

# Engineering artificial 5' regulatory sequences for thermostable protein expression in the extremophile *Thermus thermophilus*

Che Fai Alex Wong<sup>1,†</sup>, Shizhe Zhang<sup>1,2,†</sup>, Lisa Tietze<sup>1</sup>, Gurvinder Singh Dahiya<sup>2</sup>, Rahmi Lale<sup>1,2,\*</sup>

<sup>1</sup>Department of Biotechnology and Food Science, Faculty of Natural Sciences, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway

<sup>2</sup>Syngens AS, NO-7089, Trondheim, Norway

\*Corresponding author. Department of Biotechnology and Food Science, Faculty of Natural Sciences, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway. E-mail: [rahmi.lale@ntnu.no](mailto:rahmi.lale@ntnu.no)

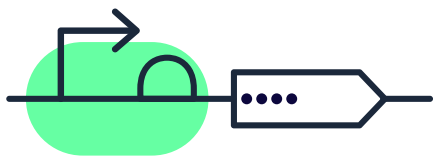
†Che Fai Alex Wong and Shizhe Zhang equal contribution

## Abstract

The utilization of biocatalysts in biotechnological applications often necessitates their heterologous expression in suitable host organisms. However, the range of standardized microbial hosts for recombinant protein production remains limited, with most being mesophilic and suboptimal for certain protein types. Although the thermophilic bacterium *Thermus thermophilus* has long been established as a valuable extremophile host, thanks to its high-temperature tolerance, robust growth, and extensively characterized proteome, its genetic toolkit has predominantly depended on a limited set of native promoters. To overcome this bottleneck, we have expanded the available regulatory repertoire in *T. thermophilus* by developing novel artificial 5' regulatory sequences (ARESs). In this study, we applied our Gene Expression Engineering platform to engineer 53 artificial ARES in *T. thermophilus*. These ARES, which comprise both promoter and 5' untranslated regions, were functionally characterized in both *T. thermophilus* and *Escherichia coli*, revealing distinct host-specific expression patterns. Furthermore, we demonstrated the utility of these ARES by achieving high-level expression of thermostable proteins, including  $\beta$ -galactosidase, a superfolder citrine fluorescent protein, and phytoene synthase. A bioinformatic analysis of the novel sequences has also been carried out indicating that the ARES possess markedly lower Guanine (G) and Cytosine (GC) content compared to native promoters. This study contributes to expanding the genetic toolkit for recombinant protein production by providing a set of functionally validated ARES, enhancing the versatility of *T. thermophilus* as a synthetic biology chassis for thermostable protein expression.

## Graphical Abstract

### Engineering Artificial 5' Regulatory Sequences for *Thermus thermophilus*



**Keywords:** extremophiles; *Thermus thermophilus*; *Escherichia coli*; artificial promoter; thermostable protein expression

## Introduction

Evolution has generated a vast array of biocatalysts across diverse biological systems. Recent advances in metagenomics have begun to reveal this largely untapped diversity, opening new avenues for biotechnological applications [1]. However, to harness these

biocatalysts it is often necessary to express them heterologously in host organisms that can produce them in sufficient quantities while maintaining their functional integrity [2]. Although several microbial hosts have been standardized for industrial and environmental applications [3], the repertoire is predominantly

Received: 7 May 2025; revised 7 October 2025; accepted 8 October 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [reprints@oup.com](mailto:reprints@oup.com) for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com).

**Table 1.** Summary of *Thermus thermophilus* promoters reported in the literature

Type	Promoter(s)	Notes/Reference
Constitutive	15 promoters	Early large-scale promoter screen [10]
	3 promoters	Reporter system enabling precise promoter activity measurement [11]
	13 promoters	Identified during construction of efficient chassis strains [12]
	1 promoter	New plasmid and integrative vectors with inducible and constitutive expression [13]
Regulated/Toolkit	$P_{slpA}$ , $P_{nqo}$	Modular vector toolkit with tailored thermosensors [14]
Inducible	$P_{sip}$	Silica-inducible promoter [15]
	$P_{nar}$	Promoter from respiratory nitrate reductase [16]
	$P_{dnaK}$ , $P_{arg}$ , $P_{scs-mdh}$	Stress- and metabolism-related promoters [17]
	$P_{pilA4}$	Temperature-dependent pilin promoter [18]

limited to mesophilic hosts. This limitation is particularly pronounced when expressing proteins that require the robust folding and stability provided by high-temperature environments [4].

The thermophilic bacterium *Thermus thermophilus* is a unique extremophile. This Gram-negative organism, which has a genomic GC content of ~69% [5], thrives at temperatures between 55°C and 80°C. Its thermostable enzymes such as DNA polymerase are already widely used in biotechnological processes [6]. In addition, nearly 20% of its proteins have been structurally characterized by X-ray crystallography, making it one of the most well-studied thermophiles. Key advantages include high cell densities in simple media, an aerobic metabolism that simplifies laboratory handling, and relatively high natural transformation efficiencies [7].

The repertoire of well-characterized promoters in *T. thermophilus* remains limited (Table 1). To overcome this bottleneck, we employed our Gene Expression Engineering (GeneEE) platform to create artificial 5' regulatory sequences (ARES) that do not rely on prior knowledge of native promoter architecture [8]. In our previous work, we demonstrated the feasibility of the GeneEE approach in seven different microorganisms [8, 9]. In this study, we report the establishment of 53 ARES in *T. thermophilus* and characterization in both *T. thermophilus* and *Escherichia coli*. We further validate these regulatory elements by driving the high-level expression of thermostable enzymes, thereby underscoring the potential of these synthetic parts for enhanced expression relevant for biotechnological applications.

## Materials and methods

### Bacterial strains and growth conditions

*Escherichia coli* DH5 $\alpha$  was used for both cloning and expression. Cultures were grown in Lysogeny Broth (LB; tryptone 10 g/l, yeast extract 5 g/l, NaCl 5 g/l) supplemented with 50 mg/l kanamycin when necessary. For the screening of ARES and expression studies, a knockout strain of *Thermus thermophilus* HB27 with high natural transformation efficiency (HB27 $\Delta$ ago, genome accession number: AE017221) was used. *T. thermophilus* cultures were grown in Thermus Broth (TB; bactotryptone 8 g/l, yeast extract 4 g/l, NaCl 3 g/l), prepared in commercial mineral water (Farris, Ringnes AS, Oslo), and supplemented with 30 mg/l kanamycin when selection was required. Overnight cultures of *E. coli* were incubated at 37°C, whereas *T. thermophilus* cultures were grown at 65°C with agitation at 150 rpm. For solid medium cultivation, TB agar plates were incubated in a sealed, damp plastic box to prevent drying. Due to occasional colony merging on solid TB, only isolated colonies were selected for further experiments. For storage, an overnight culture of *T. thermophilus* was centrifuged at 5000  $\times$  g for 5 min at room temperature, the supernatant removed, and the cell pellet stored at -20°C. Prior to experiments, pellets were thawed at room

temperature, resuspended in 1 ml of TB, and then inoculated into 20 ml of pre-warmed TB in 100–150 ml baffled Erlenmeyer flasks.

For transformation into *T. thermophilus*, we followed the protocol provided by Dr José Berenguer's group. Briefly, an overnight culture was diluted 1:50 in fresh pre-warmed TB and incubated at 65°C until reaching log phase (OD<sub>550</sub> ~0.4). Then, 0.8 ml of culture was transferred to a 12 ml tube and mixed with ~300 ng of plasmid DNA. After a further 4-h incubation at 65°C with shaking, the cells were plated on pre-warmed selective TB agar. Colonies typically appeared overnight; however, plates were maintained for additional nights to allow slower-growing transformants to emerge.

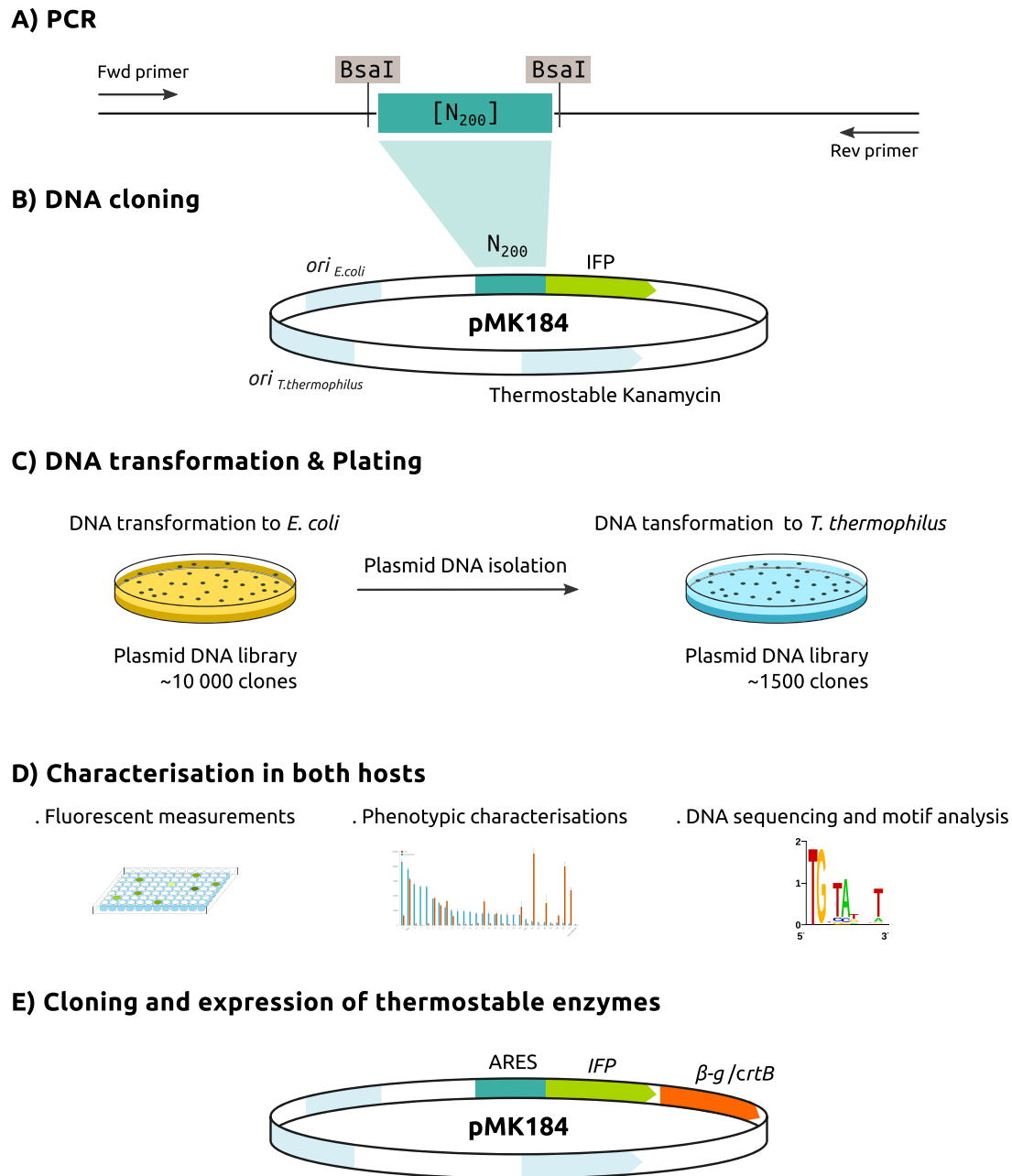
### Plasmid DNA library construction

A shuttle plasmid, pMK184 [19], was used for constructing the DNA libraries. The plasmid comprises four parts: (i) a dual replicon backbone; (ii) a kanamycin resistance marker; (iii) a reporter gene encoding a superfolder citrine fluorescent protein (IFP) from pMoTK110 [14]; and (iv) a GeneEE segment that is 200 nt long with random nucleotide composition [8]. These segments are flanked by primer binding sites (BioBrick Prefix and Suffix, see Supplementary Table S1 for primer sequences) which enable conversion from single-stranded oligonucleotides (synthesized by Integrated DNA Technologies, Coralville, Iowa, USA) to double-stranded DNA and facilitate cloning via Gibson or Golden Gate assembly (Fig. 1A).

Two plasmids were constructed as positive controls: one with the strong promoter  $P_{slpA}$ , and one with the medium-strength promoter  $P_{nqo}$ , both driving IFP expression [14] (plasmid maps are available both in the Supplementary and in our GitHub repository). The backbone pMK184 was linearized using back-to-back PCR with primers Tt1 and Tt2.  $P_{slpA}$  was amplified from pMK184 using primers Tt3 and Tt4, and IFP was amplified from pMoTK110 with primers Tt5 and Tt6. After DpnI digestion to remove parental plasmid and purification (QIAGEN PCR Purification Kit), the fragments were assembled using Gibson cloning [20]. The resulting plasmid (pMK184-Pslp-IFP) was verified by DNA sequencing. To facilitate subsequent Golden Gate cloning, the BsaI sites in pMK184-Pslp-IFP were removed via overlapping PCR (primers Tt7–Tt10) and reassembled by Gibson cloning.

The second control plasmid, pMK184-Pnqo-IFP, was constructed from the BsaI-free pMK184-Pslp-IFP by replacing the  $P_{slpA}$  promoter with  $P_{nqo}$  (amplified using primers Tt11 and Tt12, with exclusion of  $P_{slpA}$  via primers Tt13 and Tt14).

For ARES library construction, pMK184-Pslp-IFP was PCR amplified with primers Tt15 and Tt16 to excise the  $P_{slpA}$  region and introduce BsaI sites. After DpnI digestion and purification, the randomized GeneEE segments (75 ng total) were ligated into the backbone via Golden Gate assembly (30 cycles of alternating 37°C



**Figure 1.** Schematic overview of the study. The plasmid library was constructed by cloning a 200-nucleotide random sequence upstream of the thermostable reporter gene, superfolder citrine fluorescent protein (IFP). Plasmid DNA library was pooled from *E. coli* and transformed to *T. thermophilus*. Approximately 1500 *T. thermophilus* clones, arrayed on 96-well plates, were screened under UV light, and 53 fluorescent clones were selected for further characterization. The thermostable enzymes were transcriptionally fused to the IFP CDS. Plasmid DNAs were subsequently transformed into *E. coli* for comparative expression analysis.

and 16°C for 5 min each, final inactivation at 65°C for 20 min) [21] (Fig. 1B). The resulting library was transformed into *E. coli* DH5 $\alpha$ , yielding ~10 000 colonies. Plasmids were then isolated and used for transformation into *T. thermophilus*; however, due to colony merging on TB agar, only 1500 individual *T. thermophilus* colonies could be recovered (Fig. 1C).

### Selection and characterization of ARES

Using IFP as a reporter, an initial screening based on qualitative visual inspection of fluorescence under UV light was performed on the 1500 *T. thermophilus* clones (Fig. 1D). After 48 h of growth, all the transformants appeared and were immediately screened. Fifty-three clones displaying fluorescence were selected (yielding

a 3.5% hit rate) and cultured overnight in 1.5 ml TB in 96 deep-well plates at 65°C. The next day, a 96-pin replicator was used to transfer cultures into separate plates for storage at -20°C and for plasmid isolation. A negative control plasmid, pMK184 lacking IFP CDS, was used as background control.

For quantitative assessment, selected 53 *T. thermophilus* transformants were individually grown in 1.5 ml TB medium at 65°C overnight (three biological replicates). About 1 ml of the cell culture was centrifuged, and cells were pelleted and resuspended in an equal volume of phosphate-buffered saline (PBS). One hundred microlitres of each suspension was transferred to 96-well black microtitre plates. Fluorescence (excitation: 485 nm, emission: 520 nm) and OD<sub>600</sub> were measured using a TECAN SpectraMax reader,

and fluorescence values were normalized to cell density. Plasmids from the 53 clones were then isolated (QIAGEN protocol) and Sanger sequenced using primer Tt17 to determine their ARES sequences, and bioinformatic analysis of the ARES sequences predicting promoter and transcription factor binding sites was performed [22] (See [Supplementary Table S2](#)).

The same 53 plasmids were transformed into *E. coli* for cross-characterization. Cultures (three biological replicates) were grown overnight in 1.5 ml LB at 37°C, and 1 ml culture was washed and resuspended in equal volume of PBS, and then analysed for normalized fluorescence.

### Cloning and expression of thermostable $\beta$ -galactosidase and phytoene synthase using ARES

To evaluate the functionality of the isolated ARES for recombinant protein production, two thermostable enzymes— $\beta$ -galactosidase and phytoene synthase—were selected as model proteins [14]. The gene variant encoding  $\beta$ -galactosidase (denoted as  $\beta$ -g) and the gene for phytoene synthase (*crtB*) (see [Supplementary Table S3](#)) were cloned by transcriptionally fusing to the IFP gene (Fig. 1E) as follows. Initially, the  $\beta$ -g gene was amplified to include BamHI and PvuI restriction sites using primers Tt18 and Tt19, while the *crtB* gene was amplified with HindIII and PvuI sites using primers Tt20 and Tt21. A pooled plasmid preparation, consisting of all 53 plasmids each harbouring a unique ARES, was then digested with either BamHI-HF and PvuI-HF or HindIII-HF and PvuI-HF to establish the corresponding ARES libraries. The digested  $\beta$ -g and *crtB* fragments were ligated into the appropriate ARES library backbones and transformed into *E. coli*. After scraping all *E. coli* transformants from the agar plates, the plasmid DNA was isolated and subsequently transformed into *T. thermophilus*. For each gene, 12 transformants exhibiting fluorescence under UV illumination were selected for further experiments.

### Measurement of $\beta$ -galactosidase activity

Cell-free extracts were prepared using CellLytic B (Sigma-Aldrich) with 0.2 mg/ml lysozyme. Briefly, 1.5 ml of *T. thermophilus* culture (OD<sub>550</sub> ~0.6) was centrifuged at maximum speed for 2 min. The pellet was resuspended in 400  $\mu$ l of 2 $\times$  CellLytic B solution and vortexed for 10 min. After centrifugation (5 min at max speed), the supernatant was used for enzyme assays.

$\beta$ -Galactosidase activity was measured using *o*-nitrophenyl- $\beta$ -D-galactopyranoside (ONPG) as a substrate. A 5 mM ONPG solution in 50 mM sodium phosphate buffer (pH 7.0) was prepared, and 20  $\mu$ l of cell extract was mixed with 80  $\mu$ l ONPG in a transparent 96-well plate. After incubation at 65°C for 15 min, the reaction was terminated by cooling on ice for 5 min and adding 100  $\mu$ l of 0.5 M Na<sub>2</sub>CO<sub>3</sub>. The reaction product *o*-nitrophenol was quantified by measuring absorbance at 420 nm, with protein concentration normalized using absorbance at 280 nm.

### Extraction and absorbance measurements of carotenoids

For phytoene synthase expression, 3 ml of *T. thermophilus* culture (OD<sub>550</sub> ~3.0) (three biological replicates) was centrifuged at full speed for 5 min. The pellet was resuspended in 1 ml acetone and sonicated for 5 min, followed by centrifugation (5 min at full speed). An absorbance scan from 400 to 500 nm was performed using 100  $\mu$ l of the acetone extract in an Infinite M200 Pro TECAN fluorimeter. The absorbance at 452 nm was recorded as an indicator of carotenoid production [23].

### IFP fluorescence measurements of clones with enzyme expression

IFP fluorescence was measured in *T. thermophilus* transformants carrying plasmids for enzyme expression. Individual colonies were grown overnight in 4 ml LB (three biological replicates). The next day, 1 ml of culture was centrifuged at maximum speed, and the pellets were resuspended in an equal volume of PBS. Both fluorescence (excitation: 485 nm, emission: 520 nm) and OD<sub>600</sub> were measured, and the values were normalized to assess expression levels.

### Statistical analysis

One-tailed Student's t-tests were performed to compare fluorescence values and enzymatic activities among different ARES. Significance is indicated in the figure legends with asterisks denoting p-values: \*\*\*(*P* <.001) and \*\*(*P* <.05) when compared to the positive control, and +++ or ++ when compared to the negative control.

### Genomic data analysis

Genomic datasets for *E. coli* DH5 $\alpha$  (accession number CP080399) and *Thermus thermophilus* HB27 (Accession number GCA\_000008125) were downloaded from The European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute website. GenBank files were parsed using BioPython to extract coding sequences (CDSs) and the upstream intergenic regions (IRs). These sequences were stored in parquet format and made available in our [GitHub repository](#). GC contents were calculated for both CDSs and IRs.

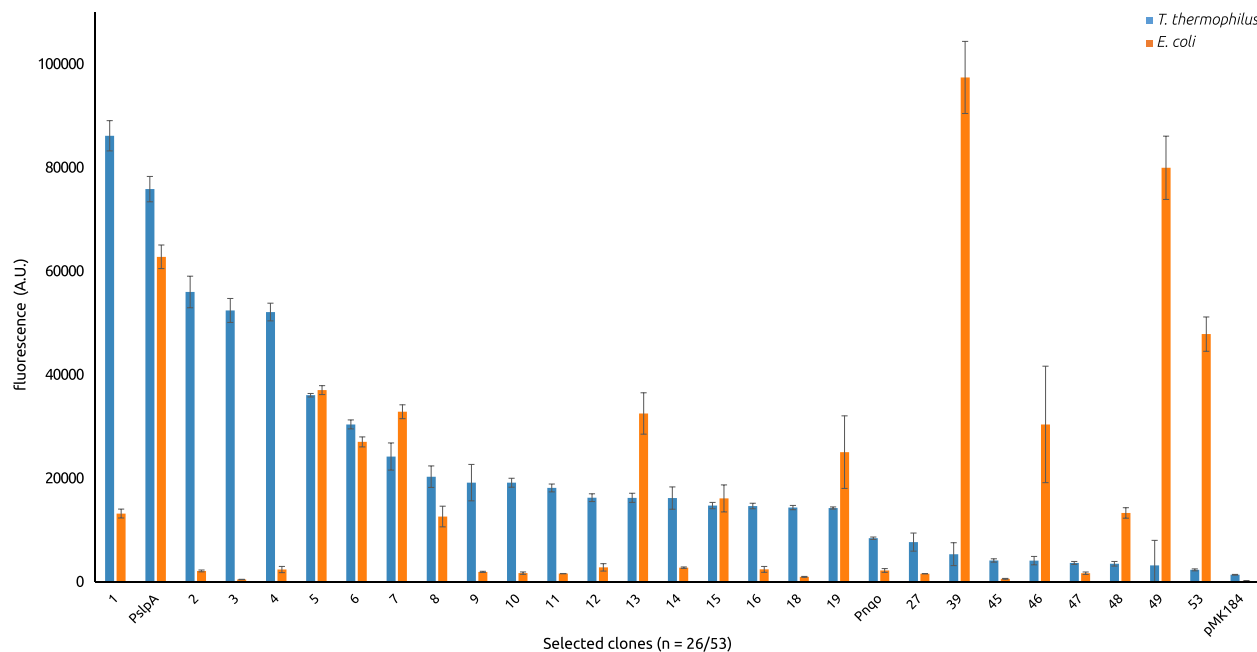
Additionally, RNA-seq datasets (for *E. coli* DH5 $\alpha$ : Run Accession SRR10995619-21, and for *T. thermophilus*: Run Accession SRR1038512-17) were processed using Salmon [24] to obtain normalized Transcripts Per Million (TPM) values. CDS, IR sequences, and TPM values were merged into a dataframe, and genes were categorized into high (>90% TPM), medium (50%–90% TPM), and low (<50% TPM) expression buckets for further analysis.

## Results

### Isolation of 53 artificial regulatory sequences in *T. thermophilus*

To generate artificial 5' regulatory sequences in *T. thermophilus*, we replaced the native promoter of the reporter gene, superfolder citrine fluorescent protein (IFP), with a 200-nucleotide GeneEE fragment (Fig. 1A and B) in the shuttle plasmid, pMK184 [19]. The initial ARES plasmid DNA library, generated in *E. coli*, comprised ~10 000 clones. After transformation of the library into *T. thermophilus*, 1500 individual colonies were obtained. Visual screening under UV light identified 53 fluorescent clones (3.5% hit rate), which were then cultured in 96-deep-well plates for further analysis (Fig. 1C and D).

To benchmark expression strength, two native *T. thermophilus* promoters were used as controls: (i) *P*<sub>slpA</sub>, a strong cross-species promoter; and (ii) *P*<sub>nqo</sub>, a medium-strength promoter with reduced activity in *E. coli*. In *T. thermophilus*, the ARES clones exhibited a broad range of fluorescence intensities (Fig. 2). The strongest ARES clone 1 exceeded the fluorescence of the strong *P*<sub>slpA</sub> control, while the majority (49/53) showed expression levels within  $\pm$ 1-fold of *P*<sub>nqo</sub>. Notably, ARES clones 1–4 displayed minimal background expression in *E. coli*, which is advantageous when cloning potentially growth inhibiting proteins in *E. coli*. When comparing the normalized fluorescence intensity values between the two



**Figure 2.** Normalized IFP fluorescence of the isolated ARES in *T. thermophilus* (light) and *E. coli* (dark). Fluorescence values represent mean  $\pm$  SD of  $n = 3$  biological replicates. The isolated ARES exhibit a range of expression strengths, with the strongest clone (ARES 1) exceeding the fluorescence of the strong control  $P_{slpA}$ , in *T. thermophilus*. Most ARES (49/53) display expression levels similar to the medium control  $P_{nqo}$ . In *E. coli*, a similar trend is observed; however, clones 1–4 show notably lower expression, which is advantageous for cloning toxic genes. Only 25 representative clones are shown in the figure; for all clones and their phenotypes see [Supplementary Table S2](#). Positive controls:  $P_{slpA}$ , and  $P_{nqo}$ . Negative control: empty plasmid pMK184.

hosts, 45 of 53 clones exhibited stronger expression in *T. thermophilus* than in *E. coli*, underscoring the benefits of performing a functional screening directly in the target thermophilic host.

## DNA sequence analysis

Sanger sequencing was performed on all 53 ARES clones (see [Supplementary Table S2](#) for the DNA sequences). An analysis of the GC content revealed that, despite the high genomic GC content of *T. thermophilus* (~69%, [Table 1](#)), regulatory regions are typically AT-enriched likely to facilitate DNA strand separation during transcription. In our study, while the native promoters  $P_{nqo}$  and  $P_{slpA}$  have GC contents of ~60%, the ARES had an average GC content of 44% ( $\pm 6\%$ ), ranging from 29% (ARES 52) to 73% (ARES 41). In line with other GeneEE studies, these results underscore the flexibility of the bacterial transcriptional machinery in accommodating a wide variety of regulatory sequences. Moreover, none of the ARES did yield significant hits in BLAST searches (no significant hits found using `tblastn` as of April 2025), suggesting that they represent novel sequence solutions that are rarely or if at all encountered in this bacterium.

## Characterization of enzyme expression using ARES

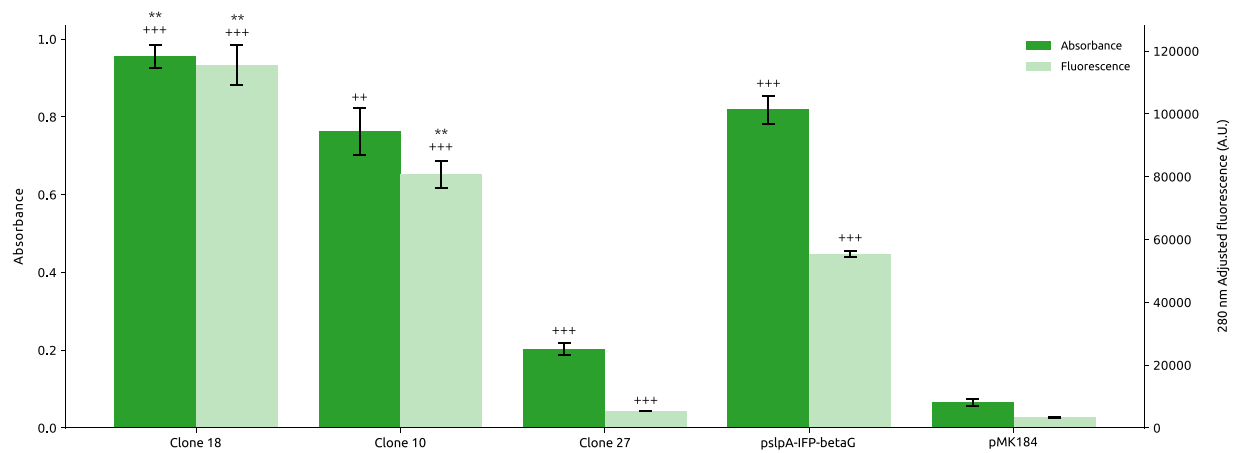
To validate the functionality of the ARES for driving expression of biotechnologically relevant thermostable enzymes, we transcriptionally fused thermostable  $\beta$ -galactosidase and phytoene synthase by placing them downstream of the IFP CDS ([Fig. 1E](#)). For each enzyme, 12 fluorescent clones were selected and subsequently characterized for enzyme expressions. Clones exhibiting enzymatic activities were Sanger sequenced to identify the corresponding ARES sequences (see [Supplementary Table S2](#)).

Initially, IFP fluorescence intensities were measured on enzyme-expressing transformants. In the transcriptional fusion setup ([Figs 3 and 4](#)), clones showing higher fluorescence generally

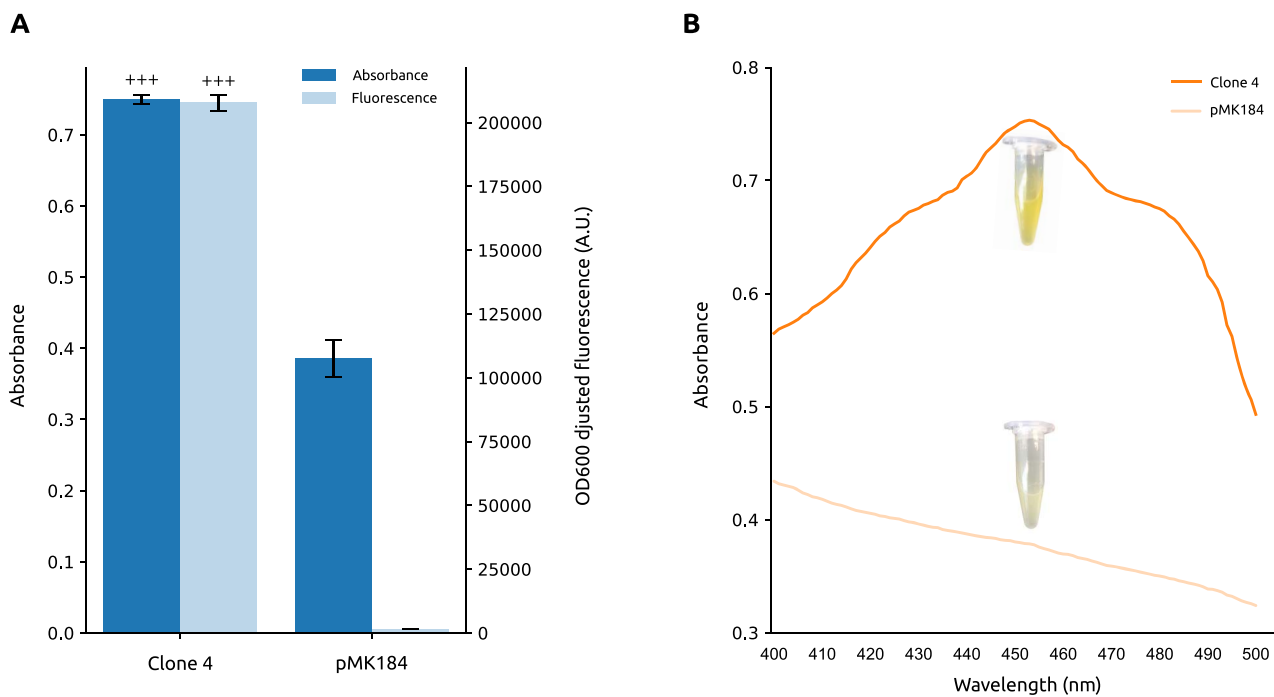
also exhibited higher enzymatic activity, such as  $\beta$ -galactosidase clone 18 and clone 10. This correlation supports the utility of the ARES in recombinant gene expression with transcriptionally-fused thermostable proteins. However, it was observed that fusion of the IFP with the enzyme CDS altered the fluorescence profiles compared to when IFP was expressed alone ([Fig. 2](#)), indicating that protein expression controlled by the same ARES was affected by the downstream CDS. For instance, ARES 18 and ARES 10, which drove  $\beta$ -galactosidase expression in clones 18 and 10 respectively, showed the highest fluorescence intensities after fusion, despite exhibiting relatively lower expression in the IFP-only setup. In contrast, ARES 4 driving expression of *CrtB* in clone 4 consistently showed high fluorescence in both configurations. These differences are due to the influence of the downstream CDS on gene expression and protein synthesis, where changes in mRNA structure and amino acids possibly affect translation elongation in the fusion setup [[25, 26](#)].

$\beta$ -Galactosidase activity was assessed via ONPG conversion. Three clones exhibited detectable enzymatic activity: clone 18 showed significantly higher activity than the positive control ( $P_{slpA}$ ), clone 10 was comparable to the positive control, and clone 27 had lower activity ([Fig. 3](#)). These differences confirm that the ARES can be used to obtain a range of expression levels even when fused to a fluorescent protein.

For phytoene synthase, enzyme function was evaluated both visually and spectrophotometrically ([Fig. 4](#)). The culture of clone 4 displayed an orange colour, in contrast to the pale yellow of the negative control (pMK184) ([Fig. 4B](#)). An acetone extract of clone 4 showed an absorbance peak at 452–453 nm ([Fig. 4B](#)), consistent with carotenoid production and indicative of a  $\beta, \beta$ -carotene type chromophore [[23, 27](#)]. These findings demonstrate that the isolated ARES are suitable for both reporter and functional enzyme expression in *T. thermophilus*.



**Figure 3.** Absorbance and fluorescence measurements of  $\beta$ -galactosidase and IFP expression in *T. thermophilus* clones. Clones 18, 10, and 27 displayed  $\beta$ -galactosidase activity using the substrate ONPG. These clones represent *T. thermophilus* colonies transformed with ligated plasmids carrying three different ARES for driving  $\beta$ -galactosidase expression. The positive control (pslpA-IFP-betaG) is a transformant harbouring a plasmid with the strong promoter  $P_{slpA}$ , while the negative control (pMK184) contains the native plasmid backbone without an inserted promoter. Absorbance values represent mean  $\pm$  SD of  $n = 3$  biological replicates. \*\*\*indicates a  $P$ -value  $< .001$  compared to the positive control; \*\*indicates a  $P$ -value  $< .05$  compared to the positive control; +++ indicates a  $P$ -value  $< .001$  compared to the negative control; ++ indicates a  $P$ -value  $< .05$  compared to the negative control. Fluorescence measurements of IFP expression in the same clones showed a similar pattern. Clone 18 exhibited higher IFP fluorescence than the positive control, indicating elevated expression; clone 10 also demonstrated higher expression, while clone 27 showed reduced IFP fluorescence. Fluorescence values represent mean  $\pm$  SD of  $n = 3$  biological replicates. \*\*\*indicates a  $P$ -value  $< .001$  compared to the positive control; +++ indicates a  $P$ -value  $< .001$  compared to the negative control.

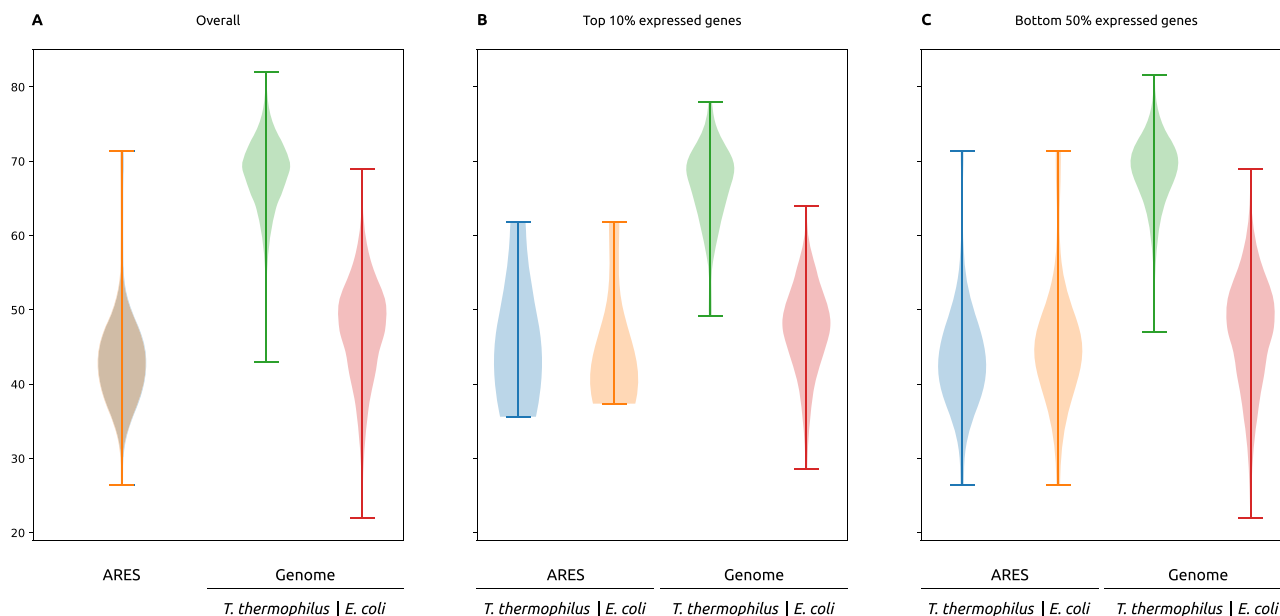


**Figure 4.** Absorbance and fluorescence assessments of phytoene synthase and IFP expression in *T. thermophilus* clone. Clone 4 represents *T. thermophilus* colony transformed with a ligated plasmid carrying an ARES driving both IFP and phytoene synthase expressions, while pMK184 as the negative control represents the transformant carrying the empty plasmid backbone. The expression of phytoene synthase is represented by the presence of carotenoids from absorbance measurements (A and B) and visual evaluation (B). Bar graphs (A) display absorbance measured at 452 nm and fluorescence values for clone 4 and pMK184. Clone 4 shows higher absorbance and fluorescence compared to pMK184, indicating the expression of both phytoene synthase and IFP. Absorbance and fluorescence values represent mean  $\pm$  SD of  $n = 3$ . \*\*\* indicates a  $P$ -value  $< .001$  compared to the negative control. The colour of the culture and the distinct absorption peak at 452–453 nm (B) further indicate carotenoids production in clone 4.

## Bioinformatic analysis

To contextualize the performance of ARES in *T. thermophilus* versus *E. coli*, we conducted a comparative bioinformatic analysis of both genomes, focusing on overall genomic features, CDS length, intergenic region (IR, refers to the DNA sequence located between two

adjacent CDSs, which may contain regulatory elements such as promoters, terminators, and/or untranslated regions) length, and GC content. Table 2 summarizes key statistics, including genome size, number of genes, total CDS length, IR length, and average GC content for each organism.



**Figure 5. GC content distributions of ARES and genomic CDSs.** Violin plots show the distribution of GC content for ARES and genomic CDSs in *T. thermophilus* and *E. coli*. Panel A shows all sequences, panel B shows the top 10% most highly expressed genes, and panel C shows the bottom 50% expressed genes (based on TPM). Sample sizes were as follows: Genome, *T. thermophilus* ( $n = 1969$  overall, 197 top 10%, 984 bottom 50%) and *E. coli* ( $n = 4205$  overall, 421 top 10%, 2102 bottom 50%); ARES, both hosts ( $n = 53$  overall, 6 top 10%, 27 bottom 50%). Internal boxplots indicate the interquartile range; thin lines show the full data (min, max) range.

**Table 2.** Comparative genomic features of *E. coli* and *T. thermophilus*

	<i>E. coli</i>	<i>T. thermophilus</i>
Genome size	4 490 755	1 894 877
Number of CDSs	4205	1969
Sum of CDS length	3 961 399	1 796 235
Sum of IRs	529 356	98 642
CDS density (%)	88.21	94.79
IR density (%)	11.79	5.21
GC content (%)		
Genome	50.74	69.44
CDSs	51.80	69.59
IRs	47.35	68.76
IFP	41.94	

CDS density (%) represents the percentage of the genome occupied by annotated genes, whereas IR density (%) denotes the fraction occupied by intergenic regions. GC content is presented for the entire genome, genes, and IRs.

As shown in Table 2, *E. coli* has a larger genome (4.49 Mb) with more genes (4205) than *T. thermophilus* (1.89 Mb, 1969 genes). Correspondingly, *E. coli* exhibits lower CDS density (88.21%) and higher IR density (11.79%) compared to *T. thermophilus* (94.79% and 5.21%, respectively), indicating a more compact genome in the thermophile. Notably, the average GC content is significantly higher in *T. thermophilus* (69.44%) than in *E. coli* (50.74%), a likely consequence of adaptations to high-temperature environments.

Figure 5 shows the GC content distributions for each organism's overall gene set at the genome level (Fig. 5A), as well as for the top 10% (Fig. 5B) and bottom 50% of expressed genes (Fig. 5C). *T. thermophilus* (green) retains its high-GC signature across all categories, whereas *E. coli* (red) clusters around 50%–51%. Notably, the synthetic ARES (blue and orange) have lower GC content than the native genomic regions, consistent with the observation

that many functional promoters and regulatory elements are AT-enriched, potentially aiding transcriptional initiation (the GitHub repository contains a notebook with all data and scripts used for figure generation and data analysis.).

Together the analyses on the distributions of CDS (Figure S1) and of IR lengths (Figure S2) in *T. thermophilus* and *E. coli*, underscore the compact, high-GC nature of the *T. thermophilus* genome compared to that of *E. coli*. They also demonstrate that our reported ARES diverge significantly from native promoters in both length and nucleotide composition (as confirmed by BLAST analyses). Such divergence is advantageous in synthetic biology, as it broadens the sequence space for uncovering novel, potentially stronger or more finely tunable transcriptional control elements.

## Discussion

In this study, we generated and characterized 53 ARES functional in both *T. thermophilus* and *E. coli*. Our results demonstrate that the GeneEE platform can efficiently deliver a diverse set of regulatory elements without requiring prior knowledge of native promoter architecture in *T. thermophilus*. By cloning fully randomized sequences, the selection process inherently ensures that functional ARES provide both transcription initiation and translation of the downstream gene of interest. Thus, the recovered sequences represent complete regulatory units rather than partial or spurious motifs. This feature is central to the GeneEE approach, as it enables the unbiased discovery of regulatory elements that couple transcriptional and translational control, thereby expanding the functional landscape beyond native promoter architectures. The observed host-specific differences where most ARES drive stronger expression in *T. thermophilus* than in *E. coli* further highlight the importance of evaluating synthetic parts in the intended host context.

Notably, the ARES exhibit an average GC content of about 44%, markedly lower than that of the native promoters  $P_{slpA}$

and  $P_{ng0}$  (60%–69%). The BLAST analysis revealed no significant matches to known natural promoters, indicating that our artificial design accesses a seldom-explored region of sequence space in the thermophile. By avoiding homology to existing genomic elements, these non-natural sequences expand the available promoter repertoire for extremophilic hosts and pave the way for de novo regulatory engineering.

The functional characterization of thermostable  $\beta$ -galactosidase and phytoene synthase further validates the utility of these ARES. The ability to modulate expression levels over a broad range is critical for optimizing industrial processes. For example, the production of thermostable enzymes such as  $\beta$ -galactosidase is important in the dairy industry for the production of lactose-free products, while phytoene synthase is a key enzyme in carotenoid biosynthesis, a pathway of growing interest due to its applications in food colourants, nutraceuticals, and pharmaceuticals. Moreover, the potential to express heterologous pathways in thermophilic hosts could simplify downstream processing in high-temperature industrial processes.

For functional characterization of the entire library, it is important to consider both the practical and theoretical constraints of the screening process. While cytometry-based analyses could, in principle, offer a more detailed view of the expression distribution across individual library members, such approaches need to be interpreted in the context of the theoretical library size ( $4^{200} \approx 10^{120}$ ). Any experimental screen, including cytometry, can only probe a minute fraction of this vast sequence space and therefore provides, at best, an indication of the spread of accessible variants rather than exhaustive coverage. Our strategy is thus intentionally pragmatic, focusing on identifying representative functional regulatory elements rather than attempting comprehensive library characterization, which is neither technically feasible nor conceptually meaningful at this scale.

In summary, our study enriches the synthetic biology toolkit for extremophiles by delivering a panel of novel, functionally validated ARESs, thereby laying the groundwork for both metabolic engineering strategies and the scalable production of high-value thermostable proteins in industrial settings.

## Acknowledgements

We thank Dr José Berenguer (Universidad Autónoma de Madrid) who kindly provided the *T. thermophilus* strain used in this study. Open access was supported by the NTNU library.

## Author contribution

C.F.A.W. and R.L. conceived the study and involved in the design of experiments. C.F.A.W. and S.Z. carried out the experiments and wrote the initial draft. L.T. involved in plasmid DNA library construction experiments in *E. coli*. G.S.D. performed the bioinformatics analysis on the ARES and genomes of *E. coli* and *T. thermophilus*. All authors contributed to manuscript revision, read, and approved the submitted version.

## Supplementary data

Supplementary data is available at SYN BIO online.

## Conflicts of interest

G.S.D. and R.L. are founders of Syngens AS, and S.Z. is partially employed by the company. These authors hold a financial interest

in Syngens AS; however, the work was carried out independently of any commercial or financial influences that might constitute a conflict of interest. All other authors declare no competing interests.

## Funding

We acknowledge the funding from EU, grant number 685474 and 101081957; and The Research Council of Norway, grant number 316129.

## Material availability

The materials used and reported in this study are available from the corresponding author, R.L., upon reasonable request.

## Data availability

The data and scripts underlying this article are available in a GitHub repository at [https://github.com/LaleLab/Publication\\_ARES\\_Thermus/](https://github.com/LaleLab/Publication_ARES_Thermus/).

## References

- Schloss PD, Handelsman J. Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* 2003; **14**:303–10. [https://doi.org/10.1016/S0958-1669\(03\)00067-3](https://doi.org/10.1016/S0958-1669(03)00067-3)
- Lewin A, Lale R, Wentzel A Expression platforms for functional metagenomics: emerging technology options beyond *Escherichia coli*. In: Charles T, Liles M, Sessitsch A. (eds), *Functional Metagenomics: Tools and Applications*. Cham: Springer, 2017, 13–44. [https://doi.org/10.1007/978-3-319-61510-3\\_2](https://doi.org/10.1007/978-3-319-61510-3_2)
- de Lorenzo V, Krasnogor N, Schmidt M. For the sake of the bioeconomy: define what a synthetic biology chassis is!. *New Biotechnol* 2021; **60**:44–51. <https://doi.org/10.1016/j.nbt.2020.08.004>
- Hwang S, Joung C, Kim W et al. Recent advances in non-model bacterial chassis construction. *Curr Opin Syst Biol* 2023; **36**:100471. <https://doi.org/10.1016/j.coisb.2023.100471>
- Henne A, Brüggemann H, Raasch C et al. The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nat Biotechnol* 2004; **22**:547–53. <https://doi.org/10.1038/nbt956>
- Niehaus F, Bertoldo C, Kähler M et al. Extremophiles as a source of novel enzymes for industrial application. *Appl Microbiol Biotechnol* 1999; **51**:711–29. <https://doi.org/10.1007/s002530051456>
- Cava F, Hidalgo A, Berenguer J. *Thermus thermophilus* as biological model. *Extremophiles* 2009; **13**:213–31. <https://doi.org/10.1007/s00792-009-0226-6>
- Lale R, Tietze L, Fages-Lartaud M et al. A universal approach to gene expression engineering. *Synth Biol* 2022; **7**:ysac017. <https://doi.org/10.1093/synbio/ysac017>
- Tietze L, Mangold A, Hoff MW et al. Identification and cross-characterisation of artificial promoters and 5' untranslated regions in *Vibrio natriegens*. *Front Bioeng Biotechnol* 2022; **10**. <https://doi.org/10.3389/fbioe.2022.826142>
- Maseda H, Hoshino T. Screening and analysis of DNA fragments that show promoter activities in *Thermus thermophilus*. *FEMS Microbiol Lett* 1995; **128**:127–34. <https://doi.org/10.1111/j.1574-6968.1995.tb07511.x>
- Fujita A, Sato T, Koyama Y et al. A reporter gene system for the precise measurement of promoter activity in *Thermus thermophilus* HB27. *Extremophiles* 2015; **19**:193–201. <https://doi.org/10.1007/s00792-015-0789-3>

12. Liang Y, Motawaa M, Bu X et al. Construction of primary chassis cells with efficient protein expression in *Thermus thermophilus*. *Microb Cell Factories* 2025; **24**:163. <https://doi.org/10.1186/s12934-025-02785-y>
13. Kayser K, Kwak J, Park H et al. Inducible and constitutive expression using new plasmid and integrative expression vectors for *Thermus* sp. *Lett Appl Microbiol* 2001; **32**:412–8. <https://doi.org/10.1046/j.1472-765X.2001.00933.x>
14. Verdú C, Sanchez E, Ortega C et al. A modular vector toolkit with a tailored set of thermosensors to regulate gene expression in *Thermus thermophilus*. *ACS Omega* 2019; **4**:14626–32. <https://doi.org/10.1021/acsomega.9b02107>
15. Fujino Y, Goda S, Suematsu Y et al. Development of a new gene expression vector for *Thermus thermophilus* using a silica-inducible promoter. *Microb Cell Factories* 2020; **19**:126. <https://doi.org/10.1186/s12934-020-01385-2>
16. Moreno R, Zafra O, Cava F et al. Development of a gene expression vector for *Thermus thermophilus* based on the promoter of the respiratory nitrate reductase. *Plasmid* 2003; **49**:2–8. [https://doi.org/10.1016/S0147-619X\(02\)00146-4](https://doi.org/10.1016/S0147-619X(02)00146-4)
17. Park H-S, Kilbane JJ. Gene expression studies of *Thermus thermophilus* promoters PdnaK, Parg and Pscs-mdh. *Lett Appl Microbiol* 2004; **38**:415–22. <https://doi.org/10.1111/j.1472-765X.2004.01512.x>
18. Kirchner L, Müller V, Averhoff B. A temperature dependent pilin promoter for production of thermostable enzymes in *Thermus thermophilus*. *Microb Cell Factories* 2023; **22**:187. <https://doi.org/10.1186/s12934-023-02192-1>
19. Cava F, Laptenko O, Borukhov S et al. Control of the respiratory metabolism of *Thermus thermophilus* by the nitrate respiration conjugative element nce. *Mol Microbiol* 2007; **64**:630–46. <https://doi.org/10.1111/j.1365-2958.2007.05687.x>
20. Gibson DG, Young L, Chuang RY et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 2009; **6**:343–5. <https://doi.org/10.1038/nmeth.1318>
21. Potapov V, Ong JL, Kucera RB et al. Comprehensive profiling of four base overhang ligation fidelity by T4 DNA ligase and application to DNA assembly. *ACS Synth Biol* 2018; **7**:2665–74. <https://doi.org/10.1021/acssynbio.8b00333>
22. Umarov RK, Solovyev VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* 2017; **12**:e0171410. <https://doi.org/10.1371/journal.pone.0171410>
23. Takano H, Kondo M, Usui N et al. Involvement of CarA/LitR and CRP/FNR family transcriptional regulators in light-induced carotenoid production in *Thermus thermophilus*. *J Bacteriol* 2011; **193**:2451–9. <https://doi.org/10.1128/JB.01125-10>
24. Patro R, Duggal G, Love MI et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017; **14**:417–9. <https://doi.org/10.1038/nmeth.4197>
25. Tietze L, Lale R. Importance of the 5' regulatory region to bacterial synthetic biology applications. *Microb Biotechnol* 2021; **14**:2291–315. <https://doi.org/10.1111/1751-7915.13868>
26. Forsberg A, Pavitt G, Higgins C. Use of transcriptional fusions to monitor gene expression: a cautionary tale. *J Bacteriol* 1994; **176**:2128–32. <https://doi.org/10.1128/jb.176.7.2128-2132.1994>
27. Yokoyama A, Sandmann G, Hoshino T et al. Thermozeaxanthins, new carotenoid-glycoside-esters from thermophilic eubacterium *Thermus thermophilus*. *Tetrahedron Lett* 1995; **36**:4901–4. [https://doi.org/10.1016/00404-0399\(50\)0881-C](https://doi.org/10.1016/00404-0399(50)0881-C)