Synthetic and Systems Biotechnology 7 (2022) 1046–1055

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

Sequence and thermodynamic characteristics of terminators revealed by FlowSeq and the discrimination of terminators strength

Weiji Zhai [a,b], Yanting Duan [a,b], Xiaomei Zhang [c,d], Guoqiang Xu [a,b], Hui Li [c,d], Jinsong Shi [c,d], Zhenghong Xu [a,b], Xiaojuan Zhang [a,b,*]

[a] Biotechnology of Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi, China
[b] National Engineering Research Center for Cereal Fermentation and Food Biomanufacturing, Jiangnan University, Wuxi, China
[c] School of Life Science and Health Engineering, Jiangnan University, Wuxi, 214122, China
[d] Jiangsu Provincial Engineering Research Center for Bioactive Product Processing, Jiangnan University, 1800 Lihu Avenue, Wuxi, 214122, China

ARTICLE INFO

Keywords:
Intrinsic terminator
Transcription termination
Machine learning
Structure-activity relationship
FlowSeq
Free energy

ABSTRACT

The intrinsic terminator in prokaryotic forms secondary RNA structure and terminates the transcription. However, leaking transcription is common due to varied terminator strength. Besides of the representative hairpin and U-tract structure, detailed sequence and thermodynamic features of terminators were not completely clear, and the effect of terminator on the upstream gene expression was unclearly. Thus, it is still challenging to use terminator to control expression with higher precision. Here, in *E. Coli*, we firstly determined the effect of the 3′-end sequences including spacer sequences and terminator sequences on the expression of upstream and downstream genes. Secondly, terminator mutation library was constructed, and the thermodynamic and sequence features differing in the termination efficiency were analyzed using the FlowSeq technique. The result showed that under the regulation of terminators, a negative correlation was presented between the expression of upstream and downstream genes (r=−0.60), and the terminators with lower free energy corelated with higher upstream gene expression. Meanwhile, the terminator with longer stem length, more compact loop and perfect U-tract structure was benefit to the transcription termination. Finally, a terminator strength classification model was established, and the verification experiment based on 20 synthetic terminators indicated that the model can distinguish strong and weak terminators to certain extent. The results help to elucidate the role of terminators in gene expression, and the key factors identified are crucial for rational design of terminators, and the model provided a method for terminator strength prediction.

1. Introduction

The terminator mainly performs the function of preventing downstream gene transcription from reading through, thus it is an indispensable element of the gene expression circuit [1,2]. The strength of the terminator determines the leakage level of downstream genes, enabling controlled expression in the densely packed genomes of bacteria [3]. There are two mechanisms of transcriptional termination, one is ρ factor-dependent termination, which needs the help of ρ-factor to terminate the transcription [4,5], another is the intrinsic termination (ρ factor-independent termination), which relied on the intrinsic terminator identified on the genome of prokaryotes to perform functions. The intrinsic terminator is composed with the RNA sequence formed the secondary structure: a hairpin rich in GC bases, followed by a U-tract structure, and A-tract structure often appears at the upstream of the hairpin which is usually related to the encoding of a bidirectional intrinsic terminator [6–8]. The hairpin and U-tract are common characteristics of the intrinsic terminator, and the variation in the sequence will lead to varied RNAP pausing time and the dissociation rate of TEC complex, and eventually affected the terminator efficiency [9].

Many previous works are devoted to the systematic analysis of the factors that affect the termination strength of the terminator. Chen et al. [10] characterized the termination efficiency of 582 natural and synthetic terminators, established a kinetic model to predict the termination efficiency, and also calculated the contributions of each structure features devoted to the terminator strength based on thermodynamic

Peer review under responsibility of KeAi Communications Co., Ltd.
* Corresponding author. Biotechnology of Ministry of Education, School of Biotechnology, Jiangnan University, Wuxi, China.
E-mail address: zhangxj@jiangnan.edu.cn (X. Zhang).

https://doi.org/10.1016/j.synbio.2022.06.003
Received 15 March 2022; Received in revised form 11 June 2022; Accepted 11 June 2022
Available online 20 June 2022

parameters. Cambray et al. proposed a linear model to predict the strength of the terminator on the basis of a stepwise regression method, and found that 5′-extended sequence upstream of terminators may enhance termination efficiency (TE) by partially hybridizing with the U-tract [11]. Cui et al. comprehensively analyzed 75 natural intrinsic terminators from *Bacillus subtilis* and revealed that hairpin thermodynamics and A-U pairing length cooperatively contributed to the TE [12]. The above studies provided information for understanding the structure-activity relationship of the terminator. However, more detailed sequence features were not fully studied and the discrimination model can be further optimized via comprehensive features selection to integrate the sequence characteristics and thermodynamics of terminators.

One option to obtain the discriminant characteristics of terminator from vast dataset is to correlate phenotype and genotype information of variants in the mutant library through FlowSeq. FlowSeq is the technique that combines cell sorting by flow cytometry and high-throughput sequencing, which provided a large number of sequence information of variants with different phenotypes, such as protein expression [13,14]. At the same time, as a method of processing big data, machine learning has been successfully applied in the field of computational biology [15–19]. Evfratov et al. used the random forest algorithm to establish a classification model for sequences in 5′-UTR (5′-Untranslated Region) with different translation rates that generated by FlowSeq technology [14]. A sequence-based promoter strength prediction model finished by XGBoost algorithm was established by Zhao et al. on the base of 3665 synthetic promoters [20]. Wang et al. designed a new AI framework for de novo design of promoters in *E. coli*, among them, up to 70.8% of AI-designed promoters were experimentally proven to be effective [21]. These all provide methods for solving the problem of identifying the strength of terminators.

Meanwhile, as one of the essential elements of gene expression, it is known that the terminator can efficiently inhibit the expression of downstream gene, but the effect of terminator on the upstream gene expression was unclearly in prokaryotes. The hairpin structure in RNA can regulate gene expression, for example, a stem-loop structure at the 3′-end of mRNA, which is the transcript from the terminator sequence, can protect mRNA from degradation to certain extent [22,23] and the structure can also protect mRNA from cleavage with RNA degradosome. Therefore, it would be interesting if the terminator can both terminate transcription and still increase the expression of upstream genes.

In this study, based on a dual-reporter gene expression cassette, the effect of terminators on the upstream and downstream gene expression was firstly comprehensively analyzed by comparing the insertion of sequences with and without terminator features. Secondly, in order to obtain detailed characteristics of terminators associated with terminator strength, phenotypes and genotypes of terminator variants from a mutant library were collected through the FlowSeq technique. The correlation between the upstream and downstream fluorescence intensity of each terminator were analyzed. Finally, the discriminant features of terminators analyzed here, including both sequence and thermodynamic characteristics, were applied to construct a terminator strength classification model on the basis of machine learning. The efficacy of the model was further verified by a set of synthetic terminators. The characteristics associated with the terminator strength, and the model established here extended the application of terminators in forward design of gene expression circuits.

2. Materials and methods

2.1. Strains, plasmids and culture conditions

All strains and plasmids involved in this article were listed in Table S1. *Escherichia coli* JM109 was the host strain used for plasmid construction and characterization of the termination efficiency. The terminator-probe plasmid PTK-EGFP-mRFP1 was used as the vector backbone to construct plasmids containing natural terminators, spacer sequences, terminator mutation library and synthetic terminators. The strains were cultured in LB medium (peptone 10.0 g/L, yeast powder 5.0 g/L, NaCl 10.0 g/L) at 37 °C. When a solid medium was required, 2% agar powder was added to the above medium. The concentration of the antibiotic was as followed:50.0 mg/L kanamycin. The 0.5 mmol/L Isopropyl β-D-thiogalactoside (IPTG) was added when reporter genes needed to be expressed.

2.2. Construction of plasmids containing natural terminator, spacer sequence, synthetic terminator and terminator mutation library respectively

All primers were listed in Table S2. Restriction enzymes *Hin*d III, *Eco*R I, *Nhe* I, *Sal* I and T4 ligase were purchased from TAKARA Bio Inc. (Beijing, China). The oligonucleotides were synthesized from GENEWIZ Bioscience Ltd (Suzhou, China). Firstly, the corresponding synthesized forward and reverse oligonucleotide dry powders were diluted to 20 μM. Then 10 μL of each sample were taken, mixed well and placed in the PCR machine, carried out the following procedure: 95 °C for 5 min, 94 °C for 1 min, 93 °C for 1 min, 92 °C 1 min, 91 °C 1 min, 90 °C 1 min. After the reaction over, the sample was put in boiling water immediately and let cool to room temperature. Finally, the finished sample was placed at −20 °C for later use. T4 polyphosphorylase (NEB, the USA) was used to phosphorylate the 5′-end of the sample formed by the above reaction for the next step of the ligation reaction. The linearized plasmid acquired by double digested with restriction enzymes *Hind* III and *EcoR* I, then the linearized plasmid was linked with the fragment overnight at 4 °C using T4 DNA ligase. The recombinant plasmid was transformed into the host *Escherichia coli* JM109 by the calcium chloride procedure. Next, the transformants were picked from the resistant plate and inoculated into LB medium. The sangar sequencing was applied to confirm that the sequence was inserted successfully and correct strains were placed at −80 °C for later use.

The method of constructing the mutant terminator library was as followed. After the two synthetic complementary mutant primers gm-F and gm-R annealed and phosphorylated according to the above steps, the fragment was linked with the linearized plasmid obtained by double digestion with *Nhe* I and *Sal* I. 10 μL of the ligation product was added to 100 μL of competent cells. The cells were heated at 42 °C for 1 min and then 800 mL of SOC medium (peptone 20.0 g/L, yeast powder 5.0 g/L, glucose 5.0 g/L, NaCl 0.5 g/L, KCl 0.186 g/L, $MgSO_4$ 1.2 g/L) was added. After 220 r/min shaking for 1 h at 37 °C, the bacterial solution was directly added to 10 mL of LB liquid medium supplemented with kanamycin, then the cells were cultured with 220 r/min shaking at 37 °C for 12 h. Then the cultured cells received by centrifugation were placed at −80 °C for later use.

2.3. Flow Cytometry analysis

Before performing fluorescence detection with the help of flow cytometry, the sample was needed to prepare for loading. Strains were firstly inoculated into 2 mL LB medium in a 24-well plate at 2% of the inoculum volume, cultured with 600 r/min shaking at 37 °C for 12 h. Then the strain was delivered into 2 mL LB medium added with 1% (v/v) of IPTG and cultured with 600 r/min shaking at 37 °C for 3 h. At last, 1 mL of bacterial solution was taken out and cells were obtained by centrifugation. The collected cells were washed twice with PBS buffer (NaCl 8.0 g/L, KCl 0.2 g/L, Na_2HPO_4 1.44 g/L, KH_2PO_4 0.24 g/L, pH 7.4) and resuspended in an appropriate volume of PBS buffer to an OD_{600} nm between 0.1 and 0.2.

The flow cytometer (Becton Dickinson FACSAria III) was excited at 561 nm and emitted at 610/20 nm to detect mRFP1 fluorescence, and excited at 488 nm and emitted at 530/30 nm to detect EGFP fluorescence. The software FlowJo v10 was used to analyze the data from the flow cytometer and calculate the average fluorescence intensity.

2.4. The measurement of the mRNA level

Firstly, the selected strains were inoculated into a 2 mL LB medium in a 24-well plate and the subsequent culture method was as the same with the procedure in Method 2.3. 1 mL of induced cultured cells was taken out and TaKaRa MiniBEST Universal RNA Extraction Kit (TAKARA Bio Inc) was used to extract RNA from the sample. Then the single strand cDNA was synthesized by the reverse transcription using random primers by PrimeScript™ RT reagent Kit (TAKARA Bio Inc). The Quantitative Real-time PCR (qPCR) was performed following the protocol of the TB Green® Fast qPCR Mix (TAKARA Bio Inc). The primer used to measure the mRNA level of mRFP1 were the mRFP1_q –F and mRFP1_q –R listed in Table S2. Additionally, a custom 16S rRNA primer was used as an endogenous control gene and the calculation method of relative mRNA levels was from Δ*C*t numbers.

2.5. Flow cytometric sorting

The strain JM109-PTK-EGFP-mut_terminator_library-mRFP1 containing terminator mutation library, strain JM09-PTK-EGFP and JM109-PTK-mRFP1 that contained only a single fluorescent reporter gene were respectively inoculated from the sterile tube to LB medium at an inoculum of 2%. The culture method was as the same with the procedure in Method 2.3 and the appropriate concentration of the bacterial solution resuspended in PBS buffer was used for loading. The single positive tubes that only expressed mRFP1 and EGFP were applied to adjust the compensation of fluorescence. Then cells were divided into 7 subgroups according to the difference of the expression of two reporter genes. After removing the cell debris in bin1, the cells in the remaining 6 bins were collected.

2.6. High-throughput sequencing

The collected cells in the 6 subgroups were inoculated into 100 mL LB medium supplemented with kanamycin, cultured with 220 r/min shaking at 37 °C for 8 h, and about 1.5 mL bacterial solution in each bin was collected, and 2 × 250 bp paired-end high-throughput sequencing was carried out by illumine Miseq in Sangon Biotech, Inc. (Shanghai, China).

2.7. Data analysis

The process of extracting the terminator sequence from the fastq file was as followed. The cutadapt (1.2.1) software was used to excise the linker sequence, and prinseqlite (0.19.5) software was used for quality control to ensure that the quality value of each base was greater than 30. The seqkit (0.10.1) software was employed to complete the task of sequence extraction. Meanwhile, a sequence with at least one T in the first three bases of the U-tract site was considered as a terminator sequence and retained. The RNAfold and RNAeval software in the vienna RNA package (2.5.0) toolkit was used to predict the secondary structure and received the free energy ΔG、ΔG_H、ΔG_L of each terminator sequence [10,24,25]. The calculation method of ΔG_U, $\Delta G_U'$, $\Delta G_H/L_H$ and U_{score} was listed in supplementary Note 1. Welch's *t*-test, Student's *t*-test, and Mann-Whitney test were implemented by Python packages scipy (1.6.0) and statsmodels (0.12.2). Machine learning modeling was finished by Python packages numpy (1.19.2), pandas (1.1.3), imblearn (0.0), xgboost (1.3.3) and scikit-learn (0.23.2).

2.8. Characterization of the termination efficiency of synthetic terminator

The characterization method of termination efficiency was referred to previous research [26]. Strains containing different synthetic terminators were cultured according to Method 2.3. Then 150 μL cells were taken out and resuspended by 1 mL PBS buffer to measure the fluorescence intensity and cell-density (OD_{600}). The excitation and emission wavelengths of the EGFP were 488 and 517 nm, and the excitation and emission wavelengths of the mRFP1 were 560 and 650 nm respectively. Cell-density and fluorescence intensity were measured using Infinite 200 Pro multimode microplate reader (Tecan Company, Switzerland), and the fluorescence of each sample was divided by OD_{600} for normalization.

The calculation method of termination efficiency was as follows:

$$\text{termination efficiency}(TE) = 1 - \frac{mRFP1_{ter}/mRFP1_{ref}}{EGFP_{ter}/EGFP_{ref}}$$

It was considered that when a spacer sequence was inserted between upstream and downstream genes as a reference sequence, the ratio of the fluorescence intensity of the downstream gene mRFP1 to the fluorescence intensity of the upstream gene EGFP was 1, in our study, spacer sequence *spacer6* was selected as the reference sequence. When determining the termination strength of the terminator, the background level needed to be subtracted to ensure the accuracy of the termination strength determination.

3. Results and discussion

3.1. The effects of terminator and non-terminator sequences on upstream and downstream gene expression

In order to quantify the effect of terminators on both upstream and downstream gene expression, 9 intrinsic terminators with varied strength, reported by Chen et al. [10] were inserted into terminator-probe plasmids containing two reporter genes (Fig. 1) constructed previously [26]. These terminators were truncated by the same fashion to iliminate the effect of the gene context: 1) the hairpin structure was retained; 2) the A-tract region was truncated to 8 bases before the hairpin structure; 3) and the U-tract region with 12 bases after the terminator hairpin structure was retained. Ten spacer sequences, *spacer1-spacer10*, which were previously designed and tested by Cambray et al. [11], were inserted into probe plasmid instead of terminators, in order to explore how sequences with no terminator characteristics impacted on upstream gene and downstream gene.

We first investigated the effect on upstream gene expression when different 3′-end sequences were inserted, as illustrated in Fig. 2A, although no significant difference of the reporter gene expression was observed between two groups (p value, 0.55822), the variation of upstream gene expression was huge within the terminator group or spacer group (the highest and lowest expression was over 2 times). We further calculated the free energy (ΔG) of the 19 sequences and divided them by the length of sequence to obtain the free energy of per base in each sequence (ΔG/L), smaller the value was, the sequence was easily to form the secondary structure. The result showed in Fig. 2B revealed that there existed an inverse correlation between upstream gene expression and ΔG/L (r=−0.57). This result demonstrated that the hairpin structure might improve the expression of upstream gene. The degradosome was responsible for the majority of 3′ to 5′ exonuclease activity [23,27]. RNase E is an endonuclease that binds unstructured mRNA regions, cleaves mRNA, and recruits the RNA degradosome to the site of cleavage. Therefore, the binding between mRNA and RNase E is a rate-limiting step to mRNA degradation, and mRNA with more stable structure (lower free energy) should have higher capacity to protect mRNA from being degraded. The mRNA sequence and structural determinants related with RNase E binding has been comprehensively studied by Cetnar and his colleagues [28].

The downstream gene expression of the strain with the plasmid inserted with spacer sequence showed varied amount of leaking expression, ranging from 60.9 to 216 a.u., while the mRFP1 expression regulated by the terminators were tightly controlled below 55 a.u., except for the weakest terminator *recA*. These results indicated that downstream gene expressions were significantly different between the terminator group and spacer group (p value, 0.0006), and the insertion

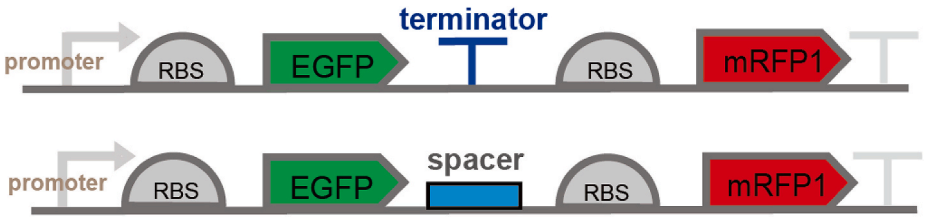

Fig. 1. Schematic diagram of the terminator-probe plasmids. The expression level of upstream gene EGFP and downstream gene mRFP1, with terminators or spacers inserted, were quantified.

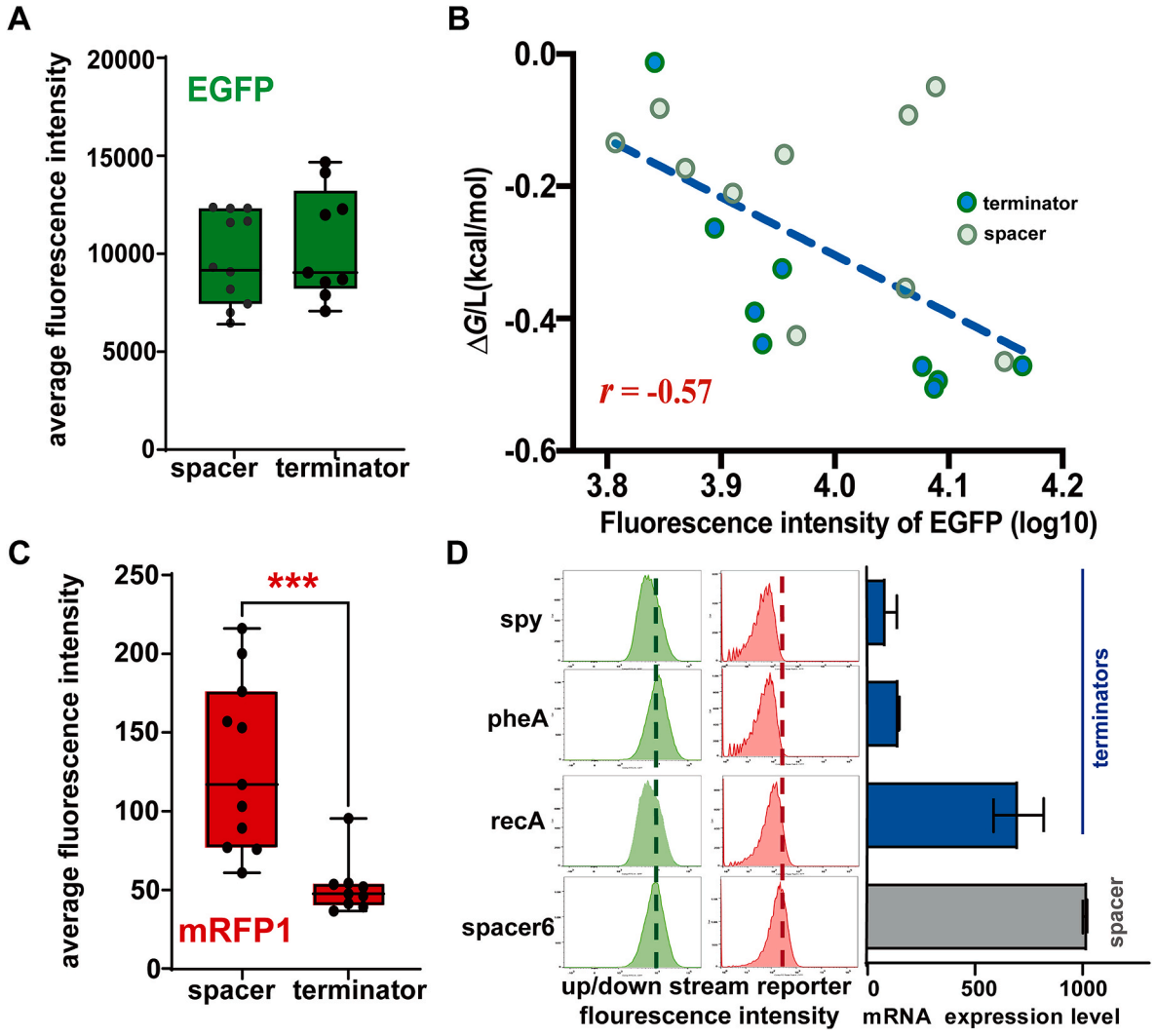

Fig. 2. Comparison of gene expression that located at upstream and downstream of varied intrinsic terminator sequences and spacer sequences. (A and C) The fluorescence intensity of upstream and downstream genes. ***P < 0.001 indicated significant difference (Welch's *t*-test). (B)The correlation analysis between expression of upstream gene and Δ*G*/L. (D) The left figure showed distribution of fluorescence intensity of up/down stream reporter gene regulated by three terminators and a spacer sequence. The right figure indicated the relative transcriptional levels of downstream gene regulated by terminators and spacer sequence.

of the terminator sequence had an obvious inhibition on the expression of downstream gene.

We next quantified the transcription level of downstream gene regulated by three terminators, which were *spy* (strong), *pheA* (medium), *recA* (weak) (Fig. 2D). With the assistance of flow cytometry, the fluorescence of each cell was measured, and the distribution of the cells with the same terminator variant were analyzed. We found that all terminators resulted in decreased downstream gene expression compared to the control (*spacer6*), and the amount of mRNA decreased with the terminator strength. According to our results, the leaking expression of downstream gene was common, even with the strong terminator inserted.

3.2. Rational design and flow cytometric sorting of terminator mutation library

The generic structure of the intrinsic terminator consists with a hairpin structure immediately following the polyuracil structure at 3′-end [10]. A strong terminator *pyrBI* was applied as a template, and M9-M13 and M20-M29 position were partially randomized to create a

terminator mutation library (Fig. 3A). All terminator variants in this study were predicted to form a 6 nt stem and 6 nt loop, or 7 nt stem and 4 nt loop structure. The beginning 3 bases and rear 2 bases of U-tract region were also mutated to create a series of variants with imperfect U tract.

The constructed terminator library was transformed and expressed in *E. Coli* and the cells were further sorted by flow cytometric on the basis of fluorescence intensity of both upstream (EGFP) and downstream (mRFP1) reporter genes of each cell. As can be seen from Fig. 3B, majority of cells showed high fluorescence of the upstream gene while varied fluorescence intensity of the downstream gene, and distributed in quadrant 4 (Supplementary Fig. S3). These cells were sorted to bin 2 to bin 7 differing in EGFP/mRFP1 ratio. A part of cells sparsely distributed on quadrant 2 and 3 that showed significantly low fluorescence intensity of the upstream gene, especially in bin1, percentage of cells was 0.65%. These cells might be the cell debris, and therefore were not taking account for further analysis. The cells in bin 2 to bin 7 were collected and subjected to high-throughput sequencing. The sequences generated by NGS(Next Generation Sequencing) were further screened according to several criteria, including: 1) as the most basic sequence characteristics of terminators summarized by Lesnik et al., the beginning 3 bases in front of U tract must possess one U base [29]; 2) the sequences with only one readdetected was removed. Finally, a total of 808 terminator sequences that meet the criteria were retained, and the average reads number of each sequence was 4.3.

It would be interesting to evaluate if there were any correlations between the up-and down-stream gene expression amongst all terminator variants. Since in lots of cases, the sequences of the same terminator variant were showed up in multiple bins, the fluorescence intensity of each terminator variant was quantified based on its distribution over fractions [13] (weighted score). We calculated the weighted EGFP and mRFP1 fluorescence intensity ($EGFP_t$ and $mRFP1_t$) regulated by each terminator variants. The calculation formula was as follows:

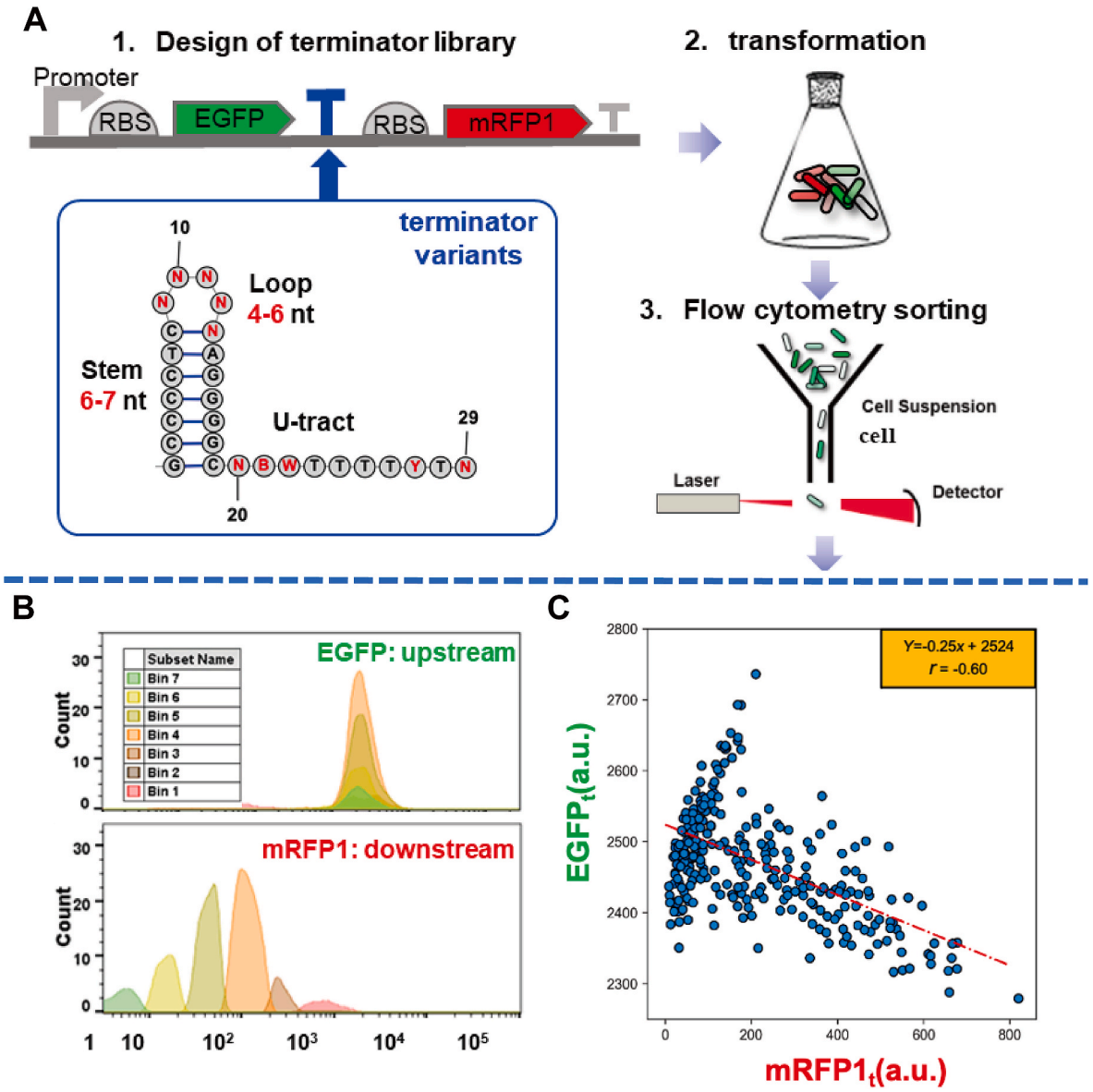

Fig. 3. Schematic diagram of construction of terminator mutation library and the result of flow cytometric sorting. (A) The design of terminator mutation library and the FlowSeq process to obtain the phenotype and genotype characteristics of terminator variants, including the design of library, transformation and cell sorting. The red letter represented the mutant base and the number represented the nucleotide position. (N: A/T/C/G; B:C/G/T; W: A/T; Y:C/T) (B) The distribution of fluorescence of the up/down-stream reporter gene in each bin. (C) The weighted fluorescence intensity of the upstream and downstream genes regulated by each terminator variant.

$$EGFP_t = \sum_{n=2}^{7} \frac{N_{bin\ n}}{N} \times EGFP_{bin\ n}$$

$$mRFP1_t = \sum_{n=2}^{7} \frac{N_{bin\ n}}{N} \times mRFP1_{bin\ n}$$

$N_{bin\ n}$ indicated the sequence number in bin n. $EGFP_{bin\ n}$ and $mRFP1_{bin\ n}$ represented the average fluorescence intensity of EGFP and mRFP1 in bin n.

$EGFP_t$ and $mRFP1_t$ of 225 variants with read number greater than five were analyzed. It can be seen from Fig. 3C that the expression of up and down-stream genes of each variant showed negative correlation (r=−0.60). We speculated that the efficient terminating of downstream transcription prevented unintended transcription of flanking gene sequences, and further accelerated RNA polymerase recycling for subsequent rounds of transcription [30–33]. Meanwhile, in prokaryotes, the process of transcription and translation is highly coupled [34], and therefore, the termination of transcription can avoid the unnecessary occupation of ribosomes. Overall, timely termination of the transcription is beneficial for upstream gene expression due to the higher recycling efficiency of both RNA polymerase and ribosome.

3.3. The sequence and secondary structure of terminators differing in terminating efficiencies

In order to identify the factors determining the terminator efficiency, we screened the variants with significant bin-distribution preference to unravel the sequence characteristics associated with the termination efficiency. Specifically, the variants with majority reads (≥90%) fell into bin 5–bin 7 were considered to be strong terminators, while weak terminators were variants with more than 90% reads fell into Bin 2-Bin 4. Totally, 214 terminator sequences, including 147 strong terminators and 67 weak terminators were obtained for further analysis.

As illustrated in Fig. 4A and B, in the strong terminator group, the first three positions of the U-tract region (M20-M22) showed obvious compositional bias towards T bases, especially the M22 position, the proportion of T bases reached 75.5%, significantly higher than that of the weak terminator group, which was 50.8% (Fig. 4C and D). Beside of M27 and M29, T enrichment was observed in the strong terminator group. At position M13, the nucleotide composition determines if this position is the beginning of the stem or the end of a loop. Specifically, if M13 was a G nucleotide, it would hybridize with the C nucleotide at M8, and contributed to a longer stem. A significant prevalence of A or G were characteristics for weak or strong terminators, respectively, indicating a stable and longer stem structure may favor the termination efficiency.

The stem length of the strong and weak terminators was compared. As described in Fig. 4E, the variants with 6 nt stem accounted for 88.1% of total variants in the weak terminator group, and only 11.9% variants possess 7 nt stem. In contrast, in the strong terminator group, the terminators with 7 nt stem accounted for 34%, significantly higher than that of the weak terminator group. When the stem length of the terminator mutant was 7 nt, the corresponding terminator loop was a tetraloop, which was favor to the stability of the hairpin structure in RNA [35]. The content of T base in the U-tract region also showed bias between two groups. Variants in strong terminator group preferred higher T content in U-tract region, and the variants with over 90% T in U-tract region fell into strong terminator group without any exceptions. Overall, these results indicated that terminators with longer stem and perfect U-tract structure was more likely to efficiently shut down downstream gene expression.

3.4. The influence of thermodynamic parameters on the termination efficiency

The free energy ΔG_T of 214 terminator sequences were analyzed (Fig. 5A), and the terminator variants with relatively low free energy ΔG_T range (−17 to −14 kcal/mol) were more likely gathered in strong terminator group. Overall, the ΔG_T distribution of strong and weak terminators showed significant difference (see Fig. 5B).

The free energy of three structural elements of each variant were further analyzed, including the free energy of the hairpin structure (ΔG_H), the U-tract structure (ΔG_U) and the loop structure (ΔG_L). It can be seen from Fig. 5 C-D, that the free energy of the hairpin structure and loop structure of the strong terminator were significantly lower than those of the weak terminator group. Together, these thermodynamic parameters showed the stability of hairpin structure had a positive effect on terminator efficiency. This result was consistent to the results reported by Chen et al. [10], that there was a correlation between terminator strength (*Ts*) and the free energy of hairpin, and the low ΔG_L of terminators reflected ease of loop formation.

The free energy of U-tract structure was analyzed based on the nearest-neighbor thermodynamic model to quantify the stability of RNA/DNA complex [36]. According to the terminating mechanism

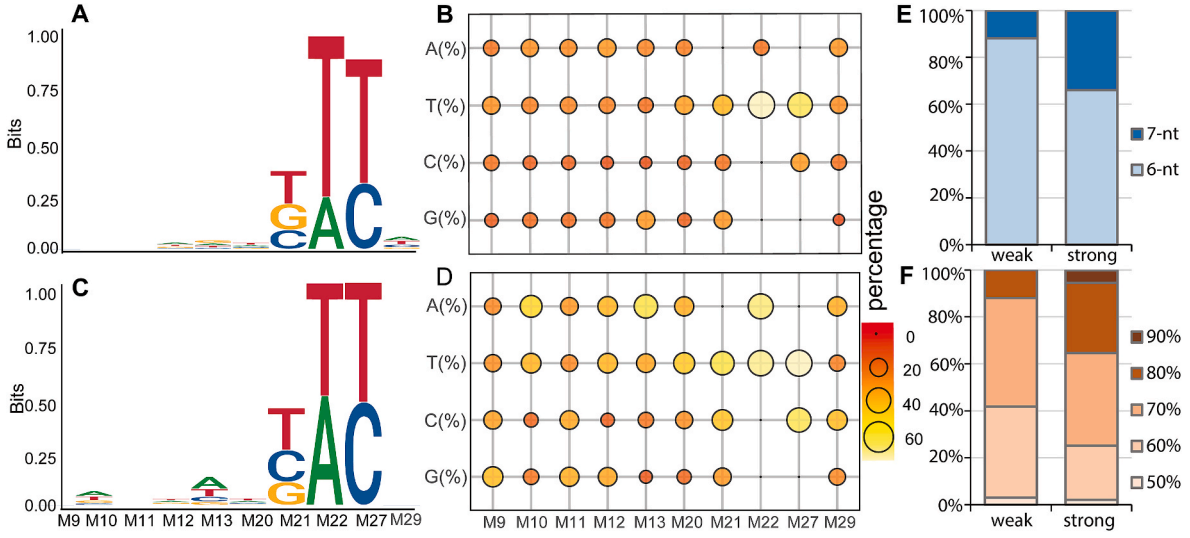

Fig. 4. Terminators with longer stem and more T bases in the U-tract region were more likely to become strong terminators. (A and C) The sequence conservation analysis at each position of variants in strong terminator and weak terminator group. (B and D) The frequency of each nucleotide in each position in strong and weak terminator groups. The dots with larger size and brighter color indicated the higher frequency. The proportion of 6- and 7-nt stem (E) and the content of T base in the U-tract region (F) in weak and strong terminator group.

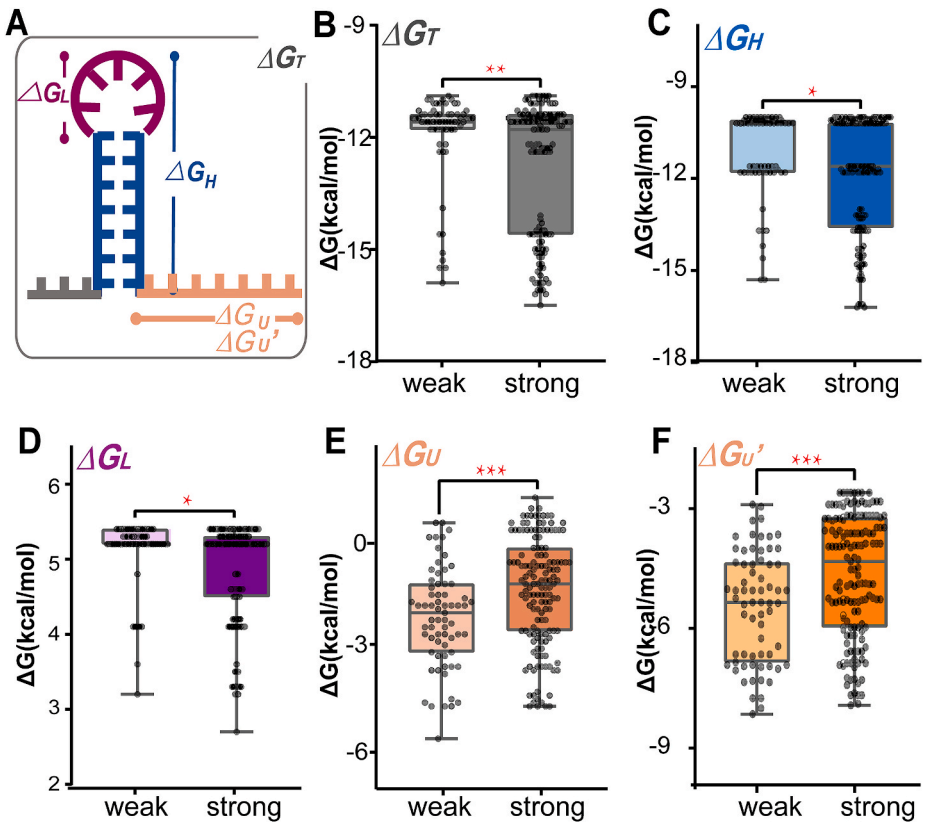

Fig. 5. Thermodynamic parameters of the terminators between weak and strong terminator groups. ΔG_T represented the free energy of entire terminator. ΔG_H, ΔG_L *and* ΔG_U represented the free energy of hairpin structure, loop structure and U-tract of terminator, respectively. $\Delta G_U'$ reflected the weighted free energy of the 15 bp broad U-tract region included U-tract and its downstream 7 bp. Mann-Whitney test was used for statistical analysis, and $P < 0.05$ was considered to be significant (*$P < 0.05$, **$p < 0.01$, ***$p < 0.001$).

described as hybrid shearing model, the transcription termination process started with the transcription of the U-tract [37], and the RNA polymerase was in a suspended state in response to the formation of U-tract structure. The pause time allowed the formation of terminator hairpin secondary structure [38], resulted in the dissociation of transcription elongation complex (TEC), and eventually mRNA released from the templated DNA strand. In this process, high energy of the RNA-DNA hybrid facilitated the dissociation of coding DNA strand from the newly generated RNA strand [2,23,39]. Comparing with the weak terminator group, the overall ΔG_U of the strong terminator group was higher (Fig. 5E). Combined with the difference of T content between two terminator groups (Fig. 4C), it can be concluded that the higher T content in U tract region associated with the higher free energy, and resulted in ease of disassociation of RNA/DNA complex, is beneficial for transcription termination.

According to the results reported by wan et al. [40], the U enrichment in 5′-region in U-tract is more important than that in 3′-region. However, the ΔG_U may not reflect the unequal contribution of each position. Therefore, a weighted free energy of U region, noted as $\Delta G_U'$, with position information of U-nucleotide in the broad U-tract region (a total length of 15bp, with U-tract and downstream 7 bp) taking into consideration. The calculation method was described in Supplementary Note 1. Similar with ΔG_U, strong terminator had higher $\Delta G_U'$ than weak terminator, indicating that the T base in the U-tract region of strong terminators was closer to the 5′-end, which facilitated the dissociation of RNA from template DNA strands.

3.5. *The establishment and validation of terminator strength prediction model*

In this study, since more characteristics describing the sequence and secondary structure of terminators were found to be significantly different from terminators with varied strength, a sophisticated classification model was constructed, the features used to train the model included both thermodynamic and sequence parameters. Five thermodynamic features obtained above, including ΔG_T, ΔG_H, ΔG_L, ΔG_U and $\Delta G_U'$ were applied as input features. Three sequence features, including stem-length, T-ratio in U-tract and the nucleotide in M22 were applied as input features. Two specially generalized features were added to the feature set to improve the prediction accuracy, including $\Delta G_H/L_H$ (the ratio of free energy of the hairpin structure and the length of sequence) and U_{score} (the U distribution over the first 15 bp after the hairpin, supplementary Note 1). Meanwhile, the 3-mers frequency in terminator sequence was also taken into consideration (totally there are 64 3-mers, such as AAA, AAT, …, GGG) [41].

The terminator is part of the 5′-UTR of downstream gene, therefore, the influence of terminator sequence on the translation of downstream gene was needed to be verified. RBS calculator [42] (version 2.1) was used to predict the translational initiation rates of the downstream gene that varying the intergenic region inserted. 60 terminators were randomly chosen to predict the translation initiation rate (supplement Note 3). The results showed that in our system, the TIR of all these terminators showed the exact same value. These results indicated that probably due to the long distance between the terminator and the start codon of the downstream gene, the insertion of the terminator sequence had no effect on the translation initiation of the downstream reporter gene.

To solve the problem of sample-imbalance, the comprehensively sampling method based on the SMOTE and Tomek Links algorithms [43,44] was adopted to obtain the input samples, which is an expanded data set of initial input data set (in this case, it is the weak terminators). The resampled data was used to establish the model, the main process was as followed: the XGBoost algorithms combined with random search was used to find the best hyperparameters for modeling and the method of 5-fold cross validation was utilized for the model evaluation (Fig. 6). Finally, in order to further verify the efficacy of the model in discriminating terminator strength, we designed 20 new terminator variants based on the template of the terminator variant showed in Fig. 3A and

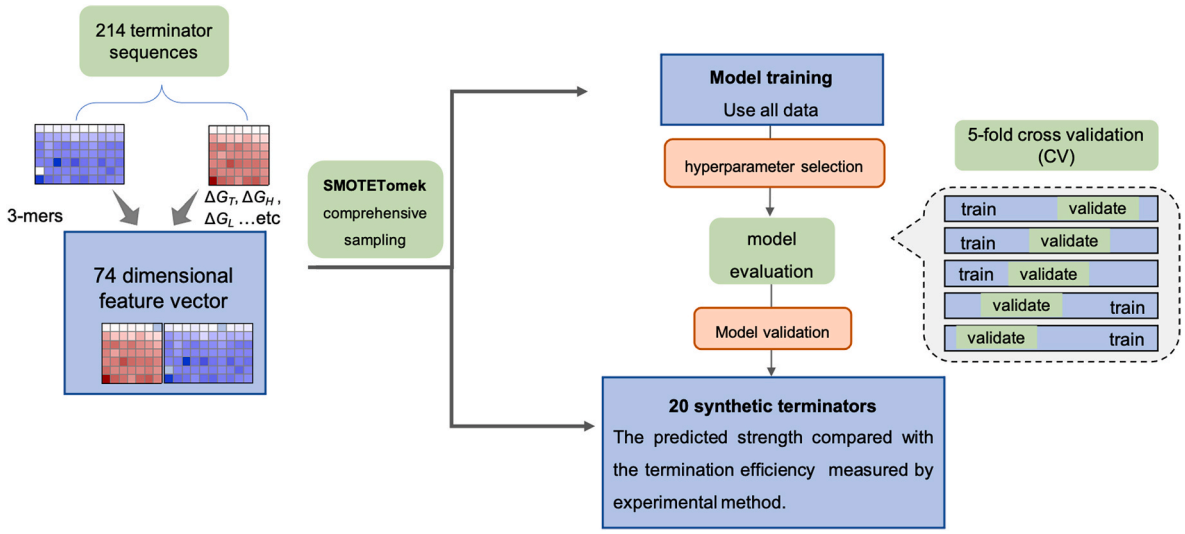

Fig. 6. Schematic diagram of establishing the model.

the predicted strength was further compared with the actual termination efficiency.

As shown in Fig. 7B, the average accuracy, precision, recall, f1-score of the classification model in 5-fold cross validation was 0.956, 0.90, 1.0 and 0.946 respectively. The results demonstrated that the model have good predictive performance to discriminated between the weak and strong terminators from our dataset. The feature importance ranking returned by the algorithm was exhibited in supplementary Note 3. According to the above established model, 8 and 12 terminators were classified into weak and strong terminator classes respectively. The fluorescence intensity of up/down stream reporter genes and termination efficiency of each terminator variant were quantified experimentally and the results showed that the negative correlation between up and downstream gene expression were also observed ($r = -0.617$). Student' s *t*-test showed that the significant differences were present between the predicted weak and strong terminator groups as to the expression of downstream reporter gene and termination efficiency. Among them, the fluorescence intensity of the downstream reporter gene of sequences in the strong terminator group was generally low, and the termination efficiency was at high state in general, indicating that our model had a certain advantage in the generalization of the terminator design. However, an overlap between the predicted weak and

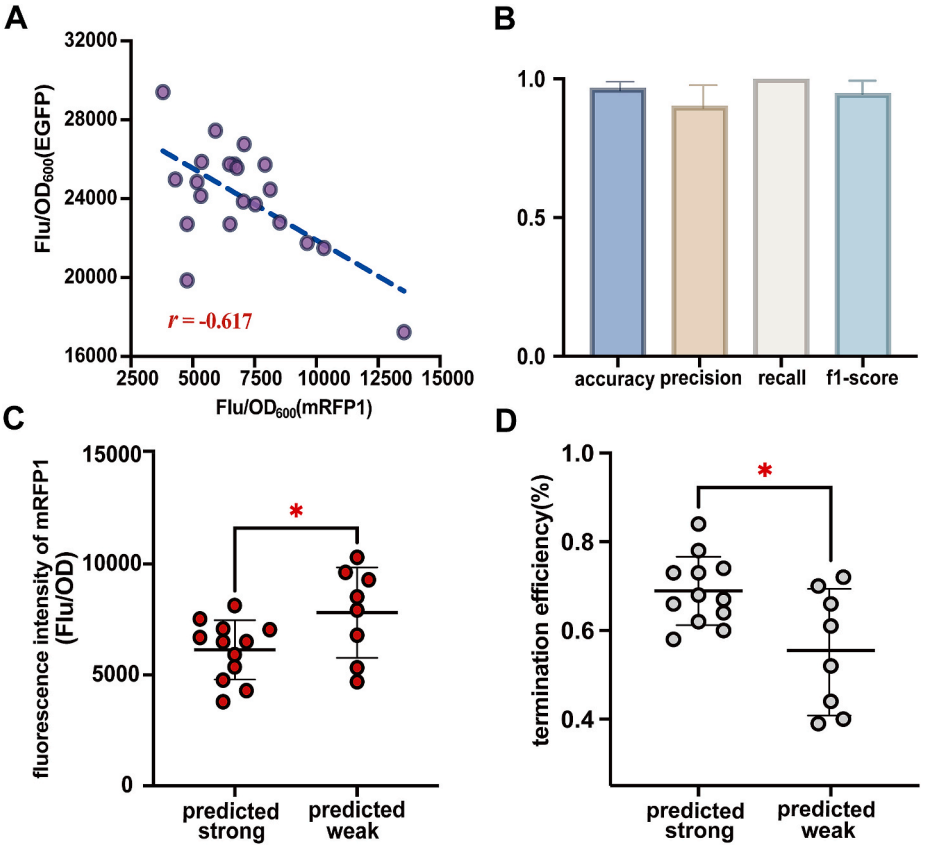

Fig. 7. The results of the terminator strength classification model and the experimental verification. (A) The correlation analysis of relative fluorescence intensity of upstream and downstream reporter genes of 20 synthetic terminators. (B)The average accuracy, precision, recall and f1-score of 5-fold cross validation trained by the XGBoost algorithm. (C and D) The comparison of downstream reporter gene ($p = 0.0387$, two independent sample T-test) and termination efficiency ($p = 0.0109$, two independent sample T-test) between terminator groups that classified by the model as strong or weak. $P < 0.05$ was considered to have significant difference.

strong terminators of termination efficiency were observed, especially in the range of 60–70%, reflecting that it was still challenging in discriminateing the terminators with moderate strength. The TransTermHP [45] was further used to predict the 20 sequences we designed here, and the results showed that 18 and 2 sequences were identified as terminator and non-terminator sequences respectively. According to the prediction results of the classification model established here, among the above 18 sequences identified as terminators based on the TransTermHP model, 12 of them were identified as strong terminator sequences and 6 of them were identified as weak terminator sequences. The other two sequences identified as non-terminators were also predicted as weak terminator sequences based on the model established here. Overall, the output of the two models are mostly consistent. (Supplementary Note 5).

It is worth to mention that the distance between the stop codon of the upstream gene and the terminator significantly affects the termination efficiency of the terminator [30]. In this study, the terminators are located at 12 base pairs downstream of the upstream gene. It is possible that by adjusting the terminators more far away from the stop codon of the upstream gene may improve the termination efficiency.

Overall, the method used for classifying the strength of terminators is a useful tool, and according to the prediction model established here, we can roughly discriminate the strong and weak terminators. Compared with binary classification model, a multi-class model for terminator strength prediction would be more useful. However, significantly larger dataset would be necessary for multi-class modeling. Moreover, minimizing the influence of varied physiological status of the cells (variants), and the noise of the fluorescence measurement would significantly improve the accuracy of phenotype information collected, which would further facilitate advanced model to discriminate terminator strength. Beside of the discrimination model established here, the features used for model establishment include sequence features in addition to thermodynamic features is an improvement compared with that reported in the previous studies [10], and these key features of terminators identified here should be paid close attention to for rational design of terminators.

4. Conclusions

The availability of terminators with varied strength, and the understanding of the sequence-activity relationship of terminators are important for forwards design of gene expression operons. Here, we first compared the up and down stream gene expression regulated by the sequences with and without terminator structure features, and confirmed that the hairpin structure present in 3′-end can increase the expression of upstream genes, while the insertion of terminator sequences can further inhibit the expression of downstream genes. On the basis of mutant terminator library, thermodynamic and sequence features differing in the terminating efficiency were analyzed using the FlowSeq technique. An inversely proportional relationship was displayed between the expression levels of upstream and downstream genes, revealing that the shutdown of downstream gene was benefit for the efficient expression of upstream gene. A machine learning model based on XGBoost algorithm was established for discriminating the terminator strength based on the features analyzed above, and the efficacy of the prediction model was further validated by a new set of synthetic terminators. The information about the characteristics associated with the terminator strength deepened the understanding of terminator functionality and facilitate the rational design of terminators.

CRediT authorship contribution statement

Weiji Zhai: Data curation, Formal analysis, Software, Visualization, Writing – original draft. **Yanting Duan:** Validation. **Xiaomei Zhang:** Funding acquisition, Reviewing. **Guoqiang Xu:** Reviewing. **Hui Li:** Reviewing. **Jinsong Shi:** Investigation, Resources. **Zhenghong Xu:** Project administration. **Xiaojuan Zhang:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare no competing financial interest.

Acknowledgements

This study was supported by the National Key Research and Development Program of China [2018YFA0900300], National Natural Science Foundation of China [32171421].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2022.06.003.

References

[1] Madigan, M. T.; Martinko, J. M.; Parker, J., Brock biology of microorganisms/M.T. Madigan, J.M. Martinko, J. Parker.
[2] Santangelo TJ, Artsimovitch I. Termination and antitermination: RNA polymerase runs a stop sign. Nat Rev Microbiol 2011;9(5):319–29.
[3] Scheer H, De Almeida C, Sikorska N, Koechler S, Gagliardi D, Zuber H. High-Resolution mapping of 3' extremities of RNA exosome substrates by 3' RACE-seq. Methods Mol Biol 2020;2062:147–67.
[4] Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R. Rho directs widespread termination of intragenic and stable RNA transcription. Proc. Natl. Acad. Sci. U.S.A 2009;106(36):15406–11.
[5] Graham JE, Richardson JP. Rut sites in the nascent transcript mediate rho-dependent transcription termination in vivo. J Biol Chem 1998;273(33):20764–9.
[6] d'Aubenton Carafa Y, Brody E, Thermes C. Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. J Mol Biol 1990;216(4):835–58.
[7] Rosenberg M, Court D. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu Rev Genet 1979;13:319–53.
[8] Wilson KS, Vonhippel PH. Transcription termination at intrinsic terminators - the role of the RNA hairpin. Proc. Natl. Acad. Sci. U.S.A 1995;92(19):8793–7.
[9] Larson MH, Greenleaf WJ, Landick R, Block SM. Applied force reveals mechanistic and energetic details of transcription termination. Cell 2008;132(6):971–82.
[10] Chen Y-J, Liu P, Nielsen AAK, Brophy JAN, Clancy K, Peterson T, Voigt CA. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat Methods 2013;10(7):659–+.
[11] Cambray G, Guimaraes JC, Mutalik VK, Lam C, Quynh-Anh M, Thimmaiah T, Carothers JM, Arkin AP, Endy D. Measurement and modeling of intrinsic transcription terminators. Nucleic Acids Res 2013;41(9):5139–48.
[12] Cui W, Lin Q, Hu R, Han L, Cheng Z, Zhang L, Zhou Z. Data-driven and in silico-assisted design of broad host-range minimal intrinsic terminators adapted for bacteria. ACS Synth Biol 2021;10(6):1438–50.
[13] Sauer C, van Themaat EVL, Boender LGM, Groothuis D, Cruz R, Hamoen LW, Harwood CR, van Rij T. Exploring the nonconserved sequence space of synthetic expression modules in Bacillus subtilis. ACS Synth Biol 2018;7(7):1773–84.
[14] Evfratov SA, Osterman IA, Komarova ES, Pogorelskaya AM, Rubtsova MP, Zatsepin TS, Semashko TA, Kostryukova ES, Mironov AA, Burnaev E, Krymova E, Gelfand MS, Govorun VM, Bogdanov AA, Sergiev PV, Dontsova OA. Application of sorting and next generation sequencing to study 5 '-UTR influence on translation efficiency in Escherichia coli. Nucleic Acids Res 2017;45(6):3487–502.
[15] Komarova ES, Chervontseva ZS, Osterman IA, Evfratov SA, Rubtsova MP, Zatsepin TS, Semashko TA, Kostryukova ES, Bogdanov AA, Gelfand MS, Dontsova OA, Sergiev PV. 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 2020;13(4):1254–61.
[16] Osterman IA, Chervontseva ZS, Evfratov SA, Sorokina AV, Rodin VA, Rubtsova MP, Komarova ES, Zatsepin TS, Kabilov MR, Bogdanov AA, Gelfand MS, Dontsova OA, Sergiev PV. Translation at first sight: the influence of leading codons. Nucleic Acids Res 2020;48(12):6931–42.
[17] Yoon B-J. Hidden markov models and their applications in biological sequence analysis. Curr Genom 2009;10(6):402–15.
[18] Hu S, Ma R, Wang H. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. PLoS One 2019;14(11).
[19] Ding N, Yuan Z, Zhang X, Chen J, Zhou S, Deng Y. Programmable cross-ribosome-binding sites to fine-tune the dynamic range of transcription factor-based biosensor. Nucleic Acids Res 2020;48(18):10602–13.
[20] Zhao M, Yuan Z, Wu L, Zhou S, Deng Y. Precise prediction of promoter strength based on a de novo synthetic promoter library coupled with machine learning. ACS Synth Biol 2022;11(1):92–102.
[21] Wang Y, Wang H, Wei L, Li S, Liu L, Wang X. Synthetic promoter design in Escherichia coli based on a deep generative network. Nucleic Acids Res 2020;48(12):6403–12.

[22] Rauhut R, Klug G. mRNA degradation in bacteria. FEMS Microbiol Rev 1999;23(3):353–70.
[23] Ray-Soni A, Bellecourt MJ, Landick R. Mechanisms of bacterial transcription termination: all good things must endKornberg RD, editor. Annual Review of Biochemistry 2016;85:319–47.
[24] Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U.S.A 2004;101(19):7287–92.
[25] Lorenz R, Bernhart SH, Siederdissen CHZ, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithm Mol Biol 2011;6.
[26] He Z, Duan Y, Zhai W, Zhang X, Shi J, Zhang X, Xu Z. Evaluating terminator strength based on differentiating effects on transcription and translation. Chembiochem 2020;21(14):2067–72.
[27] Hui MP, Foley PL, Belasco JG. Messenger RNA degradation in bacterial cellsBassler BL, editor. Annual Review of Genetics 2014;48:537–59.
[28] Cetnar DP, Salis HM. Systematic quantification of sequence and structural determinants controlling mRNA stability in bacterial operons. ACS Synth Biol 2021;10(2):318–32.
[29] Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic Acids Res 2001;29(17):3583–94.
[30] Li R, Zhang Q, Li J, Shi H. Effects of cooperation between translating ribosome and RNA polymerase on termination efficiency of the Rho-independent terminator. Nucleic Acids Res 2016;44(6):2554–63.
[31] Liu X, Jiang H, Gu Z, Roberts JW. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proc. Natl. Acad. Sci. U.S.A 2013;110(29):11928–33.
[32] Berkemer SJ, Maier LK, Amman F, Bernhart SH, Wortz J, Markle P, Pfeiffer F, Stadler PF, Marchfelder A. Identification of RNA 3 ' ends and termination sites in Haloferax volcanii. RNA Biol 2020;17(5):663–76.
[33] Hudson AJ, Wieden H-J. Rapid generation of sequence-diverse terminator libraries and their parameterization using quantitative Term-Seq. Synthetic Biology 2019;4(1).
[34] McGary K, Nudler E. RNA polymerase and the ribosome: the close relationship. Curr Opin Microbiol 2013;16(2):112–7.
[35] Varani G. Exceptionally stable nucleic-acid hairpins. Annu Rev Biophys Biomol Struct 1995;24:379–404.
[36] Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M, Sasaki M. Thermodynamic parameters to predict stability of rna/dna hybrid duplexes. Biochemistry 1995;34(35):11211–6.
[37] Gusarov I, Nudler E. The mechanism of intrinsic transcription termination. Mol Cell 1999;3(4):495–504.
[38] Toulokhonov I, Landick R. The flap domain is required for pause RNA hairpin inhibition of catalysis by RNA polymerase and can modulate intrinsic termination. Mol Cell 2003;12(5):1125–36.
[39] Peters JM, Vangeloff AD, Landick R. Bacterial transcription terminators: the RNA 3 '-End chronicles. J Mol Biol 2011;412(5):793–813.
[40] Wan XF, Xu D. Intrinsic terminator prediction and its application in Synechococcus sp. WH8102. J Comput Sci Technol 2005;20(4):465–82.
[41] Pan XY, Yang Y, Xia CQ, Mirza AH, Shen HB. Recent methodology progress of deep learning for RNA-protein interaction prediction. Wiley Interdisciplinary Reviews-Rna 2019;10(6).
[42] Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol 2009;27(10):946–50.
[43] Prati RC, Batista G, Monard MC. Learning with class skews and small disjunctsBazzan ALC, Labidi S, editors. Advances in Artificial Intelligence - Sbia 2004 2004;3171:296–306.
[44] Sun Y, Castellano CG, Robinson M, Adams R, Rust AG, Davey N. Using pre & post-processing methods to improve binding site predictions. Pattern Recogn 2009;42(9):1949–58.
[45] Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol 2007;8(2).