A universal approach to gene expression engineering

Rahmi Lale1, Lisa Tietze1, Maxime Fages-Lartaud1, Jenny Nesje1, Ingerid Onsager1, Kerstin Engelhardt1, Che Fai Alex Wong1, Madina Akan1,3, Niklas Hummel1,4,5, Jörn Kalinowski2, Christian Rückert2, and Martin Frank Hohmann-Marriott1,6

1Department of Biotechnology, Faculty of Natural Sciences, Norwegian University of Science and Technology, Trondheim 7491, Norway
2Center for Biotechnology, Bielefeld University, Bielefeld 33615, Germany
3Department Microbiology and Biochemistry of Geisenheim University, Geisenheim 65366, Germany
4Feedstocks Division, Joint BioEnergy Institute, Emeryville 94608, CA, USA
5Department of Plant Biology, University of California, Davis 95616, CA, USA
6United Scientists CORE (Limited), Dunedin 9014, New Zealand

*Corresponding authors: E-mails: rahmi.lale@ntnu.no and martin.hohmann-marriott@united-scientists.org

Abstract

In this study, we provide a universal approach to Gene Expression Engineering (GeneEE) for creating artificial expression systems. GeneEE leads to the generation of artificial 5′ regulatory sequences (ARES) consisting of promoters and 5′ untranslated regions. The ARES lead to the successful recruitment of RNA polymerase, related sigma factors and ribosomal proteins that result in a wide range of expression levels. We also demonstrate that by engaging native transcription regulators, GeneEE can be used to generate inducible promoters. To showcase the universality of the approach, we demonstrate that 200-nucleotide (nt)-long DNA with random composition can be used to generate functional expression systems in six bacterial species, Escherichia coli, Pseudomonas putida, Corynebacterium glutamicum, Thermus thermophilus, Streptomyces albus and Streptomyces lividans, and the eukaryote yeast Saccharomyces cerevisiae.

Key words: artificial promoters; 5′ untranslated region; transcription; translation

1. Introduction

Successful heterologous gene expression and recombinant protein production are among the major goals in biotechnological applications, which are accelerating progress in academic research and in the pharmaceutical, agricultural, food and chemical industries (1, 2). The selection of a suitable microbial host is essential for achieving high-level recombinant protein production (3). A variety of factors influence the host selection, including growth characteristics, recombinant protein production capacity, secretion systems, posttranslational modifications, the desired biological activity of the final product and compatibility with the industrial processes (2). Therefore, a thorough understanding of gene expression is required to successfully genetically engineer microbial hosts for recombinant protein production. The majority of genetic engineering efforts are aimed at optimizing cellular processes and metabolism to increase recombinant protein production rates, yields, quality and functionality. Successful genetic engineering efforts, on the other hand, must overcome a number of hurdles related to gene expression, host specificity and cell fitness (2).

Transcription and translation, together with DNA replication, are the most fundamental processes shared by all life on Earth. In bacteria, the RNA polymerase core enzyme (α2ββ′ω) interacts with sigma (σ)-factors to recognize promoters, unwind the DNA
double helix and activate transcription (4). In eukaryotes, RNA polymerase II associates with transcription factors to activate the transcription of central promoters, but tighter control is guaranteed by a complex regulation dependent on chromatin state, distal enhancer regions, transcription factors and cofactors (5). Bacterial promoters contain conserved sequences that $\sigma$-factors recognize, as well as UP elements that interact with the RNA polymerase $\alpha$-subunit. The predominant bacterial $\sigma$-factor ($\sigma^{70}$) is known to recognize two motifs separated by 17 nt, the $\sim$35 (TTGACA) and the $\sim$10 boxes (TATAAT), in addition to the $\sim$10 extended element (TGN$_{10}$) and $\sim$6 to $\sim$8 discriminator region (GGG) (4, 6). Across all life, genetic information in mRNA is translated by ribosomes into functional proteins. Translation is strongly affected by mRNA secondary structures formed within the 5′ untranslated region (UTR) and at the start of the coding sequences (CDS). Efficient translation involves a trade-off between mRNA stability, unfolding and the accessibility of ribosome entry sites (such as the Shine–Dalgarno ‘SD’ sequence in bacteria and the Kozak sequence in eukaryotes) (7–12). In addition, a multitude of additional proteins and RNA molecules participate in the modulation of transcription and translation.

Conventional genetic engineering approaches employ native or modified versions of existing promoters and 5′ UTRs (6), in a plug-and-play fashion which often disregard the context-dependency issues and may lead to unpredictable expression (13). Other techniques such as combinatorial libraries (14), semi-rational engineering (15–18) or computational design (19,19–22) have been effective in constructing artificial promoters and/or 5′ UTRs. However, these techniques frequently fail to address the complete spectrum of gene expression complexity, host specificity, cell fitness and intermediate metabolite toxicity. The variety and complexity of the processes involved are reflected in the nucleotide content of promoter motifs and translation initiation regions that varies greatly within and across species (23,24). This nucleotide variety is also a result of evolutionary processes to achieve effective gene expression while avoiding gene expression noise and metabolite toxicity to reach an optimum cellular fitness (25–28). However, suboptimal gene expression caused by a lack of specificity in sequence recognition might also provide an evolutionary advantage for acquiring additional functions or establishing precise regulation under various contexts (29,30).

To investigate promoter evolution mechanisms, Wolf et al. showed that sequences of 100–150 nt of random DNA composition were able to replace natural promoters in a relatively high proportion of sequences (29). These artificial promoters resulted in a wide range of transcription levels in Escherichia coli. In another study, Yona et al. cloned 40 randomly generated sequences in front of the lacZYA operon, which resulted in 10% of sequences directly providing functional promoters, and about 60% were only one mutation away from transcription activation (30). Yim et al. combined a complete promoter randomization approach with a semi-random 5′ UTR, containing an AGGA SD sequence, to successfully generate functional expression in Corynebacterium glutamicum (31). Similar approaches consisted of using a stretch of DNA with random composition containing only a few fixed nucleotide matching $\sim$35, $\sim$10 promoter boxes and/or 5′ UTRs (32–34). In eukaryotes, the mechanisms regulating transcription and translation present a higher complexity. Therefore, Gene Expression Engineering (GeneEE) using random DNA sequences remains reliant on fixed motifs such as upstream activating elements (UASs), core promoters or 5′ UTRs (35). De Boer et al. and Kotopka et al. successfully engineered yeast promoters by nucleotide randomization while conserving UASs or core promoter motifs (36,37).

Interestingly, the random DNA stretch led to the generation of UAS-binding sites responsible for transcriptional regulation and thus, synthetic promoters. Similarly, Cuperus et al. used a 50-nt-long DNA sequence to investigate eukaryotic 5′ UTR in yeast that led to the generation of constructs exhibiting a wide range of translation efficiency and provided optimal Kozak sequences (38). Although all of the aforementioned ways proved effective in creating promoters and/or 5′ UTRs of varying strengths in bacterial or eukaryotic cells, they all rely on known DNA patterns, which limits the application in other microorganisms (39). As a result, we set out to create a method that avoids relying on predefined DNA motifs, allowing the approach to be applied across species.

In this study, we present a universal approach to Gene Expression Engineering (GeneEE) for constructing artificial 5′ regulatory sequences (ARES), consisting of promoters and 5′ UTR, suited to the gene of interest (Gol) in the host of interest (Hol). A GeneEE DNA segment is composed of 200 random nucleotides that are placed directly upstream of a CDS to construct random DNA libraries. The screening of random DNA libraries results in the identification of artificial promoters that recruit the host’s native transcription machinery and the formation of functional 5′ UTRs that are customized to the CDS to initiate translation. One advantage of generating a wide set of ARES is that it allows the identification of DNA constructs that result in expression levels that are appropriate for the cellular metabolic milieu, therefore avoiding undesirable expression levels that may affect cell fitness. As a result, the nucleotide diversity of a random DNA library provides Gol-tailored regulatory components that address both transcriptional and translational processes, ensuring that DNA sequences are context-dependent (13). Overall, GeneEE can be used as a universal approach for constructing regulatory sequences, eliminating the need for previously known promoters and 5′ UTRs and resulting in the generation of artificial expression systems that can be adapted to any Hol.

2. Methods
2.1 The composition of the single-stranded random nucleotide oligo

For the construction of GeneEE DNA libraries, two different versions of single-stranded Random Nucleotide Oligos (ss-RaNuQ) were ordered from Integrated DNA Technologies, Inc. (Belgium) as a four nanomole ultramer: (A) containing the random stretch of 200 nt, N$_{200}$, named ss-RaNuQO(−SD) (Figure 1A, Figure S1A) and (B) containing the defined GGAG sequence in between the two stretches of random nucleotide, N$_{100}$ and N$_{100}$, named ss-RaNuQO(+SD) (Figure 1B, Figure S1B). In the last mentioned version, the defined GGAG sequence serves as the minimal bacterial consensus SD sequence, and the −SD naming indicates where the oligo harbors the SD sequence (+SD) or not (−SD).

The ss-RaNuQOs have adapters on both ends. The Adapter1 contains the BioBrick Prefix (40), and the Adapter2 contains the BioBrick Suffix (40) sequences (Figures S1A and S1B). Each adapter also contains a type II restriction enzyme recognition sequence for BsaI (5′-GGTCTC(N1)/(N5)-3′), with overhang sequences ACGG in Adapter1 and NATG in Adapter2. These adapters also serve two additional functions: first, they are used for the functional immobilization of the oligo during the chemical synthesis and second, they are used to generate the complementary DNA strand by polymerase chain reaction (PCR) leading to double-stranded random nucleotide DNA (GeneEE segments). The BsaI overhang sequence in the Adapter2 has been designed to contain an N nucleotide in the 3′ overhang sequence, NTAC, to ensure that all possible 4nt
Figure 1. Overview of the GeneEE approach. A) The composition of the different GeneEE segments with and without the SD sequence and B) the plasmid map depicting the different combinations of GoI used in the identification of ARES in the hosts. Panels C and D depict the steps used in the selection and characterization of the clones that carry functional ARES.

2.2 Generation of GeneEE segments for BsaI-based restriction cloning

The double-stranded GeneEE segment was generated by using either of the ss-RaNuqO(±SD) as the DNA template and the BBa-Prefix-F and BBa-Suffix-R as primers (Table S1) in a 10-cycle PCR leading to GeneEE(−SD)BsaI (Figure S1A) and GeneEE(+SD)BsaI (Figure S1B). The low number of cycles was chosen to prevent a possible bias during the PCR amplification that can lead to reduced sequence diversity.

2.3 Generation of GeneEE segments for Gibson cloning

The GeneEE(±SD) segments (Figures S1C and S1D) were generated by using either of the ss-RaNuqO(±SD) as the DNA template and the BBa-Prefix-delBsaI-F and BBa-Suffix-R as primers (Table S1) in 10 cycles of PCR. The BBa-Prefix-delBsaI-F carries a single nucleotide mismatch, eliminating the BsaI recognition sequence upon PCR amplification (Figures S1C and S1D). The resulting PCR product was digested by BsaI removing the Adapter2 and leaving the 3′-NTAC overhang. In this study, three linkers were ligated to generate three versions of the GeneEE(±SD) segments each carrying a specific linker for the mCherry, Km or trp genes, introducing the complementary fragment to be used in Gibson cloning.

The ligation of gene-specific linkers. Three complementary oligo pairs for each gene were ordered from Sigma-Aldrich (Germany) as unphosphorylated oligos. The oligos were designed so that upon annealing they would carry the required 5′-NATG overhang (Table S2). First, each oligo was phosphorylated by T4 Polynucleotide Kinase (NEB) and afterward annealed to each other.
The resulting double-stranded linkers with sticky ends were ligated to the Bsal digested GeneEE±SD segments by T4 Quick DNA Ligase (NEB) leading to GeneEE+(SD)mCherry, GeneEE+(SD)KmT (Figure S1C) and GeneEE(−SD)Trp (Figure S1D). And finally, all three GeneEE segments were PCR-amplified using primer pair Bba-Prefix-F and the corresponding Upper DNA Strand (UpS) oligo (Table S2) of the Gol.

2.4 Growth conditions

**Escherichia coli.** Escherichia coli cells were grown in lysogeny broth (LB, 10 g/l tryptone, 5 g/l yeast extract and 5 g/l NaCl) or lysogeny agar (LA, LB + 15 g/l agar) at 37°C, supplemented with 50 μg/ml kanamycin or varying levels of ampicillin as specified in the text.

**Pseudomonas putida.** Pseudomonas putida cells were grown in LA or LB at 30°C, supplemented with 50 μg/ml kanamycin when selection was required as described in the text.

**Thermus thermophilus.** Thermus thermophilus cells were grown in Thermus broth (TB, 8 g/l lactotryptone, 4 g/l yeast extract and 3 g/l NaCl, all dissolved in mineral water) or Thermus agar (TA, TB + 15 g/l agar). A range of kanamycin (30, 60 and 90 μg/ml) were added to the growth medium when selection was required as described in the text. TA plates were incubated in a moisturized plastic container to avoid agar plates from drying out due to the high incubation temperature.

A stock pellet of T. thermophilus from −20°C was thawed at room temperature and was inoculated to 1 ml of TB and subsequently was transferred to a 125-ml flask containing 20 ml of TB. The strain was cultivated under constant mild agitation (150 rpm) at 65°C.

**Corynebacterium glutamicum.** Corynebacterium glutamicum cells were grown in Brain Heart Infusion Broth (BHIB, 37 g/l Brain Heart Infusion mix [Difco] and 91 g/l sorbitol) or Brain Heart Infusion Agar (BHIA, 37 g/l Brain Heart Infusion mix [Difco] and 10 g/l agar) at 30°C, supplemented with 15 μg/ml chloramphenicol when selection was required as described in the text.

**Streptomyces albus and Streptomyces lividans.** The Streptomyces strains were grown in Yeast Extract Tryptone medium (YET, 16 g/l Tryptone, 10 g/l yeast extract and 5 g/l NaCl). For sporulation, Streptomyces albus J1074 was grown at 30°C on Soy Flour Mannitol Agar (SFMA, 20 g/l soy flour, 20 g/l mannitol and 20 g/l agar), and Streptomyces lividans TK24 was grown at 30°C on ISP4 agar (BD Difco ISP medium 4). Conjugation reactions were performed using the cognate media supplemented with 10 mM MgCl2 at 30°C.

**Saccharomyces cerevisiae.** Saccharomyces cerevisiae cells were grown in Yeast Extract Peptone Dextrose Broth (YEPDB, 20 g/l bacto peptone, 10 g/l yeast extract and 20 g/l dextrose) or YEPD agar (YEPDA, 50 g/l YEPDB and 15 g/l agar) (Sigma-Aldrich) at 30°C. For the functional screening of GeneEE segments controlling the expression of tryptophan gene the drop-out media (0.68 g/l yeast nitrogen base powder [Sigma-Aldrich], 0.5 g/l glucose and 1.92 g/l of Yeast Synthetic Drop-Out Media Supplements without tryptophan [Sigma-Aldrich]) was used.

2.5 Construction of GeneEE plasmid DNA libraries in **E. coli**

For the functional screening of GeneEE segments in *E. coli* two different genes were used: an antibiotic selection marker, β-lactamase, conferring ampicillin resistance; and a fluorescent reporter, *mCherry*, a red fluorescent protein variant.

For the construction of GeneEE plasmid DNA libraries with the β-lactamase gene, pUC19-Bba-Km was used (Table S3). This plasmid carries a functional ampicillin and kanamycin resistance marker (with their native promoter and coding sequence); the KmT gene was cloned into the multiple cloning site of pUC19. The β-lactamase coding sequence carries an internal Bsal recognition site; to eliminate this site the entire plasmid was amplified by the primer pair delBsaI-F and delBsaI-R (Table S1), by following the Overlap Extension PCR cloning method (41), and the resulting linear PCR product was digested with DpnI, transformed into *E. coli* via chemical transformation and grown on LA plates supplemented with 50 μg/ml ampicillin. The primers carry a single nucleotide point mutation, hence eliminating the Bsal recognition sequence. The resulting plasmid, pRL101, was then amplified by the primer pair BsaI-Blu-F and BsaI-Blu-R (Table S1), amplifying the entire plasmid excluding the native β-lactamase promoter. Both the PCR product and the GeneEE+(−SD)BsaI were digested with BsaI and used for cloning as a vector and insert, respectively. The ligation mix was made using 1:5 molar vector:insert ratio in a total volume of 20 μl using the T4 Quick Ligase (NEB). The mixture was then transferred into the cloning host *E. coli* via chemical transformation.

For the construction of GeneEE plasmid DNA libraries with the *mCherry* gene, a mini-RK2-based broad host range plasmid pH100 (Table S3) was used. An *E. coli* codon-optimized variant of the *mCherry* gene (a gift from Yanina R. Sevastyanovich, University of Birmingham) was PCR-amplified using primer pair mCherry-Ndel-F and mCherry-BamHI-R (Table S1) and digested with Ndel and BamHI. The digested fragment was then used to replace the luciferase gene in pH100, resulting in pLT101. pLT101 replicates both in *E. coli* and *P. putida* and also carries a functional kanamycin antibiotic marker and the mCherry gene. For the construction of pLT101-based GeneEE plasmid DNA libraries, Gibson cloning method was used (42). The pLT101 was amplified by the G-mCherry-F and G-mCherry-R (Table S1) that lead to the amplification of the plasmid excluding the native mCherry promoter. The reverse primer mCherry-R carries the BioBrick Prefix (40), as a flanking sequence, that serves as a homologous sequence in Gibson cloning. This amplicon was used as the backbone and the GeneEE(+SD)mCherry, carrying the mCherry linker, as the insert. The Gibson ligation mixture was made using equimolar amounts of purified vector and insert (totaling 100 ng DNA) in a total volume of 5 μl, added to a volume of 15 μl of Gibson assembly mix. The mixture was then transferred into the cloning host *E. coli* via chemical transformation and grown on LA plates supplemented with 50 μg/ml kanamycin. The resulting library consisted of ~5000 transformants.

2.6 Functional screening of GeneEE plasmid DNA libraries in **E. coli**

**Ampicillin screening.** For the functional screening of GeneEE plasmid DNA libraries for clones that confer ampicillin resistance, the library was plated out both on LA plates supplemented with 50 μg/ml kanamycin and LA plates supplemented with 10, 30, 50, 75 and 100 μg/ml ampicillin. Among the obtained clones, 20 clones were randomly selected (eight colonies growing on 50 μg/ml ampicillin, eight colonies growing on 100 μg/ml ampicillin and four colonies growing on 50 μg/ml kanamycin) and grown overnight at 37°C; their respective plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen) and were Sanger sequenced (GATC Biotech, Germany) using the primer pRL101-Blu-Seq (Table S1). With these sequencing results, all the 20 constructs were confirmed to carry the cloned GeneEE segment.
Figure 2. Characterization of the clones identified in E. coli expressing mCherry. A plasmid DNA library was constructed using the mCherry gene, encoding for a red fluorescence protein, as a reporter for the identification of clones that lead to GeneEE-mediated transcription and translation in E. coli. The WebLogo image depicting the random nucleotide distribution of ARES expressing mCherry at varying levels (A). The defined SD sequence (GGAG) and the start codon of mCherry (ATG) are depicted in a larger font size below the alignment. The total number of mCherry transcripts (B), relative fluorescence intensities measured from each library colonies (C) and the number (D) and position (E) of transcription start site (TSS) identified within each ARES. Sequences grouped based on mCherry transcript abundance show distinct consensus sequences (motifs) for interaction with $\sigma$-factors (F–K). The sequences were realigned based on the identified motifs, cut down to the core motifs including 3 nt up- and downstream and binned based on relative ‘transcription strength’, i.e. the number of mapped reads starting at each TSS as well as the presence or absence of the $-10$ extended motif $Tg$. For each bin, the median distance of the motifs was also calculated, the resulting numbers are given in the ‘spacers’. The numbers on the x-axis are based on these median distances. Motifs of ‘weak’ promoters (10–99 mapped reads) without (F) and with (G) −10 extended motif. Motifs of ‘medium’ promoters (100–999 mapped reads) without (H) and with (I) −10 extended motif. Motifs of ‘strong’ promoters (1000 mapped reads and more) without (J) and with (K) −10 extended motif. The number of sequences used in each alignment is depicted with ‘n’.
mCherry screening. From the ~5000 transformants, three colonies were randomly picked and plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen) for Sanger sequencing (GATC Biotech, Germany) using the primer plT101-mCherry-Seq (Table S1). Sequencing confirmed the presence of GeneEE segments in each clone. Upon this confirmation, 192 visibly red colonies were randomly picked, possessing a varying degree of intensity (weak to strong red) to the naked eye, and inoculated into 2x 96-well plates containing LB supplemented with 50 µg/ml kanamycin. The plates were then incubated overnight under constant agitation, 800 rpm. The fluorescent measurements were made (excitation and emission wavelength of 584 and 620 nm, respectively) using an Infinite 200 PRO fluorescence microplate reader (Tecan). The 192 colonies from the 96-well plates were also replica plated into 14-cm agar plates containing LA supplemented with 50 µg/ml kanamycin and were high-throughput sequenced (see subsection 2.8 below). The identified ARES with their corresponding phenotypes (mCherry fluorescence intensity measurements) are listed in the supplementary spreadsheet. Based on these data, a WebLogo (45) image was generated (due to the multiple alignment parameters only the ARES with 211-nt-long sequences were used), indicating the high-sequence variation within GeneEE DNA libraries (Figure 2A).

2.7 Determination of the relative presence of clones with functional artificial regulatory sequences

To determine what fraction of a GeneEE DNA library is harboring clones with functional ARES, we have created two libraries which were based on the plasmid plRL101 (Table S3) that carries the antibiotic marker kanamycin on the backbone along with the β-lactamase gene that confers ampicillin resistance. plRL101 was PCR-amplified by the primer pair BsaI-Blal-F and BsaI-Blal-R (Table S1) amplifying the entire plasmid excluding the native β-lactamase promoter. Both the PCR product and the GeneEE(±SD)BsaI inserts were digested with BsaI and DpnI overnight. Both vector and insert were purified, and the vector was calf intestinal alkaline phosphatase (CIP)-treated (NEB) at 37 °C for 1 h and purified. The ligation mix was made using 1:10 molar vector:insert ratio in a total volume of 30 µl using the T4 Quick Ligase (NEB). A total of 10 µl of three aliquots were then transferred into E. coli via chemical transformation. Both libraries along with a control without insert were plated out both on LA plates supplemented with 50 µg/ml kanamycin and LA plates supplemented with 50, 500 and 1000 µg/ml ampicillin that resulted in libraries consisting of ~62 000 (plRL101 + GeneEE(−SD)) and ~37 000 (plRL101 + GeneEE(+SD)) transformants. The relative numbers of clones are listed in Table 1.

2.8 High-throughput sequencing of the GeneEE library with mCherry as reporter

To determine ARES resulting from the GeneEE segments, a multiplexing strategy was applied. Two 96-well plates were arranged as one 192-well plate with 16 rows (A–P) and 12 columns (1–12). For each row and each column, one pool of cells was established, resulting in a total of 28 pools. For each pool, a sequencing library was created by two rounds of PCR. In the first PCR round, the GeneEE segment was PCR-amplified using the primers Illumina-fwd and Illumina-rev (Table S1) in 15 cycles. In the second round of PCR, indexed sequencing libraries were constructed using the standard Illumina TruSeq primers. The indexed libraries were quantified using a BioAnalyzer (Agilent Technologies, Germany), pooled in equimolar amounts and sequenced in a 2x 300-nt run on an Illumina MiSeq sequencer. The reads from each library were joined with FLASH (44) using the following parameters: –max-overlap 450, –min-overlap 200, –max-mismatch-density 0.65 and –allow-outies. From the 57 672 to 156 052 joined reads per library, ARES were identified via search for the flanking regions. By counting the number of occurrences of each sequence in each library, the most abundant sequence found for a row/column pair was assigned as ARES in the corresponding clones.

2.9 Determination of transcription start sites

To obtain RNA suitable for sequencing, all cell materials from 192 E. coli GeneEE clones were combined. With this pool, a 10 ml overnight pre-culture in LB was inoculated. From this culture, the main culture was inoculated with a starting OD600 nm of 0.2, which was then grown until mid-log phase. For RNA isolation, samples of this culture were harvested by quick centrifugation, removal of the supernatant, flash-freezing in liquid nitrogen and were stored at −80 °C. RNA was isolated and a TSS library was prepared as described previously (45), with one modification: Instead of the unspecific loop adapter, the Illumina-TSS oligo (Table S1) was used for reverse transcription. Sequencing was done in a 75-nt run on illumina MiSeq sequencing machine, resulting in 6 085 833 reads. Mapping of the reads was done using bowtie2 (46) with standard parameters, using the ARES sequences, each extended by the mCherry coding sequence between the ATG and the Illumina-TSS oligo, as references. From the 309 794 mapped reads (122 680 unique and 187 114 multiple), counting of read starts, identification of potential TSS and extraction of the region upstream of identified TSS were performed with custom Perl scripts (see supplementary material). To determine the threshold for a potential TSS, a simple statistical analysis was performed by randomly distributing a number of hits in an array with the length of all reference sequences combined. The average number of reads per position and the standard deviation were calculated, and the threshold number of reads for a ‘true’ TSS was set to this average + 3 standard deviation. Using this approach, a total number of 375 TSSs was detected (Table S5). The raw sequence reads are uploaded to The National Center for Biotechnology Information (NCBI) and available under the BioProject, PRJNA853256.

2.10 Functional screening of GeneEE plasmid DNA libraries in P. putida

For the selection of functional GeneEE sequences in P. putida, using mCherry as the reporter, the same library that was established

| Antibiotic markers and concentrations (µg/ml) | Ap (50) | Ap (500) | Ap (1000) |
|---------------------------------------------|---------|----------|-----------|
| pRL101 + GeneEE(−SD)                       | 43.1    | 5.0      | 2.2       |
| Percentage functional clones (%)           |         |          |           |
| Standard deviation (%)                     | 5.5     | 0.4      | 0.7       |
| pRL101 + GeneEE(+SD)                       | 33.2    | 4.7      | 1.7       |
| Percentage functional clones (%)           |         |          |           |
| Standard deviation (%)                     | 6.3     | 1.8      | 0.4       |
| pRL101 (No insert control)                 |         |          |           |
| Percentage false positives (%)             | 6.8     | 3.4      | 0         |

*The percentage of positive clones was calculated from three biological replicas by comparing colonies obtained on ampicillin with the total library selected on kanamycin. The libraries consisted of ~62 000 (pRL101 + GeneEE(−SD)) and ~37 000 (pRL101 + GeneEE(+SD)) transformants.
in E. coli was used. A volume of 10 ml of LB was inoculated with 100 μl of E. coli cells from the frozen culture harboring the GeneEE plasmid DNA libraries. Cells were grown for 2 h at 37 °C, and the plasmid library was purified with QIAprep Spin Miniprep Kit (Qiagen) and quantified with Nanodrop (ThermoFisher). A total of 25 ng of plasmid was electroporated (2.5 kV for 2-mm gap cuvette; 200 Ω, 25 μF) into electrocompetent P. putida with the Gene Pulser Xcell Electroporation System (BioRad) using a previously established method (47). The electroporated cells were then added to 1 ml of pre-warmed LB and were incubated for 1.5 h at 30 °C. Afterward, cells were plated on LA supplemented with 50 μg/ml kanamycin. Among the obtained colonies (Figure 3), several colonies with visibly red color to the naked eye were picked, and plasmid DNA was isolated with the QIAprep Spin Miniprep Kit (Qiagen). These plasmids were then Sanger sequenced (GATC Biotech, Germany) using the sequencing primer pLT101-mCherry-Seq (Table S1). The identified ARES with their corresponding phenotypes (mCherry readouts) are listed in the supplementary spreadsheet.

2.11 Motif analysis of the artificial promoters
To identify potential promoter motifs in the regions upstream of the identified TSS, the tool Improbizer/Ameme (48) was used. The regions 50 nt upstream of the TSS were ordered in descending order based on the amount of read starts for the given TSS, and Improbizer was run with standard parameters except that the parameter for ‘constrainer’ was set to 100. The results of the motif search are listed in Table S5.

As the identified consensus motifs, TTGTGT and TATAAT, strongly resembled the known E. coli σ70 binding site, the motifs identified were parsed and split into six bins. The prevalence of σ70 motifs in most of the constructs may be due in part to the sampling taking place during the mid-log growth phase. The binning approach was used since the importance of the position weights would be obscured if all of the samples were aggregated into a single bin. These bins were created based on two categories: First, based on the rough equivalent of the promoter strength, i.e. the number of read starts, and second, based on the presence of an extended −10 motif, i.e. a TGN directly upstream of the −10 box. The resulting bins were visualized via WebLogo (43), and the results are displayed in Figure 2F–K.

2.12 Construction of GeneEE plasmid DNA libraries for functional screening of inducible gene expression in E. coli
For this construction, the xylS gene with its native promoter was PCR-amplified from pHH100 by the primer pair XylS-EcoRI-F and XylS-XbaI-F (Table S1), introducing the recognition sequences for EcoRI and XbaI, respectively. EcoRI and XbaI digested vector, pRL101, and insert carrying xylS were ligated using 1:3 molar vector:insert ratio in a total volume of 20 μl using the T4 Quick Ligase (NEB). The mixture was then transferred into E. coli via chemical transformation and grown on LA plates supplemented with 50 μg/ml kanamycin. The xylS coding sequence carries a BsaI recognition sequence, and to eliminate this site the entire plasmid was amplified by the primer pair XylS-delBsaI-F and XylS-delBsaI-R (both primers carry a single nucleotide point mutation, hence eliminating the BsaI recognition sequence, Table S1), by following the Overlap Extension PCR cloning method (41). After DpnI treatment, the resulting linear PCR product was transformed into E. coli and grown on LA plates supplemented with 50 μg/ml kanamycin. The resulting plasmid, pJN101, was then PCR-amplified by the primer pair BsaI-Bla-F and BsaI-Bla-R (Table S1), amplifying the entire plasmid excluding the native β-lactamase promoter, followed by DpnI digestion. Both the PCR product and the GeneEE(−SD)BsaI were digested with BsaI and used in cloning as a vector and insert, respectively. The ligation mix was made using 1:5 molar vector:insert ratio in a total volume of 20 μl using the T4 Quick Ligase (NEB). The mixture was then transferred into E. coli via chemical transformation and grown on LA plates supplemented with 50 μg/ml kanamycin. The resulting library was consisting of ~30,000 transformants.
2.13 Functional screening for artificial inducible gene expression in *E. coli*

For the functional screening of clones with inducible gene expression phenotype, the library, consisting of 30,000 clones, was plated out on two sets of LA plates, one set with inducer (1 mM m-toluic acid) and another set without inducer. Each set consisted of nine LA plates supplemented with 2, 3, 4, 5, 6, 7, 8, 9 and 10 mg/ml ampicillin. A 96-well microtiter plate, containing LB supplemented with 50 µg/ml kanamycin, was inoculated by picking random colonies grown on plates with inducer and ampicillin concentrations ranging from 6 to 10 mg/ml. The plate was incubated overnight at 37°C under 800 rpm constant agitation. The cells were then transformed, using a 96-pin replicator, to two sets of plates: one with (1 mM m-toluic acid) and one without inducer, with each plate supplemented with 2, 3, 4, 5, 6, 7, 8, 9 or 10 mg/ml ampicillin. In total, 27 clones were found to be showing an inducible phenotype, and all 27 clones were Sanger sequenced (GATC Biotech, Germany) using the primer pRL101-Bla-Seq (Table S1). This led to seven unique constructs A5, A7, C12, H3, C4, F2 and F11 (Table 2, Figure 4). To confirm the XylS-dependent induction, the xylS coding sequence with its native promoter was deleted from the corresponding constructs by digesting with XbaI and NdeI, blunting the linear DNA with T4 DNA polymerase (NEB) and ligated using the T4 Quick Ligase (NEB). The mixture was then transferred into the cloning host *E. coli* via chemical transformation and grown on LA plates supplemented with 50 µg/ml kanamycin. As for the final confirmation, all 16 clones were grown in a 96-well microtiter plate containing LA supplemented with 50 µg/ml kanamycin, and after overnight growth they were replica plated onto two sets of plates with (1 mM m-toluic acid) and without the inducer, each set with LB plates supplemented with ampicillin concentrations ranging from 2 to 10 mg/ml (Table 2).

| Table 2. The ampicillin resistance phenotypes of the clones with inducible phenotype identified in *E. coli*. |
|---|---|---|---|
| Clones | Plasmids with XylS<sup>a</sup> | Plasmids without XylS<sup>b</sup> |
| | − inducer<sup>c</sup> | + inducer<sup>d</sup> | − inducer | + inducer |
| A5 | 2 | 4 | 2 | 4 |
| A7 | 2 | 4 | 2 | 4 |
| C12 | 2 | 4 | 2 | 4 |
| H3 | 2 | 4 | 2 | 4 |
| C4 | NG<sup>e</sup> | 4 | NG | NG |
| F2 | 2 | 6 | NG | 2 |
| F11 | 2 | 6 | NG | 2 |
| Control<sup>f</sup> | NG | NG | NG | NG |

<sup>a</sup>Indicates the condition where the xylS gene is present in the plasmids.
<sup>b</sup>Indicates the condition where the xylS gene is absent from the plasmids.
<sup>c</sup>In the absence of inducer.
<sup>d</sup>In the presence of inducer, 1 mM m-toluic acid.
<sup>e</sup>The numbers indicate the highest ampicillin concentrations at which growth was observed. The range of concentrations used was 2, 3, 4, 5, 6, 7, 8, 9 and 10 mg/ml.
<sup>f</sup>No growth.
<sup>g</sup>The control construct is the re-circularized plasmid with no insert.

2.14 Construction and functional screening of GeneEE plasmid DNA libraries in *T. thermophilus*

For the construction of GeneEE plasmid DNA libraries, an *E. coli*-T. thermophilus shuttle vector, pMK184 (Table S3), was used, providing the thermostable kanamycin resistance as the reporter. The plasmid was PCR-amplified using the primer pair Km<sup>T</sup>-F and Km<sup>T</sup>-R (Table S1) excluding the native P<sub>slpA</sub> promoter controlling the thermostable kanamycin marker. The reverse primer Km<sup>T</sup>-R carries the BioBrick Prefix as a flanking sequence that serves as a homologous sequence in Gibson cloning. This amplicon was used as the insert. The Gibson ligation mixture was made using equimolar amounts of purified vector and insert (totaling 100 ng DNA). The mixture was then chemically transformed into *E. coli* via chemical transformation and grown on LA plates supplemented with 50 µg/ml kanamycin. As for the final confirmation, all 16 clones were grown in a 96-well microtiter plate containing LA supplemented with 50 µg/ml kanamycin, and after overnight growth they were replica plated onto two sets of plates with (1 mM m-toluic acid) and without the inducer, each set with LB plates supplemented with ampicillin concentrations ranging from 2 to 10 mg/ml (Table 2).

For the functional screening of the GeneEE plasmid DNA libraries in *T. thermophilus*, the purified plasmids from the *E. coli* transforms were transformed, using natural competence, to *T. thermophilus* as previously described (49). A fresh pre-warmed TB was re-inoculated with an overnight culture of *T. thermophilus* with a 1:50 dilution ratio. The strain was cultivated in a 65°C shaking incubator until an OD<sub>600</sub> of 0.4 was reached. From this culture, 0.8 ml were aliquoted into a new tube and 200 ng of the GeneEE plasmid DNA libraries isolated from the *E. coli* clones were added. The mixture was incubated for 4 h at 65°C and afterward was plated on selective TA supplemented with 30, 60 and 90 µg/ml kanamycin. The plates were incubated overnight at 65°C. Several clones were randomly picked from the TA plates, and plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen) and quantification with Nanodrop (ThermoFisher).

For the functional screening of the GeneEE plasmid DNA libraries in *C. glutamicum* for the construction of the GeneEE plasmid DNA libraries to be screened in *C. glutamicum*, the pXM19 plasmid was amplified by the primer pair pXM19-Chl-F and pXM19-Chl-R (Table S1), leading to the amplification of the entire pXM19 plasmid excluding...
the native ‘chloramphenicol’ promoter. This primer pair also introduces BsaI restriction sites within the overhangs used for the BsaI-based restriction cloning. After the PCR, the PCR product was digested with DpnI and BsaI at 37°C for 3 h and purified with the QiaQuick PCR purification kit (Qiagen). The GeneEE(−SD) segment was used as the insert and ligated with the pXM19 plasmid amplicon in a 1:7 vector:insert ratio, overnight at 16°C with T4 DNA ligase (NEB). The ligation mixture was first heat-inactivated at 65°C for 20 min and then transformed to chemically competent E. coli cells, leading to ~1000 transformants/10μl of ligation mix. The final library consisted of ~2000 E. coli clones. The resulting transformants were scraped from the agar plates, and plasmid DNA was isolated with the QIAprep Spin Miniprep Kit (Qiagen).

For the functional screening of the GeneEE plasmid DNA libraries in C. glutamicum, the purified plasmids from the E. coli transformants were electroporated to C. glutamicum, as described previously (45). The electroporated cells were plated on LA plates supplemented with 15 μg/ml chloramphenicol, leading to a library of ~200 transformants in C. glutamicum. Among the obtained clones, 10 clones were randomly selected and plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen) and were Sanger sequenced (GATC Biotech, Germany) using the primer pXM19-Chl-Seq (Table S1). The identified ARESs are listed in the supplementary spreadsheet.

2.16 Construction and functional screening of GeneEE plasmid DNA libraries in S. albus and S. lividans

For the functional screening of the GeneEE plasmid DNA libraries in S. albus and S. lividans, a shuttle expression plasmid based on pKC1218 (Table S3) was modified as follows: the pKC1218 backbone was amplified using primers pKC-Km-F and pKC-Km-R (Table S1) introducing 40-nt homology overhangs to a 1100 bp Sall fragment of pHH100 containing the aph(3′)′ kanamycin resistance gene, and the cassette was cloned between oriT and SCP2 replicon by in vivo homologous recombination in E. coli (50) yielding pKE101.

pKE101 was linearized by PCR using primers BsaI-Am5-F and BsaI-orf3-R (Table S1) introducing BsaI overhangs on both ends. The product was phosphorylated and re-ligated, and the circularized product was digested with BsaI. GeneEE(−SD)-BsaI was digested with BsaI and ligated directly upstream of the promoterless aac(3′)IV gene, leading to GeneEE library in pKE101. This library was then transferred to turbo competent E. coli C2984 cells (NEB) by chemical transformation. Transformants were selected on agar plates containing LA supplemented with 50 μg/ml kanamycin. The entire population of the obtained transformants (lawn on plate) was pooled, and the plasmid DNA was isolated with the QIAprep Spin Miniprep Kit (Qiagen). One microgram of plasmid DNA was then transformed to chemically competent E. coli S17-1 cells for conjugal transfer of the library to the Streptomycetes strains. Escherichia coli S17-1 transformants were selected on LA supplemented with 50 μg/ml kanamycin. For the GeneEE sequences that lead to functional ARES both in E. coli and Streptomyces strains, transformants were also selected on LA supplemented with 50 μg/ml apramycin.

For the functional screening of GeneEE plasmid DNA libraries based on apramycin resistance phenotype, the GeneEE plasmid DNA library established in E. coli was conjugated from E. coli S17-1 to S. albus and S. lividans. Conjugation reactions were performed as described previously (51) with minor modifications. S17-1 library transformants (~20000 transformants for both libraries) were pooled by adding 3 ml of LB to a 14-cm agar plate, and the cells were brought into suspension using a sterile glass rod. A volume of 100 μl of the obtained suspensions was used to inoculate 25 ml of LB and incubated at 37°C for ~2.5 h until the cells had reached an OD600 nm of 0.4. The cells were centrifuged at 2000 g for 5 min at room temperature, and the obtained pellet was re-suspended in 2 ml of fresh LB and plated on ice. S. albus and S. lividans were prepared using 4 ml of sDH2O, and the obtained suspensions were filtered through sterile cotton wool. A volume of 50 μl of these spore suspensions were added to 500 μl 2x YT medium and incubated at 50°C for 5 min to induce germination. The spore suspensions were cooled under running water before 500 μl of E. coli suspension was added, and the resulting suspension was mixed by inversion before spreading it onto two 9-cm agar plates (SFMA + 10 mM MgCl2 for S. albus, and ISP4 + 10 mM MgCl2 for S. lividans). Conjugation plates were incubated at 30°C for 16 h before overlaying them with antibiotic solutions, yielding final concentrations of 30 μg/ml nalidixic acid and 50, 250 and 500 μg/ml apramycin in agar media. The plates were then further incubated at 30°C until exconjugants appeared (3 days). Single exconjugant colonies were transferred to fresh plates supplemented with the corresponding concentrations of apramycin.

Single colonies of exconjugants were colony-PCR-amplified with the primer pair 5511-F and 234-R (Table S1) amplifying a 710-bp fragment surrounding the ARES sequence region. Single colonies were picked into 100 ml of 200 mM lithium acetate and 1% sodium dodecyl sulfate and incubated at 70°C for 5 min. A total of 300 μl of 96% ethanol were added, and the suspension vortexed and centrifuged at 15 000x g for 3 min. The pellet was washed with 70% ethanol and dried and dissolved in 20 ml of sterile deionized water. One milliliter of the resulting solution was used as template for PCR using Taq polymerase (NEB). Amplicons of the expected size were extracted from 0.8% agarose gels and purified using the QiaQuick Gel Extraction Kit (Qiagen) and were Sanger sequenced (GATC Biotech, Germany) with the sequencing primers 5654-F-Seq or Am-Sc-R-Seq (Table S1). The identified ARES with their corresponding phenotypes (apramycin resistance levels) are listed in the supplementary spreadsheet.

2.17 Construction and functional screening of GeneEE plasmid DNA libraries in S. cerevisiae

For the construction of GeneEE plasmid DNA libraries, an E. coli- S. cerevisiae shuttle vector, pENZ004 (Table S3), was used. The pENZ004 plasmid provides the homologous regions suitable for genomic integration, via gene replacement into chromosome II, and the tryptophan gene used for the functional screening in S. cerevisiae. pENZ004 was PCR-amplified using the primer pair Trp-F and Trp-R (Table S1), leading to the amplification of the entire plasmid excluding the native tryptophan promoter. The reverse primer Trp-R carries the BioBrick Prefix (40) as a flanking sequence that serves as a homologous sequence in Gibson cloning. This ampiclon was used as the backbone, and the GeneEE(−SD)Trp was used as the insert. The Gibson ligation mixture was made using equimolar amounts of purified vector and insert (totaling 100 ng DNA). The mixture was then transferred into the cloning host E. coli via chemical transformation and grown on LA plates supplemented with 50 μg/ml ampicillin. The resulting library consisted of ~10 000 transformants which were scraped, and the plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen) and later quantified with Nanodrop (ThermoFisher).

The GeneEE plasmid DNA library isolated from the scraped E. coli cells was linearized with SfiI in CutSmart Buffer at 50°C. Saccharomyces cerevisiae cells were grown in 5 ml of 2x YEPDB
on a rotary shaker at 200 rpm at 30 °C. The titer of the S. cerevisiae culture was determined by measuring OD600 (an OD600 of 1 equals to 3 x 10^3 cells/ml). Cells were inoculated to 25 ml of prewarmed 2x YEPD to a titer of 5 x 10^6 cells/ml and were grown for 4 h at 30 °C, with agitation until titer reached 2 x 10^9 cells/ml. The cells were harvested by centrifugation at 3000 g for 5 min, and the pellet was washed twice in 25 ml of sterile deionized water and resuspended in 1 ml of 0.1 M LiOAc. The DNA was transformed into S. cerevisiae using the established LiOAc protocol (52), and the resulting cells were plated on drop-out media. Plates were incubated at 30 °C for 4 days. To sequence ARES, crude yeast genomic DNA extractions were performed as described previously (53). The genomic DNA was then PCR-amplified using the primer pair Hom-F and Hom-R (Table S1). The resulting amplicon was PCR-purified with the QiAQuick PCR Purification Kit (Qiagen) and was incubated at 30 °C for 1 day.

3. Results

3.1 Proof of principle demonstration of the GeneEE approach

To test the potential of random DNA sequences to mediate gene expression in E. coli, we cloned double-stranded random DNA fragments (GeneEE segments) upstream of a β-lactamase CDS, conferring ampicillin resistance, into a plasmid, pRL101, harboring a constitutively expressed kanamycin resistance gene in the backbone.

We adopted two strategies regarding the composition of the random nucleotide libraries. The first strategy considers the prevalence of the SD sequence within the 5' UTR of mRNA in recruiting ribosomes (54). Therefore, we constructed a GeneEE(±SD) library, using double-stranded DNA composed of 200 nt random nucleotides, followed by a defined SD sequence (GGAG) and seven additional random nucleotides before the start codon (ATG) to provide a sufficient spacer length (55). To keep the fixed sequences to a minimum, the GGAG was chosen as the minimal yet functional SD sequence (56). The second DNA library, GeneEE(−SD), was meant to be more adaptable and free of constraints; therefore, it was made up of only 200 random nucleotides without any fixed sequences placed upstream of the start codon (Figure 1A). Control plasmids were created by ligating the 3’-end of the backbone and the 5’-end of the β-lactamase CDS together with no insert. After transformation of both plasmid libraries into competent E. coli cells, aliquots were plated out on agar plates supplemented with 50 µg/ml kanamycin, to calculate the library size, and on plates containing 50, 500 and 1000 µg/ml ampicillin, to determine the frequency of clones that carry functional ARES.

To our surprise, a large fraction of the GeneEE(±SD) library (33.2 ± 6.3%) contained functional ARES, as determined by the number of colonies obtained on 50 µg/ml ampicillin agar plates (Table 1). The total number of positive clones obtained were less on plates containing higher concentration of ampicillin: 4.7 ± 1.8% and 1.7 ± 0.4% on 500 and 1000 µg/ml ampicillin, respectively. Similarly, the GeneEE(−SD) library resulted in 43.1 ± 5.5% of positive clones on 50 µg/ml ampicillin and similar values as for GeneEE(±SD) on higher ampicillin concentrations. Although there is no statistical difference between the two libraries (one-tailed t-test, P-value = 0.559, 95% confidence level), given the observation that a random nucleotide sequence with no defined SD sequence is at least as likely to trigger protein production as SD-containing sequences indicates the ubiquity of alternate translation initiation mechanisms in bacteria (57, 58).

3.2 Protein expression quantification from GeneEE segments and insights into transcription machinery recruitment

To get more insight into how GeneEE segments attract the native transcription machinery and acquire quantitative information, we created a plasmid DNA library containing GeneEE(±SD) inserted upstream of a gene expressing the red fluorescent protein mCherry. From a library consisting of ~5000 clones, 192 E. coli clones expressing mCherry at various levels were picked from the agar plates and grown overnight in 96-well plates, and their respective fluorescence intensities were measured. For all clones, the functional ARES, mRNA levels (transcript per million, TPM) and TSSs were experimentally determined by DNA and RNA sequencing, respectively, that led to the identification of unique 157 constructs (the reduction from 192 to 157 constructs was due to the elimination of constructs containing ambiguous characters). The analysis of DNA sequencing data of ARES reveals the random nucleotide composition with high-sequence variation of the GeneEE(±SD) library without any conserved positions (Figure 2A). We observe a large variation in the number of mCherry transcripts (Figure 2B) and mCherry fluorescence intensities (Figure 2C). However, there is no correlation between the measured levels of mCherry fluorescence intensities and mCherry transcript abundance for the different clones based on the TPM counts in Table S5 versus fluorescent intensities listed in the spreadsheet for E.coli-mCherry constructs. The majority of ARES contain multiple TSSs (up to 11) (Figure 2D), indicated by groups of transcripts with different mRNA lengths (supplementary spreadsheet), with the TSS located mostly closer to the translational start of the mCherry CDS (Figure 2E). There is also no correlation between the number of TSS and expression levels; the construct O3 leads to one of the strongest expressions despite having only one TSS, while the construct B1 leads to one of the weakest despite having five TSSs.

To investigate the presence of potential promoter motifs within the artificial promoter sequences, the regions spanning from +1 to −50 were analyzed and led to the identification of various motifs (Table S5). In all artificial promoter sequences, conserved nucleotides in specific positions corresponding to σ-factor motifs could be identified. For example, we found the TG_N−10 extended box in a fraction of ARES across clones with different transcript levels. Interestingly, relative proportion of TG_N−10 within each population increased with expression levels. Similarly, we identified a −35 TTGNNN motif in all sequences analyzed regardless of the transcript levels. Lastly, we could establish a positive correlation between the amount of mCherry transcripts and the relative presence of a −10 TANNNT motif (Figures 2F–K). With these experimental findings, we demonstrate that GeneEE segments...
can be used for generating constitutive artificial gene expression systems in E. coli presenting a wide range of expression levels.

3.3 Potential of GeneEE for generating inducible expression systems

We were able to confirm that GeneEE segments may be used to generate functional ARES that led to constitutive gene expression using the experiments described above. We then sought to investigate whether the random DNA segments can also be used for recruiting transcription factors enabling inducible gene expression. For this study, we used the XylS/Pm inducible system as a model. XylS is a transcription factor that dimerizes upon binding of the inducer (m-toluic acid). Subsequently, the dimerized XylS binds to two DNA regions within the Pm promoter (59), one of which overlaps with the −35 box σ-factor binding site by 2 bp. This DNA interaction, in combination with σ72 or σ24, is thought to be the key in initiating transcription from Pm. To generate artificial promoters regulated by the XylS protein, we constructed a plasmid library containing the xylS CDS (with its native constitutive promoter) and GeneEE(−SD) upstream of the β-lactamase CDS. Upon transformation of plasmid DNA into competent E. coli cells, a library consisting of ∼30,000 clones was obtained. Using replica plating, transformants were inoculated onto two sets of LA plates containing increasing concentrations of ampicillin, one set with and the other without m-toluic acid. The screening resulted in the identification of 27 E. coli clones exhibiting an inducible phenotype, of which DNA sequencing revealed seven distinct ARES (the DNA sequences are provided in the supplementary spreadsheet, Table 2).

To ensure that induction is XylS-dependent, we deleted the xylS gene from the seven plasmids and characterized the phenotype of the clones harboring the plasmids without XylS. Among the seven clones, three clones had lost their inducible phenotype, indicating the reliance of induction on the presence of XylS. However, the remaining four clones still exhibited an inducible phenotype despite the absence of XylS, indicating the presence of an alternative mechanism that responds to m-toluic acid in E. coli. The DNA sequence analysis of ARES that mediate XylS-dependent induction do not reveal the known binding sites for XylS. However, conserved thymidine nucleotides were present in all ARES that led to the inducible phenotype (Figure 4). Overall, we demonstrate that GeneEE can be used to generate artificial inducible promoters responding to a defined transcription factor.

3.4 Universality of GeneEE

Finally, we wanted to look at GeneEE’s ability to generate gene expression systems in a variety of bacteria and yeast with varying genomic GC/AT ratios (Table 3). We chose two Gram-negative bacterial species, P. putida and T. thermophilus, three Gram-positive bacteria, C. glutamicum, S. albus and S. lividans and the eukaryotic yeast S. cerevisiae (Table S4). For this experimental setup, the following selection markers were used: antibiotic markers apramycin for Streptomyces strains; ampicillin for E. coli; chloramphenicol for C. glutamicum and thermostable kanamycin for T. thermophilus; mCherry for E. coli and P. putida and tryptophan for S. cerevisiae. The antibiotic markers used in this study were chosen due to their availability and extensive use in the hosts. While functioning ARES in bacterial systems relied on replicating plasmids, they have been chromosomally incorporated in yeast.

The antibiotic marker selection principle is the same as that used for the ampicillin resistance screening in E. coli, as detailed above. For antibiotic screening in T. thermophilus, the natural regulatory sequences upstream of the thermostable kanamycin were replaced with the GeneEE(−SD) segment. To find clones that lead the various expression levels, Gibson assembly was conducted and the resulting ligation mixture was plated on agar plates supplemented with 30, 60 and 90 g/ml kanamycin. DNA sequencing revealed 11 unique constructs, and their sequences are provided in the supplementary spreadsheet.

For functional screening of mCherry expressing clones in P. putida, the same library that was produced for E. coli was used since the plasmid with the mini-RK2 replicon can also replicate in P. putida (Table 3). The plasmid DNA library was electroporated into electrocompetent P. putida cells and were plated on kanamycin-containing agar plates. ARES were determined by DNA sequencing in colonies with detectable mCherry production (Figure 3), and 11 such sequences are presented in the supplementary spreadsheet.

The selection marker chloramphenicol was used for functional screening of the GeneEE plasmid DNA libraries in C. glutamicum. The chloramphenicol-carrying plasmid was PCR-amplified by removing the native regulatory regions and replacing them with the GeneEE(−SD) segment. The resulting library in E. coli consisted of around 2000 clones that were electroporated to C. glutamicum, yielding ∼200 clones that were growing on LA plates supplemented with 15 g/ml chloramphenicol. Ten growing clones were chosen at random from the growing clones, and the resulting ARES are shown in the supplementary spreadsheet.

The antibiotic marker apramycin was utilized to identify ARES in S. albus and S. lividans. The natural regulatory region upstream of the apramycin gene was removed and replaced with a GeneEE(−SD)BsaI segment in this cloning. The library was created in E. coli and conjugated to the two Streptomyces strains before being plated on agar plates containing 50, 250 and 500 g/ml apramycin. By DNA sequencing, 16 functional ARES in S. albus and eight clones in S. lividans were identified among the growing cells, and their sequences are reported in the supplementary spreadsheet.

Baker’s yeast strain S. cerevisiae CEN.PK, a tryptophan auxotroph, was employed for the final demonstration. Plasmid libraries were constructed by removing the native regulatory sequences upstream of the tryptophan gene and replacing them with the GeneEE(−SD)Trp segment. The plasmid DNA library that resulted was transformed into S. cerevisiae, and the transformants were plated out on drop-out media devoid of the amino acid tryptophan. Growth could occur only if the effectively chromosomally integrated DNA cassette contains a functional ARES that causes the tryptophan gene to be expressed, allowing growth in the drop-out medium. Similar to what was discovered in the bacterial screening efforts, functional expression in the yeast strain could be achieved using GeneEE segments. The ARES of ten such clones were determined by DNA sequencing, and the sequences are supplied in the supplementary spreadsheet.

In this study, we demonstrate that using the GeneEE approach, artificial gene expression systems can be created in all seven investigated microbes utilizing six markers/reporters (Table 3). The GeneEE approach relies on nature’s ability to detect functional sequences among a set of DNA with random DNA composition. In this aspect, the approach can be applied in any HoI, making it a universal and versatile strategy for creating artificial gene expression systems capable of functionally expressing a wide range of CDSs in a variety of host microorganisms.
Table 3. The hosts, GC contents, selection markers/reporters, plasmids, inserts and the cloning methods used for the identification of ARES.

| Selection markers/reporters | Ap | Am | Cm | Km | mCherry | Trp |
|-----------------------------|----|----|----|----|---------|-----|
| Plasmids                    | prL101 | pKE101 | pXM19 | pMK184 | pLT101 | pENZ004 |
| Inserts (GeneEE)            | (±SD) | (−SD) | (−SD) | (+SD)Km | (+SD)mCherry | (−SD)Trp |
| DNA cloning methods         | BsaI | BsaI | BsaI | Gibson | Gibson | Gibson |
| Hosts (GC contents %)       | E. coli (50.6) | P. putida (61.5) | T. thermophilus (69.5) | C. glutamicum (53.8) | S. albus (72.6) | S. lividans (72.2) | S. cerevisiae (38) |

4. Discussion

In this study, we demonstrate that DNA libraries with random nucleotide composition can be used to generate artificial gene expression systems in seven different microorganisms. The detailed analysis of artificial promoter sequences identified in E. coli demonstrates that the RNA polymerase holoenzyme can recognize a wide variety of DNA sequences for transcription initiation. Furthermore, the GeneEE segments also provide functional motifs and mechanisms for transcription and translation initiation in eukaryotes. GeneEE may also provide gene regulatory functions or noncanonical translation initiation in E. coli (57). Alternative mechanisms of translation initiation are also known from viruses, such as the T7-type translation in bacteria (60) and internal ribosome entry site-type translation in eukaryotes (61). The experimentally observed sequence diversity allows diverse transcription and translation mechanisms to participate in a wide range of expression levels. Application of the GeneEE approach alleviates the species barriers by removing the dependency on specific motifs in semi-rational genetic designs and proposing appropriate sequences for practically any microbial host. The large-scale application of GeneEE has the potential to reveal unknown DNA motifs and mechanisms for transcription and translation initiation in bacteria. GeneEE may also provide gene regulatory sequences to help with establishing novel hosts suitable for industrial applications. While the GeneEE approach is broadly applicable, it does have certain limitations because it requires that replicating plasmid(s) and/or an efficient genome-integration method and DNA transformation protocols are established for the host.

Heterologous production of recombinant proteins in host microorganisms frequently results in a cellular stress response characterized by a decline in the overall cell fitness. As a result, cell fitness (among many other factors) is an important constraint that restricts the extent of the feasible region for optimum recombinant protein production (Figure 5). When numerous GoIs are expressed using individual expression systems, they will occupy a single area in a large solution space (Figure 5, empty circle). Using GeneEE, however, a large number of constructs can be generated (Figure 5, full circles), and these constructs can explore viable solutions that go beyond what can be reached with individually constructed systems (exemplified by an empty circle). As a result, using the GeneEE approach, iterative DNA modification efforts in generating construct(s) that lead to optimum protein output while ensuring cell fitness can be minimized. By screening functional proteins, one may easily identify multiple constructs that fulfill numerous constraints, as any construct that ends up outside of the feasible region will not be represented as a viable colony.

The capacity of DNA segments with random nucleotide composition to generate functional gene expression and protein production is of importance not just for biotechnology but also for providing insights into evolutionary processes such as the acquisition of novel features via horizontal gene transfer. If a given gene confers a selective advantage to the receiving organism, the new feature can be passed down to offspring permanently. Our study indicates that the probability for initiation
of transcription and translation is likely not restricted when an organism receives a foreign gene. Therefore, an evolutionary advantage can be acquired rather easily by incorporating exogenous DNA sequence. This characteristic also reflects the prevalence of promoter sequences that the transcription machinery recognizes, indicating that the evolution of transcriptional regulation is centered on transcription inhibition (or downregulation) rather than on transcription initiation. Apart from these evolutionary implications, XylS/artificial-Pm study also demonstrates that inducible gene expression systems can be achieved with GeneEE by recruiting and identifying native transcription regulation systems. The entire repertoire of native transcription machinery, including transcription regulators/factors and small regulatory RNAs, can be recruited by the GeneEE segments.

Overall, our study demonstrates the potential of the GeneEE approach, using easily available methods for assessing selection markers and reporter proteins. GeneEE, in combination with high-throughput approaches, such as fluorescence-activated cell sorting (62) and microfluidics (63), will excel as a fast and versatile tool for gene expression and protein production engineering in a plethora of microorganisms.

Supplementary Data
Supplementary data are available at SYNBIO Online.

Data availability
All data generated or analyzed during this study are included in this article and its supplementary materials or are available from the corresponding authors upon reasonable request.

Funding
Norwegian University of Science and Technology (NTNU)-Discovery program (in part); NTNU-Biotechnology, Enabling Technologies Program [personal PhD stipend to L.T.]; Faculty of Natural Sciences at NTNU [personal PhD stipend to M.F.L.]; EU-H2020 MetaFluidics project [685474]; Research Council of Norway [244278 and 316129].

Acknowledgements
The authors thank Dr José Berenguer (Universidad Autónoma de Madrid) for providing the pMK184 plasmid and the Thermus thermophilus strain; Sara Castaño Cerezo and Gilles Truan (Institut National des Sciences Appliquées de Toulouse, INSA Toulouse, Biosystems and Process Engineering Laboratory) for providing the pENZ004 plasmid and the Saccharomyces cerevisiae strain; Sara Castaño Cerezo and Gilles Truan (Institut National des Sciences Appliquées de Toulouse, INSA Toulouse, Biosystems and Process Engineering Laboratory) for providing the pENZ004 plasmid and the Saccharomyces cerevisiae strain; Niels-Ulrik Frigaard (University of Copenhagen, Department of Biology) for the pUC19-Bba-Km plasmid and Jochen Schmid for his valuable feedback to the manuscript.

Author contributions
R.L. and M.F.H-M. conceived and designed the study. L.T. worked with Escherichia coli and Corynebacterium glutamicum and solely performed the yeast experiments; J.N. performed the inducible phenotype experiments in E. coli; I.O. involved in creation of libraries in E. coli and C. glutamicum experiments; K.E. performed the Streptomyces experiments; C.F.A.W. performed the Thermus thermophilus experiments; M.A. and N.H. involved in the establishment of the method in E. coli; J.K. and C.R. performed the DNA sequencing and TSS determination experiments and M.F.L., R.L. and M.F.H-M. wrote the manuscript with feedback from all co-authors.

Conflict of interest statement. M.F.H-M. is employed by the company United Scientists CORE (Limited). The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References
1. Voigt,C.A. (2020) Synthetic biology 2020–2030: six commercially-available products that are changing our world. Nat. Commun., 11, 6379. 10.1038/s41467-020-20122-2.
2. Bhandari,B.K., Lim,C.S. and Gardner,P.P. (2021) TISIGNER.com: web services for improving recombinant protein production. Nucleic Acids Res., 49, W654–W661. 10.1093/nar/gkab175.
3. Liu,L. et al. (2013) How to achieve high-level expression of microbial enzymes: Strategies and perspectives. Bioengineered. 4, 212–223. 10.4161/bioe.24761.
4. Browning,D.F. and Busby,S.J. (2016) Local and global regulation of transcription initiation in bacteria. Nat. Rev. Microbiol., 14, 638–650. 10.1038/nrmicro.2016.103.
5. Haberle,V and Stark,A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol., 19, 621–637. 10.1038/s41580-018-0028-8.
6. Engstrom,M.D. and Pfleger,B.F. (2017) Transcription control engineering and applications in synthetic biology. Synth. Syst. Biotechnol., 2, 76–91. 10.1016/j/sysbio.2017.09.003.
7. Shine,J. and Dalgarno,L. (1974) The 3’-terminal sequence of Escherichia coli 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites (terminal labeling/stepwise degradation/protein synthesis/suppression). Proc. Natl. Acad. Sci., 71, 1342–1346. 10.1073/pnas.71.4.1342.
8. Kozak,M. (1978) How do eucaryotic ribosomes select initiation regions in messenger RNA?. Cell, 15, 1109–1123. 10.1002/0992-8674(78)90009-9.
9. Borujeni,A.E., Channarasappa,A.S. and Salis,H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Res., 42, 2646–2659. 10.1093/nar/gkt1139.
10. Borujeni,A.E. and Salis,H.M. (2016) Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. J. Am. Chem. Soc., 138, 7016–7023. 10.1021/jacs.6b01453.
11. Borujeni,A.E. et al. (2017) Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. Nucleic Acids Res., 45, 5437–5448. 10.1093/nar/gkx061.
12. Migone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. Genome Biol., 3, 1–10. 10.1186/gb-2002-3-3-reviews0004.
13. Cardinale,S. and Arkin,A.P. (2012) Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems. Biotechnol. J., 7, 856–866. 10.1002/biot.201200085.
14. Kosuri,S. et al. (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A., 110, 14024–14029. 10.1073/pnas.1301301110.
15. Horwitz,M.S.Z. and Loeb,L.A. (1986) Promoters selected from random DNA sequences (mutagenesis/evolution). Proc. Natl. Acad. Sci. USA, 83, 7405–7409. 10.1073/pnas.83.19.7405.
16. Hammer,K., Mijakovic,I. and Jensen,P.R. (2006) Synthetic promoter libraries - tuning of gene expression. Trends Biotechnol., 24, 53–55. 10.1016/j.tibtech.2005.12.003.
56. de Smit, M. H. and van Duin, J. (1994) Translational initiation on structured messengers: Another role for the Shine-Dalgarno interaction. *J. Mol. Biol.*, 235, 173–184. 10.1016/S0022-2836(05)80024-5.

57. Chang, B., Halgamuge, S. and Tang, S.-L. (2006) Analysis of SD sequences in completed microbial genomes: Non-SD-led genes are as common as SD-led genes. *Gene*, 373, 90–99. 10.1016/j.gene.2006.01.033.

58. Vigar, J. R. and Wieden, H.-J. (2017) Engineering bacterial translation initiation—Do we have all the tools we need? *Biochim. Biophys. Acta (BBA) - General Subjects*, 1861, 3060–3069. 10.1016/j.bbagen.2017.03.008.

59. Dominguez-Cuevas, P., Marin, P., Marques, S. and Ramos, J. L. (2008) Xyls–Pm promoter interactions through two helix-turn-helix motifs: Identifying XylS residues important for DNA binding and activation. *J. Mol. Biol.*, 375, 59–69. 10.1016/j.jmb.2007.10.047.

60. Sprengart, M. L., Fuchs, E. and Porter, A. G. (1996) The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*. *EMBO J.*, 15, 665–674. 10.1002/j.1460-2075.1996.tb00399.x.

61. Cornelis, S. et al. (2000) Identification and characterization of a novel cell cycle-regulated internal ribosome entry site. *Mol. Cell*, 5, 597–605. 10.1016/S1097-2765(00)80239-7.

62. Mattanovich, D. and Borth, N. (2006) Applications of cell sorting in biotechnology. *Microb. Cell Fact.*, 5, 12. 10.1186/1475-2859-5-12.

63. Agresti, J. J. et al. (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc. Natl. Acad. Sci.*, 107, 4004–4009. 10.1073/pnas.0910781107.