microbial biotechnology

Minireview

Importance of the 5′ regulatory region to bacterial synthetic biology applications

Lisa Tietze and Rahmi Lale* (iD)

PhotoSynLab, Department of Biotechnology, Faculty of Natural Sciences, Norwegian University of Science and Technology, Trondheim, N-7491, Norway.

Summary

The field of synthetic biology is evolving at a fast pace. It is advancing beyond single-gene alterations in single hosts to the logical design of complex circuits and the development of integrated synthetic genomes. Recent breakthroughs in deep learning, which is increasingly used in *de novo* assembly of DNA components with predictable effects, are also aiding the discipline. Despite advances in computing, the field is still reliant on the availability of pre-characterized DNA parts, whether natural or synthetic, to regulate gene expression in bacteria and make valuable compounds. In this review, we discuss the different bacterial synthetic biology methodologies employed in the creation of 5′ regulatory regions – promoters, untranslated regions and 5′-end of coding sequences. We summarize methodologies and discuss their significance for each of the functional DNA components, and highlight the key advances made in bacterial engineering by concentrating on their flaws and strengths. We end the review by outlining the issues that the discipline may face in the near future.

Introduction

In synthetic biology (SynBio) applications, standardized sequences, also known as DNA parts, whether natural or synthetic, are used to control gene expression in bacteria to produce valuable compounds. Standardized regulatory DNA parts are expected to behave consistently and independently of their genetic context. Decades of research have gone into standardizing and characterizing various genetic parts to enable assembly and programming of cells such as computers (Knight, 2003; Voigt, 2006; Salis *et al.*, 2009; Na and Lee, 2010; Nielsen *et al.*, 2016; Decoene *et al.*, 2018). However, for any phenotype there are a set of interrelated factors that determine the sequence-specific characteristics of DNA, which in turn contribute to the phenotype. These interrelated factors (referred to as context) are as follows: composition-specific (e.g. DNA composition and physical integration of genetic devices); environment-specific (e.g. temperature and pH); and host-specific (e.g. cellular components and physiology, and biochemical capacity) (Cardinale and Arkin, 2012). Therefore, the context independency of standardized parts has been questioned by many researchers as these parts may behave unexpectedly both in model and in non-model organisms (Cardinale and Arkin, 2012; Kittleson *et al.*, 2012). In bacteria, transcription and translation are coupled, and translation can be initiated on the nascent mRNA. The regulation of transcription and translation relies on the unique combination of features that are present in the promoter, 5′ untranslated region (UTR) and 5′-end of the coding sequences (CDS) (Fig. 1) (Goodman *et al.*, 2013; Tuller and Zur, 2015; Cambray *et al.*, 2018; Verma *et al.*, 2019). In this review, the promoter and 5′ UTR will together be referred to as 5′ regulatory sequence (RES). Given the multiple functions associated with RESs, context dependency determines the sequence composition of each of the functional elements within them (Cardinale and Arkin, 2012).

At present, methods that allow the rational design of the entire bacterial RES are not available. In the absence of such rational design tools, two main approaches are used in the field: either the use of standardized DNA parts, as outlined above; or the generation of specific RESs for each use case. In this review, we focus on the latter and outline the different methods that can be used to create specific RESs for bacteria. We exclude other factors that influence overall gene expression such as the 3′ UTR (for a general overview over the biological process of gene expression in
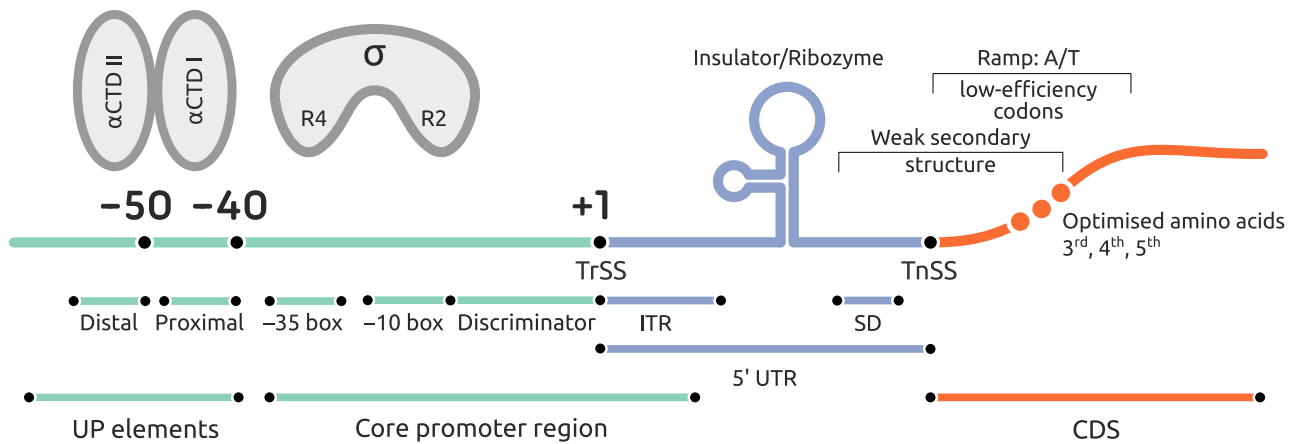
Fig. 1. Schematic representation of 5′ regulatory regions in bacteria. RNA polymerase α-subunits, C-terminal domain (CTD) I and II, interact with the proximal and distal subsites of the UP element (UP); sigma (σ)-subunit region 2 (R2) and region 4 (R4) interact with the −10 and −35 promoter motifs respectively (please note this depiction applies to σ^{70}, other σ-factors may have different binding preferences). ITR, initially transcribed region; SD, Shine–Dalgarno sequence; 5′ UTR, 5′ untranslated region; TrSS, transcriptional start site (+1); TnSS, translational start site. The regulatory regions are not to scale.

prokaryotes, please see the extensive review by Bervoets and Charlier (2019)). For the sake of clarity, we categorize the different approaches by outlining the terminology and definitions of different methods and coin umbrella terms complying with other reviews (Gilman and Love, 2016; Deaner and Alper, 2018; Jin *et al.*, 2019b).

Synthetic sequences

In this review, we define any sequence that is not naturally occurring as a synthetic sequence. Thus, we use synthetic as a broad term for the different categories of synthetic sequences described below.
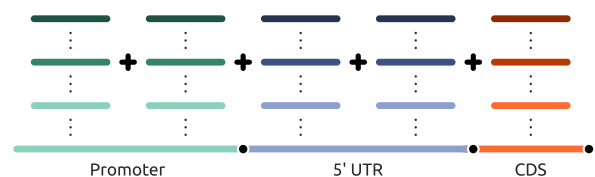
Hybrid regulatory sequences

The construction of hybrid RES relies on the combination of known DNA parts (Fig. 2A). An unlimited number of starting parts can be (re-)combined or merged, leading to the generation of combinatorial libraries. Using such combinations, a large combinatorial library can quickly grow in size based on the number of starting DNA parts. Hybrids methods are also known as shuffled, recombined or combinatorial method.
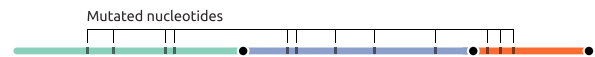
Mutated regulatory sequences

Mutagenesis is an approach that leads to the generation of novel sequences relying on existing DNA sequences (Fig. 2B). Mutagenesis introduces mutations in arbitrary positions of the DNA template, usually through a process called error-prone PCR – a method that relies on DNA polymerases that lack proofreading. Thus, the DNA
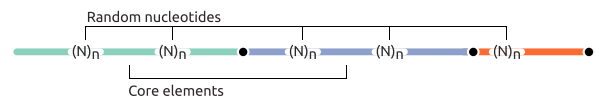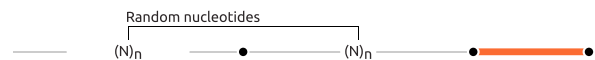
Fig. 2. Overview of methods used to create novel 5′ regulatory sequences. Promoter (green), 5′ untranslated region (blue) and CDS (orange).

polymerase can incorporate incorrect nucleotides to the growing nucleotide chain. It is possible to control the error rate by adjusting the PCR conditions.

Semi-artificial regulatory sequences

In the semi-artificial approach, known RES motifs are combined with random nucleotides (Fig. 2C). To construct semi-artificial RESs, core motif elements of a promoter and/or 5′ UTR are left intact, while the rest of the

sequence is replaced by random nucleotides (or the other way around). The nature of random nucleotides can vary from a completely random pool to partially randomized (also known as doped). This approach is also known as flanking saturation mutagenesis.

Artificial regulatory sequences

Artificial RESs are stretches of DNA sequences that are created using a random pool of nucleotides in the no predetermined order (Fig. 2D). This method leads to the generation of RESs that are not occurring naturally, hence the term artificial, and do not rely on the known RES motifs. Therefore, these artificial sequences are a valuable source for *de novo* construction of RESs when working with non-model organisms or when known RES do not satisfy the experimental needs.

5′ Regulatory sequences

Promoters

Transcription is the first step in gene expression. The RNA polymerase (RNAp) machinery recognizes promoter regions in DNA templates and generates transcripts. In bacteria, RNAp core enzymes build a complex with sigma (σ)-factors, forming holoenzymes, also called the transcription machinery. The holoenzyme recognizes and binds to various motifs within the promoter regions. Which motif the holoenzyme binds to depends on the specific σ-factor that forms the holoenzyme. For example, in *E. coli* the housekeeping σ^{70} binds to the −10 and −35 regions of a promoter (Fig. 1). RNAp contains an αC-terminal domain (CTD) subunit, which can bind UP elements (Lee *et al.*, 2012). Once assembled, the transcription machinery reads the downstream DNA and generates the transcript, mRNA, beginning from the transcription start site (TrSS), which is annotated as +1 (Engstrom and Pfleger, 2017; Bervoets and Charlier, 2019). With some exceptions, transcript abundance correlates positively with protein abundance (Liu *et al.*, 2016). Thus, for SynBio applications, controlling transcription through promoters is equal to controlling protein abundance, even though this simplified gene expression model fails to capture the complex biological processes; i.e., a strong promoter may not necessarily always lead to high protein production.

Hybrid, mutated and (semi-)artificial promoters

Hybrid, mutagenesis and (semi-)artificial approaches are valuable methods for the construction of constitutive and inducible promoters. Hybrid promoters commonly comprise a mix of two or more different promoters (Fig. 2) (Zhang *et al.*, 2012; Jiao *et al.*, 2017; Chen *et al.*,

2018b; Han *et al.*, 2020). For example, the widely used inducible *tac* promoter is composed of the naturally occurring *trp* promoter and the synthetic *lac*UV5 promoters (Boer *et al.*, 1983). Such hybrid promoters can contain multiple binding sites for DNA-binding proteins (Zong *et al.*, 2018; Jin *et al.*, 2019a; Wang *et al.*, 2019a; Wang *et al.*, 2019b). New promoter variants can also be created using mutagenesis, which has been used to vary the promoter strength of inducible and constitutive promoters (Alper *et al.*, 2005; Bakke *et al.*, 2009; Kinney *et al.*, 2010). Mutagenesis relies on the error-prone PCR method, which was established in 1996 by Zaccolo *et al.* (1996), to generate protein variants by altering the CDS. Since then, other groups also applied mutagenesis to RESs instead of the CDS and further developed the method including various alterations, such as doped mutagenesis (Winther-Larsen *et al.*, 2000), nicking mutagenesis (Wrenbeck *et al.*, 2016) or PFunkel (Firnberg and Ostermeier, 2012). Constitutive and inducible promoters can also be created using (semi-)artificial (Nair and Kulkarni, 1994; Carr *et al.*, 2017; Hwang *et al.*, 2018) and fully artificial sequences (Yona *et al.*, 2018; Lale et al., 2019; Urtecho et al., 2020). As early as 1986, Horwitz and Loeb (1986) replaced parts of a promoter region with random nucleotides, thus creating (semi-)artificial promoters. Transcriptional activities from two of the (semi-)artificial promoters led to increased protein production levels compared with the levels reached with the wild-type promoter, while being only partially homologous to it. The field of (semi-)artificial promoter design truly started when Hammer *et al.* (2006) showed its relevance not only for *E. coli* but also for the eukaryotic microorganism *Saccharomyces cerevisiae*. Nowadays, it is possible to create millions of (semi-) artificial promoter variants and screen them with high-throughput (HT) methods such as fluorescence-activated cell sorting (FACS) (Boer *et al.*, 2020).

Promoter engineering efforts usually focus on to the region where the holoenzyme binds. Different holoenzymes exist, depending on which σ-factor associates with the RNAp core enzyme. Which σ-factor associates with the RNAp core enzyme depends on environmental conditions. Different σ-factors bind to distinct promoter motifs. Hence, promoters can be created that are active in a wide spectrum of environmental and/or growth conditions, for instance, by combining the binding sites for housekeeping transcription factors such as σ^{70} with binding sites for σ-factors that are active under various stress conditions (Wang *et al.*, 2019a). In contrast, sometimes there is a need for promoters to be active only under specific conditions. Condition specificity can, for instance, be achieved by including binding sequences for non-housekeeping σ-factors in the promoter regions (Chen *et al.*, 2018a; Wu *et al.*, 2019).

Orthogonal promoter–σ-factor pairs are also highly specific, relying on the introduction of orthogonal σ-factors and binding sites from one microorganism into a new microorganism (Bervoets *et al.*, 2018; Bervoets and Charlier, 2019). Such orthogonal σ-factors will solely target the orthogonal promoter that was introduced along with it, making the system highly specific. Similar to that are chimeric promoters, which are a form of hybrid promoters that contain −10 and −35 regions of several microorganisms creating promoter variants that do not exist in nature (Rhodius *et al.*, 2013).

As outlined above, promoters created with random approaches can meet different requirements. A drawback of using hybrid and (semi-)artificial promoters is the need for fundamental knowledge about the various motifs. Rational design efforts are increasingly applied, especially with the advancement of deep learning methods. However, accurately predicting the promoter output is difficult to achieve as promoter strength varies depending on distinct features, such as the associated σ-factor (Wang *et al.*, 2019a), associated transcription factors (Ishihama *et al.*, 2016), UP elements (Lee *et al.*, 2012), competition for binding sites (Ishihama *et al.*, 2016; Han *et al.*, 2020), the sequence to which holoenzymes bind (Kinney *et al.*, 2010), and the spacer length and sequence between the motifs (Nair and Kulkarni, 1994; Jensen and Hammer, 1998; Han *et al.*, 2019). The distinct features can be a target for promoter engineering approaches; however, it is not always clear how these different features factor into promoter strength. For example, while combining multiple binding sites into one promoter has been shown to lead to increased promoter strength, as reviewed by Blazeck and Alper (2013), (Wang *et al.* (2019a) report that introduction of multiple σ-factor binding sites into one promoter can lead to decreased expression. Similarly, Sun *et al.* (2020a) found that combining multiple strong promoters in tandem did not lead to changes in expression levels. Despite the limitations, the various methods outlined above are valuable approaches that enable the generation of promoter variants with various characteristics. The above-mentioned methods provide ways to control gene expression and also enable the fundamental studies on the role of DNA sequence composition in promoter activity.

Transcription factors

Transcription factors (TFs) regulate transcription by activating or repressing promoter activities. A promoter may contain multiple transcription factor binding sites (TFBSs) and therefore might be regulated by multiple TFs (Ishihama *et al.*, 2016). The TFBSs can be located upstream and/or overlap with the σ-factor binding regions. TFs that repress promoter activity often bind downstream of the −10 region and sterically block the RNAp holoenzyme from progressing (for a detailed summary, see the review by Bervoets and Charlier (2019)). Gene expression can be modulated by targeting TFBSs. For example, Rohlhill *et al.* (2017) used mutagenesis to analyse and improve a native formaldehyde-inducible promoter in *E. coli*. The group identified several critical repressor binding sites and created promoter variants that carry point mutations that led to a lower basal and higher induced expression levels compared with the levels reached with the wild-type promoter. Mutagenesis can also be used to gain insights into the basic molecular mechanisms of how TFBSs influence transcription. For example, Urtecho *et al.* (2020) mutated different TFBSs and studied their effect on gene expression. The group found that TFBS mutations leading to increased expression clustered in the −10 and −35 regions, while mutations leading to decreased expression localized in and downstream of the −35 region. Even though such studies provide new insights into the transcriptional regulation mechanisms in *E. coli*, Urtecho *et al.* (2020) point out that it is still difficult to predict the function from DNA sequences alone. It is also unclear how the flanking regions of TFBSs influence promoter activity (Kinney *et al.*, 2010; Belliveau *et al.*, 2018; Urtecho *et al.*, 2020). As a comparison, in eukaryotes, variations in TFBS and flanking regions affect TF binding and gene expression output (Melnikov *et al.*, 2012; Sharon *et al.*, 2012; Dror *et al.*, 2015; Levo *et al.*, 2015; Li and Eisen, 2018). It is likely that in bacteria the flanking regions of TFBSs also affect TF binding and gene expression, even though it has not been extensively studied. Studies pointing at the possibility to modulate gene expression through TFBSs and their flanking regions mainly come from groups that characterized and mutated repressors or activators, such as LexA (Wertman and Mount, 1985), AraC (Hamilton and Lee, 1988), XylS (González-Pérez *et al.*, 1999) and LacI (Bahl *et al.*, 1977; Sadler *et al.*, 1983). Several of the above-mentioned research groups report that mutations in the TFBSs strongly influence promoter strength. However, the primary focus so far has been to discover TFBSs, and only a few research reports rational engineering strategies to engineer TFBSs and flanking regions to regulate protein production levels (Bulyk *et al.*, 2004; Rigali *et al.*, 2004).

Multiple transcription factor binding sites within promoters

It is still uncertain how the occurrence of multiple TFBSs within one promoter affects promoter function. TFs may compete for adjacent binding sites and thus hinder each other. Cox *et al.* (2007) studied the underlying principles

of TF competition using combinatorial libraries of hybrid promoters containing multiple TFBSs. The hybrid promoters contain fixed −10 and −35 regions with varying spacer sequence/length between them. Different TFBSs were placed upstream and downstream of the −10 and −35 regions. Two rules emerged from this study for the design of promoters containing multiple TFBSs: first, repression always overruled activation; second, repressive motifs repressed gene expression irrespective of their location in relation to the −10 and −35 regions. The second rule confirmed the findings reported in other previously published studies, outlined in a review by Ishihama on prokaryotic genome regulation (Ishihama, 2010). The first rule that repression always overrules activation was also reported in a study published by Kinkhabwala et al. (2008). Kinkhabwala and Guet (2008) combined activator and repressor elements and investigated how the different combinations affect gene expression. Activator proteins could only activate expression in the absence of a repressor protein. As soon as a repressor protein was present, gene expression decreased substantially or stopped completely. In no combinations tested, an activator protein was able to overpower the effect of a repressor protein. Furthermore, Moneteiro et al. (2018) described that the architectural order of TFBSs within one promoter influences its activity. Arranging multiple TFBSs in distinct orders can lead to varied gene expression. This also means that the effects of the individual elements do not behave additively when combined. Novel regulatory logics seem to emerge when combining repressing and activating TFBSs (Monteiro et al., 2018, 2019). It is also possible to merge multiple TFBSs into one TFBS as shown in two separate studies by the Silvia-Rocha group (Guazzaroni and Silva-Rocha, 2014; Amores et al., 2015). Merging different motifs into one TFBS creates a TFBS, which harbours a DNA-binding site for multiple TFs resulting in intricate logic gates since several TFs compete for the same DNA-binding site. This is an exciting discovery; however, it makes it difficult to predict the output of combinatorial and merged promoters without testing their performance first.

Discriminators – at the edge of promoter and 5′ UTR

One of the rate-limiting steps of transcription is the transition from initiation to elongation, also called promoter escape, which involves the formation of so-called open complexes (Reppas et al., 2006; Häkkinen et al., 2013; Dulin et al., 2018). When the holoenzyme binds to the promoter, it unwinds the DNA close to the TrSS forming an open complex. RNAp has to escape from this open complex to transition into the elongation phase of transcription, forming the elongation complex. Open complex lifetimes vary 10^5-fold (Henderson et al., 2017). The more stable the open complex is, the longer it takes RNAp to transition into the elongation phase (Dulin et al., 2018). Hence, short open complex lifetimes have been associated with strong promoters. For example, the strong rRNA promoter rrnB P1 is known for its short-lived open complexes (Barker et al., 2001). The lifetime of the open complex is influenced by the nucleotide sequence upstream of the TrSS also called the discriminator (Fig. 1). The discriminator is a non-consensus region between the −10 and +1 nucleotides and is involved in open complex formation (Sosa et al., 2020). The discriminator harbours a binding site for the RNAp-associated σ-factor (Haugen et al., 2006; Mazumder and Kapanidis, 2019). Non-optimal binding to the discriminator decreases open complex lifespan, causing rapid promoter escape (Haugen et al., 2006). Altering the discriminator sequence composition changes open complex lifetimes, RNAp escape rates, TrSS selection and promoter output as shown by several studies using mutated and artificial discriminator sequences (Josaitis et al., 1995; Pemberton et al., 2000; Haugen et al., 2006; Winkelman et al., 2016). The importance of discriminator sequence composition on gene expression has also been substantiated by a computational modelling study (Leiby et al., 2020). Leiby et al. developed a convolutional neural network (CNN) model that uses the DNA nucleotide sequence of a promoter to predict its resulting transcriptional initiation rate (TIR). The group investigated which nucleotides were the most relevant for the accurate prediction of TIR using the CNN model. First, the model was trained using a published promoter data set; afterwards, the promoters were randomized within a sliding window of varying length, and finally, promoter strength was predicted from these randomized promoter sequences. The predicted promoter strength of the randomized sequence was compared with the predicted strength of the original sequence to find nucleotide positions that contributed the most to the promoter strength. They reported that the −10, −35 and the discriminator sequence strongly contributed to σ^{70} promoter strength. Since the discriminator is involved in the rate-limiting step of open complex formation and stability, it is a suitable target for gene expression optimization at the junction between a promoter and a 5′ UTR (Fig. 1).

5′ untranslated region

The 5′ UTR is a frequently overlooked region for gene expression engineering. While several established tools enable the rational design of 5′ UTRs, such as RBS Calculator (Salis et al., 2009) and UTR designer (Seo et al., 2013), a substantial number of studies either fail to consider the 5′ UTR as a unique functional region, or simply

consider it as a part of the promoter region (Decoene *et al*., 2018; Nijs *et al*., 2020). In this review, we want to highlight the importance of the 5′ UTR since it influences both transcriptional and translational processes in bacteria. During translation initiation, ribosomes bind to mRNA and build the translation machinery together with several accessory proteins. Ribosomes can bind to the Shine–Dalgarno (SD) sequence present within the 5′ UTR part of an mRNA through complementary base pairing of the 16S rRNA (Fig. 1). There are many studies in the literature that erroneously refer to the entire 5′ UTR as ribosome binding site (RBS), a term that is synonymous with the SD sequence as specified in the Sequence Ontology (The MISO Sequence Ontology Browser). The binding rate and strength of the translation machinery to the 5′ UTR affect how often a transcript will be translated. Translation rates are affected by the 5′ UTR sequence composition and properties, which can be modified to modulate gene expression as we will outline in the following subsection.

The role of 5′ UTRs in transcript and translation

Several studies point out the importance of the 5′ UTR sequence composition on transcript abundance and translation rates. Studies by Berg *et al*. (2009) and Lou *et al*. (2012) indicate that point mutations within the 5′ UTR not only lead to changes in translation rates but also affect transcript abundance, potentially through the creation of a more stable transcript. Moreover, Le *et al*. (2020) recently reported a novel dual UTR concept benefiting from the 5′ UTRs diverse role both in transcriptional and in translational processes. In their study, the mutagenesis of a 5′ UTR in *E. coli* resulted in 5′ UTR variants that could be classified into two groups based on their distinct expression profiles: 5′ UTRs leading to increased translation (Tn group) and 5′ UTRs leading to increased transcription of the reporter gene (Tr group). Single 5′ UTRs from the two groups were then combined into a dual UTR, composed of two 5′ UTRs and a spacer DNA sequence in between them. The dual UTR leads to the discovery of a synergistic effect that results in increased expression of the reporter gene. The group hence demonstrated that dual UTRs can be used to control both transcription and translation of the reporter gene. They also report that it is crucial to place the 5′ UTR of the two different groups in the correct order to benefit from the observed synergistic effect as the expression of the reporter gene was decreased when the order of the 5′ UTRs in the dual UTR was Tn-Tr. These synergy and anti-synergy effects highlight the two distinct roles that 5′ UTRs play in bacterial gene expression, influencing translational and transcriptional processes. It also highlights how

RES composition and architecture can influence gene expression.

The 5′ UTR contains an initially transcribed region (ITS, Fig. 1), which influences promoter escape of RNAp (Heyduk and Heyduk, 2018). To study how different nucleotides in distinct positions affect RNAp promoter escape, Heyduk and Heyduk created a library of ITS variants, which covered up to a 144-fold difference in promoter escape kinetics. The first 10 nucleotides of the ITS influenced gene expression the most, requiring an A in position +1. As a follow-up experiment, the first 10 nucleotides of the ITS of different promoters were replaced with a fully randomized artificial sequence. Replacing the first 10 nucleotides of the ITS with every possible nucleotide combination, in their set-up, revealed that position +2 of the ITS requires a specific promoter-dependent nucleotide. Furthermore, nucleotide pairs of GG and GA within the ITS increased escape velocity, and high T content correlated with slow escape, specifically in the combination of (T/C)G. Similar effects were also seen in a study looking at the effect of ITS nucleotide composition at a genome-wide scale. Promoter escape seems to be hampered by T-rich ITS nucleotide sequences that can induce abortive transcription (a non-productive reiterative transcription process that occurs at most promoters leading to short transcripts in a template-dependent manner) (Imashimizu *et al*., 2020).

Another target for gene expression optimization is the junction between the promoter and 5′ UTR, and 5′ UTR and CDS. Mutalik *et al*. (2013a) report that these junctions affect gene expression and that they can be optimized and used to regulate gene expression levels. The importance of junction optimization for gene expression has also been supported through a study by Mirzadeh *et al*. (2015). The group used degenerate primers to randomize the six nucleotides upstream of the start codon and exchange codons 2 and 3 with all possible synonymous codons. Changing the nucleotides at the 5′ UTR and CDS subsequently led to differences in gene expression levels by several orders of magnitude.

The above-mentioned findings indicate that several promoter structures/functions overlap within the 5′ UTR (Mutalik *et al*., 2013b) (Fig. 1), even though some findings may stem from incorrectly annotated promoters and 5′ UTRs (Lou *et al*., 2012). Overall, controlling and studying gene expression through the 5′ UTR region presents a great potential (Ameruoso *et al*., 2019) as will be highlighted in the next subsection.

Modulating gene expression through the 5′ UTRs

Methods targeting the 5′ UTR can be used to vary gene expression by: mixing two or more different 5′ UTRs; making hybrids of sequences that ribosomes can bind to

(Isaacs *et al.*, 2004); mutating the 5′ UTR (Huang *et al.*, 2006); or replacing parts of conserved motifs or spacer regions with (semi-)artificial sequences (Min *et al.*, 1988; Zhelyabovskaya *et al.*, 2004; Zhang *et al.*, 2015; Bonde *et al.*, 2016; Oesterle *et al.*, 2017; Shi *et al.*, 2018) (Fig. 2). For example, in a study by Huang *et al.* (2006), protein production levels of *E. coli* alkaline phosphatase were improved by randomly mutating the 5′ UTR. Alkaline phosphatase activity was increased as much as sevenfold compared with the levels reached with the wild-type 5′ UTR. The mutations were studied separately, and it was concluded that the SD sequence had been affected by the mutagenesis step, creating a seemingly stronger motif. Notably, when the group studied the individual mutations, they found that the additive effects of single mutations were not enough to explain the sevenfold increase, suggesting a more intricate regulation of gene expression through the 5′ UTR than a simple addition of factors.

When modulating gene expression of metabolic pathways, there is a need to control the expression of multiple genes, for instance, in operons. In the assembly of pathways, the hybrid method is often used. For example, Pfleger *et al.* (2006) used the hybrid method to assemble a metabolic pathway, balancing gene expression by controlling the intergenic regions in an operon through the use of hairpins. The hairpins originated from a library of tuneable intergenic regions (TIGRs), which were inserted into the intergenic region of the operon. These intergenic hairpins post-transcriptionally controlled the expression of different genes in the operon, showing that hybrid 5′ UTRs can be used to tune the expression from a metabolic pathway. An interesting take on pathway assembly is the so-called Golden Mutagenesis (Püllmann *et al.*, 2019). This pathway assembly approach combines hybrid pathway assembly with mutagenesis, thus combining two different randomizing methods into one.

Hybrid, mutagenized and (semi-)artificial 5′ UTRs can be used to study the fundamental biology of gene expression

As outlined above, methods that use random approaches can be used to modulate gene expression for protein production purposes. In addition, these methods are also valuable for studying the fundamental mechanistic of gene expression regulation. For example, in 1985 Whitehorn *et al.* (1985) used the hybrid method to show that the spacer length between an SD sequence and the start codon affects gene expression levels of the human β-interferon in *E. coli*. The group also found that single nucleotide changes in the spacer region had dramatic effects, increasing accumulation of human

β-interferon to about 15% of the total cell mass. Holmqvist *et al.* (2013) studied how mutations in the 5′ UTR of the *csgG* gene would affect *cis*- and *trans*-regulation of the gene. The *csgG* mRNA is regulated by at least four small RNAs (sRNA) that repress translation by binding to specific stem–loop areas of the 5′ UTR. Introducing random mutations into the stem–loop area prevented binding of the sRNA and therefore led to increased expression, showing that random mutagenesis can also be used to regulate *trans*-effects.

Rational design approaches commonly rely on the use of known conserved motifs to introduce changes in nucleotide sequences. However, conserved motifs are not always necessary to drive gene expression. For example, the SD sequence was long thought to be necessary for ribosome binding to mRNA in all prokaryotes because the motif was found to be highly present when it was first described in *E. coli* (Shine and Dalgarno, 1974). The 16S rRNA ribosome subunit contains a conserved anti-SD sequence at its 3′-end that is thought to be essential for ribosome recruitment to the mRNA based on the complementary binding of these two sequences (Steitz and Jakes, 1975), supporting the notion of SD sequence-dependent translation. However, the SD sequence has since then been shown to be only partially conserved in other bacterial species, some species simply lacking an SD sequence for up to 88% of their genes (Chang *et al.*, 2006; Nakagawa *et al.*, 2017). Different organisms may also express genes from the so-called leaderless mRNAs lacking the 5′ UTR entirely (Zheng *et al.*, 2011). A study by Fargo *et al.* (1998) reports that several genes could successfully be expressed both in *E. coli* and in the chloroplast of *Chlamydomonas reinhardtii* (a single-celled green alga) after SD sequences were removed, indicating the existence of SD-independent translational mechanisms. Another study found that ribosomes with a defect 16S-anti-SD sequence could still bind to start codons (Saito et al., 2020). Furthermore, strong SD motifs may fail to outcompete weaker SD motifs by binding to ribosomes strongly and therefore leading to translational arrest (Komarova *et al.*, 2005).

The mRNA secondary structure around the start codon plays a crucial role in translational regulation (Kudla *et al.*, 2009; Chiaruttini and Guillier, 2020). Translation initiation is considered to be a rate-limiting step of translation (Hersch *et al.*, 2014); therefore, when ribosomes are prevented from binding to mRNA, it affects translation rates significantly (Duval *et al.*, 2015; Gualerzi and Pon, 2015). Ribosomes can be prevented from binding to mRNA through strong mRNA secondary structures. It has been shown that mRNA building weaker secondary structures correlate with higher translation rates (de Smit and van Duin, 1994a). Weak secondary

structures occur more frequently in mRNA derived from A/T-rich regions in the genome (Kudla *et al*., 2009; Allert *et al*., 2010), while GC-rich mRNAs may build more stable secondary structures. Thus, 5′ UTRs can possibly be engineered to build weak secondary structures around the start codon to increase translation rates. However, strong GC-rich secondary structures only seem to prevent ribosomes from binding to the mRNA when no SD motif is present (Sterk *et al*., 2018). It is possible that accessible SD motifs counterbalance mRNA secondary structures that would otherwise prevent ribosome binding (de Smit and van Duin, 1994b; Ma *et al*., 2002). Additionally, unstructured regions at the 5′-end of the mRNA can facilitate ribosome binding. Ribosomes can be loaded onto mRNA through non-specific binding of the 30S subunit to a 'standby' site. The ribosome can then relocate from the standby site onto the start codon (de Smit and van Duin, 2003; Sterk *et al*., 2018).

While strong secondary mRNA structures lead to increased mRNA half-life, weak secondary structures lead to short mRNA half-life (Bervoets and Charlier, 2019). It is somewhat contradictory that weaker and therefore short-lived mRNA structures have been shown to correlate with high translation levels. Hence, how translation is regulated at the mRNA level is yet to be fully understood. Overall, many different factors influence translation levels, such as the presence of specific sequences or nucleotides in various positions (context dependency) (Wu *et al*., 2018), the spacing between the motifs and the strength of mRNA secondary structures (Mortimer *et al*., 2014; Chiaruttini and Guillier, 2020).

Coding sequence

The coding sequence (CDS) influences translation and thus protein production rates. This influence has long been thought to stem from codon usage bias. Codon usage bias describes the phenomenon that synonymous codons are not equally used across organisms (Grantham, 1980). Within the CDSs of a genome, some codons occur rarely, and others occur frequently. It has been hypothesized that evolutionary processes caused highly expressed proteins to contain highly abundant codons since the corresponding tRNAs are also highly abundant, allowing fast translation (Ikemura, 1981; Chu *et al*., 2014). It is therefore surprising that we associate the presence of low-efficiency codons at the N-terminal of a protein with increased translation rates (Fredrick and Ibba, 2010; Tuller *et al*., 2010, 2011). Low-efficiency codons are codons that are recognized by rare tRNAs, while high-efficiency codons are recognized by abundant tRNAs (Tuller *et al*., 2010). In all three domains of life, transcripts of highly expressed genes seem to be adapted towards the more efficient codons, except for the first 30–50 codons of CDSs (Tuller *et al*., 2010). The efficiency of a codon correlates well with translation speed of that codon: low-efficiency codons may delay elongation progression due to the low availability of the necessary tRNA (Chiaruttini and Guillier, 2020) and other factors such as the binding strength between the tRNA and the codon (full complementary binding of tRNA to the codon vs. wobble base pairing) (Brule and Grayhack, 2017). Tuller *et al*. (2010) explain why highly abundant proteins are enriched with low-efficiency codons in their N-terminal domain by proposing the concept of a low-efficiency ramp, which decreases translation speed in the first 50 codons of a CDS. Decreased translation speed minimizes ribosome traffic jams and abortive protein synthesis. Thus, a slow elongation rate in the early elongation process may increase protein production.

The transcripts of highly expressed proteins correlate with decreased mRNA secondary structures around the start codon (Bentele *et al*., 2013). Decreased mRNA secondary structures around the start codon are known to play an important role in translation initiation (Kudla *et al*., 2009). This has led many groups to believe that the codons themselves may be less relevant than the mRNA secondary structures. Some groups hypothesize that the codon bias towards low-efficiency codons in the N-terminal domain is driven by the selection for codons that reduce mRNA folding around the start codon (Voges *et al*., 2004; Allert *et al*., 2010; Bentele *et al*., 2013; Osterman *et al*., 2020). An issue with these studies is that they focus on mRNA secondary structures and folding kinetics to explain translation rates and often assume equilibrium conditions for their models, while it would be more realistic to look at RNA folding under non-equilibrium conditions (Espah Borujeni and Salis, 2016). What exactly causes the ramp effect is still uncertain. Tuller *et al*. (2011) and other groups point out that the ramp effect is most likely a combination of adaption of the tRNA pool, amino acid charge, mRNA folding, ribosome interactions with codons and potentially more (Spencer and Barral, 2012; Espah Borujeni *et al*., 2014; Brule and Grayhack, 2017).

Methods used to determine elongation rates impose challenges in form of creating artefacts, which subsequently influences the results and conclusions drawn from the experiments. Elongation rates are commonly studied using ribosome profiling techniques. For ribosome profiling, the ribosomes are arrested while they are attached to the mRNA. The arresting position is then determined by first digesting the mRNA with RNase, a process that removes transcripts except for the part that is protected within the ribosome, usually > 23 nucleotides (Glaub et al., 2019). Then, the protected mRNA pieces are extracted and sequenced to find the position

of ribosome binding. It is assumed that the number of ribosomes per position correlate negatively with translation speed, meaning that high ribosome abundance at a specific codon indicates that the codon is translated slowly. Ribosomal profiling methods are underdeveloped in prokaryotes, and many reported artefacts are introduced through the work flow (Ingolia *et al.*, 2019). For instance, the sampling of bacterial cultures is achieved by filtering. However, intriguingly, the filtering step leads to an artificial translational pause at serine and glycine residues (Mohammad *et al.*, 2019). Another type of artefact is introduced using chloramphenicol to arrest translation. While chloramphenicol is an effective agent to stop translation, it only arrests elongating ribosomes and does not prevent ribosomes from binding and accumulating at the 5′-end of the mRNA. To address the abovementioned challenges, an improved method to study ribosomal arrest in prokaryotes has been proposed by Mohammad *et al.* (2019). Their proposed protocol includes direct freezing of the cultures in liquid nitrogen to arrest translation. Furthermore, they developed an optimized lysis buffer for the extraction step. These measures remove antibiotics and filtration from the protocol, reducing the number of artefacts caused by the procedure. This new protocol also improves the resolution of the ribosome profiling data and thus can contribute to a better understanding of the translation process in bacteria in the future.

Influence of amino acids on protein abundance

A recent study by Verma *et al.* (2019) provides insight into the mechanisms of ribosome interaction with specific amino acids (AAs) in specific positions of the CDS of a green fluorescent protein (GFP). The group replaced all nucleotides for the 3rd, 4th and 5th codon of GFP with an artificial sequence, thus creating a library of constructs that contain all possible nucleotide combinations for the three codons. The group then studied how the presence of specific AAs and codons affected translation. Translation rates were dependent on the occurrence of specific AAs, and independent of tRNA abundance, the chemical properties of the AAs and the codon in the case of synonymous codons. AAs seemed to affect the elongation but not the initiation process. The occurrence of certain AAs led to slowed translation progress, arrest and even dissociation of the translation machinery. Certain AA combinations led to abortion of translation. Whenever an abortive motif was present, for example $T_3V_4G_5$, most ribosomes stopped translating after adding the 3rd and 4th AAs to the growing peptide chain. The group also detected productive motifs that seemed to support the processivity of ribosomes. The effect was found to stem from the AAs themselves and

not from codon usage since the same AAs led to the same result independent of the nucleotide sequence. Thus, the group concluded that translation rates heavily depend on the interaction between the ribosome and the AAs and are affected less by the nucleotide sequence themselves. This study gives new insights into how protein production can be regulated for industrial applications by introducing processivity–improving AAs at the N-terminal domain of proteins. It is unclear whether the findings are universally applicable, since Verma *et al.* used only GFP in their study. However, a recent study by Moreira *et al.* (2019) supports that these findings are universally applicable. Moreira *et al.* found that naturally highly abundant native *E. coli* proteins are enriched with the AA compositions that improved protein production according to Verma *et al.* (2019). The discovery that the presence of specific AAs in specific positions strongly influences protein production contrasts with findings reported in similar studies by Kudla *et al.* (2009) and Cambray *et al.* (2018) who found that mRNA secondary structures influence translation levels more strongly than the presence of specific AAs does. This discrepancy demonstrates that we still lack a complete picture of the subtle nuances that affect transcriptional and translational control. However, it is clear that the CDS considerably impacts gene expression and protein synthesis (Zahurancik *et al.*, 2020).

Context dependency challenge

Standardization of DNA parts is one of the engineering goals of SynBio (Gilman and Love, 2016; Deaner and Alper, 2018; Jin *et al.*, 2019b). The standardized parts have to be characterized to be useful in so-called plug-and-play biology. The reliance on characterization and standardization of DNA parts, which is usually done in model hosts (also known as chassis), is a constraint of the plug-and-play concept. When model organisms are utilized as hosts for protein synthesis, such standardization efforts have shown to be successful (Salis *et al.*, 2009). However, it will be critical for the future of SynBio applications to make use of the biochemical potential of non-model or unique microbes. Context dependency remains a barrier when working with non-model species. First, even in model organisms, informed design of RESs may not be successful due to context dependency. The output that is empirically measured may differ from the output that is projected. Bonde *et al.* (2016) for example, substituted an SD motif with six random nucleotides to see how this affected GFP reporter production levels, and compared experimental results with theoretical estimates from the RBS Calculator. Based on the predicted hybridization energy between the 16S rRNA sequence and the altered SD sequences, the

empirically measured production levels did not match the predicted theoretical output (Bonde *et al*., 2016). The group tried RBS Calculator versions 1.0 and 2.0, and both versions had a mediocre correlation with the empirical measurements ($R^2 = 0.44$, $P < 10^{-13}$ for version 1.0 and $R^2 = 0.54$, $P < 10^{-17}$ for version 2.0), showing that prediction tools may fail to predict outcomes accurately even when model organisms such as *E. coli* are used. Second, our understanding of RES influencing gene expression is primarily based on research with model organisms. Because basic principles differ, it might be difficult to convey this information to other organisms. This is also known as the host-specific context (Cardinale and Arkin, 2012). When applying standardized genetic elements in a novel host-specific context, issues might arise since the standardized genetic elements are normally suited to the model organism and may fail to perform consistently in other species.

Several studies have been conducted to address the aforementioned issues, with the goal of developing strategies that cope better with host and context dependency, as we will discuss in the following subsections.

The issue with context dependency

The context dependency of DNA compositions is a challenge for SynBio applications. The strength of a promoter is determined by the DNA environment that surrounds and connects the core components. When the same RES is utilized with a different CDS, for example, gene expression levels can vary significantly (Mutalik *et al*., 2013b; Zwick *et al*., 2013). Similarly, it is recognized that the DNA sequence outside of the core motifs influences promoter strength (Urtecho *et al*., 2019), which can be used to tune the expression levels (Carr *et al*., 2017). Trans-effects can also play an important role, such as controlling gene expression through antisense RNA (Brophy and Voigt, 2016). Thus, many studies propose that promoter strength depends on the composition-specific context (Cardinale and Arkin, 2012; Leavitt and Alper, 2015; Wu *et al*., 2018; Jin *et al*., 2019a). Despite the importance of transcription and the widespread usage of promoters, context dependency remains an unsolved issue that impedes the standardization of genetic components across chassis. When working with multi-input systems, the context dependency problem intensifies. Experimental design for multi-input systems, such as genetic circuits and metabolic pathways, is still mostly based on time-consuming trial-and-error procedures. Because individual parts are removed from their natural or tested environment and placed in a new genetic context, combining numerous parts frequently produces unexpected consequences, which may be troublesome even when dealing with standardized DNA parts in model hosts (Zong *et al*., 2018; Han *et al*., 2020). When assembling pathways, some genetic elements behave relatively stable, while other parts vary depending on the genetic context (Kosuri *et al*., 2013; Mutalik *et al*., 2013b). It is difficult to predict the quantitative output when combining genetic elements, since the individual effects behave non-additively (Lou *et al*., 2012; Zwick *et al*., 2013). Predicting the appropriate order of individual genes and RESs in a pathway assembly is exceptionally difficult, because of the numerous designs in which a pathway may be assembled and balanced. Research by Smanski *et al*. (2014) looking at gene organization inside the *nif* cluster, which is critical for nitrogen fixation, yielded an unexpected discovery on operon architecture for pathway design. The researchers discovered that the *nif* cluster is organized differently in various bacteria. To investigate how the cluster's specific design affects gene expression levels, the researchers dismantled and reassembled it, enabling genes inside the cluster to assemble in varied orders and directions. The researchers next used RNA-Seq to assess transcript levels and characterize promoter strength. RNA-Seq was done on the cluster's genes and context-free controls of the promoters in isolation. When the promoter strength was averaged across the multiple settings, it was similar to the isolated promoter's promoter strength. When the genetic context of the cluster was modified, however, transcript levels from the same promoter varied significantly and unexpectedly. This was deemed a favourable impact since the combined unexpected context-dependent behaviour altered gene transcription more than context-free transcription levels could. The team was able to produce clusters that recovered up to 57% of wild-type activity, but they were unable to deduce any architectural rationale from the operons of the best functional constructs.

Tackling context dependency: randomization of RESs

Artificial RESs can be used to address context dependency in the absence of rational design tools. The use of DNA consisting of random nucleotides enables the synthesis and identification of functional DNA constructs with optimal DNA sequence composition. Yona *et al*. (2018) employed artificial promoters to drive gene expression of the endogenous *E. coli lac* operon. By exerting selection pressure, the artificial promoters were driven to evolve into functional ones. After only one round of mutagenesis, functional promoters arose, which was connected with a rise in A/T content. Surprisingly, before adding selection pressure, 10% of the artificial sequences already resulted in functional expression. A similar finding was also reported by Urtecho *et al*. (2020). The group measured promoter activity of *E. coli*

genomic promoters using RNA-Seq to determine their corresponding transcript levels. Additionally, they constructed a library of 1000 artificial promoters of 150 nt in length and determined the fraction of artificial promoters that could drive gene expression. About 7% of the artificial sequences led to functional gene expression (Urtecho *et al.*, 2020). This implies that random sequences can easily serve as RESs. This is consistent with another work that employed a stretch of 200 random nucleotides to replace both the promoter and the 5′ UTR to create large random DNA libraries that were tested in several hosts using different reporter genes (Lale *et al.*, 2019). The group identified a range of functional artificial RESs in six Gram-negative and Gram-positive bacterial species (including *Streptomyces albus* and extremophile *Thermus thermophilus*) and in Baker's yeast *Saccharomyces cerevisiae*. This study demonstrates that artificial RESs are suitable to regulate gene expression to a desired level both in model and in non-model organisms. In this study, 40% of the artificial RES library led to gene expression in *E. coli*. Analysis of the functioning artificial RESs revealed sequences that closely resembled known motifs, but other sequences diverged from anything previously described, including sequences that should not have resulted in any gene expression based on our existing understanding (Yona *et al.*, 2018; Lale et al., 2019). Another study reported an unexpectedly high number of binding sites arising from random nucleotides in *Streptococcus pyogenes* (Fontana *et al.*, 2020). A study in yeast even found that a spacer sequence specifically designed to be devoid of any binding motifs could still recruit RNA polymerase (Gertz *et al.*, 2009).

Artificial sequences can also be used as 5′ UTR. For example, Evfratov *et al.* (2017) replaced the 5′ UTR of a fluorescent reporter gene with artificial sequence stretches of 20 to 30 random nucleotides, resulting in a clone library with widely varying fluorescence intensities. The group sorted the clones based on their translation efficiency and reported that 0.03% (20 random nt) and 0.45% (30 random nt) of the clones fell into the category of maximal translation efficiency. The constructs with highest translation efficiency frequently comprised one or more SD-like sequences and only produced weak secondary mRNA structures, whereas inefficiently translated mRNA had secondary structures that overlapped with the RBS. Furthermore, highly translated mRNAs were found to be adenosine-enriched but cytosine-depleted. This latter observation was also reported by Komarova *et al.* (2020), who inserted four random nucleotides within an RBS's eight nucleotide spacer and the gene start of a reporter gene. Both studies by Evfratov and Komarova reported using artificial sequences as 5′ UTR resulted in fluorescence levels varying up to 10^4-fold.

Artificial sequences can also be used to improve model predictions. For example, Höllerer *et al.* created a clone library containing a 17 nt artificial sequence stretch upstream of the CDS and used the resulting data to train several models. The 17 nt artificial sequence stretch replaced the RBS and its flanking regions, resulting in a library of 303 503 RBSs that led to gene expression. To predict RBS strength from the DNA sequences, several models were trained with 248 451 sequences. The models were validated and evaluated with two smaller data sets. The deep learning model Sequence-Activity Prediction In Ensemble of Networks (SAPIEN) was found to be the most accurate ($R^2 = 0.927$) and outperformed other classical linear and non-linear machine learning models. The researchers discovered several characteristics that influence the strength of an RBS; strong secondary structures were associated with weak RBSs, despite the fact that mRNA folding energy was typically not a reliable predictor of RBS strength. Furthermore, SD-like motifs that occurred within appropriate spacing to the CDS had a beneficial impact on RBS strength.

The preceding examples show that artificial sequences may be used to replace the promoter and 5′ UTR. The use of artificial RESs has the benefit of already having the required context built into the regulatory sequences. The artificial method has the disadvantage of having a large sequence space that can never be fully sampled. As a result, the artificial method can only provide a glimpse into a vast range of possibilities and frequently necessitates the use of HT facilities to handle the substantial number of DNA constructs. Despite the scale challenge, artificial RESs have been shown to operate in both model and non-model species, providing a viable alternative to employing standardized genetic components when dealing with composition- and host-specific context dependency.

Tackling context dependency: insulating the mRNA from the genetic context

Another method for dealing with context dependency is to use insulators to remove the effects of genetic context. Insulators are genetic elements that facilitate an insulation effect of different genetic parts from each other. This is usually achieved by splicing mRNA and thus physically separating the RBS from the upstream genetic context (Fig. 3B). It has been shown that the use of insulators effectively removes the impact of the upstream genetic context and hence equalizes expression levels. For example, Qi *et al.* (2012) equalized expression levels by physically separation varying upstream mRNA from the RBS. The researchers produced a combinatorial plasmid library with seven promoters, two reporter genes, two RBSs and semi-artificial

sequences in the 5′ UTR upstream of the RBS. Protein production levels varied greatly among the constructs because the genetic context changed dramatically. When the semi-artificial region was removed post-transcriptionally, protein production levels of clones carrying the constructs were comparable to those carrying the control. Post-transcriptional processing was done with the endoRNase Csy4, which recognizes specific RNA CRISPR sequences and cleaves them (Fig. 3B. ii.). The Csy4 recognition sequence was placed between the RBS and upstream semi-artificial sequences. Thus, processing the mRNA with Csy4 physically separated the artificial sequence from the RBS. Physically separating the two genetic parts equalized protein production levels of the constructs. These findings confirm that the flanking region of the RBSs influences protein production levels and that physically separating the RBS from the upstream flanking region can be a useful strategy to create genetic elements with predictable output. Another type of insulators are ribozymes. Lou *et al*. (2012) used ribozyme-mediated cutting of mRNA to create genetic elements with predictable outcome in genetic circuits (Fig. 3B. i.). Using ribozymes as insulators improved the predictability of genetic circuit output with half of the tested ribozymes. This showcases insulators as a valuable genetic tool to decrease the influence of the genetic context. Ribozymes have been widely employed as standard genetic elements to control gene expression since then (Otto *et al*., 2019; Corrêa *et al*., 2020; Davey and Wilson, 2020; Kuo *et al*., 2020; Park *et al*., 2020; Shao *et al*., 2021) and have been proposed to become a part of a standard cassette design (Nielsen *et al*., 2013; Neves *et al*., 2020). The study by Lou *et al*. (2012), however, also demonstrates that insulators must be verified first, as only half of the tested ribozymes worked reliably. It is challenging to anticipate which elements will act stably and independently inside the composition-specific environment without testing.

Tackling context dependency: standardizing the N-terminal of the coding sequence

Context dependency can also be addressed by changing the N-terminal CDS to alter the secondary structure of the mRNA, since secondary mRNA structures around the start codon influence gene expression levels (Kudla *et al*., 2009). Nørholm *et al*. (2013) reported that they could not express the two difficult-to-express membrane proteins, AraH^{WT} and NarK^{WT} even when the entire CDS was codon-optimized. However, the researchers were able to synthesize the proteins after merely optimizing the N-terminal 21 and 24 codons of AraH and NarK respectively. Additionally, the group enhanced protein production by adding 84 nucleotides to the

N-terminal. These 84 nucleotides had previously been shown to improve GFP expression. Both approaches increased protein production levels, but decreased protein stability. Similarly, Kim *et al*. (2012) used a 79-AA-long leader peptide sequence fused to the N-terminal of four difficult-to-express membrane proteins to increase protein production levels. Changing the nucleotide sequence surrounding the start codon by optimizing the CDS and adding AAs to the N-terminal region alters mRNA secondary structures. As a result, these two methods can be utilized to minimize the mRNA secondary structures in that region.

N-terminal tags are frequently added to the CDS for protein purification, but they can also be employed to decrease the impact of the genetic context. For instance, Zelcbuch *et al*. (2013) insulated an RBS library from the effect of the genetic context by adding the same 5′ UTR and N-terminal His-tag as flanking regions. It is then possible to swap the gene downstream of the His-tag without affecting the protein production levels. As a result, separating flanking regions may be a method of dealing with context dependency. N-terminal tags can also influence protein stability and solubility (Gräslund *et al*., 2008). The His-tag was long assumed to be too short to affect protein structure, function or activity. Several investigations, however, show that this assumption is erroneous (Aslantas and Surmeli, 2019; Esen *et al*., 2019; Meng *et al*., 2020; Singh *et al*., 2020). For example, Booth *et al*. (2018) used differential scanning fluorimetry to study how pH, thermal and salt stability of ten different proteins changed when a His-tag was present or absent. Thermal stability was greatly influenced, with five of the ten proteins exhibiting lower thermostability when the His-tag was present and just one protein exhibiting higher thermostability. Majorek *et al*. (2014) observed that the activity of two enzymes was influenced through His-tags and concluded that His-tags may be causing non-reproducibility issues in biomedical studies.

Tackling context dependency: bicistronic operon design and translational coupling

Bicistronic operon design can be used to influence mRNA secondary structure and corresponding expression levels without adding an N-terminal tag to the gene of interest. The architecture of a bicistronic operon's 5′ UTR and CDS is as follows: first, a 5′ UTR with an RBS (SD1/RBS1); next, a leader peptide (or fore-cistron) containing a second RBS (SD2/RBS2); and last, the gene of interest (Fig. 3A. i.) (Mutalik *et al*., 2013a; Zhao *et al*., 2016; Nieuwkoop *et al*., 2019). In a bicistronic operon, the stop codon of the leader peptide overlaps with the start codon of the gene of interest, producing two distinct gene products. The leader peptide will not be part of the
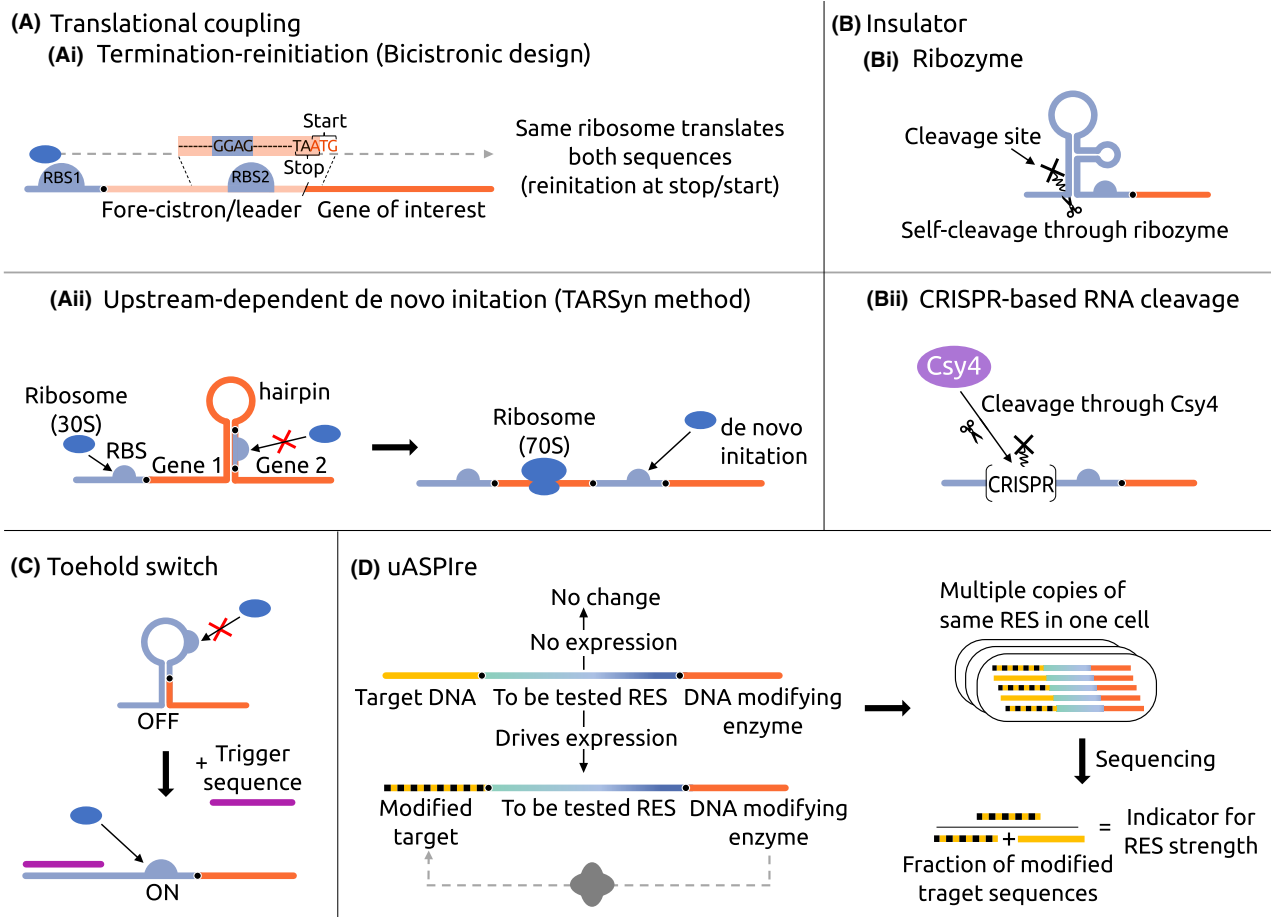
Fig. 3. Overview over methods discussed in the context dependency section. A: Translational coupling has two different mechanisms: Termination–reinitiation (i) and upstream-dependent *de novo* initiation.
A. i: Translational coupling happens when the end of a gene overlaps with the start of the next gene. This can, for example, be achieved by having the stop and the start codon overlap. The ribosome that translated the first gene will continue translating the second gene after terminating translation of the first gene. For the bicistronic design, the first gene is shortened to a leader peptide sequence (or fore-cistron). The leader sequence contains a second RBS close to the stop/start codon overlap to make the ribosome reinitiate. A. ii: Gene 1 and Gene 2 are part of the same operon. The mRNA builds a secondary structure within the second gene or in the intergenic region, which sequesters the RBS of the downstream gene. Once gene 1 is translated, the secondary structure unfolds, and the downstream RBS becomes available for the 30S subunit of a ribosome to bind to. The TARSyN method makes use of this principle by adding hairpins in between two genes in one operon. The hairpin sequesters the RBS of the downstream gene, and the secondary structure unfolds when the upstream gene is translated.
B: Insulators insulate the RBS from upstream genetic context by (i) self-cleavage through a ribozyme or (ii) externally mediated RNA cleavage such as by endoRNase Csy4, which recognizes specific CRISPR sequences in the mRNA.
C: Toehold switches sequester the RBS and start codon within a secondary structure (OFF state), which unfolds when a complementary trigger RNA binds and thus makes the RBS accessible (ON state).
D. uASPIre is a method to test regulatory sequences (RESs) and to record phenotypically if the RES leads to expression. The to-be-tested RES is upstream of a gene, which codes for a DNA-modifying enzyme. If the RES leads to expression, the DNA-modifying enzyme modifies its target. Since the system is encoded on a plasmid, multiple copies of the same plasmid are present within the cell. The plasmids are extracted and sequenced, and the fraction of modified DNA target sequences is determined. This indicates how strongly the DNA-modifying enzyme was expressed. Illustrations of the different figures were based on: A. i., A. ii. (Rennig *et al.*, 2018; Huber *et al.*, 2019), B. i. (Lou *et al.*, 2012), B. ii. (Qi *et al.*, 2012), C. (Valeri *et al.*, 2020), D. (Höllerer *et al.*, 2020).

protein of interest, but its mRNA sequence influences the secondary structure of the mRNA. Thus, the leader peptide can be used to improve gene expression of the downstream gene. The bicistronic operon design makes use of a standard bacterial operon structure, which leads to transcription of polycistronic mRNA. In polycistronic mRNAs, translation of the downstream coding sequence

happens through either independent recruitment of ribosomes (Quax *et al.*, 2013) or translational coupling (Aksoy *et al.*, 1984). Translational coupling describes the effect where the translation of the upstream gene affects translation of the downstream gene of polycistronic mRNAs. Mutalik *et al.* (2013a) report the use of translational coupling in the form of a bicistronic operon to

2304 *L. Tietze and R. Lale*

reduce the influence of flanking regions to reliably improve translation initiation. In the bicistronic operon, translational coupling happens through the 'termination–reinitiation' mode of action due to the overlapping stop and start codon (Fig. 3A. i.) (Huber *et al.*, 2019). It is unclear how exactly termination–reinitiation works mechanistically in bacteria; however, overlapping stop and start codons are prevalent throughout all prokaryotes (Huber *et al.*, 2019). The bicistronic design approach has been used in several studies to modulate gene expression and enable expression of difficult-to-produce proteins such as membrane proteins (Mutalik *et al.*, 2013a; Zhao *et al.*, 2016; Jang *et al.*, 2018; Claassens *et al.*, 2019; Nieuwkoop *et al.*, 2019; Sun *et al.*, 2020b; Martin-Pascual *et al.*, 2021; Torres-Bacete *et al.*, 2021). For instance, Claassens *et al.* (2019) used bicistronic design to constitutively express 'difficult-to-produce' membrane proteins such as AraH, which is the membrane integral part of the arabinose ABC transporter in *E. coli*. Notably, using bicistronic design they could improve the expression levels of AraH compared with the levels previously reached using an optimized inducible system. The previously designed inducible system was derived from a large library of clones that was screened in an HT configuration to identify the best producing clones. In contrast, the improved bicistronic design only required characterization of 22 constructs. Similarly, Sun *et al.* (2020b) improved the expression of six different proteins in *Corynebacterium glutamicum* using the bicistronic design with a small-scale screening. The group recombined the *E. coli tac* promoter with 24 different fore-cistron sequences and found that three constructs led to higher expression than the levels reached with the *tac* promoter on its own. The bicistronic design thus allows to influence the mRNA structure without the addition of a tag or changing the CDS.

Fluorescent reporters are often used in clone library screening to detect expression levels. The use of a fluorescent reporter to determine expression levels causes two issues. First, the reporter gene's CDS is not the same as the gene of interest's CDS. Because the genetic environment changes when the reporter gene is replaced with the gene of interest, expression levels may alter. Second, many clones in clone libraries contain non-functional RESs. Sorting out functioning from non-functioning clones requires time and effort. Translational coupling, such as utilized in the Tunable Antibiotic Resistance Devices Enabling Bacterial Synthetic Evolution and Protein Production (TARSyn) technique, can be employed to solve both difficulties (Rennig *et al.*, 2018). TARSyn is used to eliminate non-functional clones from libraries by translationally linking a gene of an antibiotic resistance marker to the gene of interest. When an antibiotic resistance gene is translationally coupled to another gene, the antibiotic resistance gene is only translated when the upstream gene of interest is translated. Clones with a non-functional RES would express neither the gene of interest nor the antibiotic resistance gene. As a result, clones with a non-functional RES would not survive and would be removed from the library pool prior to the screening stage. By removing non-functional RESs from the library pool, the library size is reduced, making it simpler to manage the quantity of clones and data produced. However, because antibiotic markers have a restricted dynamic range, employing one as a reporter gene makes determining expression intensity problematic. Rennig *et al.* addressed the dynamic range of antibiotic markers by developing RNA hairpin structures with different coupling efficiencies, extending the dynamic range of the antibiotic marker downstream of the hairpin. The RNA hairpin structures contain a sequestered SD sequence, which only unfolds and becomes accessible when the upstream gene is translated (Fig. 3A. ii.). The group demonstrated the feasibility of the TARSyn method by producing a Nanobody and an Affibody, which are products of commercial interest. Thus, TARSyn is a useful method to reduce the library size using translational coupling and can be used to produce proteins of interest.

Challenges of high-throughput screening

Our ability to generate novel RESs outpaces our ability to characterize them, particularly in multi-input systems. Multi-input systems begin with a modest collection of genetic components that can be recombined in any configuration. As a result, given a linear input of beginning pieces, the output scales exponentially, making it challenging to empirically characterizing the output space due to the vast number of alternative structures. As a result, new methodologies that do not rely primarily on experimental characterization of combinatorial libraries are required to address this scalability challenge. A study by Zong *et al.* (2018) proposes two different prediction methods for multi-input expression systems. Both prediction methods by Zong *et al.* rely on experimental data to train a predictive model; however, the amount of experimental data required to train the models increases linearly with the number of inputs, rather than exponentially. This means that a small set of experimental data is enough to train these two predictive models that forecast multi-input expression systems on an exponential scale.

Another issue with HT screening is that matching the output of a RES to its sequence frequently requires two separate experiments. Typically, we assess the fluorescence of a library in one experiment and then sequence the library in a second experiment. The outcomes of the

© 2021 The Authors. *Microbial Biotechnology* published by Society for Applied Microbiology and John Wiley & Sons Ltd., *Microbial Biotechnology*, **14**, 2291–2315

two different procedures are then matched to determine which RES caused which fluorescence. This configuration may introduce mistakes due to gating and cell-to-cell differences (Peterman and Levine, 2016; Boer *et al.*, 2020). To address the problem caused by having two separate experiments, Höllerer *et al.* (2020) proposed a new method to record successful expression within a DNA sequence instead of relying on a reporter output, which is measured in a separate experiment. The principle was termed uASPIre: ultradeep Acquisition of Sequence-Phenotype Interrelations (Fig. 3D). In uAS-PIre, the RES that is tested drives gene expression of an enzyme that modifies DNA. The DNA-modifying enzyme has a distinct DNA target that it modifies upon expression. If the DNA target and the RES-driving expression are both encoded on the same plasmid, sequencing the plasmid identifies the RES while also revealing whether the target sequence was mutated. Each cell has just one kind of RES, which is found in many copies on plasmids. Thus, the strength of expression correlates with the fraction of plasmids containing the modified target sequence: if all target sequences are modified, gene expression is high; if half of the target sequences are modified, gene expression is medium; and if none are modified, the RES does not result in functional gene expression. Thus, uASPIre removes the necessity for two separate experiments to match the output of an RES to its sequence.

Host-specific context

Every protein production host has its own distinct properties, and their usage in SynBio activities necessitates the availability of standardized genetic components that have been proven to operate in the specific host. Relying on pre-characterized components is a constraint for future SynBio applications since it leads to the usage of a few model organisms as hosts. This constraint is particularly obvious in models that predict gene expression from uncharacterized DNA sequences. Because most computational research employs expression data from model organisms to train prediction algorithms, predicting gene expression in other species is difficult.

While predicting gene expression in non-model organisms remains unresolved, the predictive power of the trained algorithms performs well for model organisms. A recent example of such a computational model is the AI-based framework from the Wang group (Wang *et al.*, 2020). The model was trained using natural *E. coli* promoters to predict promoter strength and to design *de novo* promoters. As a follow-up, *de novo* promoters were experimentally tested and 70.8% were in fact functional, including promoters that showed higher activity than the control. Many of the *de novo* promoters showed little

similarity to known promoters; thus, the model expands the knowledge on promoter design and functionality. The 5′ UTR is also a suitable target for predictive models to engineer gene expression. For example, Valeri *et al.* (2020) recently created a deep neural network (DNN) to design toehold switches, which are mRNA secondary structures that can switch translation ON and OFF. In their OFF state, toehold switches build a stem and loop that sequesters the SD motif and the start codon (Fig. 3B). Thus, the stem loop prevents ribosomes from binding and subsequently prevents translation. The secondary structure of mRNA changes when a so-called trigger is present. The trigger is a complementary RNA molecule, which binds to a part of the mRNA that would build the stem loop and hybridizes with the trigger RNA (ON state); the region of the mRNA containing the SD motif and the start codon is no longer hybridized and is therefore accessible to ribosomes, allowing translation to occur. The scarcity of natural toehold switches has long hampered the design and optimization of toehold switches. The few extant toehold switches did not provide a sufficient foundation for informing the design of future toehold switches. Angenent-Mari *et al.* (2020) have constructed an RNA switch collection with 244 000 potential trigger sequences, including 10 000 artificial sequences. About 91 000 of the sequences were measured in their ON/OFF state paired ratios using GFP as a reporter. Valeri *et al.* used this data set to train the aforementioned DNN model to improve toehold switch design and prediction, making toehold switches more accessible for the future implementations into RES design. One disadvantage of this and related models is that the majority, if not all, of the data used to train the computational models comes from a model organism. This scarcity of host-specific information leads to poor model performance in non-model species (Umarov and Solovyev, 2017; Bharanikumar *et al.*, 2018; Gilman *et al.*, 2019). However, even in model organisms, computational models can give erroneous predictions. For example, we do not know how conserved motifs fit into the equation. We presume that separate factors contribute with singular or additive effects in most prediction models (Zrimec *et al.*, 2019). However, it is unclear whether they do. In fact, conserved motifs such as the −10 and −35 region in bacteria might not have additive effects as it long was thought. Recently, Einav *et al.* (2019) improved their computational model that predicts the behaviour of novel promoters by factoring in avidity between the −10 and −35 regions. Avidity describes how the two binding sites influence each other's binding strength. Once the transcription machinery has bound to one of the binding sites, the machinery is more likely and faster to attach to the second binding site. Therefore, the binding sites should not be considered on their

own to determine the overall binding strength. According to Einav *et al.*, promoters may contain weak −10 or −35 regions but still lead to strong expression, as the overall avidity of the binding motifs is high. Furthermore, promoters containing strong binding motifs may express the downstream gene weakly due to overly tight binding of the transcription machinery to the motifs, leading to subsequent arrest of the machinery at the promoter. The improved model predicted these occurrences more accurately. This shows that models improve when we increase our fundamental knowledge about the basic molecular regulatory mechanisms. However, this study also exemplifies that there is still a lot left to discover in this area.

Methods to overcome host-specific context dependency in SynBio

It is critical to create strategies that can be employed in non-model species in addition to enhancing computer models. For example, the *in vivo* genome editing technique multiplex automated genome editing (MAGE) allows editing of a genome at multiple locations at the same time (Wang *et al.*, 2009), but is still limited in several ways: (i) the host range thus far is limited to *E. coli* relatives, (ii) the host genome needs to be modified prior to use of MAGE to express a phage recombinase enzyme necessary for the recombination and to inactivate the native methyl-directed mismatch repair (MMR) system, (iii) off-target effects are prevalent (for more insights into the MAGE technique and recombineering, see the recent review by Wannier *et al.* (2021)). Attempts have been undertaken to alter and expand MERGE to solve some of the difficulties raised above. For example, the DIvERGE approach (Directed Evolution with Random Genomic Mutations), which uses MAGE as a basis, omits the host genome alterations that are normally required to deactivate the MMR system. The MMR system is not inactivated by genome modifications, but rather by plasmid-based production of a dominant-negative mutator allele of the *E. coli* MMR protein MutL, which inhibits DNA repair transiently. DIvERGE also allows for the random alteration of both long and short segments of genomic DNA, making it an excellent tool for random mutagenesis of genomic RES to control gene expression. The method has been shown to work in Enterobacteriaceae such as *E. coli* and the non-model organism *Citrobacter freundii* but has yet to be established in other bacterial species. Another method that allows genome editing at multiple locations simultaneously is the Base Editor-Targeted and Template-free Expression Regulation (BETTER) method (Wang *et al.*, 2021). BETTER uses a catalytically impaired Cas nuclease and a nucleobase deaminase to change bases in

the genome. The method avoids the formation of double-strand breaks, which are damaging to cells and are induced by Cas nucleases. A disadvantage of the approach is that a customized translation or transcription element must be inserted into the genome upstream of the target sequence before BETTER can be utilized. So far, the approach has been demonstrated to function in *C. glutamicum* and *Bacillus subtilis*, where it was used to target the RES of up to 10 genes, successfully altering their gene expression.

Another example for overcoming host dependencies is the work by Johns *et al.* (2018), who mined genomes of 184 organisms and extracted RESs to create universal RESs. The mined RESs were 165 nt long and included everything necessary to drive gene expression. The RES library was placed upstream of a reporter gene and characterized in different microorganisms for their ability to lead to stable gene expression. The group found a spectrum of RESs that were active in multiple species. RESs that are active in multiple hosts could be useful for building microbial communities with unique expression patterns and allowing genetic parts to be reused in multiple hosts.

When working with non-model species, the choice of characterized genetic components is frequently limited. Semi-artificial sequences can be beneficial in such instances since they need less prior information to generate new functioning semi-artificial RESs. For example, Ji *et al.* (2018) used semi-artificial sequences for *Streptomyces albus* to create novel RESs since well-characterized parts for *S. albus* are scarce and re-usage of the same RES within the same pathway can lead to homologous recombination in unwanted positions. The group successfully isolated high producers among the clones containing semi-artificial RES to drive gene expression of indigoidine synthetase, which produces the blue pigment indigoidine. The RESs were analysed to gather insights for future RES design in *S. albus*. In contrast to typically utilized sequences, the spacer between the SD sequence and the start codon of the strongest clone was truncated. This might imply that a shorter spacer is preferable for initiating translation in *S. albus*, although this possibility has not been investigated further in the study. Jang *et al.* (2018) created new strong promoters for usage in *L. citreum* using the semi-artificial approach, while two other studies used semi-artificial sequences to create constitutive and inducible promoters in the non-model organism *Halomonas* sp. (Li *et al.*, 2016; Trisrivirat *et al.*, 2020). In the study by Trisrivirat *et al.* (2020), a *Halomonas* sp. was used to produce propane with the novel semi-artificial promoter that they created. Li *et al.* (2016) used *Halomonas* TD01 core promoter elements to create semi-artificial promoters and characterized them, resulting in a library

containing constructs with transcriptional activity varying over 310-fold. *Halomonas* sp. are of special interest as an emerging biotechnological chassis. Because *Halomonas* sp. is alkaliphilic and halophilic, it is suited for non-sterile fermentation processes, cutting the cost of such procedures. The success of the semi-artificial approach and the insight that it offers into basic molecular mechanism is only hampered by the fact that some prior knowledge is required about core elements of a regulatory sequence.

## Standard cassette and beyond – a guide

SynBio currently relies on standardization and the reuse of well-characterized genetic parts. We provide an overview of a typical workflow to help a new user understand the gene expression engineering process (Fig. 4). Please see the Structural Genomics Consortium's review for a summary study on the fundamentals of protein synthesis and purification (Gräslund *et al.*, 2008).

   A conventional cassette serves as the beginning point for gene expression engineering. For the standard cassettes, various sections may be assembled using a plug-and-play assembly approach and should adhere to the architecture: promoter, 5′ UTR incorporating an insulator and an RBS, bicistronic operon design, CDS, RNase III site, bidirectional terminator (Nielsen *et al.*, 2013; Neves *et al.*, 2020). The promoter, 5′ UTR and CDS have been the focus of this review. Terminators and RNase III sites are two additional genetic elements that can be included in a standard cassette that we have not discussed in this review but have been discussed by others (Engstrom and Pfleger, 2017; Deaner and Alper, 2018). Terminators are nucleotide sequences that mediate release of RNA from the transcription machinery, thus leading to transcription termination. Hence, they are a useful genetic element to isolate a cassette, operon or genetic circuit from the surrounding genetic context by preventing read-through of upstream and downstream regions. Terminators are usually leaky (Deaner and Alper, 2018); thus, additional insulation from genetic context can be useful, for example, by including an RNase III recognition site. RNase III is an endonuclease that cuts double-stranded RNA. For SynBio applications, CDS flanking regions can be included that build double-stranded RNA stem structures, which are then recognized and cut by RNase III (Engstrom and Pfleger, 2017). Different genetic elements can be found in online repositories such as the iGEM registry of standard biological parts (parts.igem.org) that contain promoters, RBS, terminators and many more DNA parts. When employing a model organism, a typical cassette will often yield the necessary outcome with only a few steps and minimum screening work. If the standard cassette fails, or if no standard cassette is available
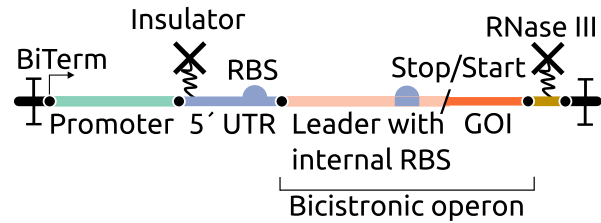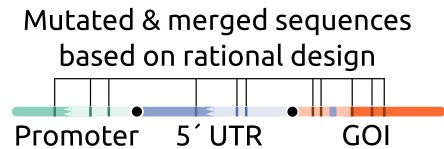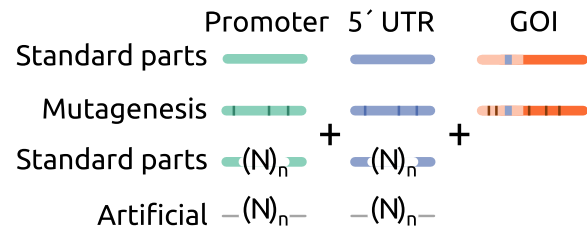

Fig. 4. Guide to gene expression engineering. The starting point (step 1) is a standard cassette with the elements: promoter, 5′ UTR including insulator and RBS, bicistronic operon, CDS, RNase III cutting site, bidirectional terminator. Many characterized parts for a standard cassette can be found in the iGEM registry. If step 1 does not lead to expression of the gene of interest, step 2 or 3 can be followed. Step 2 introduces rational design of standard and natural genetic elements. Sequences are mutated or merged; new sequences can be made through *de novo* motif discovery. Step 3 scales up the experiment to a high-throughput set-up. Here, parts are recombined and randomized. Step 1 is based on the standard cassette design suggested by Neves *et al.* (2020).

for the organism of choice, the user will be forced to experiment with alternative ways. The user can either go to step 2, the engineering technique, or to step 3, which will need the screening of hundreds of clones. Step 2 involves methodically designing and improving standardized and natural genetic elements using tools such as the RBS Calculator (Salis *et al.*, 2009), merging sequences, including point mutations based on the host requirements, or codon optimization of the N-terminal of the CDS to increase expression. These efforts will result in the creation of a small library of clones that can be tested. The rational engineering technique necessitates a thorough grasp of the system to be developed. Step 3 occurs when either steps 1 and 2 fail or when there is insufficient fundamental knowledge about the host of

choice to proceed with the rational design of step 2. In step 3, the construct is exposed to large-scale recombinatorics and randomization. Step 3 can result in library sizes of more than 10^6 clones, necessitating HT screening. Step 3 is essentially a 'brute force' strategy that generates functioning clones by generating a large number of combinations/mutations. This latter strategy should provide functioning clones, but it comes at the expense of having to screen for functioning clones, which may be a time-consuming operation that necessitates the use of specialized tools. When building a cassette, operons or genetic circuits, it is critical to avoid repeated sequences. Repeating DNA sequences are unstable and susceptible to loss by homologous recombination.

Outlook

SynBio has grown substantially. It has already made a significant contribution to how we engineer biology in its brief lifespan (Decoene *et al.*, 2018; El Karoui *et al.*, 2019). It has delivered tools to fine-tune and predict gene expression levels, specifically for model organisms (Kent and Dixon, 2020). Orthogonal systems, which enable very specialized functionalities, may now be integrated into microbial cell factories (Bervoets *et al.*, 2018), and we can design synthetic biological circuits to carry out computations (Voigt, 2006; Nielsen et al., 2016). Genomes can be re-programmed, freeing space for alternative codon usage (Kuo *et al.*, 2018) and we can incorporate xeno nucleotides into DNA (Herdewijn and Marlière, 2009; Pinheiro *et al.*, 2012) opening a completely new avenue of SynBio applications. On the industrial front, several new products are becoming accessible, ranging from fertilizers that replace synthetic fertilizers to haem proteins used in non-meat-based burgers, with yearly sales totalling $2 billion USD (Voigt, 2020).

Predictability and standardization initiatives are critical in the cases given above, as well as many more. SynBio aspires to be able to construct and programme living cells such as computers, and as a result, we anticipate that every component of the circuit will function exactly as planned. However, when we try to break down gene expression regulation and categorize distinct portions and their value, we discover that we are still missing key information. Despite the successes, we require a deeper knowledge of the underlying biological pathways in order to enhance our prediction models. We may also need to reconsider pursuing standardization as the ultimate 'one-size-fits-all' solution. A standardized genetic component or an expression cassette is an excellent place to start (see Fig. 4). Furthermore, utilizing standard cassettes enables researchers to compare expression levels reported in different studies. However, we are in need of

ways that do not rely on conventional components while y*et al*lowing us to tune gene expression to our liking. As described in this review, there are numerous promising methods available that use less rational and more random design techniques, which may be tremendously valuable when the standards fail. Random methods are a good option when working with non-model species since they require less understanding of the underlying gene expression processes.

Funding Information

Conflict of interest

All authors are employees of Jungbunzlauer Ladenburg GmbH or were so in the past.

References

Aksoy, S., Squires, C.L., and Squires, C. (1984) Translational coupling of the *trpB* and *trpA* genes in the *Escherichia coli* tryptophan operon. *J Bacteriol* **157:** 363–367.

Allert, M., Cox, J.C., and Hellinga, H.W. (2010) Multifactorial determinants of protein expression in prokaryotic open reading frames. *J Mol Biol* **402:** 905–918.

Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proc Natl Acad Sci USA* **102:** 12678–12683.

Ameruoso, A., Gambill, L., Liu, B., Villegas Kcam, M.C., and Chappell, J. (2019) Brave new 'RNA' world—advances in RNA tools and their application for understanding and engineering biological systems. *Curr Opin Syst Biol* **14:** 32–40.

Amores, G.R., Guazzaroni, M.-E., and Silva-Rocha, R. (2015) Engineering synthetic cis-regulatory elements for simultaneous recognition of three transcriptional factors in bacteria. *ACS Synth Biol* **4:** 1287–1294.

Angenent-Mari, N.M., Garruss, A.S., Soenksen, L.R., Church, G., and Collins, J.J. (2020) A deep learning approach to programmable RNA switches. *Nat Commun* **11:** 5057.

Aslantas, Y., and Surmeli, N.B. (2019) Effects of N-terminal and C-terminal polyhistidine tag on the stability and function of the thermophilic P450 CYP119. *Bioinorg Chem Appl* **2019:** e8080697.

Bahl, C.P., Wu, R., Stawinsky, J., and Narang, S.A. (1977) Minimal length of the lactose operator sequence for the specific recognition by the lactose repressor. *Proc Natl Acad Sci USA* **74:** 966–970.

Bakke, I., Berg, L., Aune, T.E.V., Brautaset, T., Sletta, H., Tøndervik, A., and Valla, S. (2009) Random mutagenesis of the Pm promoter as a powerful strategy for improvement of recombinant-gene expression. *Appl Environ Microbiol* **75:** 2002–2011.

Balzer Le, S., Onsager, I., Lorentzen, J.A., and Lale, R. (2020) Dual UTR-A novel 5′ untranslated region design

for synthetic biology applications. *Synth Biol* 5. https://doi.org/10.1093/synbio/ysaa006

Barker, M.M., Gaal, T., Josaitis, C.A., and Gourse, R.L. (2001) Mechanism of regulation of transcription initiation by ppGpp. I. Effects of ppGpp on transcription initiation *in vivo* and in vitro. *J Mol Biol* **305:** 673–688.

Belliveau, N.M., Barnes, S.L., Ireland, W.T., Jones, D.L., Sweredoski, M.J., Moradian, A., *et al.* (2018) Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc Natl Acad Sci USA* **115:** E4796–E4805.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Blüthgen, N. (2013) Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol* **9:** 675.

Berg, L., Lale, R., Bakke, I., Burroughs, N., and Valla, S. (2009) The expression of recombinant genes in *Escherichia coli* can be strongly stimulated at the transcript production level by mutating the DNA-region corresponding to the 5′-untranslated part of mRNA. *Microb Biotechnol* **2:** 379–389.

Bervoets, I., and Charlier, D. (2019) Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. *FEMS Microbiol Rev* **43:** 304–339.

Bervoets, I., Van Brempt, M., Van Nerom, K., Van Hove, B., Maertens, J., De Mey, M., and Charlier, D. (2018) A sigma factor toolbox for orthogonal gene expression in *Escherichia coli. Nucleic Acids Res* **46:** 2133–2144.

Bharanikumar, R., Premkumar, K.A.R., and Palaniappan, A. (2018) PromoterPredict: sequence-based modelling of *Escherichia coli* σ70 promoter strength yields logarithmic dependence between promoter strength and sequence. *PeerJ* **6:** e26759v2.

Blazeck, J., and Alper, H.S. (2013) Promoter engineering: recent advances in controlling transcription at the most fundamental level. *Biotechnol J* **8:** 46–58.

de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., and Regev, A. (2020) Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38:** 56–65.

de Boer, H.A., Comstock, L.J., and Vasser, M. (1983) The *tac* promoter: a functional hybrid derived from the *trp* and *lac* promoters. *Proc Natl Acad Sci USA* **80:** 21–25.

Bonde, M.T., Pedersen, M., Klausen, M.S., Jensen, S.I., Wulff, T., Harrison, S., *et al.* (2016) Predictable tuning of protein expression in bacteria. *Nat Methods* **13:** 233–236.

Booth, W.T., Schlachter, C.R., Pote, S., Ussin, N., Mank, N.J., Klapper, V., *et al.* (2018) Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega* **3:** 760–768.

Brophy, J.A., and Voigt, C.A. (2016) Antisense transcription as a tool to tune gene expression. *Mol Syst Biol* **12.**

Brule, C.E., and Grayhack, E.J. (2017) Synonymous codons: choose wisely for expression. *Trends Genet* **33:** 283–297.

Bulyk, M.L., McGuire, A.M., Masuda, N., and Church, G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli. Genome Res* **14:** 201–208.

Cambray, G., Guimaraes, J.C., and Arkin, A.P. (2018) Evaluation of 244,000 synthetic sequences reveals design

principles to optimize translation in *Escherichia coli. Nat Biotechnol* **36:** 1005–1015.

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.

Carr, S.B., Beal, J., and Densmore, D.M. (2017) Reducing DNA context dependence in bacterial promoters. *PLoS One* **12:** e0176013.

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.

Chen, X., Gao, C., Guo, L., Hu, G., Luo, Q., Liu, J., *et al.* (2018a) DCEO biotechnology: tools to design, construct, evaluate, and optimize the metabolic pathway for biosynthesis of chemicals. *Chem Rev* **118:** 4–72.

Chen, Y.e., Ho, J.M.L., Shis, D.L., Gupta, C., Long, J., Wagner, D.S., *et al.* (2018b) Tuning the dynamic range of bacterial promoters regulated by ligand-inducible transcription factors. *Nat Commun* **9:** 1–8.

Chiaruttini, C., and Guillier, M. (2020) On the role of mRNA secondary structure in bacterial translation. *WIREs RNA* **11:** e1579.

Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M.F., and von der Haar, T. (2014) Translation elongation can control translation initiation on eukaryotic mRNAs. *EMBO J* **33:** 21–34.

Claassens, N.J., Finger-Bou, M., Scholten, B., Muis, F., de Groot, J.J., de Gier, J.-W., *et al.* (2019) Bicistronic design-based continuous and high-level membrane protein production in *Escherichia coli. ACS Synth Biol* **8:** 1685–1690.

Corrêa, G.G., Lins, M.R.d.C.R., Silva, B.F., de Paiva, G.B., Zocca, V.F.B., Ribeiro, N.V., *et al.* (2020) A modular autoinduction device for control of gene expression in *Bacillus subtilis. Metab Eng* **61:** 326–334.

Cox, R.S. III, Surette, M.G., and Elowitz, M.B. (2007) Programming gene expression with combinatorial promoters. *Mol Syst Biol* **3:** 145.

Davey, J.A., and Wilson, C.J. (2020) Engineered signal-coupled inducible promoters: measuring the apparent RNA-polymerase resource budget. *Nucleic Acids Res* **48:** 9995–10012.

Deaner, M., and Alper, H.S. (2018) Promoter and terminator discovery and engineering. In *Synthetic Biology – Metabolic Engineering. Advances in Biochemical Engineering/ Biotechnology*. Zhao, H., and Zeng, A.-P. (eds). Cham: Springer International Publishing, pp. 21–44.

Decoene, T., Paepe, B.D., Maertens, J., Coussement, P., Peters, G., Maeseneire, S.L.D., and Mey, M.D. (2018) Standardization in synthetic biology: an engineering discipline coming of age. *Crit Rev Biotechnol* **38:** 647–656.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015) A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **25:** 1268–1280.

Dulin, D., Bauer, D.L.V., Malinen, A.M., Bakermans, J.J.W., Kaller, M., Morichaud, Z., *et al.* (2018) Pausing controls branching between productive and non-productive pathways during initial transcription in bacteria. *Nat Commun* **9:** 1478.

Duval, M., Simonetti, A., Caldelari, I., and Marzi, S. (2015) Multiple ways to regulate translation initiation in bacteria: Mechanisms, regulatory circuits, dynamics. *Biochimie* **114:** 18–29.

Einav, T., and Phillips, R. (2019) How the avidity of polymerase binding to the −35/−10 promoter sites affects gene expression. *Proc Natl Acad Sci USA* **116:** 13340–13345.

El Karoui, M., Hoyos-Flight, M., and Fletcher, L. (2019) Future trends in synthetic biology—A report. *Front Bioeng Biotechnol* **7:** 175.

Engstrom, M.D., and Pfleger, B.F. (2017) Transcription control engineering and applications in synthetic biology. *Synth Syst Biotechnol* **2:** 176–191.

Esen, H., Alpdağtaş, S., Çakar, M.M., and Binay, B. (2019) Tailoring of recombinant FDH: effect of histidine tag location on solubility and catalytic properties of *Chaetomium thermophilum* formate dehydrogenase (CtFDH). *Prep Biochem Biotechnol* **49:** 529–534.

Espah Borujeni, A., 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.

Espah Borujeni, A., 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.

Evfratov, S.A., Osterman, I.A., Komarova, E.S., Pogorelskaya, A.M., Rubtsova, M.P., Zatsepin, T.S., *et al.* (2017) Application of sorting and next generation sequencing to study 5′-UTR influence on translation efficiency in *Escherichia coli. Nucleic Acids Res* **45:** 3487–3502.

Fargo, D.C., Zhang, M., Gillham, N.W., and Boynton, J.E. (1998) Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in *Chlamydomonas reinhardtii* chloroplasts or in *Escherichia coli. Mol Gen Genet MGG* **257:** 271–282.

Firnberg, E., and Ostermeier, M. (2012) PFunkel: efficient, expansive, user-defined mutagenesis. *PLoS One* **7:** e52031.

Fontana, J., Dong, C., Kiattisewee, C., Chavali, V.P., Tickman, B.I., Carothers, J.M., and Zalatan, J.G. (2020) Effective CRISPRa-mediated control of gene expression in bacteria must overcome strict target site requirements. *Nat Commun* **11:** 1–11.

Fredrick, K., and Ibba, M. (2010) How the sequence of a gene can tune its translation. *Cell* **141:** 227–229.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* **457:** 215–218.

Gilman, J., and Love, J. (2016) Synthetic promoter design for new microbial chassis. *Biochem Soc Trans* **44:** 731–737.

Gilman, J., Singleton, C., Tennant, R.K., James, P., Howard, T.P., Lux, T., *et al.* (2019) Rapid, heuristic discovery and design of promoter collections in non-model microbes for industrial applications. *ACS Synth Biol* **8:** 1175–1186.

Glaub, A., Huptas, C., Neuhaus, K., and Ardern, Z. (2019) *Improving Bacterial Ribosome Profiling Data Quality.* bioRxiv 863266.

González-Pérez, M.M., Ramos, J.L., Gallegos, M.-T., and Marqués, S. (1999) Critical nucleotides in the upstream region of the XylS-dependent TOL meta-cleavage pathway operon promoter as deduced from analysis of mutants. *J Biol Chem* **274:** 2286–2290.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342:** 475–479.

Grantham, R. (1980) Working of the genetic code. *Trends Biochem Sci* **5:** 327–331.

Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., Gileadi, O., *et al.* (2008) Protein production and purification. *Nat Methods* **5:** 135–146.

Gualerzi, C.O., and Pon, C.L. (2015) Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cell Mol Life Sci CMLS* **72:** 4341–4367.

Guazzaroni, M.-E., and Silva-Rocha, R. (2014) Expanding the logic of bacterial promoters using engineered overlapping operators for global regulators. *ACS Synth Biol* **3:** 666–675.

Häkkinen, A., Tran, H., Yli-Harja, O., and Ribeiro, A.S. (2013) Effects of rate-limiting steps in transcription initiation on genetic filter motifs. *PLoS One* **8:** e70439.

Hamilton, E.P., and Lee, N. (1988) Three binding sites for AraC protein are required for autoregulation of araC in *Escherichia coli. Proc Natl Acad Sci USA* **85:** 1749–1753.

Hammer, K., Mijakovic, I., and Jensen, P.R. (2006) Synthetic promoter libraries – tuning of gene expression. *Trends Biotechnol* **24:** 53–55.

Han, L., Chen, Q., Lin, Q., Cheng, J., Zhou, L.i., Liu, Z., *et al.* (2020) Realization of robust and precise regulation of gene expression by multiple sigma recognizable artificial promoters. *Front Bioeng Biotechnol* **8:** 92.

Han, L., Cui, W., Suo, F., Miao, S., Hao, W., Chen, Q., *et al.* (2019) Development of a novel strategy for robust synthetic bacterial promoters based on a stepwise evolution targeting the spacer region of the core promoter in *Bacillus subtilis. Microb Cell Factories* **18:** 96.

Haugen, S.P., Berkmen, M.B., Ross, W., Gaal, T., Ward, C., and Gourse, R.L. (2006) rRNA promoter regulation by nonoptimal binding of σ region 1.2: an additional recognition element for RNA polymerase. *Cell* **125:** 1069–1082.

Henderson, K.L., Felth, L.C., Molzahn, C.M., Shkel, I., Wang, S., Chhabra, M., *et al.* (2017) Mechanism of transcription initiation and promoter escape by *E. coli* RNA polymerase. *Proc Natl Acad Sci USA* **114:** E3032–E3040.

Herdewijn, P., and Marlière, P. (2009) Toward safe genetically modified organisms through the chemical diversification of nucleic acids. *Chem Biodivers* **6:** 791–808.

Hersch, S.J., Elgamal, S., Katz, A., Ibba, M., and Navarre, W.W. (2014) Translation initiation rate determines the impact of ribosome stalling on bacterial protein synthesis. *J Biol Chem* **289:** 28160–28171.

Heyduk, E., and Heyduk, T. (2018) DNA template sequence control of bacterial RNA polymerase escape from the promoter. *Nucleic Acids Res* **46:** 4469–4486.

Höllerer, S., Papaxanthos, L., Gumpinger, A.C., Fischer, K., Beisel, C., Borgwardt, K., *et al.* (2020) Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nat Commun* **11:** 3551.

Holmqvist, E., Reimegård, J., and Wagner, E.G.H. (2013) Massive functional mapping of a 5′-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. *Nucleic Acids Res* 41: e122.

Horwitz, M.S., and Loeb, L.A. (1986) Promoters selected from random DNA sequences. *Proc Natl Acad Sci USA* 83: 7405–7409.

Huang, X., Zhang, X.-E., Zhou, Y.-F., Zhang, Z.-P., and Cass, A.E.G. (2006) Directed evolution of the 5′-untranslated region of the phoA gene in *Escherichia coli* simultaneously yields a stronger promoter and a stronger Shine-Dalgarno sequence. *Biotechnol J* 1: 1275–1282.

Huber, M., Faure, G., Laass, S., Kolbe, E., Seitz, K., Wehrheim, C., *et al.* (2019) Translational coupling via termination-reinitiation in archaea and bacteria. *Nat Commun* 10: 4006.

Hwang, H.J., Lee, S.Y., and Lee, P.C. (2018) Engineering and application of synthetic nar promoter for fine-tuning the expression of metabolic pathway genes in *Escherichia coli*. *Biotechnol Biofuels* 11: 103.

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 151: 389–409.

Imashimizu, M., Tokunaga, Y., Afek, A., Takahashi, H., Shimamoto, N., and Lukatsky, D.B. (2020) Control of transcription initiation by biased thermal fluctuations on repetitive genomic sequences. *Biomolecules* 10: 1299.

Ingolia, N.T., Hussmann, J.A., and Weissman, J.S. (2019) Ribosome profiling: global views of translation. *Cold Spring Harb Perspect Biol* 11: a032698.

Isaacs, F.J., Dwyer, D.J., Ding, C., Pervouchine, D.D., Cantor, C.R., and Collins, J.J. (2004) Engineered riboregulators enable post-transcriptional control of gene expression. *Nat Biotechnol* 22: 841–847.

Ishihama, A. (2010) Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol Rev* 34: 628–645.

Ishihama, A., Shimada, T., and Yamazaki, Y. (2016) Transcription profile of *Escherichia coli*: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res* 44: 2058–2074.

Jang, S.H., Cha, J.W., Han, N.S., and Jeong, K.J. (2018) Development of bicistronic expression system for the enhanced and reliable production of recombinant proteins in *Leuconostoc citreum*. *Sci Rep* 8: 8852.

Jensen, P.R., and Hammer, K. (1998) The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Appl Environ Microbiol* 64: 82–87.

Ji, C.-H., Kim, J.-P., and Kang, H.-S. (2018) Library of synthetic streptomyces regulatory sequences for use in promoter engineering of natural product biosynthetic gene clusters. *ACS Synth Biol* 7: 1946–1955.

Jiao, S., Li, X.u, Yu, H., Yang, H., Li, X., and Shen, Z. (2017) In situ enhancement of surfactin biosynthesis in *Bacillus subtilis* using novel artificial inducible promoters. *Biotechnol Bioeng* 114: 832–842.

Jin, L., Nawab, S., Xia, M., Ma, X., and Huo, Y.-X. (2019a) Context-dependency of synthetic minimal promoters in driving gene expression: a case study. *Microb Biotechnol* 12: 1476–1486.

Jin, L.-Q., Jin, W.-R., Ma, Z.-C., Shen, Q., Cai, X., Liu, Z.-Q., and Zheng, Y.-G. (2019b) Promoter engineering strategies for the overproduction of valuable metabolites in microbes. *Appl Microbiol Biotechnol* 103: 8725–8736.

Johns, N.I., Gomes, A.L.C., Yim, S.S., Yang, A., Blazejewski, T., Smillie, C.S., *et al.* (2018) Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nat Methods* 15: 323–329.

Josaitis, C.A., Gaal, T., and Gourse, R.L. (1995) Stringent control and growth-rate-dependent control have nonidentical promoter sequence requirements. *Proc Natl Acad Sci USA* 92: 1117–1121.

Kent, R., and Dixon, N. (2020) Contemporary tools for regulating gene expression in bacteria. *Trends Biotechnol* 38: 316–333.

Kim, H.S., Ernst, J.A., Brown, C., Bostrom, J., Fuh, G., Lee, C.V., *et al.* (2012) Translation levels control multi-spanning membrane protein expression. *PLoS One* 7: e35844.

Kinkhabwala, A., and Guet, C.C. (2008) Uncovering cis regulatory codes using synthetic promoter shuffling. *PLoS One* 3: e2030.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA* 107: 9158–9163.

Kittleson, J.T., Wu, G.C., and Anderson, J.C. (2012) Successes and failures in modular genetic engineering. *Curr Opin Chem Biol* 16: 329–336.

Knight, T. (2003) *Idempotent Vector Design for Standard Assembly of Biobricks*.

Komarova, A.V., Tchufistova, L.S., Dreyfus, M., and Boni, I.V. (2005) AU-Rich sequences within 5′ untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J Bacteriol* 187: 1344–1349.

Komarova, E.S., Chervontseva, Z.S., Osterman, I.A., Evfratov, S.A., Rubtsova, M.P., Zatsepin, T.S., *et al.* (2020) Influence of the spacer region between the Shine-Dalgarno box and the start codon for fine-tuning of the translation efficiency in *Escherichia coli*. *Microb Biotechnol* 13: 1254–1261.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., *et al.* (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci* 110: 14024–14029.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324: 255–258.

Kuo, J., Stirling, F., Lau, Y.H., Shulgina, Y., Way, J.C., and Silver, P.A. (2018) Synthetic Genome Recoding: new genetic codes for new features. *Curr Genet* 64: 327–333.

Kuo, J., Yuan, R., Sánchez, C., Paulsson, J., and Silver, P.A. (2020) Toward a translationally independent RNA-based synthetic oscillator using deactivated CRISPR-Cas. *Nucleic Acids Res* 48: 8165–8177.

Lale, R., Tietze, L., Nesje, J., Onsager, I., Engelhardt, K., Wong, C.F.A., *et al.* (2019) A universal method for gene expression engineering. *bioRxiv* 644989.

Leavitt, J.M., and Alper, H.S. (2015) Advances and current limitations in transcript-level control of gene expression. *Curr Opin Biotechnol* **34:** 98–104.

Lee, D.J., Minchin, S.D., and Busby, S.J.W. (2012) Activating transcription in bacteria. *Annu Rev Microbiol* **66:** 125–152.

Leiby, N., Hossain, A., and Salis, H. M. (2020) Convolutional neural net learns promoter sequence features driving transcription strength. In *Proceedings of the 12th International Conference on Bioinformatics and Computational Biology*. San Francisco, CA, pp. 163–172.

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A.C., Kalma, Y., Lotam-Pompan, M., *et al.* (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res* **25:** 1018–1029.

Li, T., Li, T., Ji, W., Wang, Q., Zhang, H., Chen, G.-Q., *et al.* (2016) Engineering of core promoter regions enables the construction of constitutive and inducible promoters in *Halomonas* sp. *Biotechnol J* **11:** 219–227.

Li, X.-Y., and Eisen, M.B. (2018) *Mutation of sequences flanking and separating transcription factor binding sites in a Drosophila enhancer significantly alter its output.* bioRxiv 379974.

Liu, Y., Beyer, A., and Aebersold, R. (2016) On the dependency of cellular protein levels on mRNA abundance. *Cell* **165:** 535–550.

Lou, C., Stanton, B., Chen, Y.-J., Munsky, B., and Voigt, C.A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol* **30:** 1137–1142.

Ma, J., Campbell, A., and Karlin, S. (2002) Correlations between shine-dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* **184:** 5733–5745.

Majorek, K.A., Kuhn, M.L., Chruszcz, M., Anderson, W.F., and Minor, W. (2014) Double trouble-Buffer selection and His-tag presence may be responsible for nonreproducibility of biomedical experiments. *Protein Sci Publ Protein Soc* **23:** 1359–1368.

Martin-Pascual, M., Batianis, C., Bruinsma, L., Asin-Garcia, E., Garcia-Morales, L., Weusthuis, R.A., *et al.* (2021) A navigation guide of synthetic biology tools for *Pseudomonas putida*. *Biotechnol Adv* **49:** 107732.

Mazumder, A., and Kapanidis, A.N. (2019) Recent advances in understanding σ70-dependent transcription initiation mechanisms. *J Mol Biol* **431:** 3947–3959.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L.i., Rogov, P., *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30:** 271–277.

Meng, L., Liu, Y., Yin, X., Zhou, H., Wu, J., Wu, M., and Yang, L. (2020) Effects of his-tag on catalytic activity and enantioselectivity of recombinant transaminases. *Appl Biochem Biotechnol* **190:** 880–895.

Min, K.T., Kim, M.H., and Lee, D.S. (1988) Search for the optimal sequence of the ribosome binding site by random oligonucleotide-directed mutagenesis. *Nucleic Acids Res* **16:** 5075–5088.

Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M.J., Elofsson, A., *et al.* (2015) Enhanced protein production in *Escherichia coli* by optimization of cloning scars at the vector-coding sequence junction. *ACS Synth Biol* **4:** 959–965.

Mohammad, F., Green, R., and Buskirk, A.R. (2019) A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *eLife* **8:** e42591.

Monteiro, L.M.O., Arruda, L.M., and Silva-Rocha, R. (2018) Emergent properties in complex synthetic bacterial promoters. *ACS Synth Biol* **7:** 602–612.

Monteiro, L.M.O., Sanches-Medeiros, A., Westmann, C.A., and Silva-Rocha, R. (2019) *Modulating Fis and IHF binding specificity, crosstalk and regulatory logic through the engineering of complex promoters.* bioRxiv 614396.

Moreira, M.H., Barros, G.C., Requião, R.D., Rossetto, S., Domitrovic, T., and Palhano, F.L. (2019) From reporters to endogenous genes: the impact of the first five codons on translation efficiency in *Escherichia coli*. *RNA Biol* **16:** 1806–1816.

Mortimer, S.A., Kidwell, M.A., and Doudna, J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15:** 469–479.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.-A., *et al.* (2013a) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods* **10:** 354–360.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Mai, Q.-A., Christoffersen, M.J., Martin, L., *et al.* (2013b) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat Methods* **10:** 347–353.

Na, D., and Lee, D. (2010) RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics* **26:** 2633–2634.

Nair, T.M., and Kulkarni, B.D. (1994) On the consensus structure within the *E. coli* promoters. *Biophys Chem* **48:** 383–393.

Nakagawa, S., Niimura, Y., and Gojobori, T. (2017) Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes. *Nucleic Acids Res* **45:** 3922–3931.

Neves, D., Vos, S., Blank, L.M., and Ebert, B.E. (2020) *Pseudomonas* mRNA 2.0: boosting gene expression through enhanced mRNA stability and translational efficiency. *Front Bioeng Biotechnol* **7:** 458.

Nielsen, A.A.K., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., *et al.* (2016) Genetic circuit design automation. *Science* **352:** aac7341.

Nielsen, A.A.K., Segall-Shapiro, T.H., and Voigt, C.A. (2013) Advances in genetic circuit design: novel biochemistries, deep part mining, and precision gene expression. *Curr Opin Chem Biol* **17:** 878–892.

Nieuwkoop, T., Claassens, N.J., and van der Oost, J. (2019) Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. *Microb Biotechnol* **12:** 173–179.

Nijs, Y.D., Maeseneire, S.L.D., and Soetaert, W.K. (2020) 5′ untranslated regions: the next regulatory sequence in yeast synthetic biology. *Biol Rev* **95:** 517–529.

Nørholm, M.H.H., Toddo, S., Virkki, M.T.I., Light, S., von Heijne, G., and Daley, D.O. (2013) Improved production of membrane proteins in *Escherichia coli* by selective codon substitutions. *FEBS Lett* **587:** 2352–2358.

Oesterle, S., Gerngross, D., Schmitt, S., Roberts, T.M., and Panke, S. (2017) Efficient engineering of chromosomal ribosome binding site libraries in mismatch repair proficient *Escherichia coli*. *Sci Rep* **7:** 1–10.

Osterman, I.A., Chervontseva, Z.S., Evfratov, S.A., Sorokina, A.V., Rodin, V.A., Rubtsova, M.P., *et al*. (2020) Translation at first sight: the influence of leading codons. *Nucleic Acids Res* **48:** 6931–6942.

Otto, M., Wynands, B., Drepper, T., Jaeger, K.-E., Thies, S., Loeschcke, A., *et al*. (2019) Targeting 16S rDNA for stable recombinant gene expression in *Pseudomonas*. *ACS Synth Biol* **8:** 1901–1912.

Park, Y., Espah Borujeni, A., Gorochowski, T.E., Shin, J., and Voigt, C.A. (2020) Precision design of stable genetic circuits carried in highly-insulated *E. coli* genomic landing pads. *Mol Syst Biol* **16:** e9584.

parts.igem.org. WWW http://parts.igem.org/Main_Page

Pemberton, I.K., Muskhelishvili, G., Travers, A.A., and Buckle, M. (2000) The G+C-rich discriminator region of the tyrT promoter antagonises the formation of stable preinitiation complexes. *J Mol Biol* **299:** 859–864.

Peterman, N., and Levine, E. (2016) Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genom* **17:** 206.

Pfleger, B.F., Pitera, D.J., Smolke, C.D., and Keasling, J.D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat Biotechnol* **24:** 1027–1032.

Pinheiro, V.b., Taylor, A.i., Cozens, C., Abramov, M., Renders, M., Zhang, S., *et al*. (2012) Synthetic genetic polymers capable of heredity and evolution. *Science* **336:** 341–344.

Püllmann, P., Ulpinnis, C., Marillonnet, S., Gruetzner, R., Neumann, S., and Weissenborn, M.J. (2019) Golden Mutagenesis: an efficient multi-site-saturation mutagenesis approach by Golden Gate cloning with automated primer design. *Sci Rep* **9:** 1–11.

Qi, L., Haurwitz, R.E., Shao, W., Doudna, J.A., and Arkin, A.P. (2012) RNA processing enables predictable programming of gene expression. *Nat Biotechnol* **30:** 1002–1006.

Quax, T.E.F., Wolf, Y.I., Koehorst, J.J., Wurtzel, O., van der Oost, R., Ran, W., *et al*. (2013) Differential translation tunes uneven production of operon-encoded proteins. *Cell Rep* **4:** 938–944.

Rennig, M., Martinez, V., Mirzadeh, K., Dunas, F., Röjsäter, B., Daley, D.O., and Nørholm, M.H.H. (2018) TARSyn: tunable antibiotic resistance devices enabling bacterial synthetic evolution and protein production. *ACS Synth Biol* **7:** 432–442.

Reppas, N.B., Wade, J.T., Church, G.M., and Struhl, K. (2006) The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol Cell* **24:** 747–757.

Rhodius, V.A., Segall-Shapiro, T.H., Sharon, B.D., Ghodasara, A., Orlova, E., Tabakh, H., *et al*. (2013) Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Mol Syst Biol* **9:** 702.

Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B., and Titgemeyer, F. (2004) Extending the classification of bacterial transcription factors beyond the helix–turn–helix motif as an alternative approach to discover new cis/trans relationships. *Nucleic Acids Res* **32:** 3418–3426.

Rohlhill, J., Sandoval, N.R., and Papoutsakis, E.T. (2017) Sort-seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated *Escherichia coli* growth on methanol. *ACS Synth Biol* **6:** 1584–1595.

Sadler, J.R., Sasmor, H., and Betz, J.L. (1983) A perfectly symmetric lac operator binds the *lac* repressor very tightly. *Proc Natl Acad Sci* **80:** 6785–6789.

Saito, K., Green, R., and Buskirk, A.R. (2020) Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *eLife* **9: e55002.**

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27:** 946–950.

Seo, S.W., Yang, J.-S., Kim, I., Yang, J., Min, B.E., Kim, S., and Jung, G.Y. (2013) Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab Eng* **15:** 67–74.

Shao, B., Rammohan, J., Anderson, D.A., Alperovich, N., Ross, D., and Voigt, C.A. (2021) Single-cell measurement of plasmid copy number and promoter activity. *Nat Commun* **12:** 1475.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., *et al*. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30:** 521–530.

Shi, F., Luan, M., and Li, Y. (2018) Ribosomal binding site sequences and promoters for expressing glutamate decarboxylase and producing γ-aminobyturate in *Corynebacterium glutamicum*. *AMB Express* **8:** 61.

Shine, J., and Dalgarno, L. (1974) The 3′-Terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA* **71:** 1342–1346.

Singh, M., Sori, H., Ahuja, R., Meena, J., Sehgal, D., and Panda, A.K. (2020) Effect of N-terminal poly histidine-tag on immunogenicity of Streptococcus pneumoniae surface protein SP0845. *Int J Biol Macromol* **163:** 1240–1248.

Smanski, M.J., Bhatia, S., Zhao, D., Park, Y.J., B A Woodruff, L., Giannoukos, G., *et al*. (2014) Functional optimization of gene clusters by combinatorial design and assembly. *Nat Biotechnol* **32:** 1241–1249.

de Smit, M.H., and van Duin, J. (1994a) Control of translation by mRNA secondary structure in *Escherichia coli*: a quantitative analysis of literature data. *J Mol Biol* **244:** 144–150.

de Smit, M.H., and van Duin, J. (1994b) Translational initiation on structured messengers: another role for the Shine-Dalgarno interaction. *J Mol Biol* **235:** 173–184.

de Smit, M.H., and van Duin, J. (2003) Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J Mol Biol* **331:** 737–743.

Sosa, R.P., Florez-Ariza, A.J., Díaz-Celis, C., Onoa, B., Cassago, A., de Oliveira, P.S.L., *et al.* (2020) *Interactions of Upstream and Downstream Promoter Regions with RNA Polymerase are Energetically Coupled and a Target of Regulation in Transcription Initiation. bioRxiv* 2020.05.13.070375.

Spencer, P.S., and Barral, J.M. (2012) Genetic code redundancy and its influence on the encoded polypeptides. *Comput Struct Biotechnol J* **1:** e201204006.

Steitz, J.A., and Jakes, K. (1975) How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli. Proc Natl Acad Sci USA* **72:** 4734–4738.

Sterk, M., Romilly, C., and Wagner, E.G.H. (2018) Unstructured 5′-tails act through ribosome standby to override inhibitory structure at ribosome binding sites. *Nucleic Acids Res* **46:** 4188–4199.

Sun, H., Yang, J., and Song, H. (2020a) Engineering mycobacteria artificial promoters and ribosomal binding sites for enhanced sterol production. *Biochem Eng J* **162:** 107739.

Sun, M., Gao, X., Zhao, Z., Li, A.n., Wang, Y., Yang, Y., *et al.* (2020b) Enhanced production of recombinant proteins in *Corynebacterium glutamicum* by constructing a bicistronic gene expression system. *Microb Cell Factories* **19:** 113.

The MISO Sequence Ontology Browser. *Shine-Dalgarno Sequence.* WWW http://www.sequenceontology.org/browser/current_release/term/SO:0000552

Torres-Bacete, J., García, J.L., and Nogales, L. (2021) A portable library of phosphate-depletion based synthetic promoters for customable and automata control of gene expression in bacteria. *Microb Biotechnol*: 1–16. https://doi.org/10.1111/1751-7915.13808

Trisrivirat, D., Hughes, J.M.X., Hoeven, R., Faulkner, M., Toogood, H., Chaiyen, P., and Scrutton, N.S. (2020) Promoter engineering for microbial bio-alkane gas production. *Synth Biol*;**5:ysaa022.**

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., *et al.* (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141:** 344–354.

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., and Ziv-Ukelson, M. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol* **12:** R110.

Tuller, T., and Zur, H. (2015) Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res* **43:** 13–28.

Umarov, R.K., and Solovyev, V.V. (2017) Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* **12:** e0171410.

Urtecho, G., Insigne, K.D., Tripp, A.D., Brinck, M., Lubock, N.B., Kim, H., *et al.* (2020) *Genome-wide Functional Characterization of Escherichia coli* Promoters and Regulatory Elements Responsible for their Function. *bioRxiv* 2020.01.04.894907.

Urtecho, G., Tripp, A.D., Insigne, K.D., Kim, H., and Kosuri, S. (2019) Systematic dissection of sequence elements controlling σ70 promoters using a genomically-encoded

multiplexed reporter assay in *E. coli. Biochemistry* **58:** 1539–1551.

Valeri, J.A., Collins, K.M., Ramesh, P., Alcantar, M.A., Lepe, B.A., Lu, T.K., and Camacho, D.M. (2020) Sequence-to-function deep learning frameworks for engineered riboregulators. *Nat Commun* **11:** 5058.

Verma, M., Choi, J., Cottrell, K.A., Lavagnino, Z., Thomas, E.N., Pavlovic-Djuranovic, S., *et al.* (2019) A short translational ramp determines the efficiency of protein synthesis. *Nat Commun* **10:** 1–15. https://doi.org/10.1038/s41467-019-13810-1

Voges, D., Watzele, M., Nemetz, C., Wizemann, S., and Buchberger, B. (2004) Analyzing and enhancing mRNA translational efficiency in an *Escherichia coli* in vitro expression system. *Biochem Biophys Res Commun* **318:** 601–614.

Voigt, C.A. (2006) Genetic parts to program bacteria. *Curr Opin Biotechnol* **17:** 548–557.

Voigt, C.A. (2020) Synthetic biology 2020–2030: six commercially-available products that are changing our world. *Nat Commun* **11:** 6379.

Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460:** 894–898.

Wang, Y.u., Cheng, H., Liu, Y., Liu, Y.e., Wen, X., Zhang, K., *et al.* (2021) In-situ generation of large numbers of genetic combinations for metabolic reprogramming via CRISPR-guided base editing. *Nat Commun* **12:** 678.

Wang, Y., Liu, Q., Weng, H., Shi, Y., Chen, J., Du, G., and Kang, Z. (2019a) Construction of synthetic promoters by assembling the sigma factor binding −35 and −10 boxes. *Biotechnol J* **14:** 1800298.

Wang, Y., Shi, Y., Hu, L., Du, G., Chen, J., and Kang, Z. (2019b) Engineering strong and stress-responsive promoters in *Bacillus subtilis* by interlocking sigma factor binding motifs. *Synth Syst Biotechnol* **4:** 197–203.

Wang, Y., Wang, H., Wei, L., Li, S., Liu, L., and Wang, X. (2020) Synthetic promoter design in *Escherichia coli* based on a deep generative network. *Nucleic Acids Res* **48:** 6403–6412.

Wannier, T.M., Ciaccia, P.N., Ellington, A.D., Filsinger, G.T., Isaacs, F.J., Javanmardi, K., *et al.* (2021) Recombineering and MAGE. *Nat Rev Methods Primer* **1:** 1–24.

Wertman, K.F., and Mount, D.W. (1985) Nucleotide sequence binding specificity of the LexA repressor of *Escherichia coli* K-12. *J Bacteriol* **163:** 376–384.

Whitehorn, E.A., Livak, K.J., and Petteway, S.R. (1985) The effects of hybrid ribosome-binding-site variants on the expression of human Interferon-β in *Escherichia coli. Gene* **36:** 375–379.

Winkelman, J.T., Vvedenskaya, I.O., Zhang, Y., Zhang, Y.u., Bird, J.G., Taylor, D.M., *et al.* (2016) Multiplexed protein-DNA cross-linking: scrunching in transcription start site selection. *Science* **351:** 1090–1093.

Winther-Larsen, H.C., Blatny, J.M., Valand, B., Brautaset, T., and Valla, S. (2000) Pm promoter expression mutants and their use in broad-host-range RK2 PLASMID VECTORS. *Metab Eng* **2:** 92–103.

Wrenbeck, E.E., Klesmith, J.R., Stapleton, J.A., Adeniran, A., Tyo, K.E.J., and Whitehead, T.A. (2016) Plasmid-based one-pot saturation mutagenesis. *Nat Methods* **13:** 928–930.

Wu, F., Zhang, Q., and Wang, X. (2018) Design of adjacent transcriptional regions to tune gene expression and facilitate circuit construction. *Cell Syst* **6:** 206–215.e6.

Wu, M.-R., Nissim, L., Stupp, D., Pery, E., Binder-Nissim, A., Weisinger, K., *et al.* (2019) A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). *Nat Commun* **10:** 1–10.

Yona, A.H., Alm, E.J., and Gore, J. (2018) Random sequences rapidly evolve into *de novo* promoters. *Nat Commun* **9:** 1–10.

Zaccolo, M., Williams, D.M., Brown, D.M., and Gherardi, E. (1996) An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J Mol Biol* **255:** 589–603.

Zahurancik, W.J., Szkoda, B.E., Lai, L.B., and Gopalan, V. (2020) Ramping recombinant protein expression in bacteria. *Biochemistry* **59:** 2122–2124.

Zelcbuch, L., Antonovsky, N., Bar-Even, A., Levin-Karp, A., Barenholz, U., Dayagi, M., *et al.* (2013) Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Res* **41:** e98.

Zhang, B., Zhou, N., Liu, Y.-M., Liu, C., Lou, C.-B., Jiang, C.-Y., and Liu, S.-J. (2015) Ribosome binding site libraries and pathway modules for shikimic acid synthesis with *Corynebacterium glutamicum*. *Microb Cell Factories* **14:** 71.

Zhang, F., Carothers, J.M., and Keasling, J.D. (2012) Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat Biotechnol* **30:** 354–359.

Zhao, Z., Liu, X., Zhang, W., Yang, Y., Dai, X., and Bai, Z. (2016) Construction of genetic parts from the *Corynebacterium glutamicum* genome with high expression activities. *Biotechnol Lett* **38:** 2119–2126.

Zhelyabovskaya, O.B., Berlin, Y.A., and Birikh, K.R. (2004) Artificial genetic selection for an efficient translation initiation site for expression of human RACK1 gene in *Escherichia coli*. *Nucleic Acids Res* **32:** e52.

Zheng, X., Hu, G.-Q., She, Z.-S., and Zhu, H. (2011) Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genom* **12:** 361.

Zong, D.M., Cinar, S., Shis, D.L., Josić, K., Ott, W., and Bennett, M.R. (2018) Predicting transcriptional output of synthetic multi-input promoters. *ACS Synth Biol* **7:** 1834–1843.

Zrimec, J., Buric, F., Muhammad, A.S., Chen, R., Verendel, V., Töpel, M., and Zelezniak, A. (2019) *Gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure*. *bioRxiv* 792531.

Zwick, F., Lale, R., and Valla, S. (2013) Combinatorial engineering for heterologous gene expression. *Bioengineered* **4:** 431–434.