Genomics 109 (2017) 196–203

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Computational prediction of functional abortive RNA in *E. coli*

Jeremy I. Marcus [a], Soha Hassoun [a,b], Nikhil U. Nair [b,*]

[a] Department of Computer Science, Tufts University, Medford, MA 02155, United States
[b] Department of Chemical and Biological Engineering, Tufts University, Medford, MA 02155, United States

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 24 February 2017
Accepted 22 March 2017
Available online 24 March 2017

Keywords:
Noncoding
RNA
Transcription
E. coli
Abortives
Regulation

ABSTRACT

Failure by RNA polymerase to break contacts with promoter DNA results in release of bound RNA and re-initiation of transcription. These abortive RNAs were assumed to be non-functional but have recently been shown to affect termination in bacteriophage T7. Little is known about the functional role of these RNA in other genetic models. Using a computational approach, we investigated whether abortive RNA could exert function in *E. coli*. Fragments generated from 3780 transcription units were used as query sequences within their respective transcription units to search for possible binding sites. Sites that fell within known regulatory features were then ranked based upon the free energy of hybridization to the abortive. We further hypothesize about mechanisms of regulatory action for a select number of likely matches. Future experimental validation of these putative abortive-mRNA pairs may confirm our findings and promote exploration of functional abortive RNAs (faRNAs) in natural and synthetic systems.

1. Introduction

Non-coding RNAs (ncRNAs) are RNA molecules that are not translated into proteins and have been shown to be involved in a variety of key cellular processes. They function as gene regulatory molecules and generally exert action by the occluding or exposing binding elements in mRNA. Due to their predictability and versatility, they have been used extensively in engineering synthetic biological systems. Well-described regulatory ncRNAs in bacterial include ribosomal RNA (rRNA), transfer RNA (tRNA), small RNA (sRNA), and anti-sense RNA (asRNA) [1–3].

Abortive RNA transcripts are a poorly-documented class of ncRNAs characterized by their small size and unique mechanism of generation during transcription. RNA transcription involves three basic stages: initiation, elongation, and termination. Once RNA polymerase (RNAP) binds to a DNA promoter during initiation, it repetitiously synthesizes and releases abortive transcripts while remaining bound to the promoter region in a process known as abortive cycling. This phenomenon has been observed to some extent in nearly all in vitro transcription reactions involving RNAPs from different species, and has also been detected *in vivo* in *E. coli* [4–7]. Different RNAPs generate abortive fragments of varying length; for example, human RNAP II and *E. coli* RNAP release transcripts of up to 8 and 15 nt (nucleotides), respectively [8,9]. It has been estimated that only 1 out of every 10 to 100 transcription reactions initiated by RNAP results in successful transition to the elongation phase [10,11]. As a result, abortive initiation cycling leads to the accumulation of short abortive RNA transcripts. Short single-stranded unstructured RNA fragments tend to be unstable and are degraded quickly; but they can form weak, transient complexes with complementary nucleotide sequences [12]. For these reasons, it was considered unlikely that abortive transcripts could serve a functional role.

However, a recent study in the T7 bacteriophage identified a role for abortive transcripts in antitermination at the *Tϕ10* terminator [13]. During early stages of infection, late gene expression is repressed by this rho-independent terminator. In late lifecycle, accumulation and binding of abortive transcripts to the upstream leg of *Tϕ10* was shown to prevent hairpin formation and subsequently prevent termination. This resulted in read-through and expression of genes downstream of the terminator. The inherent time lag between initial gene expression and the accumulation of sufficiently high concentrations of abortive transcripts resulted in delayed expression of the downstream genes, which were speculated to be instrumental in T7 phage lifecycle.

A similar novel gene regulation mechanism has yet to be identified outside of the T7 bacteriophage. Further investigation of regulatory roles of abortive transcripts in other organisms requires systematic identification of abortive transcripts and their putative targets. *E. coli* is one of the most well-studied model organisms in genetics; thus, there is a wealth of available information describing its genome and regulatory mechanisms. This makes *E. coli* an appropriate choice for exploring the regulatory roles of abortive transcripts. However, the *E. coli* genome of approximately 4.6 million base-pairs, is much larger than the 39,937 base-pair T7 bacteriophage genome. Large quantities of genetic content can be prohibitive for experimentally conducting genome-wide searches for novel regulatory mechanisms. Predictive computational

* Corresponding author.
E-mail address: nikhil.nair@tufts.edu (N.U. Nair).

http://dx.doi.org/10.1016/j.ygeno.2017.03.003
0888-7543/

models can help expedite the process by focusing the experimental search space onto a manageable subset of the genome.

In this study, we utilized computational methods to predict locations in the *E. coli* genome where abortive fragments might perform some functional role in regulation at the transcriptional- or translational-level. We identified matches occurring in functionally relevant genomic features, such as terminators and ribosomal binding sites, and ranked these matches using quantitative free energy calculations and subjected them to statistical tests to assess their significance relative to random hits. Here we suggest mechanisms of regulatory control for three of the abortive fragments returned by our analysis.

2. Results

2.1. Energetics of abortive RNA-mRNA binding

Abortive initiation in *E. coli* generally results in the release of 2 to 15 nt long abortive fragments [14,15]. We chose to focus on fragments of lengths 4–15 nt, as shorter sequences are less likely to be able to exact a physiologically relevant effect beyond transcriptional priming [16]. We assumed that all lengths of all abortives within our specified range are produced, and therefore performed our initial Watson-Crick base-pairing search over the entire range for each of the 3780 known transcriptional units in the *E. coli* genome. An additional search performed using wobble base-pairing rules was run under the same conditions. For each abortive match site found, we calculated the standard Gibbs free energy of hybridization ($\Delta G^{\circ}_{binding}$) as a measure of physiological relevance. Free energy has previously been shown to be strongly and linearly correlated with the ability of abortive transcripts to disrupt the function of an intrinsic terminator [13]. A more negative free energy value implies that a reaction is more spontaneous, implying that a given abortive fragment is more likely to bind strongly and effect a regulatory function. Therefore, we could utilize free energy as a quantitative measure of an oligomer's potential to exert function, providing a basis for comparison between individual matches. Fig. 1 shows the distributions of $\Delta G^{\circ}_{binding}$ values for all matches found using both matching paradigms over all abortive RNA lengths.

In evaluating the functional significance of abortive fragments, we limited our analysis to those matches that occurred within two types of known functional genetic regions – rho-independent terminators and ribosomal binding sites (RBSs). Tables 1 and 2 present summaries of the filtered data obtained using Watson-Crick and wobble base-pairing respectively. Fig. 2a and b display the distribution of free energy values calculated for abortive matches located in RBSs and terminating regions, respectively.

The existence of a minimum free energy requirement to achieve effective antitermination or anti-translation by antisense hybridization has not been shown. However, Lee et al. demonstrated that antitermination of abortives at the T7 *ϕ10* terminator was directly correlated to the free energy of hybridization [13]. They demonstrated that abortives with free energy of hybridization <−7 kcal/mol are able to antiterminate with >60% efficiency and those with <−9 kcal/mol are able to antiterminate with >80% efficiency. As comparison, *trans*-regulation by sRNA has been shown to be weak when hybridization energies are <−10 kcal/mol even with the aid of Hfq [17]. Thus, faRNA seem to be able to execute their function under less favorable thermodynamic conditions compared to sRNA. This may be the case because abortive fragments are produced in *cis* configuration, often in large excess compared to productive transcript, and their high localized concentration may be able to compensate for the weaker hybridization energies [18].

2.2. Statistical analyses of putative faRNA binding locations

When searching for binding sites for target sequences within a genomic sequence, one would expect to find those sites randomly scattered throughout. We expected that, on average, matches occur by random chance and are not functionally relevant. For shorter-length abortives (4–7 nt), the frequencies of actual exact matches are only slightly (<20%) more than those expected by random chance (Table 3).

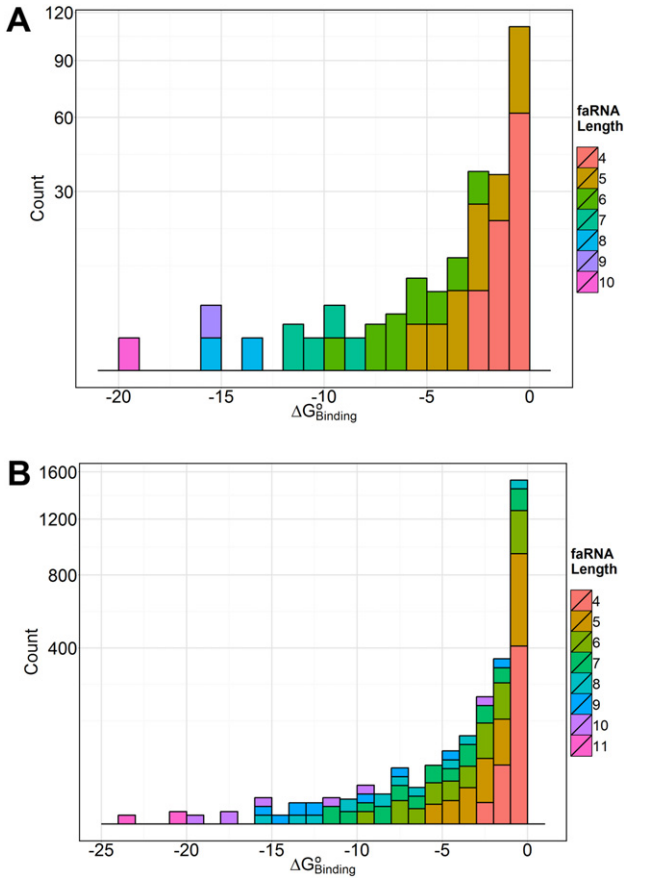

Fig. 1. Distribution of computed $\Delta G^{\circ}_{binding}$ values for predicted complexes between abortive initiation fragments and mRNAs. We only report complexes which occur within the same transcriptional unit from which the abortive initiation fragment was generated. RNA binding free energy calculations performed using UNAFold *hybrid2.pl*, with temperature range 0–100 °C, 1.0 μM concentrations of both strands. Counts displayed on a square-root scale for visual clarity. Colored bars correspond to the length of the nucleotide match.

(A) $\Delta G^{\circ}_{binding}$ distribution for complexes exhibiting exact Watson-Crick (A-U, G-C) base-pair matching.
(B) $\Delta G^{\circ}_{binding}$ distribution for complexes exhibiting exact base-pair matching under a wobble base-pairing paradigm (A-U, G-C, G-U).

Table 1
Abortive RNA hybridization sites within mRNA as matched by Watson-Crick base-pairing.

Abortive fragment length	Transcription units containing ≥ 1 match	Total match count (chance of finding[a])	Matches within terminators (chance of finding[a])	Matches within RBSs (chance of finding[a])
4	3448	26,290 (696%)	89 (2.36%)	6 (0.159%)
5	2386	6784 (180%)	26 (0.688%)	2 (0.0529%)
6	1097	1789 (47.3%)	11 (0.291)	0 (0%)
7	386	453 (12.0%)	5 (0.132%)	0 (0%)
8	123	129 (3.41%)	2 (0.0529%)	0 (0%)
9	42	42 (1.11%)	1 (0.0265%)	0 (0%)
10	14	14 (0.370%)	1 (0.0265%)	0 (0%)
11	6	6 (0.159%)	0 (0%)	0 (0%)
12	1	1 (0.0265%)	0 (0%)	0 (0%)
13	0	0 (0%)	0 (0%)	0 (0%)
14	0	0 (0%)	0 (0%)	0 (0%)
15	0	0 (0%)	0 (0%)	0 (0%)

[a] Chance of finding is % probability of finding an abortive of that length with a complimentary sequence within the same transcript = n/3780 × 100%.

Table 2
Abortive RNA hybridization sites within mRNA as matched by wobble (G-U) base-pairing.

Abortive fragment length	Transcription units containing ≥ 1 match	Total match count (chance of finding[a])	Matches within terminators (chance of finding[a])	Matches within RBSs (chance of finding[a])
4	3723	122,949 (3253%)	460 (12.2%)	54 (1.43%)
5	3484	46,962 (1242%)	191 (5.05%)	29 (0.767%)
6	2956	18,609 (492%)	86 (2.28%)	11 (0.291%)
7	2107	7167 (190%)	41 (1.09%)	6 (0.159%)
8	1313	2709 (71.7%)	12 (0.318%)	4 (0.106%)
9	719	1078 (28.5%)	10 (0.265%)	1 (0.0265%)
10	332	422 (11.2%)	7 (0.185%)	1 (0.0265%)
11	142	157 (4.15%)	3 (0.0794%)	0 (0%)
12	57	58 (1.53%)	0 (0%)	0 (0%)
13	19	19 (0.503%)	0 (0%)	0 (0%)
14	5	5 (0.132%)	0 (0%)	0 (0%)
15	2	2 (0.0529%)	0 (0%)	0 (0%)

[a] Chance of finding is % probability of finding an abortive of that length with a complimentary sequence within the same transcript = n/3780 × 100%.

Table 3
Expected number of matches in TU due to chance vs. actual matches, for Watson-Crick and Wobble (G-U) base-pairing. All counts were rounded-down to nearest integer.

Abortive fragment length	Watson-Crick base-pairing			Wobble (G-U) base-pairing		
	Expected number of matches	Total match count	Percentage increase of total matches over expected	Expected number of matches	Total match count	Percentage increase of total matches over expected
4	24,501	26,290	7.30%	121,816	122,949	0.93%
5	6161	6784	10.1%	46,753	46,962	0.45%
6	1555	1789	15.1%	18,147	18,609	2.55%
7	391	453	15.9%	6941	7167	3.26%
8	98	129	31.6%	2660	2709	1.84%
9	24	42	75.0%	1024	1078	5.27%
10	6	14	133%	394	422	7.11%
11	1	6	500%	151	157	3.97%
12	0	1	NA[a]	60	58	−3.33%
13	0	0	NA[a]	23	19	−17.3%
14	0	0	NA[a]	8	5	−37.5%
15	0	0	NA[a]	3	2	−33.3%

[a] NA refers to not applicable as the expected and actual numbers of matches were zero.

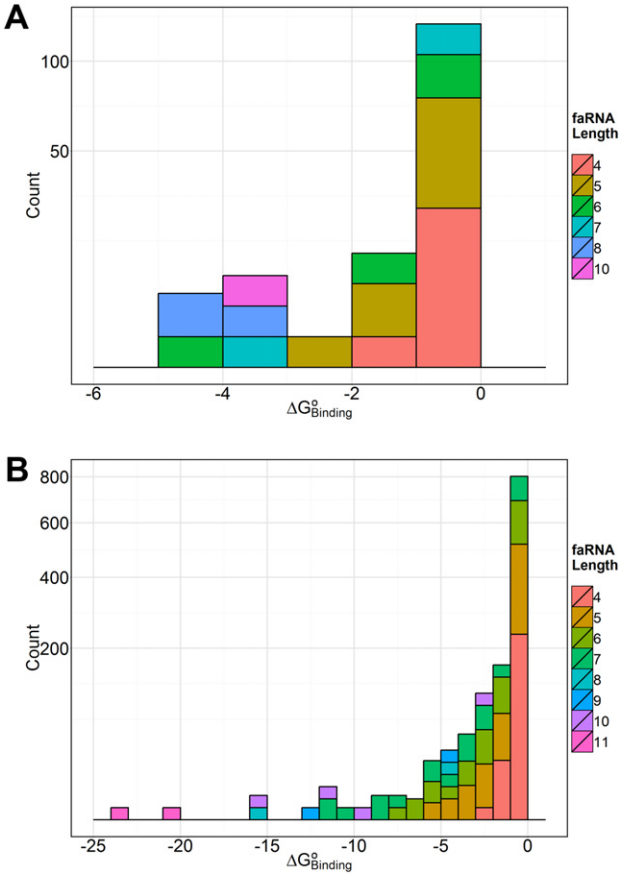

Fig. 2. Distribution of computed $\Delta G^{\circ}_{binding}$ values for predicted complexes between abortive initiation fragments and mRNAs occurring within functionally relevant genetic regions. We only report complexes which occur within the same transcriptional unit from which the abortive initiation fragment was generated. All predicted complexes exhibit exact base-pair matching under a wobble base-pairing paradigm (A-U, G-C, G-U). RNA binding free energy calculations performed using UNAFold *hybrid2.pl*, with temperature range 0–100 °C, 1.0 μM concentrations of both strands. Counts displayed on a square-root scale for visual clarity. Colored bars correspond to the length of the nucleotide match.

(A) $\Delta G^{\circ}_{binding}$ distribution for complexes with footprints overlapping ribosomal binding sites (RBSs).
(B) $\Delta G^{\circ}_{binding}$ distribution for complexes with footprints overlapping rho-independent terminators.

However, for longer abortives (8–11 nt), we find substantially more exact matches (>30%) than expected by pure chance. When taking wobble matches into consideration, we find that the frequencies of actual matches were either slightly changed (<20% change for 4–12 nt) or lower (>20% change for 13–15 nt) than those we expect to find by random chance. However, in the case of a functional nucleotide interaction arising through sequence co-evolution, one would expect to observe an unequal distribution of binding sites between regulatory and non-regulatory loci. To assess the distribution of binding site locations within each transcriptional unit, we applied Fisher's Exact Test to our terminator and RBS results (Table 4). The null hypothesis for each test was that there is no significant difference in the distribution of binding sites between regulatory and non-regulatory sequence motifs, where regulatory sequence motifs are defined as binding positions in each TU that overlap the target mRNA feature such as terminators or RBSs. There exists a small subset of all *E. coli* transcriptional units that show significant deviation ($p < 0.05$ and $p < 0.1$) from the expected distribution of binding locations. We presented data for both $p < 0.05$ and $p < 0.1$ to provide better insight into the relative distribution of significant hits. We further evaluated the significance in $\Delta G^{\circ}_{binding}$ values between the abortive matches in the TU and in the terminator and RBS regions using a Wilcoxon rank-sum test (Table 5). The null hypothesis was that there is no significant difference between the distributions of $\Delta G^{\circ}_{binding}$ values within regulatory sequence motifs and non-regulatory sequences, for each abortive length. There is a significant statistical shift ($p < 0.05$) in the $\Delta G^{\circ}_{binding}$ values for all matches within terminator regions. However, this shift does not hold in the RBS regions for abortive lengths ≥ 6 nt.

Table 4
Number of TUs that demonstrate a significant difference in the number of matches within regulatory and non-regulatory regions, as assessed by Fisher's Exact Test.

Abortive fragment length	$p < 0.1$		$p < 0.05$	
	Terminators	RBS	Terminators	RBS
4	46	24	33	14
5	46	18	22	16
6	42	9	18	7
7	32	6	17	3
8	10	4	6	4
9	8	1	7	1
10	7	1	7	1
11	3	–	3	–

Table 5
Comparison of distributions of $\Delta G^{\circ}_{binding}$ values within regulatory and non-regulatory regions using the Wilcoxon rank-sum test.

Abortive fragment length	Terminator			RBS		
	Number of abortives matches in non-terminator region	Number of abortives matches in terminator region	*p*-Values	Number of abortives matches in non-RBS region	Number of abortives matches in RBS region	*p*-Values
4	122,489	460	0.000001	122,895	54	0.0204
5	46,771	191	0.000455	46,933	29	0.000056
6	18,523	86	0.0159	18,598	11	0.0550
7	7126	41	0.00486	7161	6	0.137
8	2697	12	0.0202	2705	4	0.224
9	1068	10	0.0403	NA[a]	0	NA[a]
10	415	7	0.0476	NA[a]	0	NA[a]
11	154	3	0.00696	NA[a]	0	NA[a]

[a] NA refers to not applicable as the numbers of matches were zero.

3. Discussion

Abortive transcripts remain relatively unexplored as elements of gene regulatory mechanisms, outside of an example shown in the T7 bacteriophage [13]. To this end, we utilized a computational approach to find putative faRNA binding sites in the *E. coli* genome. This is a vital step towards determining the true potential of abortive transcripts as regulatory elements. It is our hope that by demonstrating the putative role of these RNAs in a model organism we can inspire further investigation of this often-overlooked transcriptional phenomenon. These data, coupled with future experimental results could more fully elucidate the role of abortive RNA is cellular regulation and augment our understanding of an understudied facet of gene regulation in bacteria.

In the present work, we chose to find hybridization targets only within the same transcription unit that produced the faRNA. As abortive fragments are short unstructured RNA molecules, we hypothesize that they are degraded quickly and are therefore unlikely to hybridize to targets that are spatially distal. Indeed even long antisense RNA without paired termini or extensive secondary structure are unable to exert significant regulatory effect when expressed in *trans*, spatially distant from their target locus [19]. Degradation and migration rates for abortive fragments have yet to be experimentally investigated. Without this quantitative information, it is very difficult to confidently predict the sphere of influence of individual faRNAs. Therefore we chose to search only for binding sites that followed the precedent set by the T7 bacteriophage example described by Lee et al. [13].

The full-length transcript for a given transcriptional unit contains both the abortive sequence and any binding sites, and therefore could bind to itself in a regulatory manner. However, based on the precedence of the results of the T7 bacteriophage experiments we limited our investigation to bimolecular interactions between abortive fragments and full mRNA transcripts [13]. Furthermore, the machinery required for bacterial co-translation could impede the ability of the mRNA transcript to adopt the necessary configuration for such a unimolecular reaction.

Binding sites of interest were chosen based on their association with well-defined published regulatory features, namely rho-independent terminators and ribosomal binding sites (RBSs), and their relative affinity for complementary faRNAs, measured by Gibbs free energy of hybridization ($\Delta G^{\circ}_{binding}$). Based on our statistical analysis, there are several abortive matches that are more likely to occur than expected by random chance. Further, matches within terminators and RBS regions occur at a different frequency when compared to matches within non-regulatory regions. Statistical testing demonstrates that the distribution of $\Delta G^{\circ}_{binding}$ for match sites within terminators of various length is different compared to the corresponding distribution in non-regulatory regions. This statistical difference was not evident for longer length matches within RBS regions.

We further propose mechanisms of action for several our putative binding sites, three of which are described here. These examples were chosen for discussion because are representative of a variety of different regulation mechanisms and highlight the versatility of the model we propose. Additionally, some of these examples are supported by literature, most notably the control of variable transcription rates in the *rpsA* and *ihfB* transcriptional units. Our results indicate possible roles for faRNAs in both transcriptional and translational control of *E. coli* gene expression. We provide these putative binding sites in the hope that they will help expedite experimental validation of these roles. Sequence information on the three loci discussed below are in S1 Figure and S1 and S2 Tables (Supporting information online).

3.1. Proposed mechanisms of faRNAs at different loci

3.1.1. rpsA-ihfB

The S1 ribosomal subunit protein, encoded by the gene *rpsA*, is the largest of the ribosomal proteins and an essential part of the cell's translational machinery [20]. *rpsA* lies directly upstream of and shares a transcriptional unit (TU) with *ihfB*, which codes for the β subunit of Integration Host Factor (IHF). IHF interacts directly with DNA to regulate expression of a wide variety of genes [21]. Transcription of *ihfB* gene have been shown to vary predictably based on the observed growth stage of cells [22]. During the stationary growth phase, there is an increase in production of monocistronic *ihfB* transcripts. During the translation-intensive exponential growth phase, there is an increase in production of polycistronic *rpsA-ihfB* transcripts that accompany enhanced *rpsA* expression. The exact mechanism controlling this expression pattern has not yet been described. However, disruption of a terminating hairpin structure found in the intergenic *rpsA-ihfB* region has been suggested as a part of this regulatory process [22]. Our analysis indicates that an abortive fragment produced from the *rpsAp1* promoter may be able to bind to the upstream leg of the *rpsA* terminating hairpin (two-tailed p-value = 0.0707), allowing for anti-termination and co-transcription of *rpsA* and *ihfB* (Fig. 3). This putative mechanism may explain the observed pattern of *ihfB* transcription, by associating the growth rate-dependent production of polycistronic transcripts with a growth rate-dependent concentration of faRNAs.

3.1.2. adhP

Expression of the alcohol dehydrogenase encoded by the gene *adhP* is known to be induced in the presence of ethanol [23]. However, little is known about the exact induction mechanisms. As identified in WebGesterDB, immediately downstream of the transcription start site (TSS) for the *adhP* TU lies a strong intrinsic terminator [24]. Successful transcription of the *adhP* coding region requires RNAP to pass through this region; therefore, it is likely that production of full mRNA transcripts requires some form of antitermination. We observed that the abortive fragment corresponding to the first 7 bp of the TU are an exact reverse-complement match to part of the leading stem of the terminator (two-tailed p-value = 0.0400). In the presence of sufficiently high concentrations of these *adhP* abortive fragments, the terminating hairpin may not be able to form. Upregulation of *adhP* in the presence of ethanol would require increased initiation rate by RNAP from this

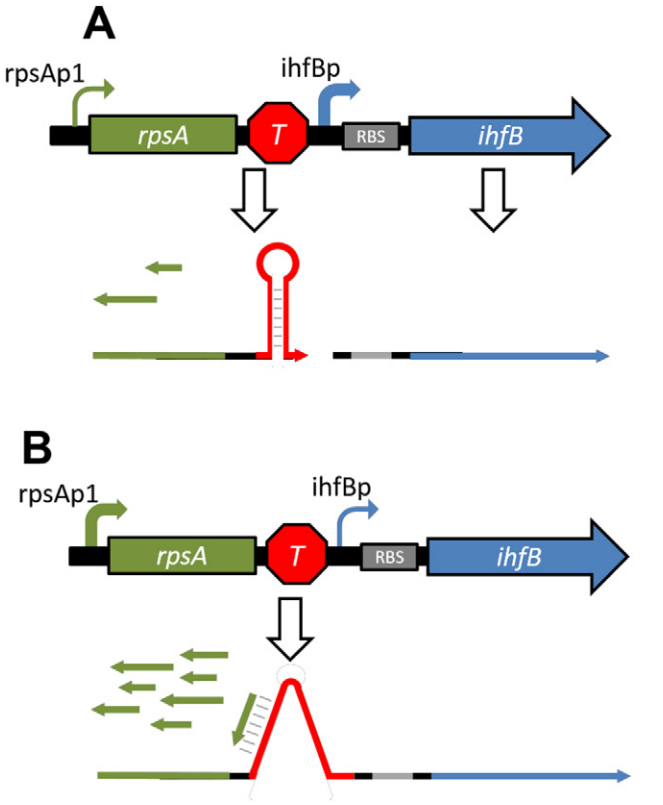

Fig. 3. Proposed mechanism of faRNA action in the *rpsA-ihfB* transcriptional unit. *rpsA* in green, *ihfB* in blue, *rpsA* ribosomal binding site (RBS) in grey, and *rpsA* terminator in red. For panels A and B, the upper portion of the panel depicts DNA while the bottom portion depicts RNA transcripts and bound/unbound abortive fragments. Curved arrows represent promoters; weight of promoter arrow represents relative rate of transcription initiation at that promoter.

(A) Proposed model of *rpsA* and *ihfB* transcript production during the stationary growth phase. Due to the lower demand for translational machinery during this phase, fewer *rpsA* transcripts are produced. This would lead to a lower concentration of abortive fragments, and a reduced binding to the *rpsA* terminator. Thus, *rpsA* transcription would be more likely to successfully terminate before the *ihfB* promoter. Initiation of *ihfB* transcription would therefore be more likely to occur at the *ihfB* promoter during the stationary phase, separately from *rpsA* transcription.

(B) Proposed model of *rpsA-ihfB* transcript production during the exponential growth phase. *rpsA* transcription rates are increased during this phase to accommodate the required increase of translation rates. Thus, higher concentration of abortive fragments increases the likelihood of antitermination at the locus indicated by our analysis for transcription reactions starting from any *rpsA* promoter. Thus, the number of *ihfB* transcripts produced from *rpsA* promoters would be greater during the exponential phase.

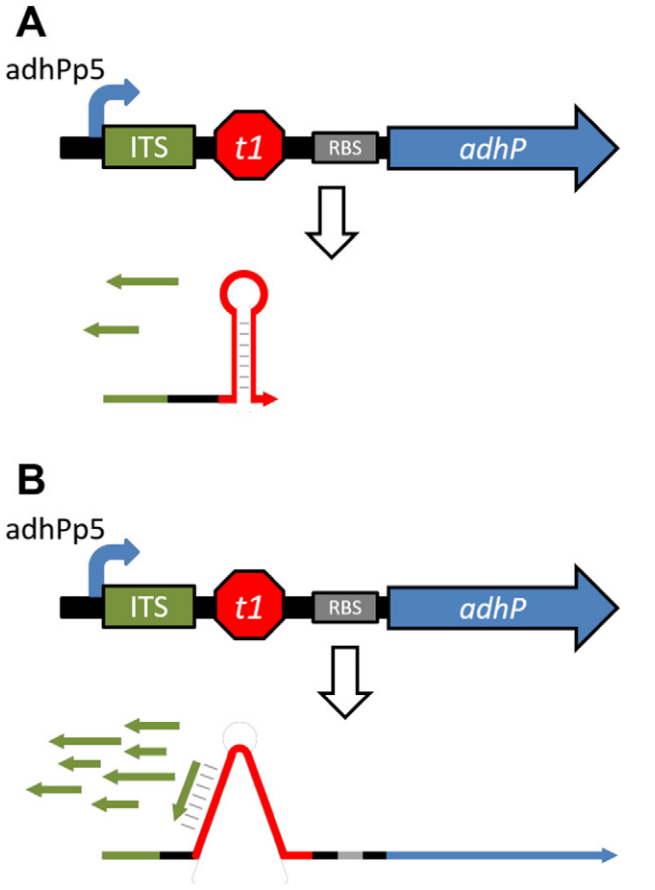

Fig. 4. Proposed mechanism of faRNA action in the *adhP* transcriptional unit. Initial transcribed sequence (ITS) in green, *adhP* in blue, *adhPt1* intrinsic terminator in red, and *adhP* ribosomal binding site (RBS) in grey. For panels A and B, the upper portion of the panel depicts DNA while the bottom portion depicts RNA transcripts and bound/unbound abortive fragments. Curved arrow represents *adhPp5* promoter.

(A) Proposed model of *adhP* transcription during non-inducing (no ethanol) conditions. Basal transcription from the *adhp5* promoter results in the production of few abortive transcripts. Under such conditions, the intrinsic terminator stops transcription ahead of *adhP* coding sequence.

(B) Proposed model of *adhP* production during inducing (with ethanol) conditions. Ethanol-induced increased transcription at *adhp5* promoter results in abortive accumulation. Excess abortive transcripts exert antitermination, enabling transcription of the *adhP* coding sequence.

promoter. The resultant increase in transcription would favor abortive synthesis rates and concurrently, the concentration of abortive RNA at this promoter – possibly to functionally relevant levels. As a result, we propose the following faRNA-mediated mechanism for *adhP* regulation: under non-inducing conditions, transcription of *adhP* is terminated by the intrinsic terminator upstream of the coding region. Under inducing conditions, accumulation of abortive fragments leads to antitermination and production of a full *adhP* transcript (Fig. 4). This proposed mechanism is of particular interest due to its similarity to the mechanism described by Lee et al. in the T7 bacteriophage, where abortive fragments exert an antitermination effect on a downstream terminating hairpin [13].

3.1.3. *fecABCDE-fecIR*

While iron is an essential nutrient for *E. coli*, it can be toxic at high concentrations and is difficult to store due to its relatively poor solubility. As a result, expression of genes pertaining to iron homeostasis is highly regulated [25]. The uptake of extracellular iron in the form of ferric citrate is mediated by the *fec* system, encoded by the *fecABCDE* and *fecIR* transcriptional units [26]. The first of these encodes for components of the ferric citrate uptake receptor, while the second encodes for a *fec*-specific sigma factor FecI and a membrane bound signaling protein FecR. Regulation of *fec* transcription has been well described. When FecA binds extracellular ferric citrate, a signal is relayed through FecR to FecI. This sigma factor then activates transcription of *fecABCDE*, resulting in increased ferric citrate uptake capabilities. Furthermore, *fecIR* is regulated by the transcriptional repressor Fur, which prevents transcription when bound to Fe^{2+}. Therefore, *fec* operon transcription is activated when intracellular ferrous concentrations are low and extracellular ferric concentrations are high, and is turned off once a sufficient amount of iron has been taken in. However, continued translation from full-length mRNA could be detrimental to the cells. Our analysis suggests that faRNAs may play a translational role to in conjunction transcriptional regulation by Fur to ensure tight regulation (Fig. 5). We

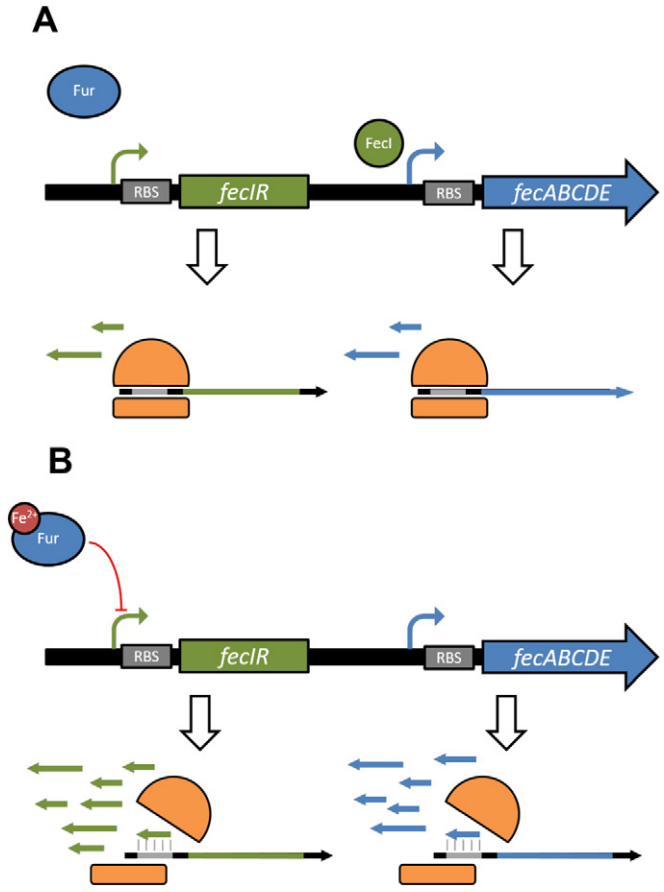

Fig. 5. Proposed mechanism of faRNA action in the *fecIR* and *fecABCDE* transcriptional units. *fecIR* in green, *fecABCD* in blue, and ribosomal binding site (RBS) in grey. For panels A and B, the upper portion of the panel depicts DNA while the bottom depicts RNA transcripts and bound/unbound abortive fragments. Curved arrows represent promoters.

(A) Predicted model of *fec* regulation under conditions of low intracellular ferrous concentration and high extracellular ferric citrate concentration. *fecIR* is de-repressed by Fur, allowing FecA to signal the presence of ferric citrate through FecI and activate *fecABCDE* transcription. Abortive initiation fragment concentrations are low, as transcription rates from both promoters were previously low. Therefore, both transcription and translation rates for both transcriptional units are high, and ferric citrate uptake increases.

(B) Predicted model of *fec* regulation following uptake of ferric citrate through the *fec* system. Previously transcription of both TU due to upregulation raises the concentration of abortive fragments, which can bind to the ribosomal binding sites for *fecA* and *fecR* and block translation of both proteins. Fur binds to Fe^{2+} and represses transcription of *fecIR*, which in turn prevents *fecABCDE* transcription. Therefore, both transcription and translation are interrupted for two key components of the system, and ferric citrate uptake is shut off.

found that abortive transcripts produced from the *fecAp* and *fecIp* promoters may be able to bind to the RBSs of mRNA of *fecA* (two-tailed *p*-value = 0.0148) and *fecR* (two-tailed *p*-value = 0.00394) respectively, occluding the binding sites and preventing translation. High levels of *fec* transcription would lead to high concentrations of faRNA, which in turn could act as time-delayed translational repressors of the *fec* operon. Thus, faRNA translational repression might therefore work in tandem with Fur-mediated transcriptional repression to more rapidly shut down iron intake, minimizing inadvertent toxicity.

3.2. Implications, conclusions, and future directions

The unique attributes of faRNAs make them a complementary addition to the corpus of known varieties of ncRNAs [2,3,19,27,28]. They extend our understanding of the roles for ncRNAs in implementation of a variety of different cellular mechanisms. Other ncRNAs have been shown to interact with both terminators and ribosomal binding sites in a variety of ways [29]. To highlight the unique attributes of faRNAs, we have included a comparison between faRNAs and other known varieties of ncRNAs in S3 Table (Supporting information online). The existence of ncRNA mechanisms somewhat akin to the ones we propose here strengthens the likelihood for faRNAs as a novel type of regulatory ncRNA. The discovery of faRNAs further adds to our understanding of the biological functionality of ncRNAs. Our analysis indicates putative biological roles for faRNAs, possibly removing these tiny transcripts from the category of transcriptional noise. Future experimental validation could provide further evidence for the presence of this regulatory system in *E. coli*, and subsequently, other organisms as well.

ncRNAs of various types have been shown to be highly prevalent across species [27,28]. However, it remains to be seen how widespread and conserved faRNAs may be. Abortive initiation has been shown in nearly all observed transcription reactions [4–6]. Therefore, it is possible that faRNAs are a ubiquitous regulatory strategy. As of now, they have only been shown in the T7 bacteriophage [13]. With validation and additional research, we hope to establish a role for faRNAs in *E. coli* and eventually other prokaryotes. This approach could also be extended to eukaryotic organisms. However, it is unclear as to how the increased complexity of eukaryotic gene expression might affect the ability of abortive transcripts to exert regulatory effects.

Our goal here was to identify a small set of putative faRNA binding sites for preliminary verification efforts. Therefore, we could make several assumptions that both simplified our search and helped limit our output. For example, we chose to allow no mismatched base pairs in our search, even though the T7 bacteriophage example indicated that imperfect matching does not prevent abortive fragment binding. We will be able to easily modify these assumptions in the future due to the inherent flexibility of computational methods. Additionally, there is currently a lack of information regarding the synthesis, degradation, and binding of abortive RNA transcripts. As this information becomes available, we can further modify our program and assumptions to reflect our growing understanding of how to find faRNAs.

The exact nature of the link between abortive initiation rates and promoter strength and structure has not been demonstrated systematically in literature. Furthermore, there are no known mechanisms by which cells regulate abortive transcription frequency. Lee et al. demonstrated that for the T7 bacteriophage gene 10 there exists a positive correlation between an increase in the number of initiation events and the antitermination efficiency through *ϕ10* terminator. While this does not directly show a correlation between transcription initiation rate and abortive concentration, it does suggest a mechanism through which faRNA-associated functions can be regulated in a physiologically-relevant manner [13]. Many genes are regulated by multiple promoters at multiple transcription start sites. Therefore, it is possible that regulation of transcription initiation rates at alternative promoters could be important for regulation of abortive-mediated function. Further elucidation of promoter characteristics and cellular mechanisms that influence abortive synthesis will be invaluable in refining our model in predicting the functional role of faRNA.

We could implement more functionally-relevant sites by identifying abortive hybridization in features such as rho-dependent terminators, attenuators, riboswitches, RNAse binding sites etc. However, consensus sequences for these features are relatively poorly defined, making it difficult to glean functional relevance. Furthermore, experimental validation of the most promising loci for anti-termination or anti-translation may be first warranted. Quantification of degradation and migration rates of short RNA fragments could inform our determination of where a given faRNA is able to bind based on spatial proximity to the appropriate transcription start site. Lee et al. showed that poly-G RNA sequences generated by slippage of RNAP during the initiation phase were able to bind more strongly to the *Tϕ10* terminating hairpin than the

abortive generated from the initially transcribed sequence [13]. Given the ability to identify factors affecting RNAP slippage and abortive production, we might be able to generate a profile of different RNA fragments produced from a single promoter and determine whether this affects the likelihood of functional binding.

Our search returned significantly fewer abortive matches within RBSs as compared to terminators. Furthermore, the largest $\Delta G^{\circ}_{binding}$ values from the RBS matches were considerably lower than those from the terminator matches (Fig. 2). This fact could be entirely coincidental, or it might reflect the difference in $\Delta G^{\circ}_{binding}$ required for occlusion of a riboprotein binding site and disruption of RNA secondary structure. Further investigation of this discrepancy may shape the way in which faRNAs are used synthetically, as well as direct future searches for faRNAs in nature.

The defining characteristics of faRNAs may be well suited to the design of novel synthetic regulatory pathways. Lee et al. showed that abortive antitermination is concentration dependent [13]. Abortive transcript production relies on the activity of a promoter; upregulation of a given gene can therefore result in an increase in the concentration of corresponding abortive fragments. This indicates that faRNAs could act as time-delayed responses to regulatory stimuli. The small size and quick degradation of abortive transcripts should limit their ability to migrate, allowing them to act as a local control mechanism. Abortive transcript production also demands lesser cellular resources than longer ncRNAs. Resultantly, faRNAs could act as concentration-dependent localized regulatory mechanism that do not impose significant metabolic burden to cells. Such a regulatory RNA could provide novel modes of gene regulation in rationally designed synthetic systems.

4. Materials and methods

4.1. Determining transcriptional unit locations and sequences

E. coli genome file version U00096.2 containing 5′-3′ forward strand sequence in GeneBank format was obtained from NCBI (http://www.ncbi.nlm.nih.gov/nuccore/U00096.2). Locations and descriptions of transcription units were derived from the "5′ and 3′ UTR sequence of TUs" file obtained from RegulonDB [30]. Based on these locations and direction of transcription, sequences for each transcription unit were extracted from the full genome file. This transcriptome data includes sequence data for loci that are transcribed by multiple promoters.

4.2. Determining abortive fragment sequences

Given the nucleotide sequence and the transcription start sites for all genes in the *E. coli* genome, we determined the sequence of each abortive fragment that is generated through abortive initiation of transcription for each gene, assuming no transcriptional errors. These sequences were determined for fragments of each length within the specified range (4–15 nt). An abortive fragment of length k is defined as the first k nucleotides of the coding strand for the given transcription unit starting from the transcription start site. Thus the abortive fragment for a forward strand transcription unit with transcription start site at n is composed of forward strand nucleotides $5'\{n, n+1, \ldots, n+k-1\}3'$, and the abortive fragment for a reverse strand transcription unit with transcription start site n is composed of reverse-strand nucleotides $5'\{n, n-1, \ldots, n-k+1\}3'$.

4.3. Abortive fragment match site search

Using the sequence of each k-length abortive fragment, our program searches for binding sites (reverse-complementary sequence matches) within the transcription unit from which it was generated. We chose to search only within the same transcription unit that generated the abortive since we hypothesize that the small unstructured nature of abortive results in quick turnover, which would limit their sphere of influence to spatially and temporally proximal RNA. Match sites are identified based on two nucleotide binding paradigms, a standard Watson-Crick base-pairing model (A-U/C-G) or a more flexible wobble base-pairing model (A-U/C-G/U-G). Allowances for mismatched base pairs under the desired binding paradigm can be made via a user-defined mismatch parameter. In this study we utilized both binding paradigms and allowed no mismatches, providing two similar outputs with different levels of selectivity. We also calculated the percent probability of finding a random hybridization target for a specified length of abortive within the transcriptome or functionally relevant sequence such as terminator or ribosomal binding site (RBS).

4.4. Calculating expected number of matches due to chance

To contextualize our results, we compared number of matches found across all TUs to the number of matches expected to be found by random chance, calculated for each abortive length as follows. First, we calculated the percent usage for each of the four nucleotides for each TU. We then calculated the probability of generating a valid binding site for the abortive fragments of each length n from this usage distribution. For n-length abortive fragment $abortive = a_1, a_2, \ldots a_n$, we produced a binding site template $binding\text{-}site = B_1, B_2, \ldots B_n$, where B_i is the set of bases that can pair with a_i under the desired base-pairing paradigm, and $P(B_i)$ is the sum of the probabilities of those bases. We were therefore able to calculate the probability of each binding site as $P(binding\text{-}site) = P(B_1) \times P(B_2) \times \ldots \times P(B_n)$. For a transcriptional unit of length k, the estimate of the expected number of matches $e = k/P(abortive)$.

4.5. Feature matching (functional relevance analysis)

To evaluate the functional relevance of abortive fragment matches, we selected the subset of fragments with binding sites overlapping gene regulatory features. For our analysis, datasets identifying the locations of terminators and ribosomal binding sites were obtained from WebGeSTer DB and RegulonDB respectively [24,30]. Our program reports all abortive fragments of a given length that coincide with a regulatory feature and the number of nucleotides of overlap.

4.6. Free energy calculations

We calculated Gibbs free energy of hybridization ($\Delta G^{\circ}_{binding}$) values for abortive-target pairs to rank the binding strength of putative faRNAs. $\Delta G^{\circ}_{binding}$ values were obtained using the *hybrid2.pl* program included in the UNAFold software suite [31]. We used the RNA sequence and energy-only options, a temperature of 25 °C, and concentrations of 1.0 μM for both nucleotide sequences [13]. Unique faRNA complexes were then ranked based on this calculated value, where complexes with more negative $\Delta G^{\circ}_{binding}$ values were ranked higher. A unique faRNA complex was defined as the longest abortive transcript derived from a specific transcription start site that was predicted to match a specific binding site.

4.7. Fisher's Exact Test calculation

Calculation of Fisher's Exact Test statistic was performed using the DendroPy Phylogenetic Computing Library [32]. For each test, the population of potential functional binding sites was defined as the number of positions for which binding would cause an overlap between an abortive fragment and the target regulatory feature. The population of potential non-regulatory binding sites was defined as the number of remaining positions in the entire transcriptional unit.

4.8. Wilcoxon rank-sum test

Calculation of the Wilcoxon rank-sum test was performed using the wilcox.test function, using the R language. To evaluate the distribution

of $\Delta G^{\circ}_{binding}$ values of matches in regulatory regions compared to the distribution in non-regulatory regions, we calculated the *p*-value. The Wilcoxon rank-sum test does not assume a normal distribution of $\Delta G^{\circ}_{binding}$, and is applicable for the given population sizes.

4.9. Manual curation

We queried the EcoCyc database with all TUs containing faRNA matches [33]. The information obtained from the database was used to investigate possible roles of faRNA binding in regulation of downstream gene products within each TU.

Abbreviations

faRNA	functional abortive RNA
ncRNA	noncoding RNA
sRNA	small RNA
asRNA	anti-sense RNA
RNAP	RNA polymerase
RBS	ribosomal binding site
TU	transcriptional unit

Acknowledgments

This work was supported by the Tufts University Summer Scholars Program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2017.03.003.

References

[1] J. Vazquez-anderson, L.M. Contreras, Charming Gene Management Styles for Synthetic Biology Applications, 10, 2013 1778–1797.
[2] F. Repoila, F. Darfeuille, Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects, Biol. Cell. 101 (2009) 117–131, http://dx.doi.org/10.1042/BC20070137.
[3] M.K. Thomason, G. Storz, Bacterial antisense RNAs: how many are there, and what are they doing? Annu. Rev. Genet. 44 (2010) 167–188, http://dx.doi.org/10.1146/annurev-genet-102209-163523.
[4] S.C. Nam, C.W. Kang, Transcription initiation site selection and abortive initiation cycling of phage SP6 RNA polymerase, J. Biol. Chem. 263 (1988) 18123–18127.
[5] A.J. Carpousis, J.D. Gralla, Cycling of ribonucleic acid polymerase to produce oligonucleotides during initiation in vitro at the lac UV5 promoter, Biochemistry 19 (1980) 3245–3253, http://dx.doi.org/10.1021/bi00555a023.
[6] B. Lescure, V. Williamson, A. Sentenac, Efficient and selective initiation by yeast RNA polymerase B in a dinucleotide-primed reaction, Nucleic Acids Res. 9 (1981) 31–45, http://dx.doi.org/10.1093/nar/9.1.31.
[7] S.R. Goldman, R.H. Ebright, B.E. Nickels, Direct detection of abortive RNA transcripts *in vivo*, Science 324 (2009) 927–928, http://dx.doi.org/10.1126/science.1169237.
[8] D.S. Luse, G.A. Jacob, Abortive initiation by RNA polymerase II in vitro at the adenovirus 2 major late promoter, J. Biol. Chem. 262 (1987) 14990–14997.
[9] L.M. Hsu, Promoter clearance and escape in prokaryotes, Biochim. Biophys. Acta Gene Struct. Expr. (2002) 191–207, http://dx.doi.org/10.1016/S0167–4781(02)00452-9.
[10] P.J. Lopez, J. Guillerez, R. Sousa, M. Dreyfus, The low processivity of T7 RNA polymerase over the initially transcribed sequence can limit productive initiation *in vivo*, J. Mol. Biol. 269 (1997) 41–51, http://dx.doi.org/10.1006/jmbi.1997.1039.
[11] N.V. Vo, L.M. Hsu, C.M. Kane, M.J. Chamberlin, In vitro studies of transcript initiation by *Escherichia coli* RNA polymerase. 3. Influences of individual DNA elements within the promoter recognition region on abortive initiation and promoter escape, Biochemistry 42 (2003) 3798–3811, http://dx.doi.org/10.1021/bi026962v.
[12] J. Houseley, D. Tollervey, The many pathways of RNA degradation, Cell 136 (2009) 763–776, http://dx.doi.org/10.1016/j.cell.2009.01.019 (Elsevier Inc.).
[13] S. Lee, H.M. Nguyen, C. Kang, Tiny abortive initiation transcripts exert antitermination activity on an RNA hairpin-dependent intrinsic terminator, Nucleic Acids Res. 38 (2010) 6045–6053, http://dx.doi.org/10.1093/nar/gkq450.
[14] R.G. Keene, D.S. Luse, Initially transcribed sequences strongly affect the extent of abortive initiation by RNA polymerase II, J. Biol. Chem. 274 (1999) 11526–11534, http://dx.doi.org/10.1074/jbc.274.17.11526.
[15] A.N. Kapanidis, E. Margeat, S.O. Ho, E. Kortkhonjia, S. Weiss, R.H. Ebright, Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism, Science 314 (2006) 1144–1147, http://dx.doi.org/10.1126/science.1131399.
[16] S.R. Goldman, J.S. Sharp, I.O. Vvedenskaya, J. Livny, S.L. Dove, B.E. Nickels, NanoRNAs prime transcription initiation *in vivo*, Mol. Cell 42 (2011) 817–825 (Elsevier Inc.) 10.1016/j.molcel.2011.06.005.
[17] D. Na, S.M. Yoo, H. Chung, H. Park, J.H. Park, S.Y. Lee, Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs, Nat. Biotechnol. 31 (2013) 170–174, http://dx.doi.org/10.1038/nbt.2461 (Nature Publishing Group).
[18] S. Oehler, B. Müller-Hill, High local concentration: a fundamental strategy of life, J. Mol. Biol. 395 (2010) 242–253, http://dx.doi.org/10.1016/j.jmb.2009.10.056.
[19] N. Nakashima, S. Goh, L. Good, T. Tamura, Multiple-gene silencing using antisense RNAs in *Escherichia coli*, in: M. Kaufmann, C. Klinger (Eds.), Methods, Springer New York, New York, NY 2012, pp. 307–319, http://dx.doi.org/10.1007/978-1-61779-424-7.
[20] M.A. Sørensen, J. Fricke, S. Pedersen, Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in *Escherichia coli in vivo*, J. Mol. Biol. 280 (1998) 561–569, http://dx.doi.org/10.1006/jmbi.1998.1909.
[21] M. Freundlich, N. Ramani, E. Mathew, A. Sirko, P. Tsui, The role of integration host factor in gene expression in *Escherichia coli*, Mol. Microbiol. 6 (1992) 2557–2563.
[22] A. Węgleńska, B. Jacob, A. Sirko, Transcriptional pattern of *Escherichia coli* ihfB (himD) gene expression, Gene 181 (1996) 85–88, http://dx.doi.org/10.1016/S0378-1119(96)00468-4.
[23] J. Shafqat, J.O. Höög, L. Hjelmqvist, U.C.T. Oppermann, C. Ibáñez, H. Jörnvall, An ethanol-inducible MDR ethanol dehydrogenase/acetaldehyde reductase in *Escherichia coli*: structural and enzymatic relationships to the eukaryotic protein forms, Eur. J. Biochem. 263 (1999) 305–311, http://dx.doi.org/10.1046/j.1432-1327.1999.00323.x.
[24] A. Mitra, A.K. Kesarwani, D. Pal, V. Nagaraja, WebGeSTer DB—a transcription terminator database, Nucleic Acids Res. 39 (2011) D129–D135, http://dx.doi.org/10.1093/nar/gkq971.
[25] S.C. Andrews, A.K. Robinson, F. Rodríguez-Quiñones, Bacterial iron homeostasis, FEMS Microbiol. Rev. (2003) 215–237, http://dx.doi.org/10.1016/S0168-6445(03)00055-X.
[26] V. Braun, S. Mahren, A. Sauter, Gene regulation by transmembrane signaling, Biometals 19 (2006) 103–113, http://dx.doi.org/10.1007/s10534-005-8253-y.
[27] Z. Qu, D.L. Adelson, Evolutionary conservation and functional roles of ncRNA, Front. Genet. 3 (2012)http://dx.doi.org/10.3389/fgene.2012.00205.
[28] T.R. Cech, J.A. Steitz, The noncoding RNA revolution - trashing old rules to forge new ones, Cell (2014) 77–94, http://dx.doi.org/10.1016/j.cell.2014.03.008.
[29] L.S. Qi, A.P. Arkin, A versatile framework for microbial engineering using synthetic non-coding RNAs, Nat. Rev. Microbiol. 12 (2014) 341–354, http://dx.doi.org/10.1038/nrmicro3244 (Nature Publishing Group).
[30] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz-Rascado, J.S. García-Sotelo, et al., RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, Nucleic Acids Res. 41 (2013) D203–D213, http://dx.doi.org/10.1093/nar/gks1201.
[31] N.R. Markham, M. Zuker, UNAFold: software for nucleic acid folding and hybridization, Methods Mol. Biol. 453 (2008) 3–31, http://dx.doi.org/10.1007/978-1-60327-429-6-1.
[32] J. Sukumaran, M.T. Holder, DendroPy: a python library for phylogenetic computing, Bioinformatics 26 (2010) 1569–1571, http://dx.doi.org/10.1093/bioinformatics/btq228.
[33] I.M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, et al., EcoCyc: fusing model organism databases with systems biology, Nucleic Acids Res. 41 (2013) http://dx.doi.org/10.1093/nar/gks1027.