Measurement and modeling of intrinsic transcription terminators

Guillaume Cambray, Joao Guimarães, Vivek K Mutalik, Colin Lam, Quynh-Anh Mai, Tim Thimmaiah, James M Carothers, Adam P Arkin, Drew Endy

Published online 19 March 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 9* ***5139–5148***
doi:10.1093/nar/gkt163

Measurement and modeling of intrinsic transcription terminators

Guillaume Cambray[1,2], Joao C. Guimaraes[1,3,4], Vivek K. Mutalik[1,2,5], Colin Lam[1,2], Quynh-Anh Mai[1,2], Tim Thimmaiah[3], James M. Carothers[5], Adam P. Arkin[1,2,3,5,*] and Drew Endy[1,6,*]

[1]BIOFAB International Open Facility Advancing Biotechnology (BIOFAB), 5885 Hollis Street, Emeryville, CA 94608, USA, [2]California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720, USA, [3]Department of Bioengineering, University of California, Berkeley, CA 94720, USA, [4]Department of Informatics, Computer Science and Technology Center, University of Minho, Campus de Gualtar, 4700 Braga, Portugal, [5]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and [6]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA

Received January 13, 2013; Revised February 19, 2013; Accepted February 20, 2013

ABSTRACT

The reliable forward engineering of genetic systems remains limited by the *ad hoc* reuse of many types of basic genetic elements. Although a few intrinsic prokaryotic transcription terminators are used routinely, termination efficiencies have not been studied systematically. Here, we developed and validated a genetic architecture that enables reliable measurement of termination efficiencies. We then assembled a collection of 61 natural and synthetic terminators that collectively encode termination efficiencies across an ~800-fold dynamic range within *Escherichia coli*. We simulated co-transcriptional RNA folding dynamics to identify competing secondary structures that might interfere with terminator folding kinetics or impact termination activity. We found that structures extending beyond the core terminator stem are likely to increase terminator activity. By excluding terminators encoding such context-confounding elements, we were able to develop a linear sequence-function model that can be used to estimate termination efficiencies ($r = 0.9$, $n = 31$) better than models trained on all terminators ($r = 0.67$, $n = 54$). The resulting systematically measured collection of terminators should improve the engineering of synthetic genetic systems and also advance quantitative modeling of transcription termination.

INTRODUCTION

The ability to rationally engineer gene expression systems underlies all cellular biotechnologies. Synthetic biology researchers, in seeking to scale the engineering of biology to genome-scale systems, are pursuing the development of self-consistent collections of well-characterized genetic components that can be reused reliably (1–4). Towards this goal, many efforts have studied libraries of natural and synthetic genetic elements regulating various aspects of gene expression, and analyzed part performance via sequence-function models [e.g. (5–7)]. However, most projects have focused on engineering elements that control transcription and translation initiation (8,9). Additional work to engineer genetic elements that regulate remaining aspects of gene expression is needed. For example, transcription terminators are known to play key roles in regulating natural genetic systems and have recently been used to implement synthetic genetic logic (10–13). Methods for measuring, modeling and standardizing terminator elements would thus support both future synthetic biology research and applications.

Transcription termination in *Escherichia coli* is known to occur via two distinct mechanisms: factor-dependent or factor-independent termination. Factor-dependent termination relies on the destabilization of transcription complexes by a regulatory protein, Rho, at Rho-dependent terminator sequences. A recent study showed that the Rho protein is responsible for ~20% of termination events in *E. coli* (14). However, the exact sequence features and steps of Rho recruitment and function are

*To whom correspondence should be addressed. Tel: +01 650 723 7027; Fax: +01 650 721 6602; Email: endy@stanford.edu
Correspondence may also be addressed to Adam Arkin. Tel: +01 510 495 2366; Fax: +01 510 486 6219; Email: aparkin@lbl.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

not understood well enough for use in synthetic genetic systems. Alternatively, factor-independent termination, which accounts for the remaining ~80% of transcription termination events in *E. coli,* occurs at defined sequence regions known as intrinsic terminators that can be encoded as reusable genetic elements (15).

Sequence features within intrinsic terminators have been well studied in *E. coli* and include a core GC-rich stem of 5–9 nt that is closed by a short 3–5 nt loop and followed by a 7–9 nt U-rich tail (Figure 1A) (10,16). A few intrinsic terminators have been extensively studied *in vitro*, resulting in mechanistic models for how individual sequence motifs contribute to overall termination efficiency (15,17). From these foundational studies, computational methods have been developed to identify putative terminator elements within natural DNA sequences. Such tools have improved the automated annotation of genome sequences and reshaped consideration of operon structure and chromosome organization (16,18–22). However, sequences that match putative terminator motifs are pervasive within natural genomes, and most computational predictions are not validated experimentally, thereby hindering the iterative development of improved terminator identification tools.

The reliability and reuse of termination efficiency measurements has also been challenged by the fact that terminator elements themselves can impact mRNA stability (23), translation initiation and translation polarity (24). Thus, a measurement for termination efficiency in one genetic context may not match a measurement obtained in another context. Furthermore, the use of diverse characterization strategies—*in vitro* (25), including single molecule approaches (26) versus *in vivo* (23) and single versus dual reporters (27)—has hindered comparison of measurements and sequence-function analyses (16). Hence, a systematic method for measuring sequence distinct terminator elements that avoids confounding effects arising from terminator elements themselves is needed. Such an approach would enable more reliable characterization and reuse of transcription terminator collections across laboratories.

MATERIALS AND METHODS

Terminator sequences

We selected 24 terminator elements identified in previous studies (Supplementary Table S1): 10 from natural expression cassettes (*crp*, *his*, *ilv*, *rnpB*, *rpoC*, *tonB*, three variants of *trp* from *E. coli* and *amyA* from *Bacillus subtilis*), four from non-protein coding RNAs in *E. coli* (*rrnA*, *rrnB*, *rrnD* and *rna1*), six from bacteriophage (T3, T7, T21, M13 and two from lambda), two from mobile genetic elements (*tet* from transposon *tn10*, and the attachment site motif from the *aadA7* integron cassette) and two synthetic terminators (BBa_B1002 and BBa_B1006). For terminators sourced from natural sequences, we included 30 nt of upstream and downstream sequence context. We also generated 11 minimal terminators from a subset of the natural elements (*crp*, *his*, *ilv* lambda, M13, *rnpB*, *rpoC*, *rrnB*, *rrnD*, *tonB* and *trp*). We designed six variants of the BBa_B1006 synthetic terminator, altering features such as U tail length and stem composition. Altogether, these seeded a diverse panel of 41 putative terminator elements. We also retained and studied 13 variants to stem or loop sequences that arose during construction. We constructed seven double terminators by concatenating some of the aforementioned elements, finalizing a set of 61 candidate terminator elements (Supplementary Table S1) that we characterized in detail via the RIIIG measurement device (later in the text). Sequence

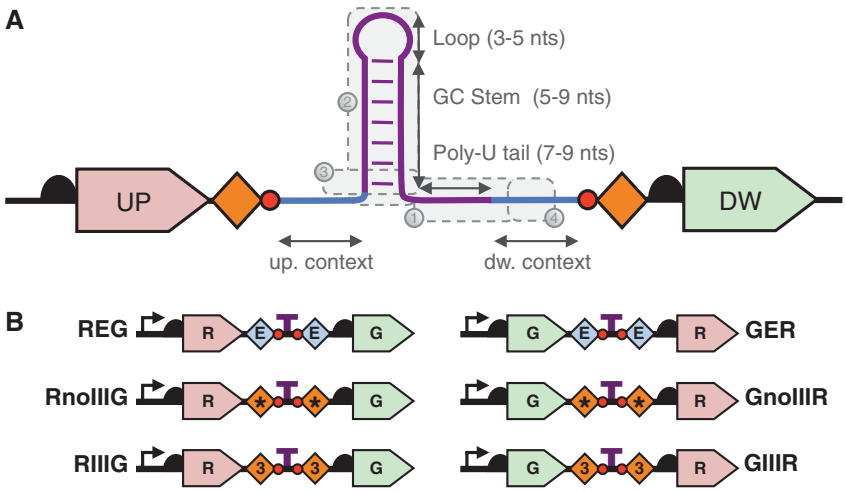

Figure 1. Architecture of a standardized genetic device for termination efficiency measurements. (**A**) Anatomy of an intrinsic terminator (purple) and generic architecture of processed mRNA originating from a terminator measurement device. RNase recognition sites (orange diamonds) are intended to standardize the 3′- or 5′-ends of processed mRNA encoding upstream (UP, red) and downstream (DW, green) reporter genes. The four features selected in our best quantitative model of termination efficiencies (main text), numbered by decreasing importance (grey regions: 1 = TTHP_utail_score; 2 = hp_norm_dg; 3 = closing_stackGC; 4 = dna_dna_pattern). (**B**) Six terminator measurement device variants tested here. Green (G, green box) and red (R, red box) fluorescent reporter coding sequences bracket a terminator (purple T) test site flanked by RNase E sites (E, blue diamonds), RNase III sites (3, orange diamonds) or non-functional RNase III sites (*, orange diamonds).

comparisons of related elements in this set are provided (Supplementary Figure S1).

Plasmids and strains

We used pFAB270 as a template plasmid for terminator library construction by inverse polymerase chain reaction (Supplementary Materials and Methods). We developed the set of candidate terminator measurement devices using the pFAB511 and pFAB512 vector backbones. Terminators propagated within pFAB270 can be moved into measurement plasmids (or any compatible plasmid) using Golden Gate cloning (28) (Supplementary Figure S2 and Supplementary Table S2). We used *E. coli* strain BW25113 for construction and testing. Specific constructs, resulting strains, primers and detailed genetic assembly procedures are given (Supplementary Tables S2–S4, Supplementary Material).

Termination efficiency calculations

Termination efficiency (TE) quantifies the fraction of arriving transcription elongation complexes that do not pass through a candidate terminator element. For example, an element that disrupted all arriving transcription complexes would have a TE of 100%. Expressed fluorescent reporter protein levels are not a direct measure of TE. Instead, fluorescence levels are used to estimate terminator read-through (TR) rates from observed ratios of downstream (F_{DW}) to upstream (F_{UP}) fluorescence intensities:

$$TR = F_{DW}/F_{UP} \quad (1)$$

Because mRNA stability, translation efficiency and the intrinsic brightness of the two reporters are different, we established a reference read-through value (TR_{REF}) using a standardized test sequence that was selected to encode maximum read-through while not itself initiating transcription (Supplementary Figure S3). We then normalized all TR measurements:

$$TR_{NORM} = TR/TR_{REF} \quad (2)$$

and estimated TE as a percentage via:

$$TE = 100 \times (1 - TR_{NORM}) \quad (3)$$

We calculated TE from single-cell fluorescence data (TE_{CELL}) and also from reconstructed population average data (TE_{BULK}) based on the same single-cell measurements (Supplementary Figure S4).

Termination efficiency measurements

We used the RIIIG (pFAB763) measurement device to observe and estimate TE values from fluorescence measurements. We screened for and established a reference control sequence that resulted in highest expression of the downstream *gfp* yet did not itself initiate transcription [Equation (2), Supplementary Figure S3]. The activity of the reference construct was observed in parallel with every assay. Cell cultures were grown to mid-exponential phase in deep 96-well plates in rich medium supplemented by kanamycin (Supplementary Material). Single-cell green and red fluorescence intensities were measured using an automated Guava® EasyCyte flow cytometer (EMD Millipore, Hayward, CA, USA). Raw data were filtered using an automated gating strategy (29) to ensure consistent distributions of TR_{CELL} ratio [Equation (1), Supplementary Figure S4]. Cell populations exhibiting multimodal fluorescence distributions were flagged, with individual colonies re-validated by sequencing and the entire assay repeated as necessary to produce consistent unimodal behavior and measurements. All terminator elements were measured in triplicate.

Terminator structure dynamic folding models

We computed the folding kinetics of nascent RNA molecules encoding terminator elements using the kinefold_long_static binary (30) on a 192 node Linux cluster (31). For each sequence, S, the predicted terminator folding frequency, f(S), was taken as the fraction of elongating transcripts with a terminator part subsequence, $T = S_jS_{j+1}\ldots S_k$ (i.e. the subsequence of S ranging from position j to position k), folding into the target structure, sT, at any given time, t. Target structures (sT) were defined as the equilibrium minimum free-energy stem and hairpin-loop secondary structure, determined from melt-and-anneal folding simulations at t = 60 s for subsequence T alone (i.e. in isolation from any upstream 5′ flanking subsequence, $F = S_iS_{i+1}\ldots S_j$); simulations at t beyond 60 s did not affect the target sT. We determined f(S) from 100 stochastic co-transcriptional folding simulations initiated with random seeds. The RNA polymerase elongation rate (kpol) was set to 25 nt s^{-1} (32,33) with a minimum helix energy of −6.346 kcal mol^{-1} (31). Within an elongating transcript, the sequence between RNA polymerase and the first translating ribosome can fold into distinct secondary and tertiary structures. F represents this RNA 'window' sequence between RNA polymerase and the first translating ribosome. Considering the elongation rate aforementioned and a translation initiation rate of 0.7 s^{-1}, F would span ~50 nt (34). Given uncertainty regarding the best window sequence lengths, we performed simulations across various lengths (F = 25, 50, 75 and 100 nt). We used six simulation times to represent pausing of the elongation complex at the U tail (17), as follows: t_1 = [length(F) + length(T)]/kpol; $t_2 = t_1 + 0.5$ s; $t_3 = t_1 + 1$ s; $t_4 = t_1 + 10$ s; $t_5 = t_1 + 20$ s; $t_6 = t_1 + 30$ s. We then defined the average of all of the folding frequencies over time and for all sizes of F as the estimated measure of folding efficiency (Supplementary Table S5).

Sequence feature modeling of terminator activity

We used a multiple linear regression model to relate measured TEs to up to 12 sequence features suspected to impact transcription termination (Supplementary Table S6):

$$TE = \beta_0 + \Sigma_{i=1..j}\,\beta_i X_i + \varepsilon \quad (4)$$

where, β_0 is the regression intercept, *i* is one of *j* sequence features, β_i and X_i are regression coefficient and value for the *i*th variable, respectively, and ε is the error term. We used stepwise regression with forward selection to find the

variables with higher explanatory power. Considering our terminator sample size ($n = 54$, full model, please see 'Results' section) and wanting to reduce the chance of overfitting, we only considered models with up to five independent variables (~10-fold less than the number of terminators considered). We generated linear models with improved explanatory power by iteratively adding the next most explanatory variable not yet in the model and re-evaluating model accuracy. We calculated to what degree each selected model could be used to predict unseen data via a cross-validation procedure in which we (i) randomly selected 80% of terminators; (ii) trained a model using this reduced subset; (iii) computed expected activities for the remaining 20%; and (iv) determined Pearson coefficient of correlation (r) between computed and observed TE values. These four steps were repeated 10^3 times for each linear model and the mean coefficient of correlation used to score model accuracy.

RESULTS

Development and validation of a terminator measurement device

RNA secondary structures encoded within terminators can differentially impact mRNA stability and thus confound measurements of TE (23). We thus sought to decouple TE measurements from the stability of the mRNA surrounding the terminators being measured. We choose to use RNase processing sites as flanking elements surrounding a terminator measurement site such that the expression of upstream and downstream reporter genes would be mediated by mRNA that do not include terminator-specific sequences (Figure 1A). We selected RNase III and RNase E sites as candidate mRNA processing elements, and green (sfGFP) and red fluorescent proteins (mRFP1) as live cell expression reporters. We constructed six candidate terminator measurement devices that explored the use of different orderings of GFP and RFP, each RNase system, and negative control devices lacking RNase-mediated normalization of reporter mRNA (Figure 1B).

We assembled a panel of 20 test sequences presumed to encode a wide range of termination efficiencies and cloned each into the six candidate measurement devices (Supplementary Table S3). If post-transcriptional RNase processing of mRNA effectively normalized both reporter mRNAs, then expression of the upstream reporter gene should remain constant across constructs, whereas that of the downstream gene should be affected only by terminator activity. As expected, variation in upstream reporter levels was lower than for the downstream reporter [0.32 versus 1.04, respectively; average coefficient of variation (CoV) across all six candidate test devices] (Figure 2A and B). The presence of functional RNase III sites reduced variation in upstream reporter expression (0.15 CoV) in comparison with constructs with RNase E (0.37 CoV) or non-functional RNase III sites (noIII, 0.43 CoV). Expression of RFP followed by GFP consistently produced less variation than GFP followed by RFP. Taken together, the RIIIG test device (*rfp* upstream of *gfp* with functional RNase III sites flanking the terminator) gave the least variation in upstream reporter expression levels, likely because of our use of two highly processive and sequence-distinct RNase III sites, R0.5 and R1.1, adapted from the early region of the bacteriophage T7 genome (35).

To compare the six candidate measurement devices in more detail, we calculated the Pearson correlations for terminator read-through [TR, Equation (1)] across all pairings of test devices. For example, we observed that switching the order of GFP and RFP produced differences in TR measurements for the RNase E devices (Figure 2C, left), whereas the RNase III devices were largely insensitive to fluorescent reporter order. More generally, TR measurements were best correlated between the RIIIG and GIIIR devices (Figure 2D) and perfectly correlated between bulk and single-cell measurements (Figure 2D, main diagonal). Taken together, our data indicated that RNase III sites provided a best practical method for standardizing measurement of termination activities, and we retained the RIIIG test device for subsequent experiments. Finally, we constructed two RIIIG variants encoding 3- and 14-fold increases in upstream promoter activity, and a third variant encoding a 2-fold increase in *rfp* translation (8). All variant RIIIG test devices maintained highly correlated measurements (average Pearson correlation ~0.9, $n = 20$, Supplementary Figure S5).

Measuring termination efficiencies across a collection of terminators

We assembled and sub-cloned an expanded set of 61 putative terminator elements into the RIIIG measurement device (see 'Materials and Methods' section and Supplementary Figure S2). We characterized each terminator in bulk culture and among single cells by measuring expression levels of the two fluorescent reporters. We rank ordered the terminators based on calculated average TEs (see 'Materials and Methods' section, Figure 3A). Of the 61 sequences tested, 17 encoded TEs >95%. Overall, the set encoded terminators sufficient to control expressed protein levels across a ~800-fold range (Figure 3B). Bulk and single-cell measurements of TEs were highly correlated ($r = 0.99$, $n = 61$, Supplementary Figure S6). We further observed that the mean and standard deviation of TEs within clonal populations were inversely correlated (Figure 3A and Supplementary Figure S7); highly active terminators exhibited little cell–cell variation, whereas the activities of weak terminators were highly dispersed among individual cells (Supplementary Figure S4).

Impact of proximal sequence context on termination efficiency

Genetic elements whose functions are encoded via RNA structures can be highly sensitive to changes in neighboring sequence context (31). For example, efficient transcription termination relies on the formation of a terminator hairpin, as the elongation complex is transiently paused at the U tail (17); the presence of competing structures upstream of a terminator core can prevent timely

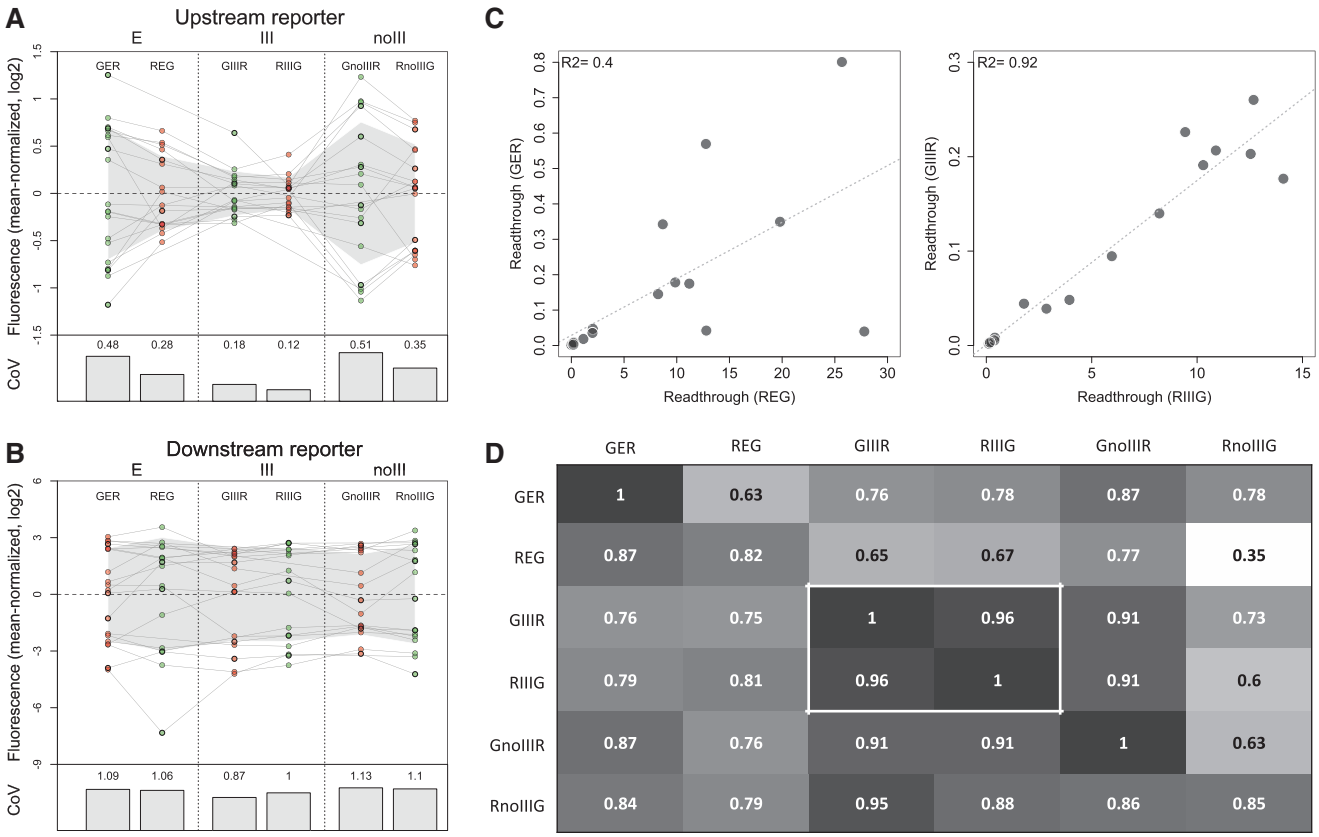

Figure 2. Testing and selection of a validated terminator measurement device. (**A**) Upstream reporter gene fluorescence data from a test panel of 20 terminator sequences cloned within six candidate terminator measurement devices; fluorescence values are normalized by the mean value obtained with each candidate measurement device. Expression levels for each terminator are connected (dotted lines). One standard deviation (shaded grey range) and coefficients of variation for expression levels (bottom bar graph) across all terminators within a given test device, as noted. (**B**) As in (A) but for a downstream reporter gene, the expression of which is expected to further vary as a result of differential termination efficiencies among the test terminator sequences. (**C**) Correlation in estimated terminator read-through measurements as upstream and downstream reporter genes are swapped. Green before red fluorescent protein versus red before green with RNase E sites (left) and with RNase III sites (right). (**D**) Pearson correlation scores for read-through measurements of the 20 terminator test panel. Correlation scores arising from comparing single cell (upper right) and bulk (lower left) measurements across the six candidate terminator measurement devices, as noted. Single-cell versus bulk correlation scores for each measurement device as given (main diagonal). Best performing (i.e. most consistent) measurement devices are bracketed (thick white line).

formation of a hairpin, thereby attenuating termination (36). To evaluate the impact of changing genetic context on TE, we compared the performance of 11 terminators in their natural genetic context with cognate minimal terminator motifs (i.e. sequences encoding only the hairpin and U tail; Figure 1). For 10 of the 11 terminator pairs, the full terminators flanked by 30 nt of native genomic context were at least as active as their cognate minimal terminators ($P = 0.04$, one-way ANOVA). Conversely, the minimal *his* terminator was ~20-fold more active than the full *his* terminator (Figure 4A).

We explored two processes that may account for some of the differences in TEs as a function of changing sequence contexts. First, co-transcription mRNA folding can dynamically constrain the formation of downstream RNA structures (37). We thus investigated whether upstream mRNA context could form competing folds that interfere with timely formation of a functional terminator. We performed kinetic folding simulations to predict the rate and frequency of correct terminator structure formation (31,30). For each terminator sequence, we assumed a constant transcription elongation rate and, allowing transcription complexes to pause at the start of the poly-U tail (17), derived frequencies of target terminator structure formation over time using 400 replicate simulations (see 'Materials and Methods' section). We found, for example, that proper folding of the full-length *his* terminator is likely prevented by a kinetically favored alternative mRNA secondary structure in which part of the upstream context associates with the first half of the terminator stem (Figure 4B, Supplementary Figure S8). In nature, the *his* terminator is part of a larger attenuation system involved in the regulation of histidine biosynthesis wherein a competing structure serves as an anti-terminator motif (38); competing structure formation is conditioned by low translation efficiency across an upstream *his* coding sequence that is not present in our test construct. Similar mRNA structural interference effects were predicted to impact the minimal versions of *rpoC* and *rnpBT1* terminators (Figure 4B). Differential folding was also predicted for the *lambda tR2* and *crp* terminators, but only in simulations corresponding to specific upstream free-mRNA window sizes (Supplementary Table S5).

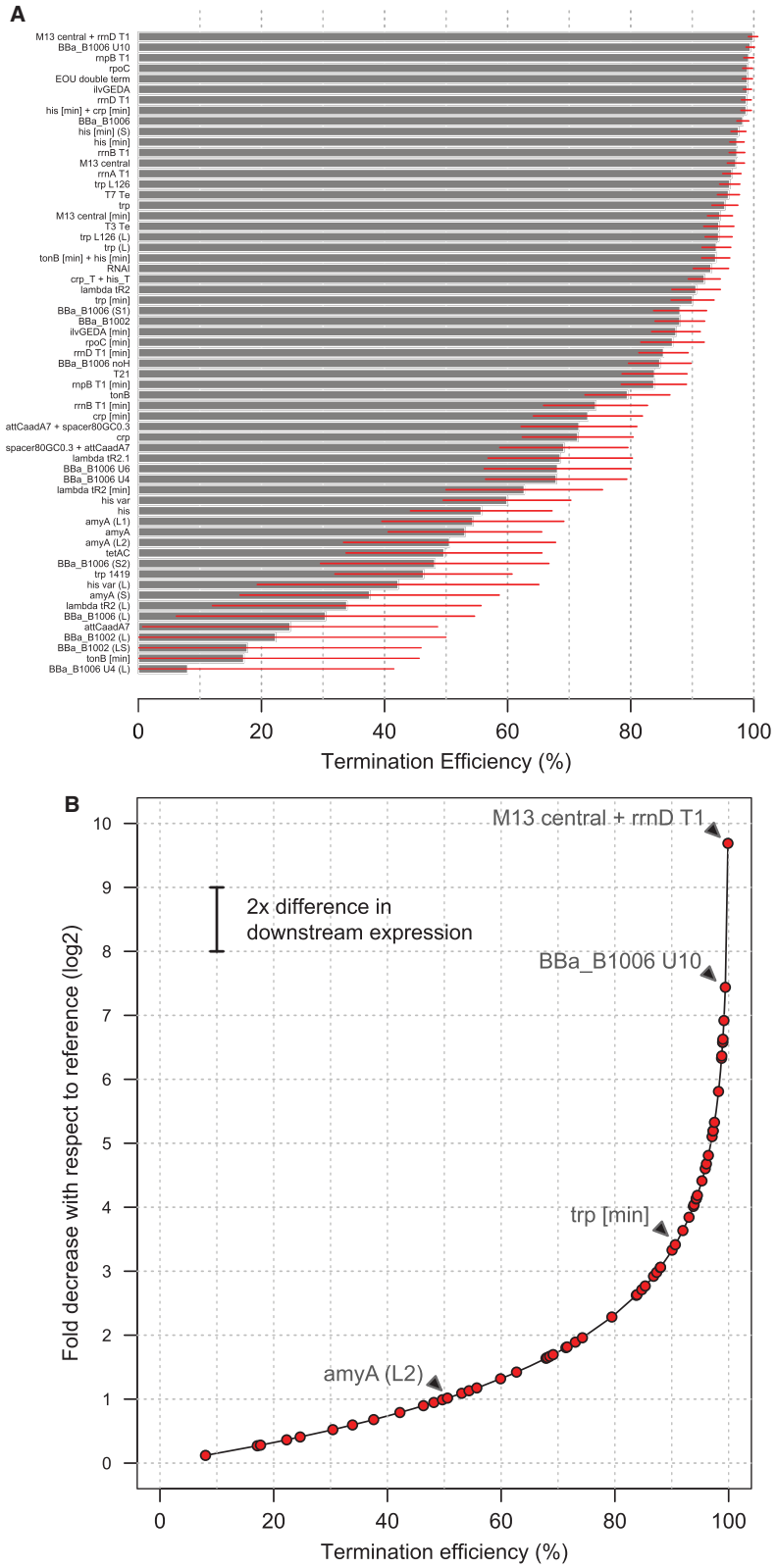

Figure 3. A wide range of termination efficiencies can be measured, enabling monotonic control of transcription read-through and downstream gene expression. (**A**) Bar chart of termination efficiencies as quantified by flow cytometry for 61 terminator sequences using the RIIIG measurement device. Error bars represent the standard deviation of TE among single cells within a population. Terminators are colored according to their functional categories (inset legend). (**B**) Mapping of termination efficiencies to transcriptional read-through and expression levels. The chart serves as a quick visual reference to determine fold expression differences arising from the terminators characterized here. For example, swapping 'amyA(L2)' (TE ~51%) with 'trp[min]' (TE ~90%) results in a ~5-fold decrease in downstream gene expression. As a second example, swapping 'BBa_B1006 U10' (TE ~99.4%) with 'M13 central + rrnD T1' (TE ~99.9%) also results in a ~5-fold decrease in downstream gene expression.

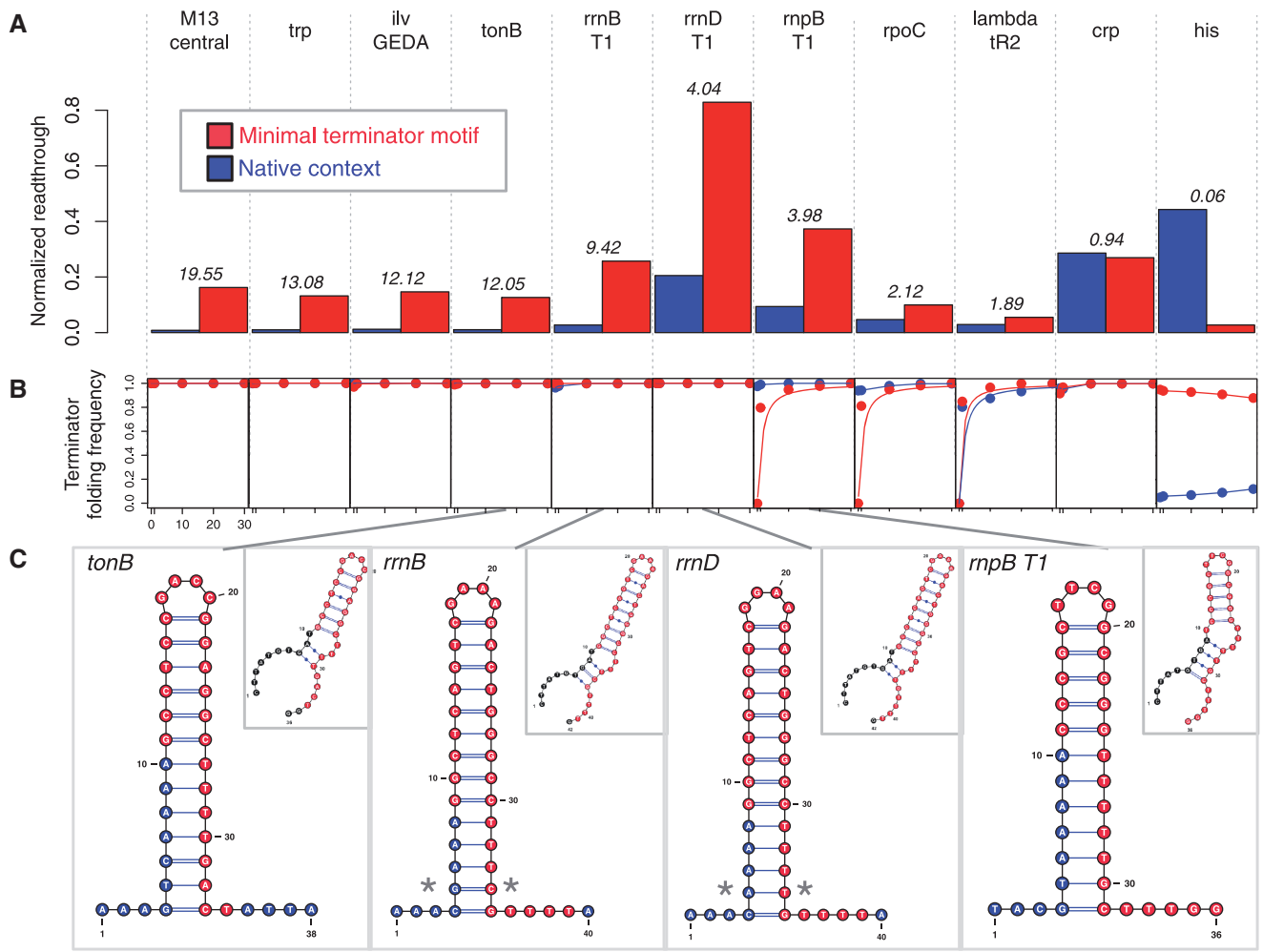

Figure 4. Immediate local sequence impacts on termination efficiencies. (**A**) Comparison of normalized transcription read-through (TR_{NORM}, 0.0–1.0) for terminators flanked by 30 nt of native upstream and downstream genomic sequence (blue) relative to minimal cognate terminators (red). Numbers above bars indicate the fold-increase in read-through for the minimal context. (**B**) Varying flanking contexts modify the predicted folding kinetics of some terminators. Each graph compares the folding frequency (0.0–1.0) for a core terminator stem over time (*x*-axes: 0, 0.5, 1, 10, 20 and 30 s) for expanded context (blue) and minimal terminators (red), as derived from co-transcriptional folding simulations (main text). (**C**) Outer terminators extending past core terminator motifs. Core terminator motifs (red bases) and native (blue, main panel) or minimal (black, insets) flanking sequences as indicated. For four terminators an extended terminator stem comprising part of the poly-U tail and closed by a GC pair could be identified in their expanded native context (main panel), but not within a minimal context (insets). Variable positions indicated at the base of the stems for paralogs *rrnB* and *rrnD* (stars).

Second, within some terminators, we observed that the sequence immediately upstream of a core stem might form an extended structure by pairing with the U tail. Such features are often thought to not impact TE, and only be required to form a U tail on the complementary strand within bi-directional terminators (39). Closer examination revealed that some extended stems are closed by G–C base pairing. In addition, when comparing the natural paralogs *rrnB* and *rrnD* within our data set, we noticed that a mutation in the upstream A-stretch is exactly complemented by a mutation in the downstream U tail, as might be expected from co-selection for base pairing (Figure 4C and Supplementary Figure S8C). These observations suggest the formation of functional outer terminators elements that extend past the core terminator motif. We found such possible nested structures for six terminators within our collection (*rnpBT1*, *tonB*, *rrnA*, *rrnB*, *rrnD* and *RNAI*; Figure 4C). Such elements within extended terminators seem likely to function sequentially, as indicated by the lower measured TEs of the minimal *rnpBT1*, *rrnB*, *rrnD* and *tonB* terminators relative to their extended counterparts (Figure 4A and C).

Sequence-activity models of termination efficiency

We defined 12 sequence features potentially involved in modulating terminator activity by reviewing the published literature and considering the roles of sequence context as noted earlier in the text (Supplementary Table S6). We developed a generic linear model for TE to select sequence features that might best account for observed TEs [Equation (4); 'Materials and Methods' section]. We found that increasing the number of predictors increased accuracy up to five predictors (Supplementary Figure S9). Overall, correlations between observed and computed TEs

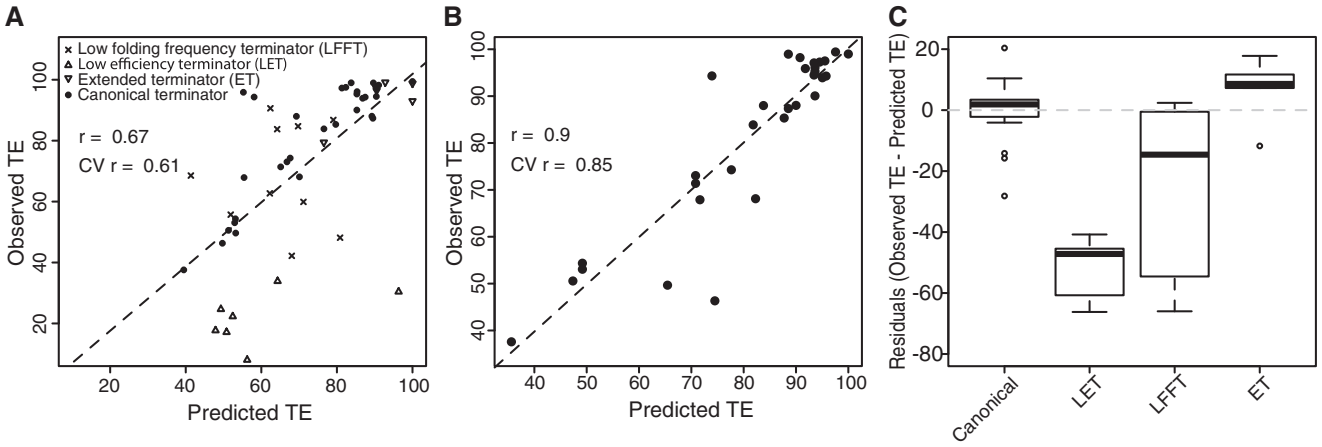

Figure 5. Quantitative sequence activity modeling of transcription termination. (**A**) Scatter plot of observed versus predicted termination efficiencies for a non-curated model that enables poor predictions compared with a model based on curated data set. (**B**) Scatter plot of observed versus predicted termination efficiencies for the 31 curated terminators used to train the model. Pearson correlation coefficient $r = 0.9$ and cross-validated (CV) $r = 0.85$ ('Materials and Methods' section). (**C**) Residual error distributions for each terminator category predicted via the curated model.

were modest ($r = 0.67$, cross-validated $r = 0.61$, $n = 54$; Figure 5A). The two features representing sequence context effects ('folding frequency' and 'ability to form extended terminators') were selected as the second and third most important variables. Additionally, we noted that terminators with very low TEs were poorly predicted. By systematically varying a TE cut-off, we found that improved correlations could be achieved by excluding seven terminators with TEs <35% (low-efficiency terminators, LET; Supplementary Figures S8B and S9). Likewise, we determined that terminators with simulated folding frequencies <90% reduced prediction quality, presumably because their TEs are not entirely encoded by core sequence features (low folding frequency terminators, LFFT; Supplementary Figures S8D and S9). We also suspected that the complex organization of extended terminators (ET) might confound model feature identification and excluded such terminators from a redacted modeling data set (Supplementary Figure S8C).

The remaining 31 canonical terminators (Supplementary Figure S8A) provided a more straightforward mapping of terminator features to TE measurements, and we expected this training set to yield more accurate models. Using the same cross-validation procedure ('Materials and Methods' section), we determined that a linear model comprising only four sequence features was highly explanatory of observed variation in TEs across this curated collection ($r = 0.9$, cross-validated $r = 0.85$, $n = 31$; Figure 5B and Supplementary Figure S9). The selected variables represented all functionally important regions of the core terminator motif, including stability of the hairpin normalized by its length (hp_norm_dg), sequence identity of the closing stack at the base of the stem (closing_stackGC) and the quality of the U tail (TTHP_utail_score) (Figure 1). The model also suggested that DNA composition bias downstream of the terminator (dna_dna_pattern) could fine-tune TE in relation to the quality of the U tail. The definitions and contributions of these features are detailed in Supplementary Table S6 and Supplementary Figure S9, and further discussed later in the text. We used the best linear model to predict the activities of the 23 terminators that had been excluded from the curated training set. Consistently, we found that predicted activities of LETs were highly overestimated, LFFTs also tended to be overestimated, whereas ETs were underestimated (Figure 5C; 'Discussion' section).

DISCUSSION

We developed a genetic testing device that uses RNase III recognition sites to create well-defined mRNA junctions surrounding intrinsic terminators. The test device reduces observed variation in upstream reporter gene expression levels while improving correlations for downstream reporter levels and resulting estimates of terminator efficiencies (TEs). Measured TEs were found to be insensitive to modest variation in fluorescent reporter transcription and translation activity. We used the validated testing device to characterize a set of sequence distinct terminators that collectively encode an ~800-fold range in observed expression levels for coding sequences downstream of terminator elements. The resulting terminator collection should enable synthetic biologists to realize fine control of transcription read-through in engineered gene expression cassettes.

We applied a simple linear model to identify possible sequence features that could be used to explain observed differences in measured TEs. We found no weighted linear combination of sequence features that could be used to perfectly estimate the observed activities for all terminators tested here. We used co-transcription folding simulations to detect potential upstream structures that might prevent the timely formation of active terminator motifs and result in decreased TEs (LFFTs). We also identified extended terminators wherein a canonical core terminator stem could be extended through the U tail to yield additional outer terminators that likely increase TEs (ETs).

Finally, we noticed that activities for a few very low-efficiency terminators were consistently overestimated (LETs). When we excluded these three types of terminators from model development, we found that a four-factor linear model could explain most of the observed differences in the activities of the remaining terminators, while maintaining a reasonable cross-validation score ($r = 0.9$, CV $r = 0.85$, $n = 31$).

Two sequence features explain most of the variation in activity within the curated terminator data set. First, 'TTHP_utail_score' is a heuristic score for the U tail that rewards both the number of uracils in the tail and their proximity to the hairpin (16). Second, 'hp_norm_dg' is the thermodynamic stability of the hairpin divided by the length of its stem. Of note, 'hp_norm_dg' performed better than either the uncorrected thermodynamic stability of the hairpin or the hairpin score used in the TransTermHP algorithm (21). 'TTHP_utail_score' and 'hp_norm_dg' together provided a reasonable quantitative model of termination efficiency ($r = 0.85$, CV $r = 0.8$, $n = 31$). Increasing the size and diversity of the training set would be necessary to fully validate the contributions of two additional features. Specifically, we found terminator hairpin stems that close with 5′-G…C-3′ base pairing (i.e. 'closing_stackGC') were associated with higher TEs. Also, 'dna_dna_pattern', a statistic tracking to what extent nucleotides downstream of a terminator core stem match sequence patterns (8), was associated with higher TEs.

The predicted activities for terminator types excluded from our curated training set were consistently biased (Figure 5C). For example, LFFTs were overpredicted with respect to measurements, as expected, given terminators with seemingly adequate core features that nevertheless fail to fold into functional structures. Such observations indirectly validate the co-transcription simulation procedure used in this work. However, it remains difficult to integrate such predictions directly into a simple linear model given that alternate structures may exert some termination activity on their own. Conversely, ETs were mostly under predicted, as expected for terminators whose true performance results from the activity of extended elements; aggregate termination from double terminators was similarly difficult to predict (Supplementary Figure S10). Finally, LETs were all strongly overpredicted. Of the seven terminators discarded for low TE, five were construction mutants with severe alterations of the stem–loop motif, one is a recombinase recognition motif ($attC_{aadA7}$) that has been suspected of termination activity but is not a canonical terminator (40) and further shows weak transcription initiation activity (Supplementary Figure S3B), and the remaining is a minimal version of a putative extended terminator identified in this work (*tonB [min]*).

Differences in termination efficiencies arising from simple changes in terminator sequences (Supplementary Figure S1) have also been observed with single molecule measurements [e.g. *his*, *tR2* (26)]. *In vivo* assays, as developed and standardized here, can now be scaled to support reliable characterization of much larger libraries of terminator mutants. We are, therefore, optimistic that high-throughput *in vivo* measurements combined with focused single molecule studies might resolve how to represent now difficult-to-model terminators, enabling *a priori* prediction of termination efficiency without consideration of terminator type.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6, Supplementary Figures 1–10 Supplementary Methods and Supplementary References [41–50].

ACKNOWLEDGEMENTS

The authors thank Morgan Price, the Joint BioEnergy Institute (JBEI) staff, and Arkin laboratory members.

FUNDING

BIOFAB [NSF Award No. EEC 0946510 plus unrestricted gifts from Genencor, Inc., Agilent, Inc. and DSM, Inc.]; Human Frontier Science Program (LT000873/2011-L) and the Bettencourt Schueller Foundation (to G.C.); Portuguese Fundação para a Ciência e a Tecnologia [SFRH/BD/47819/2008 to J.C.G.]; Synthetic Biology Engineering Research Center [NSF Award No. 04-570/0540879 to A.P.A. and D.E.]. This work was conducted at JBEI, which is supported by the Office of Science, Office of Biological and Environmental Research, U.S. 85 Department of Energy [Contract No. DE-AC02-05CH11231]. Funding for open access charge: BIOFAB project at Stanford & Cal [US National Science Foundation].

Conflict of interest statement. None declared.

REFERENCES

1. Cambray,G., Mutalik,V.K. and Arkin,A.P. (2011) Toward rational design of bacterial genomes. *Curr. Opin. Microbiol.*, **14**, 624–630.
2. Temme,K., Zhao,D. and Voigt,C.A. (2012) Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl Acad. Sci. USA*, **109**, 7085–7090.
3. Chan,L.Y., Kosuri,S. and Endy,D. (2005) Refactoring bacteriophage T7. *Mol. Syst. Biol.*, **1**, 2005.0018.
4. Canton,B., Labno,A. and Endy,D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.*, **26**, 787–793.
5. Alper,H., Fischer,C., Nevoigt,E. and Stephanopoulos,G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl Acad. Sci. USA*, **102**, 12678–12683.
6. Mutalik,V.K., Qi,L., Guimaraes,J.C., Lucks,J.B. and Arkin,A.P. (2012) Rationally designed families of orthogonal RNA regulators of translation. *Nat. Chem. Biol.*, **8**, 447–454.
7. 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.
8. Mutalik,V.K., Guimaraes,J.C., Cambray,G., Mai,Q.A., Christoffersen,M.J., Martin,L., Yua,A., Lam,C., Rodriguez,C., Bennett,G. *et al.* (2013) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods*, **10**, 347–353.

9. Mutalik,V.K., Guimaraes,J.C., Cambray,G., Lam,C., Christoffersen,M.J., Mai,Q.A., Tran,A., Paull,M., Keasling,J.D., Arkin,A.P. *et al.* (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods*, **10**, 354–360.
10. Peters,J., Vangeloff,A. and Landick,R. (2011) Bacterial transcription terminators: the RNA 3′-end chronicles. *J. Mol. Biol.*, **412**, 793–813.
11. Lucks,J.B., Qi,L., Mutalik,V.K., Wang,D. and Arkin,A.P. (2011) Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proc. Natl Acad. Sci. USA*, **108**, 8617–8622.
12. Liu,C.C., Qi,L., Yanofsky,C. and Arkin,A.P. (2011) Regulation of transcription by unnatural amino acids. *Nat. Biotechnol.*, **29**, 164–168.
13. Bonnet,J., Subsoontorn,P. and Endy,D. (2012) Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc. Natl Acad. Sci. USA*, **109**, 8884–8889.
14. Peters,J.M., Mooney,R.A., Kuan,P.F., Rowland,J.L., Keles,S. and Landick,R. (2009) Rho directs widespread termination of intragenic and stable RNA transcription. *Proc. Natl Acad. Sci. USA*, **106**, 15406–15411.
15. Reynolds,R. and Chamberlin,M.J. (1992) Parameters affecting transcription termination by *Escherichia coli* RNA. II. Construction and analysis of hybrid terminators. *J. Mol. Biol.*, **224**, 53–63.
16. d'Aubenton Carafa,Y., Brody,E. and Thermes,C. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.*, **216**, 835–858.
17. Gusarov,I. and Nudler,E. (1999) The mechanism of intrinsic transcription termination. *Mol. Cell*, **3**, 495–504.
18. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
19. Lesnik,E.A., Sampath,R., Levene,H.B., Henderson,T.J., McNeil,J.A. and Ecker,D.J. (2001) Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res.*, **29**, 3583–3594.
20. de Hoon,M.J., Makita,Y., Nakai,K. and Miyano,S. (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comp. Biol*, **1**, e25.
21. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.
22. Gardner,P.P., Barquist,L., Bateman,A., Nawrocki,E.P. and Weinberg,Z. (2011) RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res.*, **39**, 5845–5852.
23. Abe,H. and Aiba,H. (1996) Differential contributions of two elements of rho-independent terminator to transcription termination and mRNA stabilization. *Biochimie*, **78**, 1035–1042.
24. Santangelo,T.J. and Artsimovitch,I. (2011) Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Micro.*, **9**, 319–329.
25. Reynolds,R., Bermúdez-Cruz,R.M. and Chamberlin,M.J. (1992) Parameters affecting transcription termination by *Escherichia coli* RNA polymerase. I. Analysis of 13 rho-independent terminators. *J. Mol. Biol.*, **224**, 31–51.
26. Larson,M.H., Greenleaf,W.J., Landick,R. and Block,S.M. (2008) Applied force reveals mechanistic and energetic details of transcription termination. *Cell*, **132**, 971–982.
27. Rosenberg,M., Chepelinsky,A.B. and McKenney,K. (1983) Studying promoters and terminators by gene fusion. *Science*, **222**, 734–739.
28. Engler,C., Kandzia,R. and Marillonnet,S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One*, **3**, e3647.
29. Lo,K., Hahne,F., Brinkman,R.R. and Gottardo,R. (2009) flowClust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, **10**, 145.
30. Isambert,H. (2009) The jerky and knotty dynamics of RNA. *Methods*, **49**, 189–196.
31. Carothers,J.M., Goler,J.A., Juminaga,D. and Keasling,J.D. (2011) Model-driven engineering of RNA devices to quantitatively program gene expression. *Science*, **334**, 1716–1719.
32. Iost,I., Guillerez,J. and Dreyfus,M. (1992) Bacteriophage T7 RNA polymerase travels far ahead of ribosomes *in vivo*. *J. Bacteriol.*, **174**, 619–622.
33. Arnold,S., Siemann-Herzberg,M., Schmid,J. and Reuss,M. (2005) Model-based inference of gene expression dynamics from sequence information. *Adv. Biochem. Eng. Biotechnol.*, **100**, 89–179.
34. Tomsic,J., Vitali,L.A., Daviter,T., Savelsbergh,A., Spurio,R., Striebeck,P., Wintermeyer,W., Rodnina,M.V. and Gualerzi,C.O. (2000) Late events of translation initiation in bacteria: a kinetic analysis. *EMBO J.*, **19**, 2127–2136.
35. Calin-Jageman,I. and Nicholson,A.W. (2003) Mutational analysis of an RNA internal loop as a reactivity epitope for *Escherichia coli* ribonuclease III substrates. *Biochemistry*, **42**, 5025–5034.
36. Yanofsky,C. (2000) Transcription attenuation: once viewed as a novel regulatory strategy. *J. Bacteriol.*, **182**, 1–8.
37. Pan,T. and Sosnick,T. (2006) RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 161–175.
38. Chan,C.L. and Landick,R. (1993) Dissection of the his leader pause site by base substitution reveals a multipartite signal that includes a pause RNA hairpin. *J. Mol. Biol.*, **233**, 25–42.
39. Wilson,K.S. and von Hippel,P.H. (1995) Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl Acad. Sci. USA*, **92**, 8793–8797.
40. Cambray,G., Guerout,A.M. and Mazel,D. (2010) Integrons. *Annu. Rev. Genet.*, **44**, 141–166.
41. Hess,G.F. and Graham,R.S. (1990) Efficiency of transcriptional terminators in *Bacillus subtilis*. *Gene*, **95**, 137–141.
42. Mazel,D., Dychinco,B., Webb,V.A. and Davies,J. (2000) Antibiotic resistance in the ECOR collection: integrons and identification of a novel aad gene. *Antimicrob. Agents Chemother.*, **44**, 1568–1574.
43. Cheng,S.W., Lynch,E.C., Leason,K.R., Court,D.L., Shapiro,B.A. and Friedman,D.I. (1991) Functional importance of sequence in the stem-loop of a transcription terminator. *Science*, **254**, 1205–1207.
44. Edens,L., Konings,R.N. and Schoenmakers,J.G. (1975) Physical mapping of the central terminator for transcription on the bacteriophage M13 genome. *Nucleic Acids Res.*, **2**, 1811–1820.
45. Kim,S., Kim,H., Park,I. and Lee,Y. (1996) Mutational analysis of RNA structures and sequences postulated to affect 3′ processing of M1 RNA, the RNA component of *Escherichia coli* RNase P. *J. Biol. Chem.*, **271**, 19330–19337.
46. McDowell,J.C., Roberts,J.W., Jin,D.J. and Gross,C. (1994) Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. *Science*, **266**, 822–825.
47. Schollmeier,K., Gärtner,D. and Hillen,W. (1985) A bidirectionally active signal for termination of transcription is located between tetA and orfL on transposon Tn10. *Nucleic Acids Res.*, **13**, 4227–4237.
48. Lee,T.S., Krupa,R.A., Zhang,F., Hajimorad,M., Holtz,W.J., Prasad,N., Lee,S.K. and Keasling,J.D. (2011) BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *J. Biol. Eng.*, **5**, 12.
49. Musso,M., Bocciardi,R., Parodi,S., Ravazzolo,R. and Ceccherini,I. (2006) Betaine, dimethyl sulfoxide, and 7-deaza-dGTP, a powerful mixture for amplification of GC-rich DNA sequences. *J. Mol. Diagn.*, **8**, 544–550.
50. Studier,F.W. (1975) Genetic mapping of a mutation that causes ribonucleases III deficiency in *Escherichia coli*. *J. Bacteriol.*, **124**, 307–316.