequence from -100 to -15 bp upstream of the start codon.
200

201 To facilitate potential future applications of the promoter sequences in which
202 disparate DRS and RBS might be required, the 100 selected putative promoters were
203 split *in silico* into DRS and RBS parts that were subsequently flanked with type IIs
204 restriction cloning affixes (Supporting Table 2). *In vitro* cloning of the DRS and RBS
205 parts resulted in the insertion of a 4 bp scar sequence at -19 to -16 bp upstream of
206 the start codon, increasing the length of the promoters to 104 bp. The inclusion of the
207 scar sequence was empirically shown to have no statistically significant effect on
208 promoter activity for 20 out of a set of 24 characterised sequences, with significant
209 alterations in regulatory activity hypothesised to be the result of extreme alterations
210 to mRNA secondary structure (Supporting Information, Supporting Figure 2).
211

212 Of the 100 selected putative *Geobacillus* promoters, 5 *promoter::GFP* and 9
213 *promoter::mOrange* constructs could not be successfully synthesised. Furthermore,
214 11 *promoter::mOrange* constructs could not be transformed into *G.*
215 *thermoglucosidasius*; 80 sequences were therefore characterised *in vivo* upstream of
216 both reporters (Figure 2A). The characterised sequences covered a 148-fold range of
217 activity when characterised upstream of GFP, and a 107-fold range of activity when
218 characterised upstream of mOrange. 45 of the characterised promoters showed
219 expression levels for both reporter proteins that were not statistically significantly
220 greater than the negative control, *G. thermoglucosidasius* transformed with the empty
221 pS797 vector. We therefore defined these 45 sequences as inactive. 19 out of the
222 100 screened promoters showed statistically significant activity with both reporters; 3
223 sequences were active with GFP only, and 13 sequences were active with mOrange
224 only (Figure 2B). A comparison of the codon usage of the 2 reporter proteins showed
225 them to be broadly comparable (Supporting Figure 3). The discrepancies in gene
226 expression between the 2 reporters were therefore assumed to be a result of
227 promoter activity, rather than differential codon utilisation.
228

229 To identify the promoters that functioned predictably and independently of the
230 downstream CDS, K-means clustering was used to group the characterised
231 sequences into 5 clusters based on their Euclidean distance from the line of

232 equivalence between GFP and mOrange activity, $y = x$ (Figure 2C). No correlation in
233 *in vivo* activity between the two reporter proteins was observed for the majority of the
234 characterised sequences; clusters 2 and 4 contained promoters that resulted in
235 stronger GFP expression than mOrange expression, whereas clusters 3 and 5
236 resulted in stronger mOrange than GFP expression. Clustering identified 13
237 promoters (cluster 1) with activity that fell close to the line of equivalence, of which 7
238 displayed mean expression levels that were significantly greater than the negative
239 control. The characterised *Geobacillus* promoter library therefore contained 7
240 functionally composable, active sequences, covering activity levels that were
241 between 1.1 and 4.5 times greater than the *G. thermodenitrificans* *ldhA* positive
242 control.

243

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

253

254 In addition to being functionally composable, promoter sequences for
255 synthetic biology applications should ideally yield homogenous, predictable
256 expression of the protein of interest at the single-cell level⁴⁹. Flow cytometry was
257 therefore used to analyse the intra-population variation in fluorescence activity of the
258 characterised *promoter::reporter* fusions in transformed, clonal cultures. Compared to
259 the positive control, the *G. thermodenitrificans* *ldhA* promoter, 98% of the
260 characterised *promoter::GFP* fusions and 73% of the *promoter::mOrange* fusions
261 returned lower coefficients of variance, indicating that the majority of the
262 characterised sequences offered more predictable regulation of protein expression
263 than the current benchmark *Geobacillus* promoter. Furthermore, the 7 promoters that
264 functioned independently of coding sequence all returned lower coefficients of
265 variation than the positive control *ldhA* promoter (Figure 2D). Although
266 subpopulations of cells expressing the reporters were apparent for 4 of the
267 characterised promoters, the performance of these promoters was less variable and

268 therefore more predictable than that of the *ldhA* promoter which has been widely
269 used in studies with potential industrial applications^{9,11,31,35,36}.

270

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

283

284 *Sequence-function modelling*

285

286 Mathematical models with the *pre hoc* capability to determine promoter
287 function could potentially reduce the need for *in vivo* characterisation of large
288 numbers of individual *cis*-regulatory elements. Once a training set of sufficient
289 robustness is established, regulatory elements of the desired strength for a given
290 application could hypothetically be identified from the genome or designed *de novo*,
291 in a manner analogous to tools such as the RBS calculator³. To better understand
292 the basis of promoter function in *Geobacillus*, and to assess if there was a simple,
293 forward method for *in silico* prediction of promoter function, statistical learning
294 approaches were used to derive models of the design space.

295

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

301

302 *Partition modelling*

303

304 Recursive partition modelling is a powerful technique for determining the
305 relationship between a response variable and a set of independent variables without
306 the use of a mathematical model⁵⁰. Partition models were fit to both the GFP and
307 mOrange characterisation data sets. The number of times each promoter sequence
308 position caused partitions in the data set across 100 random forests was quantified;
309 the larger the number of partitions caused by a sequence position, the more
310 important that position was predicted to be in determining promoter activity.

311

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

319

320 *Artificial Neural Network & Partial Least squares sequence-function modelling*

321

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

327

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

333

334 ANNs have previously been shown to return insufficiently accurate predictions
335 when the response surface under investigation is complex and the number of
336 observations in the training data set is small⁵³. Furthermore, although the PLS
337 algorithm was specifically designed to model data sets in which the number of
338 predictor variables is greater than the number of observations in the training set⁵⁴,
339 the extreme scale of the promoter design space (there are 4¹⁰⁰ potential 100 bp
340 nucleotide sequences) compared to the number of empirically characterised

341 promoters was thought likely to result in models with limited predictive power. A
342 reduction in the dimensionality of the modelled design space was therefore deemed
343 necessary.

344

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

353

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

361

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

368

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

374

375 The optimal obtained ANN was an ensemble model that contained 2
376 constituent ANNs. Each of the constituent models used sigmoidal activation functions
377 with 5 nodes in the hidden layer, and modelled promoter activity as a function of 20

378 nucleotide sequence positions. The optimal model returned an R^2 value of 0.9746
379 when applied to the training and validation data sets, and an R^2 value of 0.9691
380 when applied to the test set, suggesting a good fit of the training data and strong
381 predictive power (Figure 4). For both ANN and PLS, models that inferred promoter
382 activity as a function of complete 100 bp sequences showed lower predictive
383 accuracy than models of reduced numbers of sequence positions (Supporting
384 Information). This result validated the use of partition modelling to reduce the size of
385 the modelled design space.

386

387 *Predicting the function of previously uncharacterised promoters*

388

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

399

400 *Future applications of promoter sequence-function modelling*

401

402 The lack of generality shown by the models derived in this investigation was
403 probably the result of the limited number of characterised promoter sequences as
404 compared to the scale of the design space, resulting in training set that does not
405 adequately capture the complexity of the response surface. Although PLS and ANN
406 promoter sequence-function models using comparatively small data sets have been
407 described⁵⁵⁻⁵⁷, the promoter libraries used in these studies contained considerable
408 sequence homology, thereby restricting the complexity of the response surface under
409 investigation. If accurate predictive models of more complex promoter design spaces
410 are to be obtained, a training data set that contains several orders of magnitude more
411 promoter sequences than the 80 sequences used here is likely necessary^{7, 26, 43}.
412 However, the scale of the required promoter libraries might be impractical in non-
413 model organisms.

414

415 Although high-throughput characterisation of libraries containing thousands of
416 genetic parts using techniques such as a combination of flow cytometry and
417 multiplexed DNA or RNA sequencing has been previously described^{7, 26, 43}, such
418 approaches require the acquisition of large numbers of transformants; approximately
419 50-fold library coverage is necessary to achieve accurate characterisation of
420 individual promoters⁴³. However, low transformation efficiencies in many non-model
421 organisms, including *Geobacillus*, preclude the production of libraries of the required
422 scale, potentially limiting the usefulness of statistical sequence-function modelling in
423 these contexts.

424

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

436

437 The models that were derived in this study were based purely on the
438 statistical likelihood of a given nucleotide occurring at a given position within the
439 promoter sequence. Measures of biophysical promoter properties, such as mRNA
440 secondary structures, AT content or the free energy barrier for promoter-RNA
441 polymerase binding were not included on the basis that unsupervised ANN models
442 could potentially learn the effect of biophysical promoter properties without specific
443 terms being explicitly defined in the model. The inclusion of biophysical terms in
444 future modelling attempts may facilitate the derivation of more accurate predictive
445 models^{26,43,58} by providing more information about promoter function than can be
446 gleaned from sequence data alone. Alternatively, the use of distance metrics⁵⁹ as
447 model terms to quantitatively define differences in nucleotide sequence between
448 promoters might also allow for more accurate mapping of the promoter sequence-
449 function design space⁶⁰.

450

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

460

461 **Conclusion**

462

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

474

475 **Materials & Methods**

476

477 *Bacterial strains & plasmids*

478

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

484

485 NEB 5-alpha (New England Biolabs, Massachusetts, United States of
486 America) chemically competent *Escherichia coli* strain (genotype: *fhuA2 D(argF-*

487 *lacZ)U169 phoA glnV44 f80D(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17)*
488 was used for microbiological cloning, storage and amplification of plasmid vectors.

489

490 *E. coli* S17-1 (genotype: *recA pro hsdRm RP4-Tc::Mu-Km::Tn7*) was used as
491 the mobilisation host for the conjugal transformation of *Geobacillus* spp. Transfer
492 genes from the RP4 plasmid are integrated into the genome of *E. coli* S17-1, allowing
493 for the conjugal transfer of plasmids containing the requisite mobilisation
494 elements^{7,61}.

495

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

503

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

506

507 *Growth media*

508

509 All complex growth media were purchased from Becton Dickson UK
510 (Berkshire, UK). *E. coli* cultures were propagated in Lysogeny Broth (LB; 10 g l⁻¹
511 tryptone, 10 g l⁻¹ NaCl, 5 g l⁻¹ yeast extract). Lennox Lysogeny Broth (LLB; 10 g l⁻¹
512 tryptone, 5 g l⁻¹ NaCl, 5 g l⁻¹ yeast extract) was used for co-culture of *E. coli* and *G.*
513 *thermoglucosidasius* during conjugal transformation of *G. thermoglucosidasius*. All
514 *Geobacillus* species were propagated in modified LB (mLB). mLB used a basal
515 composition of LLB, supplemented with 1.05 mM C₆H₉NO₆, 0.91 mM CaCl₂, 0.59 mM
516 MgSO₄ and 0.04 mM FeSO₄⁶².

517

518 For all media types, agar was supplemented as required to 15 g l⁻¹. When
519 required, *E. coli* growth media was supplemented with 100 µg ml⁻¹ ampicillin. *G.*
520 *thermoglucosidasius* growth media was supplemented with 12.5 µg ml⁻¹ kanamycin.

521

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

524

525 The genomes of 4 *Geobacillus* species, *G. kaustophilus* (DSM7263), *G.*
526 *stearothermophilus* (DSM22), *G. thermodenitrificans* (K1041) and *G.*
527 *thermoglucoasidarius* (DSM2542) were sequenced and *de novo* assembled.
528 Genomes were sequenced using an Illumina MiSeq system, using reads with 300 bp
529 paired end sequencing. The resulting raw sequencing reads were trimmed based on
530 quality score using the fastq-mcf tool⁶³ and assembled using SPAdes software
531 (Version 3.5⁶⁴). Following assembly, the genome scaffolds were annotated using
532 Prokka software (Version 1.9⁶⁵).

533

534 The GET_HOMOLOGUES software package³⁹ was used to identify gene
535 families with homologues in all 4 of the *Geobacillus* species of interest. To increase
536 calculation robustness, 3 disparate algorithms were used to cluster homologous gene
537 families: Bidirectional best-hit (BDBH), COGtriangles (COG) and OrthoMCL (OMCL).
538 In all instances, the “-t” option was used to isolate only those clusters that contained
539 single-copy proteins. All other software parameters were set as default. Only those
540 clusters that were common to all 3 algorithms were selected for further analysis.

541

542 Once identified, the core coding sequences were extracted from the 4
543 genomes. Output files were parsed, reformatted to GenBank file format and imported
544 into the Artemis genome browser⁶⁶. For each entry, the 100 bp immediately upstream
545 of the start codon was extracted. BPPROM software⁴² was subsequently used to
546 screen the extracted 100 bp sequences for the presence and nucleotide composition
547 of functional regulatory motifs. Additionally, putative promoters were screened
548 against BPPROM’s list of known Transcription Factor Binding Sites (TFBS, Supporting
549 Table 2). Any putative promoters containing TFBS were discarded.

550

551 The nucleotide sequences of the putative promoters were aligned using
552 MUSCLE software⁶⁷ and the resultant alignments were used to construct a
553 phylogenetic tree using FastTree software⁶⁸. Putative promoters were subsequently
554 manually clustered into 21 clades using FigTree software⁶⁹. Putative regulatory
555 sequences sequences were selected at random for *in vivo* characterisation from
556 these 21 clades. True randomness was achieved by using a random number
557 generator that converted atmospheric noise into numerical values⁷⁰. Initially, those
558 promoters that were selected for *in vivo* characterisation were manually checked
559 using the Artemis genome browser to ensure that they did not overlap with any

560 adjacent coding sequences. Later, to expedite this process, BEDTools intersect⁷¹
561 was used to identify those putative promoters which were non-overlapping.

562

563 Putative promoters were aligned to transcripts of each of the 4 *Geobacillus*
564 species using Bowtie 2 software⁷². Indexes of the genome files were prepared using
565 the “build” command. Putative regulatory sequences were subsequently aligned to
566 each *Geobacillus* genome using Bowtie 2, with the resultant alignments provided in
567 .sam format. The alignment .sam files were converted to .bam format, sorted and
568 indexed using SAMtools⁷³. The resultant alignments were compared against the 4
569 selected *Geobacillus* genomes using BEDTools intersect. The “-v” command was
570 used to report only those putative promoters that were non-overlapping with any
571 annotated features in the genome transcripts. Output files were provided in .bam
572 format, and were subsequently converted to FASTA format using bam2fastx
573 software⁷⁴.

574

575 *Bioinformatic identification of putative promoter sequences from bacteriophage*

576

577 The genomes of 2 bacteriophages, *Thermus* phage Phi OH2 (NC_021784)
578 and *Geobacillus* phage GBSV1 (NC_008376⁷⁵) were selected for analysis based on
579 their ready availability from the GenBank database. The retrieved GenBank files
580 were loaded into the Artemis genome browser⁶⁶ and suitable intergenic regions of at
581 least 100 bp length were manually identified. The 100 bp nucleotide sequences
582 immediately upstream of the adjacent CDS were extracted and analysed using
583 BPROM software⁴² to identify putative promoters. Putative promoter sequences were
584 screened against BPROMs list of known TFBS, and any sequences that contained
585 known TFBS were discarded.

586

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

588

589 Following bioinformatic filtering, putative promoters were synthesised and
590 independently cloned upstream of the coding sequences of 2 reporter proteins,
591 Dasher GFP and mOrange⁷⁶ (Supporting Figure 6). The *Geobacillus* promoter
592 phylogeny (Figure 1B) was used to rationally select putative regulatory sequences for
593 *in vivo* characterisation in *G. thermoglucosidasius*. To maximise the portion of the
594 design space that was empirically explored, at least 2 putative promoters were
595 selected at random from each of the 13 clades of the phylogeny that contained more
596 than 50 sequences. 2 putative promoters were also selected from each of the

597 analysed phage genomes. Initial characterisation of the bacteriophage promoters
598 showed that only 1 out of the 4 selected sequences was active in *G.*
599 *thermoglucosidasius* (Supporting Figure 7). This 1 active bacteriophage promoter
600 was added to 99 putative promoters from the *Geobacillus* phylogeny to create a set
601 of 100 putative regulatory sequences.

602

603 The 100 selected putative promoters were synthesised and cloned into the
604 pS797 vector (Supporting Figure 6). In all instances, the reporter CDS (GFP or
605 mOrange) was followed by the S718 terminator from the *G. thermodenitrificans*
606 NG80 2-oxoglutarate ferredoxin oxidoreductase subunit beta⁷⁷. Putative regulatory
607 sequences were either directly synthesised upstream of the relevant reporter CDS in
608 pS797 by ATUM (Previously DNA 2.0, California, USA), or were synthesised as
609 double stranded fragments by IDT (Illinois, USA) and cloned *in vitro* upstream of the
610 relevant reporter CDS.

611

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

619

620 For *in vitro* cloning, terminator and reporter sequences were synthesised by
621 ATUM in the pJ201 cloning vector. Cloning reactions consisted of 20 fmol of each of
622 the pS797 destination vector and the relevant cloning vectors, with 10 U BsaI
623 restriction endonuclease and 1 U T4 DNA ligase in 2 µl ligation buffer (10x Thermo
624 Scientific FastDigest buffer supplemented with 0.5 mM ATP). Final reactions were
625 made up to 20 µl with ddH₂O. Reactions were incubated for 50 cycles of 37 °C for 2
626 min then 20 °C for 5 min. This was followed by final incubation steps of 50 °C for 5
627 min then 80 °C for 5 min. 10 µl of the incubated cloning reaction mix was used to
628 transform chemically competent NEB 5-alpha *E. coli*, following the protocol described
629 below. Plasmid construction was verified by diagnostic digest, gel electrophoresis
630 and Sanger sequencing.

631

632

633 *Transformation of chemically competent E. coli*

634

635 *E. coli* S17-1 were made chemically competent using a modified version of
636 the protocol described by Hanahan⁸⁰. 5 ml overnight cultures of *E. coli* S17-1 were
637 used to inoculate 40 ml LB at a 1:1000 dilution. Inoculated cultures were incubated at
638 37 °C, with shaking at 220 rpm, until an OD₆₀₀ of 0.4-0.5 was reached. Cells were
639 harvested by centrifugation at 4,500 g for 8 min at 4 °C and resuspended in 8 ml
640 transformation buffer 1 (TF1: 150 g l⁻¹ Glycerol; 30 ml l⁻¹ 1 M CH₃CO₂K pH 7.5; 0.1 M
641 KCl; 0.01 M CaCl₂·2H₂O. Adjusted to pH 6.4 with CH₃COOH, autoclaved, then
642 supplemented with 50 ml l⁻¹ filter sterilised 1 M MnCl₂·4H₂O). Resuspended cells
643 were subsequently incubated on ice for 15 min, and harvested as above. The
644 resulting cell pellet was resuspended in 4 ml transformation buffer 2 (TF2: 150 g l⁻¹
645 Glycerol; 0.075 M CaCl₂·2H₂O; 0.01 M KCl. Autoclaved, then supplemented with 20
646 ml l⁻¹ filter sterilised 0.5 M MOPS-KOH pH 6.8). 100 µl aliquots of competent cells
647 were flash frozen in liquid nitrogen and stored at -80 °C until required.

648

649 For transformation, 100-200 ng plasmid DNA was added to chemically
650 competent *E. coli* of the relevant strain. Samples were incubated on ice for 40 min,
651 then heat shocked at 42 °C for 2 min and incubated on ice for a further 5 min. 700 µl
652 LB was added and the resulting samples were incubated at 37 °C, with shaking at
653 220 rpm, for 60 min. After incubation, samples were harvested by centrifugation at
654 4,300 g for 5 min, and 500 µl of the supernatant was removed. The cell pellet was
655 resuspended in the remaining supernatant, 200 µl of which was subsequently plated
656 out onto LB agar plates, with antibiotic selection as required. Plates were incubated
657 at 37 °C for 16 h.

658

659 *Conjugal transformation of G. thermoglucosidasius*

660

661 Approximately 5 µl of transformed *E. coli* S17-1 was collected from a
662 confluent plate-culture using a microbiological loop, suspended in 600 µl LLB and
663 centrifuged at 4,300 g for 5 min. The supernatant was removed, and the resultant
664 pellet re-suspended in a further 600 µl LLB. Approximately 10-15 µl wild-type *G.*
665 *thermoglucosidasius* was collected from a confluent plate-culture using a
666 microbiological loop, added to the *E. coli* suspension and re-suspended. The
667 resulting bacterial mix was dispensed onto LLB agar plates, in drops of
668 approximately 10 µl.

669

670 LLB plates were incubated at 37 °C for 7 h, followed by incubation at 60 °C
671 for 1 h. The resulting biomass was re-suspended in 1 ml LLB, and used to create
672 dilutions of 1:10 and 1:5 biomass to sterile LLB. 200 µl aliquots of each dilution were
673 spread onto separate mLB agar plates containing 12.5 µg ml⁻¹ kanamycin. Plates
674 were incubated at 55 °C for approximately 65 h.

675

676 *In vivo characterisation of promoter activity*

677

678 To prepare starter cultures of *G. thermoglucosidasius* for promoter
679 characterisation, transformants were picked and restreaked on mLB agar plates, with
680 antibiotic selection as required. Plates were incubated at 55 °C for 16 h. The resulting
681 biomass was subsequently re-suspended in 5 ml mLB. Bacterial suspensions were
682 then used to inoculate mLB to an OD₆₀₀ of 0.1, with antibiotic selection as required.

683

684 3 200 µl sample aliquots per transformant were loaded onto 96-well plates
685 using either a Corbett Robotics CAS-1200 (Qiagen, Netherlands) or a Gilson
686 Pipetmax 268 (Gilson Inc., Wisconsin, USA). To minimise the effect of position
687 dependant bias, to which assays performed in a 96-well plate format can be
688 susceptible⁸¹, sample aliquots were loaded in a Latin rectangle design; no
689 transformant was represented more than once on any given row or column of the
690 microplate (Supporting Figure 8). 96-well plates with lid covers have been shown to
691 suffer from significant loss of culture in the outermost wells through evaporation⁸². To
692 account for such edge effects, wells at the plate periphery were filled with 200 µl
693 aliquots of sterile growth media. Microplates were incubated using PHMP
694 Thermoshakers (Grant Instruments, UK). Incubation was at 60 °C, with shaking at
695 800 rpm.

696

697 Population-level measurements of culture absorbance and fluorescence were
698 taken using a Tecan Infinite 200 PRO microplate reader (Tecan, Switzerland). For
699 measurements of GFP activity, fluorescence excitation and emission values were
700 477 nm and 515 nm respectively. For measurements of mOrange activity, excitation
701 and emission values were 546 nm and 576 nm respectively. In both cases, the gain
702 of the instrument was set at 56. Absorbance of all cultures was measured at 600 nm.

703

704 Single-cell measurements of fluorescence activity were obtained using a BD
705 FACS Aria II Fluorescence Activated Cell Sorter (FACS), equipped with a 100 µm
706 nozzle. A sheath fluid of Phosphate Buffered Saline was used. Culture fluorescence

707 was excited at 488 nm and fluorescence intensity was recorded using a 530/30 nm
708 detector in the case of GFP fluorescence, and a 585/42 detector in the case of
709 mOrange fluorescence. 100,000 events were recorded per population.

710

711 *Promoter sequence-function modelling*

712

713 All sequence-function modelling was performed using JMP pro versions 12 &
714 13 (SAS Institute Inc., North Carolina, USA).

715

716 *Partition modelling*

717

718 100 random forest models were generated for each of the GFP and mOrange
719 characterisation data sets. In all instances, 20% of the available promoter sequences
720 were randomly selected and withheld from model training to serve as a validation set.
721 Each random forest contained a maximum of 100 decision trees, with early stopping
722 if the addition of further trees to the forest did not improve the validation statistic.
723 Each tree was trained on a data set of 26 randomly selected promoter sequence
724 positions, drawn with replacement.

725

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

737

738 *Selection of an independent test set for PLS & ANN modelling*

739

740 To provide an independent test set on which to measure the predictive power
741 of the derived models, 10 promoter sequences were selected and withheld from
742 model training and validation. So that the test set contained promoters with a range
743 of activity levels, the distribution of GFP expression levels of the 95 characterised

744 sequences was analysed. 2 sequences were subsequently selected at random from
745 the 1st distribution quartile, 5 promoters were selected from the interquartile range
746 and 3 sequences were selected from the 4th quartile.

747

748 *Partial Least squares sequence-function modelling*

749

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

763

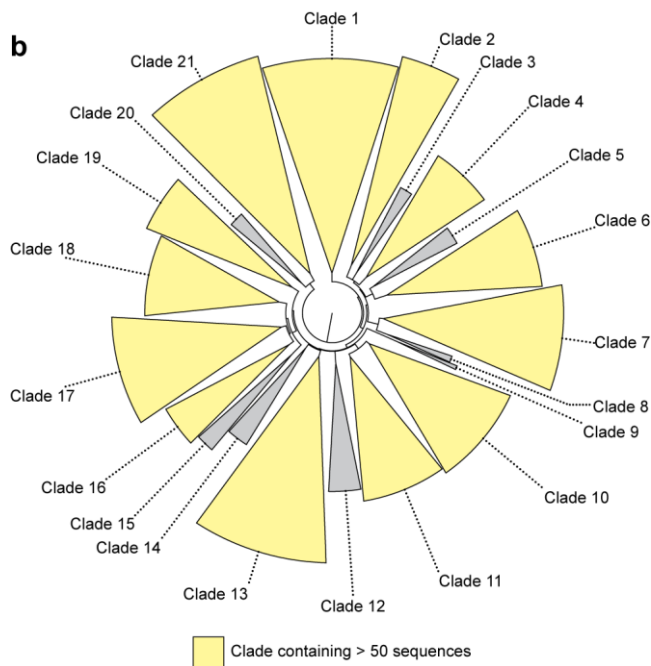
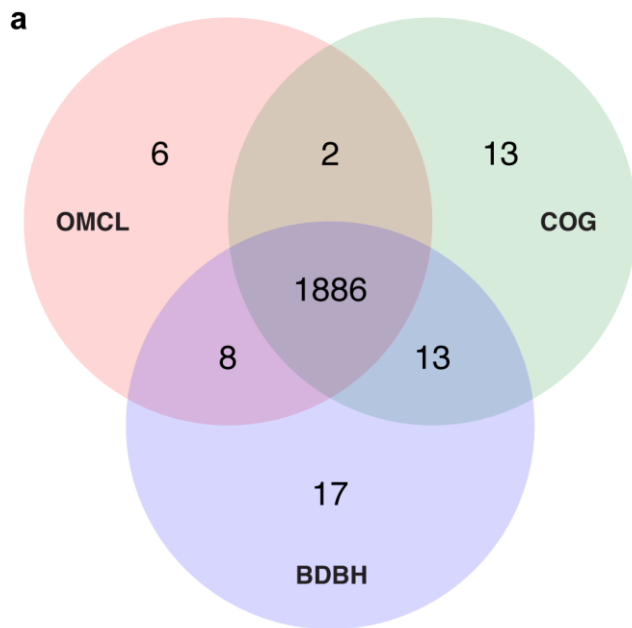
764 *Artificial Neural Network sequence-function modelling*

765

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

769

770



771

772 **Figure 1: Bioinformatic identification of putative promoter sequences.**

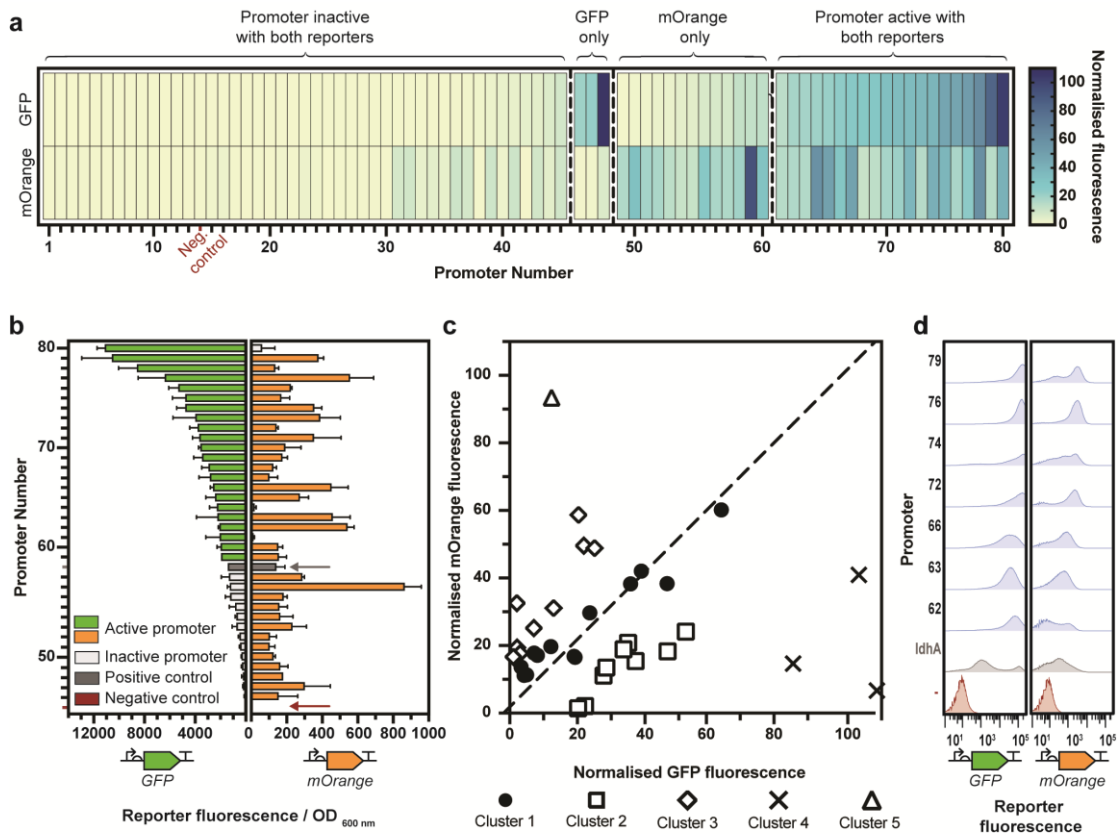
773

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

777

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

781



782

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

785

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

792

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

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

803

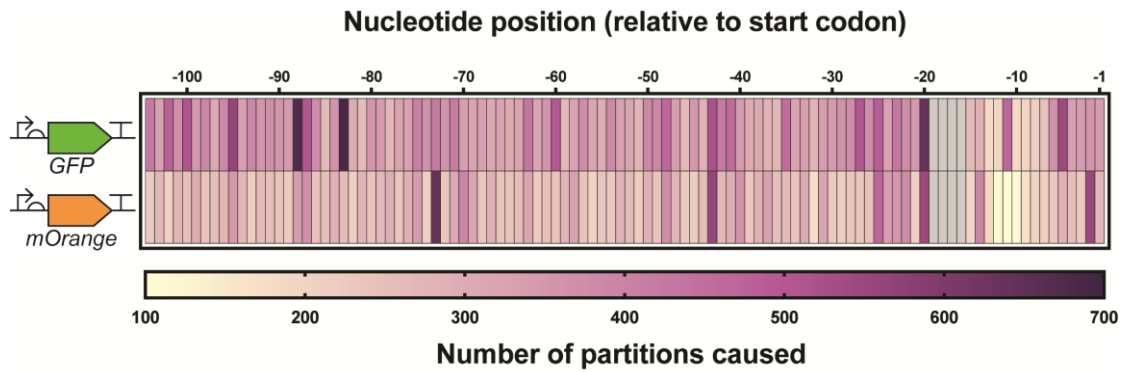
804 B) Expression levels of the promoters for which fluorescence activity was statistically
805 significant. Bars represent the mean of $n = 3$ independent starter cultures arising
806 from independent transformation events, except in the case of the negative controls,
807 where $n = 14$, and the positive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

809

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

814

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



820

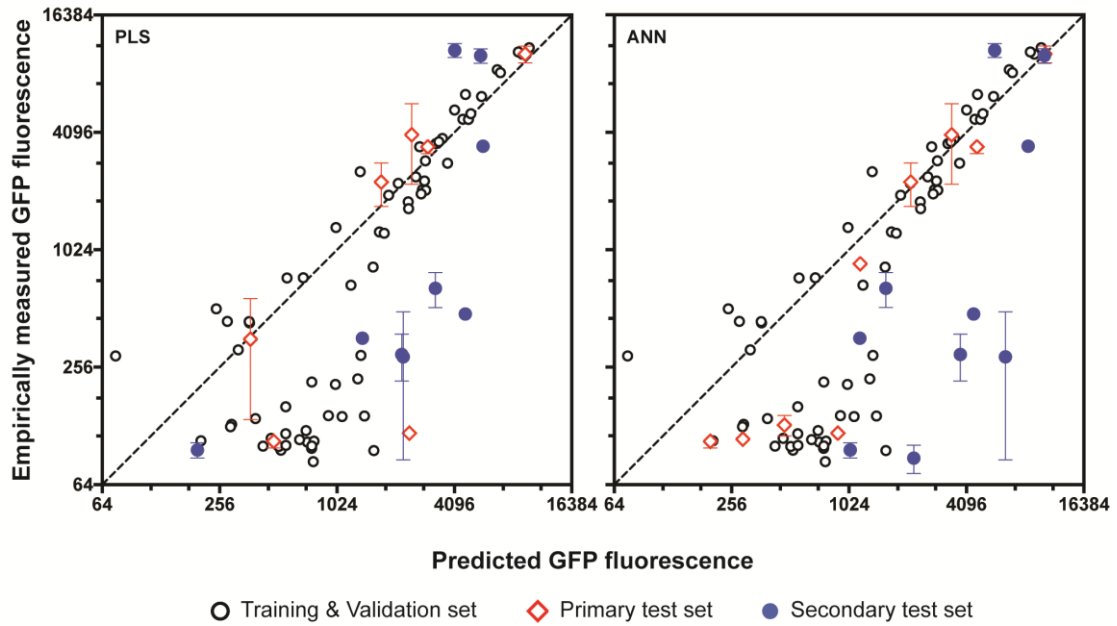
821

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

825

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

830



831

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

834

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

842

843 **List of Abbreviations**

844

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

845

846

847 **Supporting Information**

848

849 The Supporting Information file for this submission contains the following:

850

- 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.

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

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

892

893 **Author Information**

894

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

897

898 **Author Contributions**

899

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

905

906 **Acknowledgements**

907

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

911

912 **Data Availability**

913

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

917 **References**

918

- 919 1. Canton, B., Labno, A., and Endy, D. (2008) Refinement and standardization of synthetic
920 biological parts and devices. *Nat. Biotechnol.* 26, 787–793.
- 921 2. Segall-Shapiro, T. H., Sontag, E. D., and Voigt, C. A. (2018) Engineered promoters enable
922 constant gene expression at any copy number in bacteria. *Nat. Biotechnol.* 1, 1399.
- 923 3. Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome
924 binding sites to control protein expression. *Nat. Biotechnol.* 27, 946–950.
- 925 4. Nielsen, A. A. K., Der, B. S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E. A.,
926 Ross, D., Densmore, D., and Voigt, C. A. (2016) Genetic circuit design automation.
927 *Science* 352, aac7341.
- 928 5. Boyle, P. M., and Silver, P. A. (2012) Parts plus pipes: Synthetic biology approaches to
929 metabolic engineering. *Metab. Eng.* 14, 223–232.
- 930 6. Adams, B. L. (2016) The Next Generation of Synthetic Biology Chassis: Moving Synthetic
931 Biology from the Laboratory to the Field. *ACS Synth. Biol.* 5, 1328–1330.
- 932 7. Johns, N. I., Gomes, A. L. C., Yim, S. S., Yang, A., Blazejewski, T., Smillie, C. S., Smith, M.
933 B., Alm, E. J., Kosuri, S., and Wang, H. H. (2018) Metagenomic mining of regulatory
934 elements enables programmable species-selective gene expression. *Nat. Methods* 15,
935 323–329.
- 936 8. Brown, S., Loh, J., Aves, S. J., and Howard, T. P. (2018) Alkane Biosynthesis in Bacteria.
937 In *Biogenesis of Hydrocarbons* (Stams, A. J. M., and Sousa, D., Eds.), pp 1–20.
938 Springer International Publishing, Cham.
- 939 9. Reeve, B., Martinez-Klimova, E., De Jonghe, J., Leak, D. J., and Ellis, T. (2016) The
940 *Geobacillus* Plasmid Set: A Modular Toolkit for Thermophile Engineering. *ACS Synth.*
941 *Biol.* 5, 1342–1347.
- 942 10. Yan, Q., and Fong, S. S. (2017) Challenges and Advances for Genetic Engineering of
943 Non-model Bacteria and Uses in Consolidated Bioprocessing. *Front. Microbiol.* 8, 2060.
944 doi:10.3389/fmicb.2017.02060
- 945 11. Cripps, R. E., Eley, K., Leak, D. J., Rudd, B., Taylor, M., Todd, M., Boakes, S., Martin, S.,
946 and Atkinson, T. (2009) Metabolic engineering of *Geobacillus thermoglucosidasius* for
947 high yield ethanol production. *Metab. Eng.* 11, 398–408.
- 948 12. Jiang, Y., Xin, F., Lu, J., Dong, W., Zhang, W., Zhang, M., Wu, H., Ma, J., and Jiang, M.
949 (2017) State of the art review of biofuels production from lignocellulose by thermophilic
950 bacteria. *Bioresour. Technol.* 245, 1498–1506.
- 951 13. Olson, D. G., McBride, J. E., Shaw, A. J., and Lynd, L. R. (2012) Recent progress in
952 consolidated bioprocessing. *Curr. Opin. Biotech.* 23, 396–405.
- 953 14. Gilman, J., and Love, J. (2016) Synthetic promoter design for new microbial chassis.
954 *Biochem. Soc. T.* 44, 731–737.
- 955 15. Blazeck, J., and Alper, H. S. (2013) Promoter engineering: Recent advances in controlling
956 transcription at the most fundamental level. *Biotechnol. J.* 8, 46–58.
- 957 16. Brockman, I. M., and Prather, K. L. J. (2015) Dynamic metabolic engineering: New
958 strategies for developing responsive cell factories. *Biotechnol. J.* 10, 1360–1369.
- 959 17. Goldbeck, C. P., Jensen, H. M., TerAvest, M. A., Beedle, N., Appling, Y., Hepler, M.,
960 Cambray, G., Mutalik, V., Angenent, L. T., and Ajo-Franklin, C. M. (2012) Tuning
961 Promoter Strengths for Improved Synthesis and Function of Electron Conduits in
962 *Escherichia coli*. *ACS Synth. Biol.* 2, 150–159.
- 963 18. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q. A.,
964 Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P., and Endy, D. (2013) Precise and
965 reliable gene expression via standard transcription and translation initiation elements.
966 *Nat. Methods* 10, 354–360.
- 967 19. Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control
968 through promoter engineering. *PNAS* 102, 12678–12683.
- 969 20. Jensen, P. R., and Hammer, K. (1998) The Sequence of Spacers between the Consensus
970 Sequences Modulates the Strength of Prokaryotic Promoters. *Appl. Environ. Microb.*
971 64, 82–87.
- 972 21. Mordaka, P. M., and Heap, J. T. (2018) Stringency of Synthetic Promoter Sequences in
973 *Clostridium* Revealed and Circumvented by Tuning Promoter Library Mutation Rates.
974 *ACS Synth. Biol.* 7, 672–681.
- 975 22. Zhang, S., Liu, D., Mao, Z., Mao, Y., Ma, H., Chen, T., Zhao, X., and Wang, Z. (2018)

- 976 Model-based reconstruction of synthetic promoter library in *Corynebacterium*
977 *glutamicum*. *Biotechnol Lett.* 40, 819–827.
- 978 23. McWhinnie, R. L., and Nano, F. E. (2013) Synthetic Promoters Functional in *Francisella*
979 *novicida* and *Escherichia coli*. *Appl. Environ. Microb.* 80, 226–234.
- 980 24. DeLorenzo, D. M., Rottinghaus, A. G., Henson, W. R., and Moon, T. S. (2018) Molecular
981 Toolkit for Gene Expression Control and Genome Modification in *Rhodococcus opacus*
982 PD630. *ACS Synth. Biol.* 7, 727–738.
- 983 25. Blazeck, J., Garg, R., Reed, B., and Alper, H. S. (2012) Controlling Promoter Strength and
984 Regulation in *Saccharomyces cerevisiae* Using Synthetic Hybrid Promoters.
985 *Biotechnol. Bioeng.* 109, 2884–2895.
- 986 26. Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018) Evaluation of 244,000 synthetic
987 sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat.*
988 *Biotechnol.* 36, 1005–1015.
- 989 27. Kosuri, S., and Church, G. M. (2014) Large-scale de novo DNA synthesis: technologies
990 and applications. *Nat. Methods* 11, 499–507.
- 991 28. Bartosiak-Jentys, J., Hussein, A. H., Lewis, C. J., and Leak, D. J. (2013) Modular system
992 for assessment of glycosyl hydrolase secretion in *Geobacillus thermoglucosidasius*.
993 *Microbiology* 159, 1267–1275.
- 994 29. Bezuidt, O. K., Pierneef, R., Gomri, A. M., Adesioye, F., Makhalanyane, T. P., Kharroub,
995 K., and Cowan, D. A. (2015) Genomic analysis of six new *Geobacillus* strains reveals
996 highly conserved carbohydrate degradation architectures and strategies. *Front.*
997 *Microbiol.* 6, 430 doi:10.3389/fmicb.2015.00430.
- 998 30. Zhou, J., Wu, K., and Rao, C. (2016) Evolutionary Engineering of *Geobacillus*
999 *thermoglucosidasius* for Improved Ethanol Production. *Biotechnol. Bioeng.* 113, 2156–
1000 2167.
- 1001 31. Kananavičiūtė, R., and Čitavičius, D. (2015) Genetic engineering of *Geobacillus* spp. *J.*
1002 *Microbiol. Meth.* 111, 31–39.
- 1003 32. Studholme, D. J. (2015) Some (bacilli) like it hot: genomics of *Geobacillus* species.
1004 *Microb. Biotechnol.* 8, 40–48.
- 1005 33. Blanchard, K., Robic, S., and Matsumura, I. (2014) Transformable facultative thermophile
1006 *Geobacillus stearothermophilus* NUB3621 as a host strain for metabolic engineering.
1007 *Appl. Microbiol. Biot.* 98, 6715–6723.
- 1008 34. Suzuki, H., Murakami, A., and Yoshida, K. I. (2012) Counterselection System for
1009 *Geobacillus kaustophilus* HTA426 through Disruption of pyrF and pyrR. *Appl. Environ.*
1010 *Microb.* 78, 7376–7383.
- 1011 35. Bartosiak-Jentys, J., Eley, K., and Leak, D. J. (2012) Application of pheB as a Reporter
1012 Gene for *Geobacillus* spp., Enabling Qualitative Colony Screening and Quantitative
1013 Analysis of Promoter Strength. *Appl. Environ. Microb.* 78, 5945–5947.
- 1014 36. Lin, P. P., Rabe, K. S., Takasumi, J. L., Kadisch, M., Arnold, F. H., and Liao, J. C. (2014)
1015 Isobutanol production at elevated temperatures in thermophilic *Geobacillus*
1016 *thermoglucosidasius*. *Metab. Eng.* 24, 1–8.
- 1017 37. Pogrebnyakov, I., Jendresen, C. B., and Nielsen, A. T. (2017) Genetic toolbox for
1018 controlled expression of functional proteins in *Geobacillus* spp. *PLoS ONE* 12,
1019 e0171313.
- 1020 38. Jensen, T. Ø., Pogrebnyakov, I., Falkenberg, K. B., Redl, S., and Nielsen, A. T. (2017)
1021 Application of the thermostable β -galactosidase, *BgaB* from *Geobacillus*
1022 *stearothermophilus* as a versatile reporter under anaerobic and aerobic conditions.
1023 *AMB Express* 7, 169. doi:10.1186/s13568-017-0469-z.
- 1024 39. Contreras-Moreira, B., and Vinuesa, P. (2013) GET_HOMOLOGUES, a Versatile
1025 Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl.*
1026 *Environ. Microb.* 79, 7696–7701.
- 1027 40. Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B.,
1028 Jimenez-Jacinto, V., Salgado, H., Juárez, K., Contreras-Moreira, B., Huerta, A. M.,
1029 Collado-Vides, J., and Morett, E. (2009) Genome-Wide Identification of Transcription
1030 Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*. *PLoS ONE* 4,
1031 e7526.
- 1032 41. Davis, J. H., Rubin, A. J., and Sauer, R. T. (2011) Design, construction and
1033 characterization of a set of insulated bacterial promoters. *Nucleic Acids Res.* 39, 1131–
1034 1141.
- 1035 42. Solovyev, V., and Salamov, A. (2011) Automatic Annotation of Microbial Genomes and

- 1036 Metagenomic Sequences. In *Metagenomics and its Applications in Agriculture,*
1037 *Biomedicine and Environmental Studies* (Li, R. W., Ed.), pp 61–78, Nova Science
1038 Publishers, New York.
- 1039 43. Kosuri, S., Goodman, D., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Edny, D., and
1040 Church, G. M. (2013) Composability of regulatory sequences controlling transcription
1041 and translation in *Escherichia coli*. *PNAS* *110*, 14024–14029.
- 1042 44. Mutalik, V. K., Guimaraes, J. C., Cambray, G., Mai, Q. A., Christoffersen, M. J., Martin, L.,
1043 Yu, A., Lam, C., Rodriguez, C., Bennett, G., Keasling, J. D., Endy, D., and Arkin, A. P.
1044 (2013) Quantitative estimation of activity and quality for collections of functional genetic
1045 elements. *Nat. Methods* *10*, 347–353.
- 1046 45. Zong, Y., Zhang, H. M., Lyu, C., Ji, X., Hou, J., Guo, X., Ouyang, Q., and Lou, C. (2017)
1047 Insulated transcriptional elements enable precise design of genetic circuits. *Nat.*
1048 *Commun.* *8*, 52. doi:10.1038/s41467-017-00063-z.
- 1049 46. Shine, J., and Dalgarno, L. (1974) The 3'-Terminal Sequence of *Escherichia coli* 16S
1050 Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites.
1051 *PNAS* *71*, 1342–1346.
- 1052 47. Lou, C., Stanton, B., Chen, Y.J., Munsky, B., and Voigt, C. A. (2012) Ribozyme-based
1053 insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.* *30*, 1137–
1054 1142.
- 1055 48. Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A., and Arkin, A. P. (2012) RNA processing
1056 enables predictable programming of gene expression. *Nat. Biotechnol.* *30*, 1002–1006.
- 1057 49. Gasser, B., Steiger, M. G., and Mattanovich, D. (2015) Methanol regulated yeast
1058 promoters: production vehicles and toolbox for synthetic biology. *Microb. Cell Fact.* *14*,
1059 196. doi:10.1186/s12934-015-0387-1.
- 1060 50. Baltagi, Y., and Kussener, F. (2014) Advantages of Bootstrap Forest for Yield Analysis.
1061 SAS Institute Inc, Cary.
- 1062 51. Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K.,
1063 and Gourse, R. L. (1993) A Third Recognition Element in Bacterial Promoters: DNA
1064 Binding by the α Subunit of RNA Polymerase. *Science* *262*, 1407–1413.
- 1065 52. Estrem, S. T., Gaal, T., Ross, W., and Gourse, R. L. (1998) Identification of an UP
1066 element consensus sequence for bacterial promoters. *PNAS* *95*, 9761–9766.
- 1067 53. Bataineh, M., and Marler, T. (2017) Neural network for regression problems with reduced
1068 training sets. *Neural Networks* *95*, 1–9.
- 1069 54. Wold, S., Sjöström, M., and Eriksson, L. (2001) PLS-regression: a basic tool of
1070 chemometrics. *Chemom. Intell. Lab. Syst.* *58*, 109–130.
- 1071 55. De Mey, M., Maertens, J., Lequeux, G. J., Soetaert, W. K., and Vandamme, E. J. (2007)
1072 Construction and model-based analysis of a promoter library for *E. coli*: an
1073 indispensable tool for metabolic engineering. *BMC Biotechnol* *7*, 34.
- 1074 56. Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., and Wold, S. (1993) Quantitative
1075 sequence-activity models (QSAM)-tools for sequence design. *Nucleic Acids*
1076 *Res.* *12*, 733–739.
- 1077 57. Meng, H., Wang, J., Xiong, Z., Xu, F., Zhao, G., and Wang, Y. (2013) Quantitative Design
1078 of Regulatory Elements Based on High-Precision Strength Prediction Using Artificial
1079 Neural Network. *PLoS ONE* *8*, e60288.
- 1080 58. Li, J., and Zhang, Y. (2014) Relationship between promoter sequence and its strength in
1081 expression. *Eur. Phys. J. E* *37*, 86.
- 1082 59. Chen, B., and Yin, H. (2018) Learning category distance metric for data clustering.
1083 *Neurocomputing* *306*, 160–170.
- 1084 60. Li, D., and Tian, Y. (2018) Survey and experimental study on metric learning methods.
1085 *Neural Networks* *105*, 447–462.
- 1086 61. Simon, R., Priefer, U., and Pühler, A. (1983) A broad host range mobilization system for *in*
1087 *vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Nat.*
1088 *Biotechnol.* *1*, 784–791.
- 1089 62. Zeigler, D. R. (2001) Media for growth of *Geobacillus* strains. In *The Genus Geobacillus -*
1090 *Introduction and Strain Catalog*, 7th ed., pp 20. Bacillus Genetic Stock Center.
- 1091 63. Aronesty, E. (2013) Comparison of Sequencing Utility Programs. *The Open Bioinformatics*
1092 *Journal* *7*, 1–8.
- 1093 64. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V.
1094 M., Nikolenko, S. I., Pham, S., Pribelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi,
1095 N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012) SPAdes: A New Genome

- 1096 Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.*
1097 19, 455–477.
- 1098 65. Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30,
1099 2068–2069.
- 1100 66. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and
1101 Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16,
1102 944–945.
- 1103 67. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high
1104 throughput. *Nucleic Acids Res.* 32, 1792–1797.
- 1105 68. Price, M. N., Dehal, P. S., and Arkin, A. P. (2009) FastTree: Computing Large Minimum
1106 Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* 26, 1641–
1107 1650.
- 1108 69. Rambaut, A. (Ed.). FigTree. Institute of Evolutionary Biology, University of Edinburgh.
1109 Available online: <http://tree.bio.ed.ac.uk/software/figtree/> (accessed Jan. 22, 2019).
- 1110 70. Haahr, M., and Haahr, S. (Eds.) (1998) RANDOM.ORG. Available online:
1111 <https://www.random.org/> (accessed Jan. 22, 2019).
- 1112 71. Quinlan, A. R., and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing
1113 genomic features. *Bioinformatics* 26, 841–842.
- 1114 72. Langmead, B., and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat.*
1115 *Methods* 9, 357–359.
- 1116 73. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
1117 G., Durbin, R., 1000 Genome Project Data Processing Subgroup. (2009) The
1118 Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- 1119 74. Kim, D., Perte, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013)
1120 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions
1121 and gene fusions. *Genome Biol.* 14, R36.
- 1122 75. Bin Liu, Zhou, F., Wu, S., Xu, Y., and Zhang, X. (2009) Genomic and proteomic
1123 characterization of a thermophilic *Geobacillus* bacteriophage GBSV1. *Res. Microbiol.*
1124 160, 166–171.
- 1125 76. Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., and
1126 Tsien, R. Y. (2004) Improved monomeric red, orange and yellow fluorescent proteins
1127 derived from *Discosoma* sp. red fluorescent protein. *Nat. Biotechnol.* 22, 1567–1572.
- 1128 77. Feng, L., Wang, W., Cheng, J., Ren, Y., Zhao, G., Gao, C., Tang, Y., Liu, X., Han, W.,
1129 Peng, X., Liu, R., and Weng, L. (2007) Genome and proteome of long-chain alkane
1130 degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil
1131 reservoir. *PNAS* 104, 5602–5607.
- 1132 78. Engler, C., Kandzia, R., and Marillonnet, S. (2008) A One Pot, One Step, Precision
1133 Cloning Method with High Throughput Capability. *PLoS ONE* 3, e3647.
- 1134 79. Kirchmaier, S., Lust, K., and Wittbrodt, J. (2013) Golden GATEway Cloning - A
1135 Combinatorial Approach to Generate Fusion and Recombination Constructs. *PLoS*
1136 *ONE* 8, e76117.
- 1137 80. Hanahan, D. (1985) DNA Cloning: A Practical Approach (Glover, D. M., Ed.). IRL Press,
1138 Oxford.
- 1139 81. Liang, Y., Woodle, S. A., Shibeko, A. M., Lee, T. K., and Ovanesov, M. V. (2013)
1140 Correction of microplate location effects improves performance of the thrombin
1141 generation test. *Thromb. J.* 11, 12. doi:10.1186/1477-9560-11-12.
- 1142 82. Chavez, M., Ho, J., and Tan, C. (2016) Reproducibility of high-throughput plate-reader
1143 experiments in synthetic biology. *ACS Synth. Biol.* 6, 375–380.
- 1144 83. Ho, T. K. (1995) Random Decision Forests. In *Proceedings of the Third International*
1145 *Conference on Document Analysis and Recognition*, pp 278–282, Montreal.
- 1146
- 1147