

# Programmable cross-ribosome-binding sites to fine-tune the dynamic range of transcription factor-based biosensor

Nana Ding<sup>1,2</sup>, Zhenqi Yuan<sup>3,4</sup>, Xiaojuan Zhang<sup>1,2</sup>, Jing Chen<sup>3,4</sup>, Shenghu Zhou<sup>1,2,\*</sup> and Yu Deng<sup>1,2,\*</sup>

<sup>1</sup>National Engineering Laboratory for Cereal Fermentation Technology (NELCF), Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214122, People's Republic of China, <sup>2</sup>Jiangsu Provincial Research Center for Bioactive Product Processing Technology, Jiangnan University, People's Republic of China, <sup>3</sup>School of Internet of Things Engineering, Jiangnan University, 1800 Lihu Road, Wuxi, Jiangsu 214122, People's Republic of China and <sup>4</sup>Engineering Research Center of Internet of Things Technology Applications, Ministry of Education, Wuxi 214122, People's Republic of China

Received April 08, 2020; Revised August 20, 2020; Editorial Decision September 07, 2020; Accepted September 09, 2020

## ABSTRACT

Currently, predictive translation tuning of regulatory elements to the desired output of transcription factor (TF)-based biosensors remains a challenge. The gene expression of a biosensor system must exhibit appropriate translation intensity, which is controlled by the ribosome-binding site (RBS), to achieve fine-tuning of its dynamic range (i.e. fold change in gene expression between the presence and absence of inducer) by adjusting the translation level of the TF and reporter. However, existing TF-based biosensors generally suffer from unpredictable dynamic range. Here, we elucidated the connections and partial mechanisms between RBS, translation level, protein folding and dynamic range, and presented a design platform that predictably tuned the dynamic range of biosensors based on deep learning of large datasets cross-RBSs (cRBSs). In doing so, a library containing 7053 designed cRBSs was divided into five sub-libraries through fluorescence-activated cell sorting to establish a classification model based on convolutional neural network in deep learning. Finally, the present work exhibited a powerful platform to enable predictable translation tuning of RBS to the dynamic range of biosensors.

## INTRODUCTION

Biosensors have gained major attention in the field of biotechnology (1) especially for monitoring metabolite formation (2,3). The existing biosensors included RNA- and protein-based biosensors (4). Wherein, RNA aptamers were

used as building blocks for designing small molecule RNA-based biosensors (5). And we focused on the protein-based biosensors in this work. Genetically encoded biosensors derived from small-molecule inducer responsive transcription factors that produce fluorescence intensity proportional to the target metabolite concentration in the detection range have attracted substantial research attention (3,6). However, the existing genetically encoded biosensors generally have the drawback of inappropriate dynamic range (i.e. fold change in gene expression between the presence and absence of inducer) (7–11). Dynamic range is an important indicator for fine-tuning biosensors, and a high dynamic range can help to distinguish the small difference in the inducer concentrations. The gene expression in biosensor systems driven by small molecule responsive transcription factors can achieve the desired output at appropriate translation level. One of the key elements to regulate the translation level is the ribosome-binding site (RBS), which tunes the dynamic range of the biosensor by adjusting the translation level and protein folding (12) of the transcription factor and reporter. Promoter, terminator, and plasmid copy number are also important factors in defining gene expression. They mainly focus on regulating the transcriptional level or stability of mRNA, far away from translation and protein folding. Thus, we chose RBS as the target in this work. However, the existing genetically encoded biosensors usually suffer from unpredictable translation tuning of regulatory elements to dynamic range. Many attempts have been made to tune the dynamic range of biosensors. For instance, Levin-Karp *et al.* used six RBSs ranging from strongest to weakest to achieve 20–200-fold dynamic range of protein expression (13). Wang *et al.* tuned the dynamic range of device input and output using five various-strength RBSs (RBS30–RBS34) from the Registry of Standard Biological

\*To whom correspondence should be addressed. Tel: +051085328539; Email: dengyu@jiangnan.edu.cn  
Correspondence may also be addressed to Shenghu Zhou. Tel: +051085328539; Email: zhoush@jiangnan.edu.cn

Parts, and showed that RBS could be used as a linear amplifier to regulate protein expression levels (14). Although these methods might help to regulate the dynamic range of gene expression, the dynamic range of regulatory elements involved in gene expression could not be predicted. For example, if the RBS was changed, then obtaining the appropriate dynamic range of gene expression required time-consuming and laborious research.

Establishment of a predictable and robust method can quickly achieve tuning of the biosensor dynamic range. In a previous report, Salis *et al.* calculated the Gibbs free energy difference ( $\Delta G_{\text{tot}}$ ) between the initiation and termination states of protein translation initiation based on a thermodynamic model, and presented RBS calculator for designing and synthesizing the RBSs of genes of interest, ensuring the rational control of protein expression levels (15). This significant contribution had accelerated the construction and optimization of complex genetic systems as well as promoted the development of synthetic biology. However, synthesis of the RBS through the calculation of free energy lacked experimental support. Therefore, design of the RBS by using a large amount of experimental data could make research on the RBS synthesis more robust. However, a large RBS database must rely on powerful analysis tools for better utilization of their application value, which can be solved by using mathematical models such as deep learning. Deep learning is an algorithm that uses artificial neural networks as a framework to characterize and learn databases. Deep learning models based on sequence levels have broad application prospects in the field of synthetic biology. For example, Chen *et al.* established Selene, a PyTorch-based deep learning library, which enables researchers to easily train the existing models to process biological problems of interest based on new databases and can be applied to any biological sequence data, including DNA, RNA, and protein sequences (16). Nielsen and Voigt used a deep learning based convolutional neural network (CNN) containing 42 364 plasmid DNA sequences datasets from Addgene to predict the lab-of-origin of a DNA sequence, and achieved 70% prediction accuracy and rapid analyses of DNA sequence information to guide the attribution process and understand the measures (17). While these studies provide a window for translation tuning of the RBS to biosensor dynamic range, the ability to design biosensors with reasonable dynamic ranges still remains a challenge (18–20).

In general, the RBS controls the translation level of a protein (15). Therefore, in the study of biosensors, the RBS tunes the dynamic range of biosensors by regulating the expression of reporter and regulatory protein. In this study, the glucarate biosensor was used as an example to explore the regulation mechanisms of RBS. In doing so, the RBS design principles for carbohydrate diacid activator (*cdar*) and ‘superfolder’ green fluorescent protein (*sfgfp*) (21) in glucarate biosensors were established. Subsequently, a library containing 12 000 cross-RBSs (cRBSs, combining RBSs of *cdar* and *sfgfp* in glucarate biosensors) was constructed by using DNA microarray, which was divided into five sub-libraries through fluorescence-activated cell sorting (FACS). Finally, a CNN on the cRBSs libraries was trained and a classification model between cRBSs and average dynamic range of each sub-library was developed and

was termed CLM-RDR, which performed well in predicting biosensors dynamic range (Figure 1). The CLM-RDR used large RBS library data to provide a knowledge base for precise adjustment of biosensors dynamic range, thus helping researchers to better characterize biosensors dynamic range by using RBS datasets. Given the availability of a large number of RBSs library, the CLM-RDR classification model can be extended to other similar biosensors to fine-tune their dynamic ranges, thereby significantly simplifying the workload of the design–build–test–learn cycle for designing biosensors with moderate dynamic ranges in bacteria and accelerating intelligent fine-tuning of biosensor dynamic range.

## MATERIALS AND METHODS

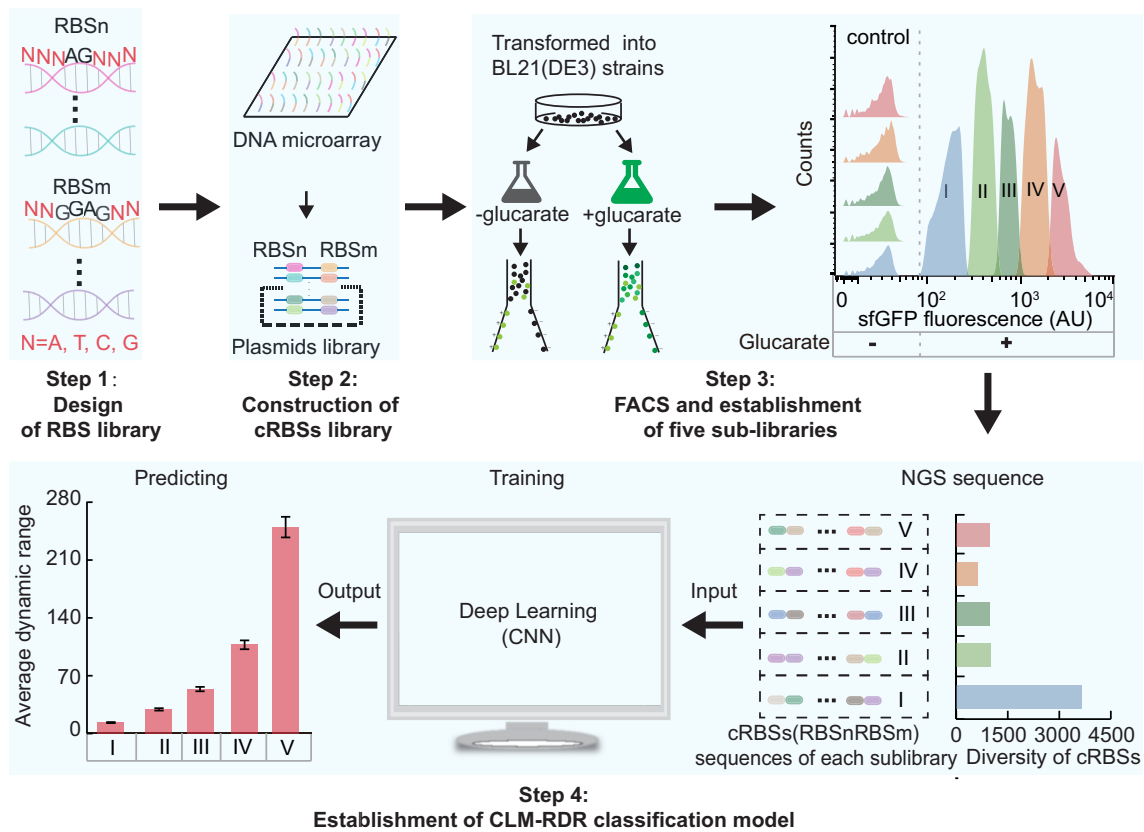
### Strains and culture conditions

All strains used in this study are listed in Supplementary Table S1. *Escherichia coli* JM109 and *E. coli* BL21 (DE3) cells were used for plasmid cloning and protein expression, respectively. M9 minimal medium, consisting of  $\text{Na}_2\text{HPO}_4$  (6.78 g/l),  $\text{KH}_2\text{PO}_4$  (3.0 g/l),  $\text{NaCl}$  (0.5 g/l),  $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$  (0.5 g/l),  $\text{CaCl}_2$  (0.011 g/l),  $\text{NH}_4\text{Cl}$  (1.0 g/l) and glucose (5 g/l), was used for fluorescence intensity assessment. The final concentrations of ampicillin, kanamycin, and spectinomycin employed in this study were 100, 50 and 50  $\mu\text{g/ml}$ , respectively. The final concentration of isopropyl  $\beta$ -D-thiogalactoside (IPTG) was 1 mM.

### Plasmid construction

All plasmids and primers used in this study are listed in Supplementary Tables S1 and S2, respectively. The pJKR-H-*cdar* plasmid for glucarate biosensor was purchased from Addgene (#62557). Wherein, the *cdar* (Gene ID, 944860) is originally from *E. coli* K12 strain (22). In addition to RBS and g10RBS, we selected seven RBSs: RBS3, RBS7, RBS8, MCD2, MCD10, BBa\_J61100 and BBa\_J61106 (Supplementary Table S3). The primer design was based on the different RBS sequences, and the pJKR-H-*cdar* plasmid was used as the template for plasmid PCR. Plasmids pJKR-H-RBSs-*cdar*-RBSs (RBSs are represented as R, R3, R7, R8, G10, M2, M10, BJ00 or BJ06) were constructed through *DpnI* digestion, and the digestion products were introduced into *E. coli* JM109 cells for screening by colony PCR and Sanger sequencing. The plasmids libraries NGS-RBSn-RBSm-I, NGS-RBSn-RBSm-II, NGS-RBSn-RBSm-III, NGS-RBSn-RBSm-IV and NGS-RBSn-RBSm-V were constructed through *XbaI/SpeI* digestion and T4 ligation. The plasmids pJKR-H-R-*cdar*-G10-*lacZ*-his, pJKR-H-R-*cdar*-M10-*lacZ*-his, pJKR-H-R-*cdar*-R8-*lacZ*-his, and pRSF-*groEL-groES* were constructed using Gibson assembly (23). The plasmid pHS-AVC-LW1125 was synthesized by Beijing Syngentech Co., Ltd in china. through DNA microarray technique.

Plasmids containing the glycolate biosensor pUC-*glcC*-ffs and arabinose biosensor pUC-*araC* were constructed through Gibson assembly methods. In both of the biosensors, the *rnnB* strong terminator, antibiotic resistance gene, and origin of replication were derived from the glucarate



**Figure 1.** Workflow of CLM-RDR development. First, the dynamic range of biosensors and the sequences of their related cRBSs were analyzed to establish an RBS design principle (Step 1). Based on this principle, a cRBSs library was designed and synthesized (Step 2) using DNA microarray. Subsequently, the library was divided into five sub-libraries (I–V) based on the fluorescence intensity of sfGFP measured by FACS (Step 3). Finally, to predict the dynamic range of biosensors with the given cRBSs, NGS and CNN model were employed to analyze the sequences of cRBSs in sub-libraries I–V and establish the CLM-RDR, respectively (Step 4). RBSn (NNNAGNNN), RBSs of *cdaR*; RBSm (NNGGAGNN), and RBSs of *sfgfp*; N = A, T, C, G.

biosensor (pJKR-H-*cdaR*) (6). All the sequences of transcriptional regulators and their promoters are provided in Supplementary Table S3. To evaluate the general performance of the CLM-RDR, we randomly selected eight RBSs to engineer three biosensors using plasmid PCR method: RBS<sub>*cdaR*</sub> (R<sub>c</sub>) and g10RBS derived from the glucarate biosensor; BBa\_J61104 (BJ04) and BBa\_J61108 (BJ08) obtained from the Anderson RBS library; MCD10 generated from the monocistronic design by Mutalik, *et al* (24); RBS<sub>*glcC*</sub> (R<sub>g</sub>) obtained from the glycolate biosensor; and RBS<sub>p<sub>RSF</sub></sub> (R<sub>pR</sub>) and RBS<sub>p<sub>Trc99a</sub></sub> (R<sub>pT</sub>) derived from plasmids pRSF and pTrc99a, respectively (Supplementary Table S3). The plasmid construction methods for each biosensor had been described earlier.

#### ANOVA model for cRBSs:glucarate combinatorial datasets

To understand the contribution and interaction between cRBSs and glucarate in the precise regulation of biosensors, we performed ANOVA (24) on the following linear model, using fluorescence data from sfGFP (21)

$$\text{Fluorescence}_{ijk} = \mu + C_i + G_j + (C : G)_{ij} + \varepsilon_{ijk}$$

for  $i = (1 - 81); j = (1 - 12)$

where Fluorescence<sub>*ijk*</sub> is the fluorescent output signal measured from the translation element, C<sub>*i*</sub>, and induced substrate glucarate, G<sub>*j*</sub>; (C:G)<sub>*ij*</sub> represents any interaction between the *i*th translational element and *j*th concentration of glucarate;  $\mu$  is the overall average signal; and  $\varepsilon_{ijk}$  is the error term for each C:G combination. The analysis output is presented in Supplementary Table S4.

#### $\beta$ -Galactose activity assays

The process of gene deletion in *E. coli* BL21 (DE3) cells was performed as described by Jiang *et al.* (25). The sgRNA of *lacZ* is shown in Supplementary Table S3. An appropriate amount of fermentation broth was centrifuged at 8000  $\times$  g for 10 min at 4°C, the supernatant was discarded, and the cells were collected. The cells were washed twice with cold lysis buffer (Tris–HCl; 0.01 M, pH 7.5). Then, the cells were resuspended in 2.5 ml of 0.01 mol/l Tris–HCl buffer (pH 7.5), and glass beads (26) and 50  $\mu$ l of PMSF stock solution were added to the cell culture. The cell culture was oscillated six times at high speed for 15 s each and placed on ice intermittently. Subsequently, 2.5 ml of Tris–HCl buffer were added to the culture, and the supernatant collected after centrifugation at 8000  $\times$  g for 15 min at 4°C was

the crude enzyme solution. Next, 1 mM *o*-nitrophenyl- $\beta$ -D-galactopyranoside (oNPG) solution was prepared with 50 mM oNPG. Approximately 10  $\mu$ l of the diluted crude enzyme solution and 20  $\mu$ l of the oNPG solution were added to 70  $\mu$ l of Z-buffer (16.1 g/l Na<sub>2</sub>HPO<sub>4</sub>·7H<sub>2</sub>O, 5.5 g/l NaH<sub>2</sub>PO<sub>4</sub>·H<sub>2</sub>O, 0.75 g/l KCl, 0.246 g/l MgSO<sub>4</sub>·7H<sub>2</sub>O and 2.7 mL  $\beta$ -mercaptoethanol; pH 7.0, stored at 4°C) for 10 min at 30°C. Then, 120  $\mu$ l of 1 mol/l pre-cooled Na<sub>2</sub>CO<sub>3</sub> were immediately added to stop the reaction and develop color. Finally, the absorbance was measured with a spectrophotometer at a wavelength of 420 nm. One unit of enzyme activity was defined as the amount of enzyme catalyzing the production of 1  $\mu$ mol *o*-nitrophenol (oNP) per minute (27,28).

Bovine serum albumin (BSA) was dissolved in Z-buffer at different dilutions (0.0–0.2 mg/ml BSA), and standard curves were generated. Crude enzyme (20  $\mu$ L) was added to 200  $\mu$ L of Bradford reagent, mixed, and its absorbance was determined at a wavelength of 595 nm. The crude enzyme concentration was calculated with a standard curve. The formula for calculating the enzyme activity was as follows. U/mg protein = OD<sub>420</sub>  $\times$  1.7/(0.0045  $\times$  protein content  $\times$  crude enzyme volume  $\times$  time), where OD<sub>420</sub> is the optical density of the product *o*-NP at 420 nm, coefficient 1.7 is the corrected value of the reaction volume, coefficient 0.0045 is the optical density (OD) of 1 mM oNP solution, protein content is expressed in mg/ml, crude enzyme volume is expressed in ml, and time is shown in min.

### Fluorescence assays

The cells were grown overnight to stationary phase before being diluted into fresh LB medium at a ratio of 1:100 and incubated at 250 rpm and 37°C. After 3 h, 100  $\mu$ l of log-phase cells were transferred to 96-well plates and stock inducers were respectively added to achieve the desired induction concentrations. Different concentrations of glucarate, glycolate, and arabinose were obtained by diluting 100 g/l glucarate, 1 M glycolate, and 1 M arabinose mother liquor in 96-well plates. Before measurements, the cultures were diluted into 0.01 M phosphate buffered saline (PBS; pH 7.4) to ensure that the OD<sub>600</sub> value was about 0.5. Measurements were performed using a Biotek HT plate reader (Winooski, VT, USA) under excitation wavelength of 485/20 nm and emission wavelength of 528/20 nm at 37°C and rapid shaking. Fluorescence intensity was measured in arbitrary units (AU), and the OD was determined by absorbance. For a given measurement, normalized fluorescence was determined by dividing the fluorescence by OD. The ratio of fluorescence to absorbance at 600 nm was used to compensate for the changes in cell density over time and between experiments (AU/OD).

*Escherichia coli* BL21 (DE3) cells containing the plasmid libraries were cultured to saturation, and then incubated at a concentration of 1% into 250-ml flasks containing LB medium at 250 rpm and 37°C. After 2 h, inducers were added to the desired final concentration, and incubation was resumed for 12 h. The induced cultures were diluted into cold PBS and kept on ice until evaluation with a BD FACS AriaII cell sorter (Becton Dickinson) (29). At least

100 000 events were captured for each sample. BD FACS-Diva software was used to divide the gate for sfGFP (21) (bandpass filter, 530/30 nm; blue laser, 488 nm). The mis- or un-folded sfGFP would be repaired to a correct folded protein by chaperonin GroEL/S. Wherein, repair rate is calculated as: (Flu (GroEL/S+) – Flu (GroEL/S–))/Flu (GroEL/S+). Where Flu (GroEL/S+) indicates fluorescence intensity with GroEL/S protein; Flu (GroEL/S–) indicates fluorescence intensity without GroEL/S protein. GroEL/S and sfGFP were expressed upon the addition of 1 mM IPTG and 20 g/l glucarate, respectively.

### Design of the RBS library

The cRBSs with the dynamic ranges over 2 were selected, which were R, G10, R7, M2, M10, BJ00 for CdaR and R3, R8, G10, M2 and M10 for sfGFP. Then, multiple sequence alignments were performed for RBSs of CdaR and sfGFP by using ClustalW in MEGA7.0 (30), whose processed files were uploaded to WebLogo 3 for analyzing the conservativeness, preference, and base frequency of the RBSs sequence. Sequence logos provided a precise description of sequence similarity and could rapidly reveal significant features of the alignment (31). Each logo consisted of stacks of bases (31). The overall height of each stack was proportional to the degree of sequence conservation at that position (measured in bits), whereas the height of each base within the stack was proportional to the relative frequency of the corresponding nucleotide at that position (31,32).

### Sorting of the RBS library

In total, 12 000 cRBS sequences were synthesized using DNA microarray, amplified by PCR, and were cloned into a glucarate biosensor plasmid backbone (pHS-BVC-LW274 and pHS-BVC-LW276) via two-step Golden Gate assembly (23) (completed by Synbiotic Gene Company) to obtain the glucarate biosensor plasmid library. Next, the plasmid library was transformed into *E. coli* BL21 (DE3) cells, which were cultured for 8 h in LB medium with or without 20 g/l glucarate supplementation. Then, the cells induced with 20 g/l glucarate were divided into five non-adjacent sub-libraries (I–V). To ensure the reliability of fluorescence intensity, cell adhesion was removed by executing FSC-A/FSC-H and SSC-A/SSC-H operation. Finally, the cells from each sub-library were obtained.

### NGS library preparation, sequencing and data processing

Cells from each sub-library were collected and their plasmids were extracted. The distance between the two RBSs in the glucarate biosensor was 2208 bp. However, the NGS was able to measure only up to 250 bp. Therefore, the isoschizomers *Xba*I/*Spe*I and T4 ligase were used to modify the plasmids of the five sub-libraries. The modified sub-libraries contained 152 bp between two RBSs (Supplementary Figure S3B), and the mixed PCR products of the five modified sub-libraries were linked with 10 barcodes (pink marked in Supplementary Table S2) and sequenced by NGS. The sequencing library was prepared following the

manufacturer's protocol (TruSeq DNA PCR-Free Sample Preparation Kit for Illumina). Library quality assessment and quantification was performed with Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA) and Q-PCR. Finally, all sub-libraries were pooled together and Nova-Seq 6000 Sequencer were used for pairing and sequencing the read length of 150 bp.

Raw reads for all sequenced sub-libraries were quality controlled using fastp v0.20.1 with default settings. After production of clean data, fastq-multx v0.20.1 with default settings was used to split the data according to barcode. Then, pairs of paired-end data were merged by FLASH (33) v1.2.11 with default setting and the polymorphism statistics were performed on merged reads of each sub-library. Subsequently, Raw reads were processed from bulk FASTQ data. Sequences for the read counts of 1 in each sub-library were eliminated. Next, the cRBSs were extracted from the raw reads by removing their sounding sequences. Furthermore, we removed the cRBS sequences that not include in the 12 000 libraries. For repeated cRBSs in different sub-libraries, only maintain the reads that have the highest counts. If the repeated cRBSs have the same counts in different sub-libraries, all these cRBSs were removed. Finally, 7053 cRBSs were obtained (Supplementary Table S9).

### Deep learning

First, 7053 cRBSs and corresponding seven features, i.e. the frequency of GC, A, T, C, G of cRBS, GC of SDn and GC of SDm, were combined to create datasets for subsequent deep learning. Then, the fluorescence intensity was divided into five levels for evaluating the biosensor corresponding to the RBS. To classify the RBS sequences, one-hot coding was initially employed. A neural network model (34,35) consisting of three convolutional layers and three full connection layers was proposed to accurately classify the RBS sequences. The convolutional layers comprised stride 1 and the pooling layers were non-overlapping. The convolution layer included two functions: feature extraction and feature mapping. On the one hand, the input of each neuron was connected to the local receptive field of the previous layer, and the local features were extracted. After the local features were extracted, the positional relationships between them and other features were also determined. On the other hand, each computing layer of the network was composed of multiple feature maps, each feature maps into a plane, and all the neurons on the plane exhibited the same weight. The feature map used the ReLU function with a small kernel of the influence function as the activation function of the convolution network, so that it had an invariance of displacement. Deep learning was performed with SciPy (1.0.0), NumPy (1.14.0) and TensorFlow (1.9.0) Python packages.

### Statistics

All statistical *T* tests are paired and two-tailed (36). Details about the statistical tests are described in the corresponding figure legends. AUC of receiver operating characteristic (ROC) curves was calculated using the metrics.auc function of the 'sklearn' python package.

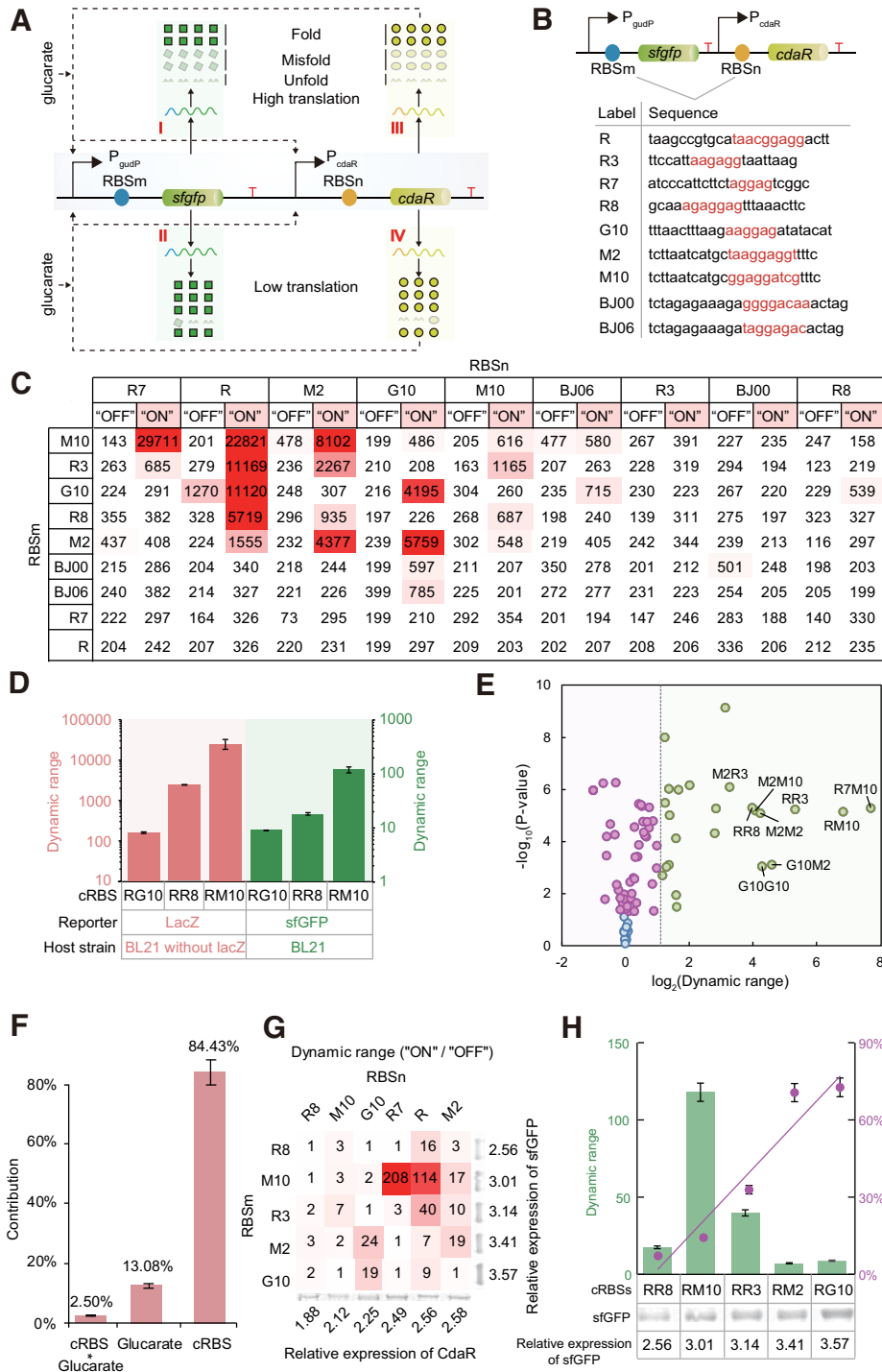
## RESULTS

### RBS plays a crucial role in the regulation of biosensor dynamic range

Although recent advances in synthetic biology have shed light on the importance of fine-tuning of biosensor dynamic range in various fields, the ability to design biosensors with moderate dynamic ranges remains limited (11,37–39). To investigate the key factors in biosensor dynamic range regulation, we used glucarate biosensor and explored its response strength by employing diverse concentrations of glucarate for induction (Supplementary Figure S1A and B). A carbohydrate diacid activator (CdaR) is inactivated in the absence of glucarate, making the biosensor to be in the 'OFF' state. In the presence of glucarate, CdaR is activated and simultaneously increases the expression level of its own and  $P_{\text{gudP}}$  controlled genes and making the biosensor to be in the 'ON' state (22,40) (Supplementary Figure S1A).

Addition of 20 g/l glucarate biosensor presented the highest nine-fold dynamic range. However, the fluorescence intensity presented a downward trend when the glucarate concentration exceeded 20 g/l (Supplementary Figure S1B). Similar observations have also been noted for other biosensors, such as *acuR*-based 3-hydroxypropionate biosensor (3), which also exhibited downward trend of fluorescence intensity when cerulenin concentration exceeded a certain threshold value. This phenomenon may be owing to the rapid translation and transcription of sfGFP, which not only cause metabolic burden (slow growth) (Supplementary Figure S1C) to the living cells, but also affect the natural folding of sfGFP (41), thus resulting in low fluorescence intensity. Faure *et al.* indicated that the occurrence of misfolding proteins increases with the increasing translation speed (12). Thus, although the amount of expressed sfGFP increased (Supplementary Figure S1D), the fluorescence intensity per protein molecule significantly decreased when glucarate concentration exceeded 20 g/l, owing to excessive misfolding. Therefore, it can be assumed that the most critical challenge for fine-tuning the dynamic range of biosensors might be to balance the translation level of regulator and reporter to simultaneously achieve the desired total fluorescence intensity with the highest fluorescence intensity per protein molecule (Figure 2A). These findings suggested that RBS might probably be a key element affecting the dynamic range of biosensors.

To investigate the correlation between RBS and biosensor dynamic range, nine RBSs covering a wide range of translation level from weak to strong were randomly selected for combinatorial replacement of the RBSs of *cdaR* and *sfgfp* (Figure 2B). The nine RBSs selected were RBS (R) and G10RBS (G10) derived from the plasmid pJKR-H-*cdaR* (6); RBS3 (R3), RBS7 (R7), and RBS8 (R8) designed with an RBS calculator (15); MCD2 (M2) and MCD10 (M10) derived from the monocistronic design by Mutalik *et al.* (24); and BBa\_J61100 (BJ00) and BBa\_J61106 (BJ06) obtained from the Anderson RBS library. Finally, 81 cRBS glucarate biosensors were obtained and their response strength and dynamic range were significantly improved when induced with various concentrations of glucarate (Figure 2C, Supplementary Table S5). In doing this,



**Figure 2.** Effects of cRBSs on biosensor dynamic range. (A) The regulation mechanisms of glucarate biosensor. Strong RBSs facilitate protein translation but could result in more mis- and un-folding proteins (I and III); Weak RBSs slower protein translation but benefit for protein folding (II and IV). Dash arrows represent feedback activation. (B) Nine RBS sequences derived from various libraries were obtained to replace the RBSs of glucarate biosensor. (C) The fluorescence intensity of 81 cRBS glucarate biosensors under ‘OFF’ and ‘ON’ state. cRBSs are defined as the RBS combination of *cdaR* (RBSn) and *sfgfp* (RBSm); for example, RM10 (R represents the RBSn of *cdaR*, M10 denotes the RBSm of *sfgfp*). The numbers in heatmap indicate the means of the fluorescence intensity of each cRBS in ‘ON’ and ‘OFF’ state. (D) The dynamic range of glucarate biosensors using LacZ and sfGFP as reporters. (E) Volcano plot demonstrates the dynamic range and differential expression of random selected 81 cRBSs. Blue circles represent cRBSs without significant in fluorescence ( $P > 0.05$ ). Pink and green circles represent cRBSs with significant in fluorescence ( $P < 0.05$ ). The cRBSs with a dynamic range higher than 2 are represented by green circles. The dynamic range of cRBSs marked on the green circle is higher than 8.  $P$  value, two-tailed  $T$ -test. (F) Contributions of cRBS and glucarate on dynamic range. (G) Effect of expression levels of CdaR and sfGFP on the dynamic range of glucarate biosensor. (H) The relationship between dynamic range, repair rate, and translation level. The correlation coefficient square ( $R^2$ ) of the fitted curve of the repair rate was 0.93. Relative expression was calculated using ImageJ software; green columns represent the dynamic range; pink circles indicate the repair rate of sfGFP. Data represents the mean and standard deviation for three replicates.

the detection deviation would significantly influence the calculated dynamic range, especially when the fluorescence at a very low level. To eliminate the detection deviation, we defined that the cRBS is non-functional when the ON state fluorescence was lower than 2-fold of that of wild type *E. coli* BL21 (DE3), whose dynamic range was defined as 1. In the cRBSs of R7M10 and RM10, 208-fold and 114-fold dynamic ranges were observed, respectively, when induced by 20 g/l glucarate, which were higher than that of the naturally existed cRBS RG10 (9-fold), indicating that the RBS played a very important role in fine-tuning biosensor dynamic range.

To validate whether the effect of cRBSs on the biosensor dynamic range was independent of reporter genes, we selected three cRBS biosensors with distinct dynamic ranges (RG10, RR8 and RM10) to replace *sfgfp* with *lacZ*. Finally, we found that the three cRBSs showed the same dynamic range trend regardless of the reporter gene (*sfgfp* or *lacZ*) (Figure 2D). This finding indicated that the cRBSs could consistently fine-tune the dynamic range of biosensor irrespective of the reporter. Subsequently, we analyzed the datasets with and without 20 g/l glucarate to assess the significance of differential expressions of genes with 81 cRBSs. We found that 63% of the 81 cRBSs had significant ( $P < 0.05$ ), and that 24.7% of the cRBSs showed significant differential expression of sfGFP (Figure 2E). To verify whether RBS was the most critical factor affecting the dynamic ranges of glucarate biosensors, we performed analysis of variance (ANOVA) on cRBSs and glucarate datasets (Figure 2F). The results suggested that cRBSs and glucarate contributed 84% and 13% to biosensor fine-tuning, respectively. In addition, an interaction (2%) between the two factors was also noted (Supplementary Table S4, see methods). These results indicated that the RBS is a key element for tuning the dynamic range of biosensors. However, it is still unclear on how the RBS fine-tunes the biosensor dynamic range.

### The RBS fine-tunes biosensor dynamic range by controlling protein translation and folding

To explore the relationship between translation level and dynamic range, the actual translation levels of the two variables, RBSn and RBSm, were respectively analyzed by SDS-PAGE. Under the same RBSn, the optimal translation levels of RBSm produced the highest biosensor dynamic range, and similar trend was also found for the translation level of RBSn under the same RBSm (Figure 2G, Supplementary Figure S2A, B), suggesting that the maximum dynamic range can be achieved at optimal protein translation level. However, translation level higher or lower than the optimal translation level could cause low biosensor dynamic range, which could be due to the rapid or slow expression of sfGFP resulting in misfolding or unfolding, thus affecting the natural folding of sfGFP (12,29). Therefore, we hypothesized that the RBS could affect protein folding by regulating the translation level of protein.

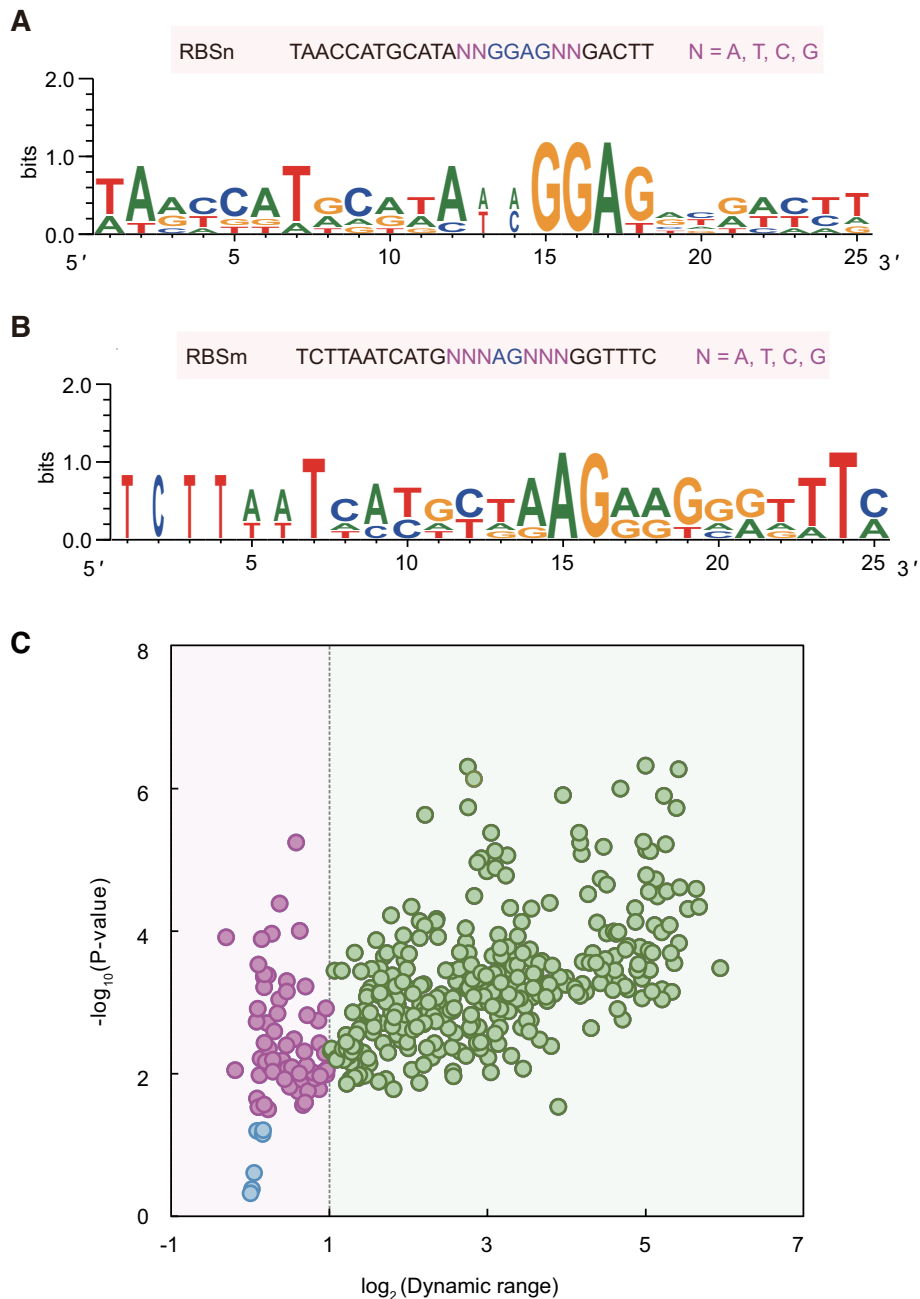
To examine the relationship between dynamic range and protein folding, the reported wild-type chaperone ring complex, GroEL/S, which has the ability to assist in the folding of heterologous protein in *E. coli* (42), was used to ver-

ify the effect of the RBS on sfGFP folding. Five cRBSs (RR8, RM10, RR3, RM2 and RG10) with different translation levels were used to investigate the misfolding and repair of sfGFP. SDS-PAGE revealed that the increase in fluorescence intensity of each cRBS was not caused by different expression levels of sfGFP, but was caused by GroEL/S repairing misfolded or unfolded sfGFP to a natural folded state (Supplementary Figure S2C). The fluorescence changes with and without GroEL/S overexpression were explored by FACS upon addition of 20 g/l glucarate (Supplementary Figure S2D, E). Furthermore, the repair rate (which was defined as the enhancement rate of fluorescence intensity after the chaperone GroEL/S (42) repair the mis- or un-folded sfGFP into a folded state.), dynamic range, and sfGFP expression levels were calculated, which indicated that sfGFP expression was positively correlated with repair rate, while optimal translational level of sfGFP was more beneficial for achieving higher biosensor dynamic range (Figure 2H, Supplementary Figure S2D–F). In addition, CdaR had a similar trend (Supplementary Figure S2G). This finding was consistent with our hypothesis, implying that strong RBSs have high translation level, which results in high misfolding rate and repair rate of sfGFP. Although dynamic range is a comprehensive phenomenon indicating the amounts and folding state of sfGFP, it is difficult to establish a quantitative equation to define the relationship between the RBS, translation level, folding, and dynamic range, which severely hinders the development of rational design of biosensors.

### Design of the RBS library to fine-tune biosensor dynamic range

Owing to the lack of quantitative relation between the RBS, translation level, folding, and dynamic range, it is possible to simulate and predict the biosensor dynamic range by mathematical models. As an alternative method, deep learning could predict complex biological relationships with simple neural network models, thereby circumventing the steps to understand the complicated biological mechanisms and achieving the expected effects of simulation and prediction. To obtain large data to train CNN model, we first accomplished designing of the RBS library and further tuned the dynamic range of the biosensor. On the basis of the 81 cRBSs datasets, the conserved sequences of the RBSs in *cdar* and *sfgfp* were generated by using the online software WebLogo (31) (see methods). The engineered RBSs included a consensus sequence defined as upstream and downstream of the Shine-Dalgarno (SD) sequence (RBSn: TAACCATGCATA-SDn-GACTT for *cdar*; RBSm: TC TTAATCATG-SDm-GGTTTC for *sfgfp*) and an SD preference sequence (SDn: NNGGAGNN for *cdar*; SDm: NN NGANN for *sfgfp*; N = A, T, C, G) (Figure 3A, B).

To evaluate the reliability of this design of RBS library, we randomly constructed 400 cRBSs (20 × 20 RBSs, 20 RBSs of *cdar* and *sfgfp*) (Supplementary Table S6). The fluorescence intensity and dynamic range of the 400 cRBSs biosensors with glucarate inducer showed a significant improvement, when compared with those without the inducer (Supplementary Table S7). In addition, the cRBSs biosensors presented an improved dynamic range upon addition



**Figure 3.** Design of RBSs library. WebLogo analyzed the design principle of RBSn (RBSs of *cdaR*) library (A) and RBSm (RBSs of *sfgfp*) library (B). (C) Volcano plot demonstrates the dynamic range and differential expression of rational designed 400 cRBSs. Blue circles represent cRBSs with no difference in fluorescence ( $P > 0.05$ ), and pink and green circles represent cRBSs with significant in fluorescence ( $P < 0.05$ ). The cRBSs with a dynamic range of higher than 2 are represented by green circles.  $P$  value, two-tailed  $t$ -test.

of 20 g/l glucarate (Supplementary Table S7). These findings implied that design of cRBSs library was more reliable and robust in improving the biosensor dynamic range. We further analyzed the datasets with and without glucarate to assess the differential expression of sfGFP, and found that up to 98% of the 400 cRBSs had significant ( $P < 0.05$ ) and 85.3% of the cRBSs showed significant differential expression of sfGFP (Figure 3C). These results indicated that the design of cRBSs library considerably contributed to the improvement of biosensor dynamic range.

#### Establishment of CLM-RDR for precise prediction of biosensor dynamic range

To further extend the dataset for CNN model training, we constructed a much larger cRBS library through the RBS library, and generated 100 RBSs for *cdaR* and 120 RBSs (Supplementary Table S6) for *sfgfp* (Figure 3A, B). Then, a combinatorial library of 12,000 cRBSs as oligonucleotides was developed with DNA microarray (see methods). To verify the homogeneity of the 12 000 cRBSs, next-



generation sequencing (NGS) was performed. The coverage of the 12 000 cRBSs was 100%, and the 10-fold variation reached a quality control value of 99.92% (Supplementary Figure S3A, Supplementary Table S8, NCBI Accession No. SRR9301216). This cRBS library was used in the following pooled screening experiment to characterize the dynamic range of the glucarate biosensor.

The 12 000 cRBS plasmid libraries were transformed into *E. coli* BL21 (DE3) cells, which were cultured for 8 h in Luria–Bertani (LB) medium supplemented with 0 or 20 g/l glucarate. Then, by using FACS, we divided the cells induced with 20 g/l glucarate into five non-adjacent sub-libraries I–V according to the expression intensity of sfGFP (Figure 4A). Five non-overlapping bins were chosen with a gap to reduce cross contaminations between the bins (29). Subsequently, the average single cell fluorescence intensity and average dynamic range of the sub-library I–V with and without glucarate were calculated, and a 13-fold, 29-fold, 53-fold, 106-fold and 247-fold average dynamic range were accomplished for the sub-libraries I–V, respectively (Figure 4B). These results further demonstrated that the cRBS library was highly effective in tuning the dynamic range of the glucarate biosensor, and helped to establish a high-quality element library in synthetic biology and construct an approach for designing complex genetic circuits to fine-tune gene expression (43–45).

To determine the cRBS sequences of the glucarate biosensors in each sub-library, we first obtained the assorted biosensor plasmids of the five sub-libraries. Then, the mixed PCR products of the five modified sub-libraries were linked with 10 barcodes and sequenced by NGS (46) (NCBI Accession No. SRR12384447–SRR12384461). Box plots showed the distribution of each cRBS count of five sub-libraries, and separate points indicated that the cRBS counts ranged from 160 to  $10^6$  (Figure 4C, Supplementary Table S9). In addition, the diversity of cRBSs in each sub-library was analyzed, and there were 3592, 980, 944, 596 and 941 cRBSs in sub-libraries I–V, respectively (Figure 4D). Finally, the sequenced 7053 cRBSs were used as the data sources for further data processing.

Although the cRBSs sequences of each sub-library were obtained, it was extremely crucial to determine the functional relationships between the cRBSs sequences and average dynamic range of glucarate biosensor. Functional relationships could help to quickly analyze the dynamic range of a corresponding cRBS biosensor, which could reduce the burden of the design–build–test–learn cycle. Therefore, CNNs of deep learning was chosen to establish a classification model between cRBSs and the average dynamic range of each sub-library (CLM-RDR). First, 85% of the cRBSs and their sequence characteristics in each sub-library were selected as datasets to train the CNN model (Supplementary Figure S4). Next, we evaluated how well CLM-RDR predicted the average dynamic range of the glucarate biosensor from the remaining 15% of cRBSs sequences and their sequence characteristics in each sub-library (Figure 4E). The results indicated that CLM-RDR predicted the dynamic range of the glucarate biosensor with high accuracy, yielding an area under the curve (AUC) of 0.81, 0.91, 0.64, 0.75 and 0.77 for sub-libraries I–V, respectively, and an average AUC of 0.86.

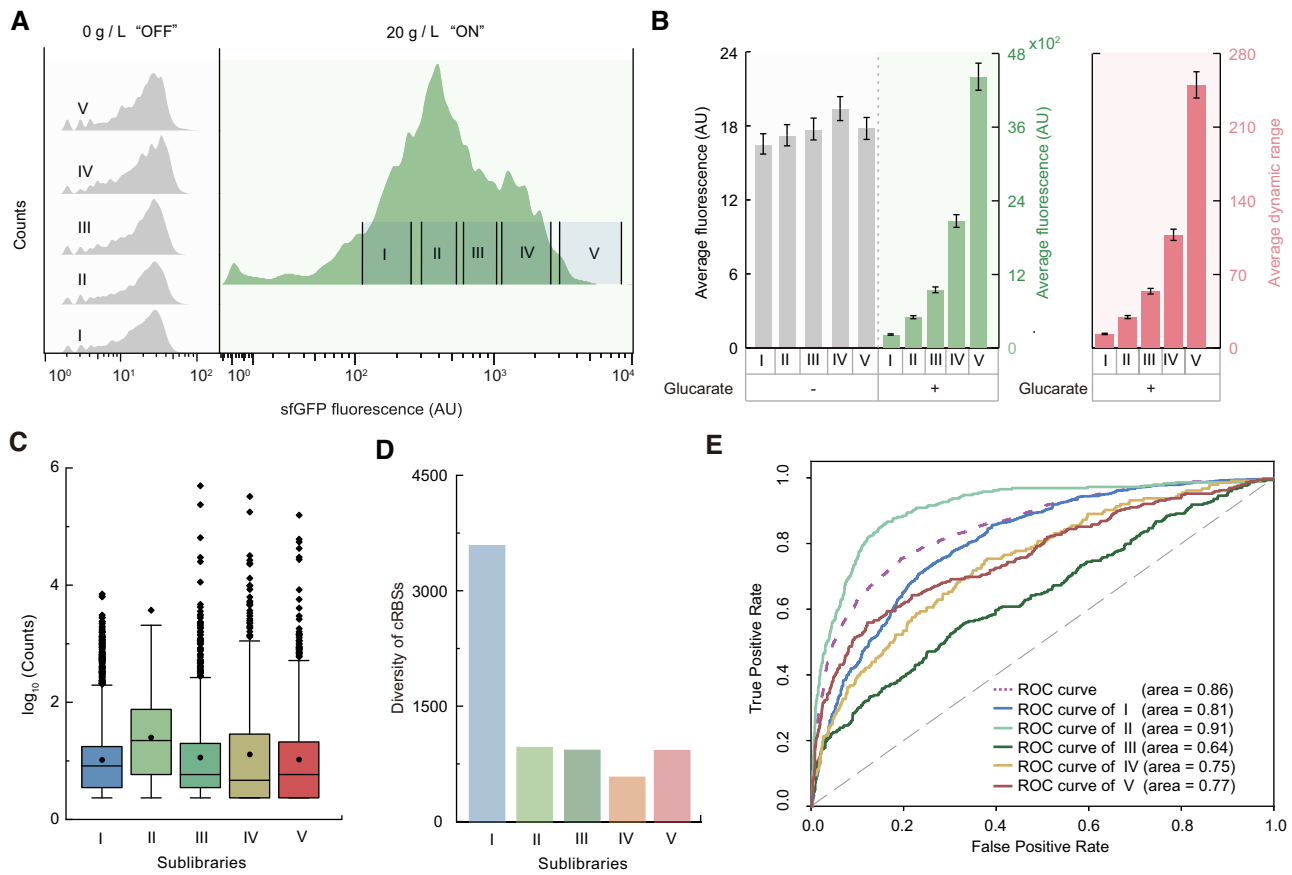
## Applications of the CLM-RDR to other biosensors

The CLM-RDR is expected to tune the dynamic range of different biosensors. Therefore, to further evaluate the performance of the CLM-RDR, we randomly selected 18 cRBSs to modify the glucarate biosensor, glycolate biosensor, and arabinose biosensor (see methods). We first predicted the average dynamic range of 18 cRBSs by using CLM-RDR and then performed an experiment to detect the dynamic ranges of the biosensors via FACS (Supplementary Figure S5). By analyzing the predicted and experimentally observed dynamic ranges, CLM-RDR was found to have good predictive performance for three biosensors. Predicted accuracy rates of 72.2% (Figure 5A), 61.1% (Figure 5B) and 50% (Figure 5C) were obtained for glucarate, arabinose (Figure 5D), and glycolate (Figure 5E) biosensors, respectively. These results indicated that the CLM-RDR had a certain degree of universality in predicting the dynamic ranges of biosensors in *E. coli*. The CLM-RDR can probably be further improved by providing additional training datasets.

## DISCUSSION

Genetically encoded biosensors derived from transcription factors responding to small-molecule inducers are receiving increasing research attention (3). The currently available genetically encoded biosensors usually have the major problem of inappropriate dynamic range (8,10). Although many valuable works, such as promoter modification studies, have attempted to tune the dynamic range of biosensors, universality may be difficult to achieve owing to small datasets and insufficient analysis tools. Therefore, fine-tuning of the biosensor dynamic range remains a huge challenge (7,20). In general, RBS controls the translation rate (15,24) of regulatory proteins and reporters, which can control the dynamic range of biosensors. Previous reports had indicated that the dynamic ranges of device input or output were not well tuned by replacing the RBS (13), mainly because the RBS datasets were limited. Therefore, to fine-tune the dynamic range of biosensors, in the present study, we established the design principle of the RBS in biosensors through ANOVA and online WebLogo processing. Accordingly, 12 000 cRBSs were designed based on the library, and five average dynamic ranges were calculated by dividing the cRBSs into five sub-libraries using FACS. Most importantly, we developed CLM-RDR, a classification model between cRBSs and average dynamic range of five sub-libraries. The CLM-RDR showed accurately predictive performance and was able to quickly determine the average dynamic range of a biosensor corresponding to a cRBS and their sequence characteristics. In addition, the CLM-RDR also had good predictive ability toward glycolate and arabinose biosensors, thus indicating that this model can be extended to other biosensors. Besides, the developed model significantly simplified the workload of the design–build–test–learn cycle of fine-tuned biosensor dynamic range in bacteria and accelerated intelligent fine-tuning of biosensor dynamic range.

RBSs play a role in fine-tuning genetic components and determining the translation level of proteins (15,24). Proteins usually present tight and loose structures. The mRNA structure affects the translation rate of a protein, and fast

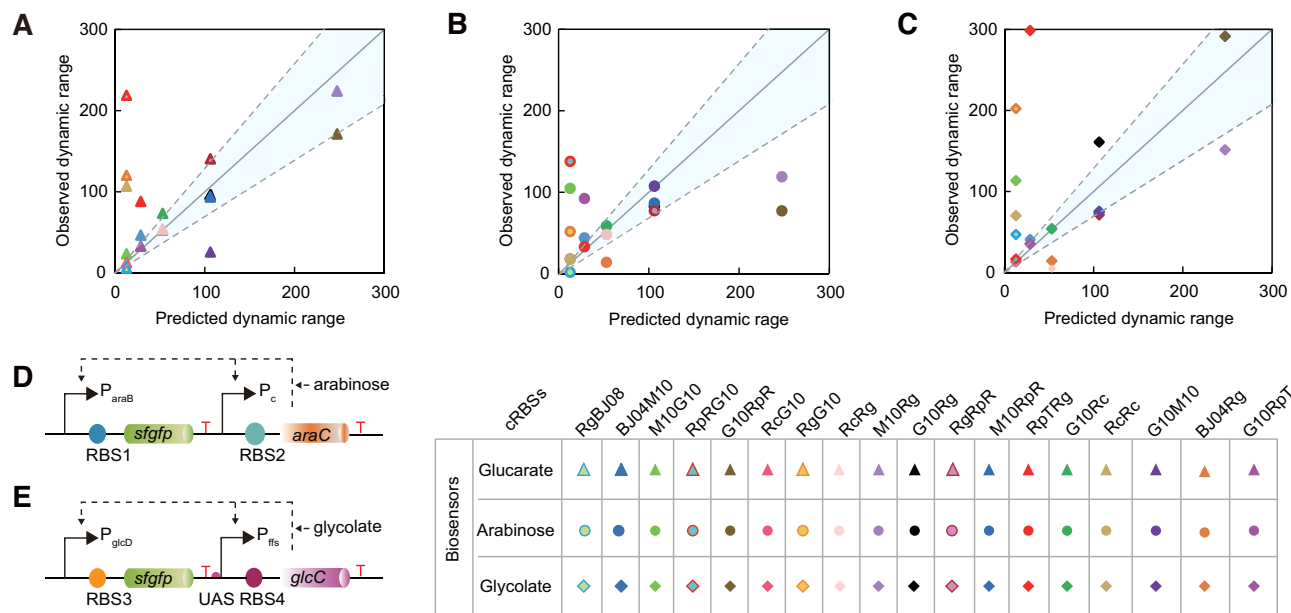


**Figure 4.** Accurate prediction of the dynamic range of glucarate biosensor from cRBS sequences by deep learning model. (A) A larger cRBSs library was formed than the original libraries. The cells induced with 20 g/l glucarate were sorted into five non-adjacent sub-libraries (I–V) by FACS. (B) Analysis of average fluorescence intensity in ‘ON’ state (green column), ‘OFF’ state (gray column), and average dynamic range (red column) of each sub-library. (C) The counts of each cRBS of the five sub-libraries were obtained by NGS. (D) Diversity of cRBSs of five sub-libraries. (E) Establishment of CLM-RDR model based on 7053 cRBS sequences. Receiver operating characteristic (ROC) curves for cRBSs of sub-libraries I–V (solid lines of various colors) and total library (pink dotted line). Biosensor dynamic ranges with five test-positive samples were used to classify. Data represents the mean and standard deviation for three replicates.

translation prevents the formation of compact structures, which affects protein folding (12). Thus, we hypothesized that the RBS might also affect the conformations of proteins by controlling translation level, thereby achieving fine-tuning of gene expression. Low translation level results in the low expressions of CdaR and sfGFP, causing the low dynamic range. Although the high translation level results in high protein expression, too many mis- or un-folded proteins are caused by the too fast translation rate, resulting in less correct-folded CdaR and sfGFP, causing low dynamic range. To further explore the relationship between translation level, protein folding, and biosensor dynamic range, a wild-type chaperone GroEL/S, which could assist in the folding of recombinant sfGFP in *E. coli*, was combined with a set of constructed biosensors (42). When compared with optimal protein expression, low and high protein expression produced more misfolded proteins, which in turn resulted in a higher repair rate of sfGFP by GroEL/S (Figure 2A, H). In addition, a positive correlation trend was observed between the expression level of sfGFP and repair rate. Therefore, appropriate protein expression level and protein folding state achieved the optimal biosensor dynamic range,

thus further implying that RBS is one of the key factors affecting the dynamic range of biosensors.

Sequence-based deep learning models had been reported to show good predictive performance for biological phenotypes (16,34,35). Deep learning models can accurately establish the correspondence between genotypes and phenotypes through large datasets, thus making investigations more universal. The present study found that one of the key factors affecting the dynamic range of biosensors was RBS. However, the mechanism of the RBS tuning the dynamic range of biosensors was complex (Figure 2A), not only requiring exploration of the mechanism of RBS tuning translation and folding of regulators and reporter, but also examination of the binding mechanism of regulators and operator sites and further investigation of the effects on downstream reporter transcription. Therefore, analysis of these mechanisms using current technology is a huge challenge. However, deep learning models do not require understanding of specific mechanisms to establish the relationship between RBS and biosensor dynamic range, and can be extended to other biosensors research. Hence, to develop a universal tool to fine-tune the dynamic range of biosensors,



**Figure 5.** CLM-RDR verification for three genetically encoded biosensors. 18 cRBSs were randomly selected for biosensor modification and comparison of the observed and predicted dynamic ranges. The CLM-RDR performed well in predicting the dynamic ranges of glucarate biosensor (A), arabinose biosensor (B), and glycolate biosensor (C). The grey diagonal denotes  $y = x$ . The light blue areas represent the predicted dynamic range has a  $\pm 30\%$  error. (D) Structure of  $P_{\text{araB}}$ -based arabinose sensor.  $P_c$  represents the constitutive promoter that controls transcription of the regulatory protein AraC.  $P_{\text{araB}}$  is an inducible promoter containing the AraC-binding DNA sequence. Dash arrows represent activation with inducer arabinose. (E) Structure of  $P_{\text{glcD}}$ -based glycolate sensor.  $P_{\text{fts}}$  (48) indicates the constitutive promoter that controls transcription of the reporter sfGFP. In the absence of glycolate, GlcC remained as a non-functional regulatory protein, whereas in the presence of glycolate, the regulatory protein GlcC and glycolate bound to the activator GlcC-glycolate, which in turn bound to the upstream activation site (UAS) of the promoter  $P_{\text{glcD}}$ , thus enhancing transcription and expression of sfGFP. Dash arrows represent activation with inducer glycolate. Solid triangle: glucarate biosensor; solid circle: arabinose biosensor; solid diamond: glycolate biosensor.

we developed CLM-RDR, a classification model based on deep learning between cRBSs and average dynamic range. The CLM-RDR showed good prediction performance for the dynamic range of the biosensor using only 7053 cRBSs datasets. More importantly, it could be extended to other biosensors, achieving the same prediction effects, implying that CLM-RDR has certain universality in predicting the dynamic range of biosensors in *E. coli*. Nevertheless, it is difficult for most mathematical models built with large datasets to achieve 100% prediction accuracy (46). So, inaccurate results will occur when using the model to predict the dynamic range of non-functional cRBSs. It should be noted that the present study only examined the effect of the RBS on biosensor dynamic range. The results of this study, along with further research on promoters, plasmid copy numbers, and regulatory protein evolution, could propel fine-tuning of the dynamic range of biosensors into the era of intelligence.

## DATA AVAILABILITY

Raw data of NGS for DNA microarray and cRBSs of five sub-libraries have been deposited to the NCBI Short Read Archive, with Accession No. BioProject: SRR9301216 (<https://dataview.ncbi.nlm.nih.gov/object/PRJNA548649>) and SRR12384447-SRR12384461 (<https://dataview.ncbi.nlm.nih.gov/object/PRJNA650172>), respectively.

To encourage experimental biologists to use CLM-RDR, we uploaded the model to GitHub, which converted an RBS sequence directly into biosensor dynamic range. The

code for predicting biosensor dynamic range can be found at <https://github.com/YuDengLAB/CLM-RDR>. Flow cytometry data for this study has also been deposited at Flow Repository (47), where it is directly accessible at <http://flowrepository.org/id/FR-FCM-Z2TW>.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

National Key R&D Program of China [2019YFA0905502]; National Natural Science Foundation of China [21877053, 31900066]; Jiangsu Province Science Foundation for Youths [BK20150159]; Postgraduate Research & Practice Innovation Program of Jiangsu Province [KYCX20.1813]. Funding for open access charge: Jiangnan University. *Conflict of interest statement.* None declared.

## REFERENCES

- Prindle, A., Samayoa, P., Razinkov, I., Danino, T., Tsimring, L.S. and Hasty, J. (2012) A sensing array of radically coupled genetic 'biopixels'. *Nature*, **481**, 39–44.
- Eggeling, L., Bott, M. and Marienhagen, J. (2015) Novel screening methods—biosensors. *Curr. Opin. Biotechnol.*, **100**, 30–36.
- Rogers, J.K. and Church, G.M. (2016) Genetically encoded sensors enable real-time observation of metabolite production. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 2388–2393.
- Carpenter, A.C., Paulsen, I.T. and Williams, T.C. (2018) Blueprints for biosensors: Design, limitations, and applications. *Genes*, **9**, 375.

5. Pham, H.L., Wong, A., Chua, N., Teo, W.S., Yew, W.S. and Chang, M.W. (2017) Engineering a riboswitch-based genetic platform for the self-directed evolution of acid-tolerant phenotypes. *Nat. Commun.*, **8**, 411.
6. Rogers, J.K., Guzman, C.D., Taylor, N.D., Raman, S., Anderson, K. and Church, G.M. (2015) Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Res.*, **43**, 7648–7660.
7. 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.
8. Nguyen, N.H., Kim, J.R. and Park, S. (2018) Application of transcription factor-based 3-hydroxypropionic acid biosensor. *Biotechnol. Bioproc. E.*, **23**, 564–572.
9. Skjoedt, M.L., Snoek, T., Kildegaard, K.R., Arsovska, D., Eichenberger, M., Goedecke, T.J., Rajkumar, A.S., Zhang, J., Kristensen, M., Lehka, B.J. *et al.* (2016) Engineering prokaryotic transcriptional activators as metabolite biosensors in Yeast. *Nat. Chem. Biol.*, **12**, 951–958.
10. Cheng, F., Tang, X.L. and Kardashliev, T. (2018) Transcription factor-based biosensors in high-throughput screening: Advances and applications. *Biotech. J.*, **13**, 1700648.
11. Kasey, C.M., Zerrad, M., Li, Y., Cropp, T.A. and Williams, G.J. (2017) Development of transcription factor-based designer macrolide biosensors for metabolic engineering and synthetic biology. *ACS Synth. Biol.*, **7**, 227–239.
12. Faure, G., Ogurtsov, A.Y., Shabalina, S.A. and Koonin, E.V. (2016) Role of mRNA structure in the control of protein folding. *Nucleic Acids Res.*, **44**, 10898–10911.
13. Levin-Karp, A., Barenholz, U., Bareia, T., Dayagi, M., Zelcbuch, L., Antonovsky, N., Noor, E. and Milo, R. (2013) Quantifying translational coupling in *E. coli* synthetic operons using RBS modulation and fluorescent reporters. *ACS Synth. Biol.*, **2**, 327–336.
14. Wang, B., Kitney, R.I., Joly, N. and Buck, M. (2011) Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nat. Commun.*, **2**, 508.
15. 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.
16. Chen, K.M., Cofer, E.M., Zhou, J. and Troyanskaya, O.G. (2019) Selene: A PyTorch-based deep learning library for sequence data. *Nat. Methods*, **16**, 315–318.
17. Nielsen, A.A. and Voigt, C.A. (2018) Deep learning to predict the lab-of-origin of engineered DNA. *Nat. Commun.*, **9**, 3135.
18. Westbrook, A.M. and Lucks, J.B. (2017) Achieving large dynamic range control of gene expression with a compact RNA transcription–translation regulator. *Nucleic Acids Res.*, **45**, 5614–5624.
19. Doong, S.J., Gupta, A. and Prather, K.L.J. (2018) Layered dynamic regulation for improving metabolic pathway productivity in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 2964–2969.
20. Zhang, F. and Keasling, J. (2011) Biosensors and their applications in microbial metabolic engineering. *Trends Microbiol.*, **19**, 323–329.
21. Pédelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. and Waldo, G.S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.*, **24**, 79–88.
22. Monterrubio, R., Baldoma, L., Obradors, N., Aguilar, J. and Badia, J. (2000) A common regulator for the operons encoding the enzymes involved in D-galactarate, D-glucarate, and D-glycerate utilization in *Escherichia coli*. *J. Bacteriol.*, **182**, 2672–2674.
23. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison III, C.A. and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.
24. Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P. and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*, **10**, 354–360.
25. Jiang, Y., Chen, B., Duan, C., Sun, B., Yang, J. and Yang, S. (2015) Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Appl. Environ. Microbiol.*, **81**, 2506–2514.
26. Ramanan, R.N., Ling, T.C. and Ariff, A.B. (2008) The performance of a glass bead shaking technique for the disruption of *Escherichia coli* cells. *Biotechnol. Bioproc. E.*, **13**, 613–623.
27. Liu, P., Wang, W., Zhao, J. and Wei, D. (2019) Screening novel  $\beta$ -galactosidases from a sequence-based metagenome and characterization of an alkaline  $\beta$ -galactosidase for the enzymatic synthesis of galactooligosaccharides. *Protein Express. Purif.*, **155**, 104–111.
28. Schaefer, J., Jovanovic, G., Kotta-Loizou, I. and Buck, M. (2016) Single-step method for  $\beta$ -galactosidase assays in *Escherichia coli* using a 96-well microplate reader. *Anal. Biochem.*, **503**, 56–57.
29. Sauer, C., Themaat, E.V.L.V., Boender, L.G.M., Groothuis, D., Cruz, R., Hamoen, L.W., Harwood, C.R. and Rij, T.V. (2018) Exploring the nonconserved sequence space of synthetic expression modules in *Bacillus subtilis*. *ACS Synth. Biol.*, **7**, 1773–1784.
30. Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, **33**, 1870–1874.
31. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.
32. Seghezzi, N., Amar, P., Koebmann, B., Jensen, P.R. and Vioille, M.-J. (2011) The construction of a library of synthetic promoters revealed some specific features of strong *Streptomyces* promoters. *Appl. Microbiol. Biot.*, **90**, 615–623.
33. Magoč, T. and Salzberg, S.L. (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.
34. Sundaram, L., Gao, H., Padigepati, S.R., Mcrae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A. and Shon, J. (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.*, **50**, 1161–1170.
35. Zhou, J. and Troyanskaya, O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
36. Yuter, E., Yang, J.S., Sogues, A. and Serrano, L. (2017) A reporter system coupled with high-throughput sequencing unveils key bacterial transcription and translation determinants. *Nat. Commun.*, **8**, 368.
37. Kim, S.K., Kim, S.H., Subhadra, B., Woo, S.G., Rha, E., Kim, S.W., Kim, H., Lee, D.H. and Lee, S.G. (2018) A genetically encoded biosensor for monitoring isoprene production in engineered *Escherichia coli*. *ACS Synth. Biol.*, **7**, 2379–2390.
38. Wang, B., Barahona, M. and Buck, M. (2015) Amplification of small molecule-inducible gene expression via tuning of intracellular receptor densities. *Nucleic Acids Res.*, **43**, 1955–1964.
39. Raman, S., Rogers, J.K., Taylor, N.D. and Church, G.M. (2014) Evolution-guided optimization of biosynthetic pathways. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 17803–17808.
40. Sampaio, M.M., Chevance, F., Dippel, R., Eppler, T., Schlegel, A., Boos, W., Lu, Y.J. and Rock, C.O. (2004) Phosphotransferase-mediated transport of the osmolyte 2-O- $\alpha$ -mannosyl-D-glycerate in *Escherichia coli* occurs by the product of the *mngA* (*hrsA*) gene and is regulated by the *mngR* (*farR*) gene product acting as repressor. *J. Biol. Chem.*, **279**, 5537–5548.
41. Ceroni, F., Boo, A., Furini, S., Gorochowski, T.E., Borkowski, O., Ladak, Y.N., Awan, A.R., Gilbert, C., Stan, G.B. and Ellis, T. (2018) Burden-driven feedback control of gene expression. *Nat. Methods*, **15**, 387–393.
42. Wang, J.D., Herman, C., Tipton, K.A., Gross, C.A. and Weissman, J.S. (2002) Directed evolution of substrate-optimized GroEL/S chaperonins. *Cell*, **111**, 1027–1039.
43. Nielsen, A.A., 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.
44. Nielsen, A.A., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., Ross, D., Densmore, D. and Voigt, C.A. (2016) Genetic circuit design automation. *Science*, **352**, aac7341.
45. Moon, T.S., Lou, C., Tamsir, A., Stanton, B.C. and Voigt, C.A. (2012) Genetic programs constructed from layered logic gates in single cells. *Nature*, **491**, 249–253.
46. Guo, J., Wang, T., Guan, C., Liu, B., Luo, C., Xie, Z., Zhang, C. and Xing, X.H. (2018) Improved sgRNA design in bacteria via genome-wide activity profiling. *Nucleic Acids Res.*, **46**, 7052–7069.
47. Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N. and Brinkman, R.R. (2012) FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytom. Part A*, **81A**, 727–731.
48. Zhou, S., Ding, R., Chen, J., Du, G., Li, H. and Zhou, J. (2017) Obtaining a panel of cascade promoter-5'-UTR complexes in *Escherichia coli*. *ACS Synth. Biol.*, **6**, 1065–1075.