Computational Prediction of Functional Abortive RNA in *E. coli*

An Honors Thesis for the Department of Computer Science

Jeremy I. Marcus

Tufts University 2016

Acknowledgments

Thank you to the Tufts Summer Scholars initiative for providing funding during the initial phase of research in the summer of 2014. And of course, special thanks to the advisors of this project, Dr. Soha Hassoun from the Department of Computer Science and Dr. Nikhil U. Nair from the Department of Chemical and Biological Engineering, without whom this thesis would not have been possible.

Table of Contents

Abstract

Failure by RNA polymerase to break contacts with promoter DNA results in release of bound RNA and re-initiation of transcription. These abortive RNA 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 3,780 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 will confirm our findings and promote exploration of functional abortive RNAs (faRNAs) in natural and synthetic systems.

Introduction

Non-coding RNAs (ncRNAs) are RNA molecules that are not translated into proteins. These single-stranded RNA molecules have been shown to be involved in a variety of functions ranging from mRNA translation to polynucleotide degradation. Many of their functions are involved in important gene regulatory mechanisms, such as up- and down-regulation of translation and the occlusion of binding sites for other regulatory molecules, and they have been used extensively in designing 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. Upon proceeding to 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 caused 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; as a result, 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, at approximately 4.6 million base pairs in length, 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 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. Here we suggest mechanisms of regulatory control for three of the abortive fragments returned by our analysis.

Results

Abortive RNA and Target Identification

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 equally likely, and therefore performed our initial Watson-Crick base pairing search over the entire range for each of the 3,780 known transcriptional units in the *E. coli* genome. Results for abortives that have matches within intrinsic terminators and ribosomal binding sites (RBSs) are also shown. The results of the search are presented in Table 1. Table 2 presents the results of a similar search performed with allowance for G-U wobble base pairing.

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

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	3,448	26,290 (696%)	89 (2.36%)	6 (0.159%)
5	2,386	6,784 (180%)	26 (0.688%)	2 (0.0529%)
6	1,097	1,789 (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) basepairing.

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	3,723	122,949 (3253%)	460 (12.2%)	54 (1.43%)
5	3,484	46,962 (1242%)	191 (5.05%)	29 (0.767%)
6	2,956	18,609 (492%)	86 (2.28%)	11 (0.291%)
7	2,107	7,167 (190%)	41 (1.09%)	6 (0.159%)
8	1,313	2,709 (71.7%)	12 (0.318%)	4 (0.106%)
9	719	1,078 (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%.

Top Ranking Abortive-Target Pairs and their Proposed Mechanism of Action

For each abortive match located in a functionally important part of a transcription unit (terminator or RBS), we calculated the Gibbs free energy of hybridization ($\Delta G°_{binding}$) of the reaction 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 were able to utilize free energy as a quantitative measure of an oligomer's potential to exert function, providing us with a basis for comparison between individual matches. Fig 1 and 2 display the distribution of free energy values calculated for abortive matches located in terminating regions and ribosomal binding sites respectively.

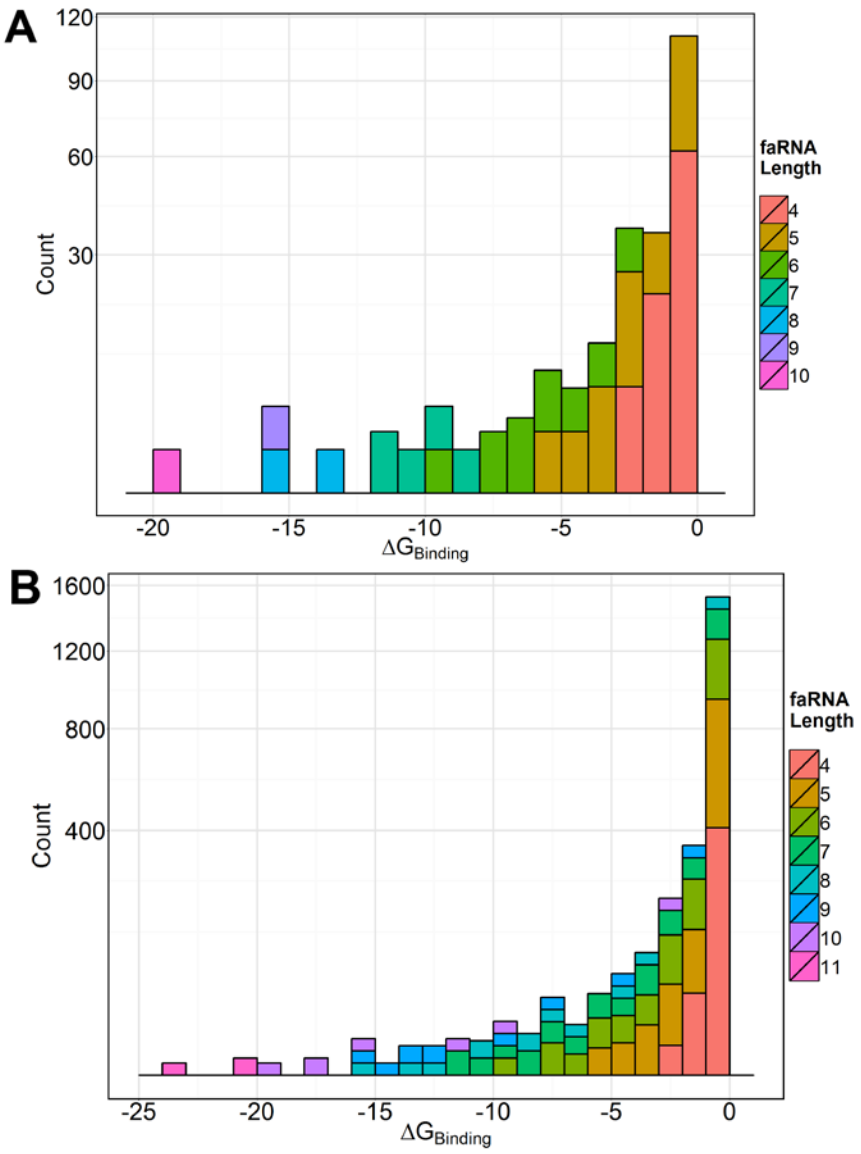

Fig 1. Distribution of computed $\Delta G°_{binding}$ values for predicted complexes between abortive initiation fragments and mRNAs. We only report complexes which occur within the same transcriptional unit from which the abortive initiation fragment was generated. RNA binding free energy calculations performed using UNAFold *hybrid2.pl*, with temperature range 0-100ºC, 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) $\Delta G°_{binding}$ distribution for complexes exhibiting exact Watson-Crick (A-U, G-C) base pair matching. 135 predicted complexes were identified with $\Delta G°_{binding} \geq -19.79$ kcal/mol.

(B) $\Delta G°_{binding}$ distribution for complexes exhibiting exact base pair matching under a wobble base-pairing paradigm (A-U, G-C, G-U). We identified 810 predicted pairs with $\Delta G°_{binding} \geq -23.04$ kcal/mol.

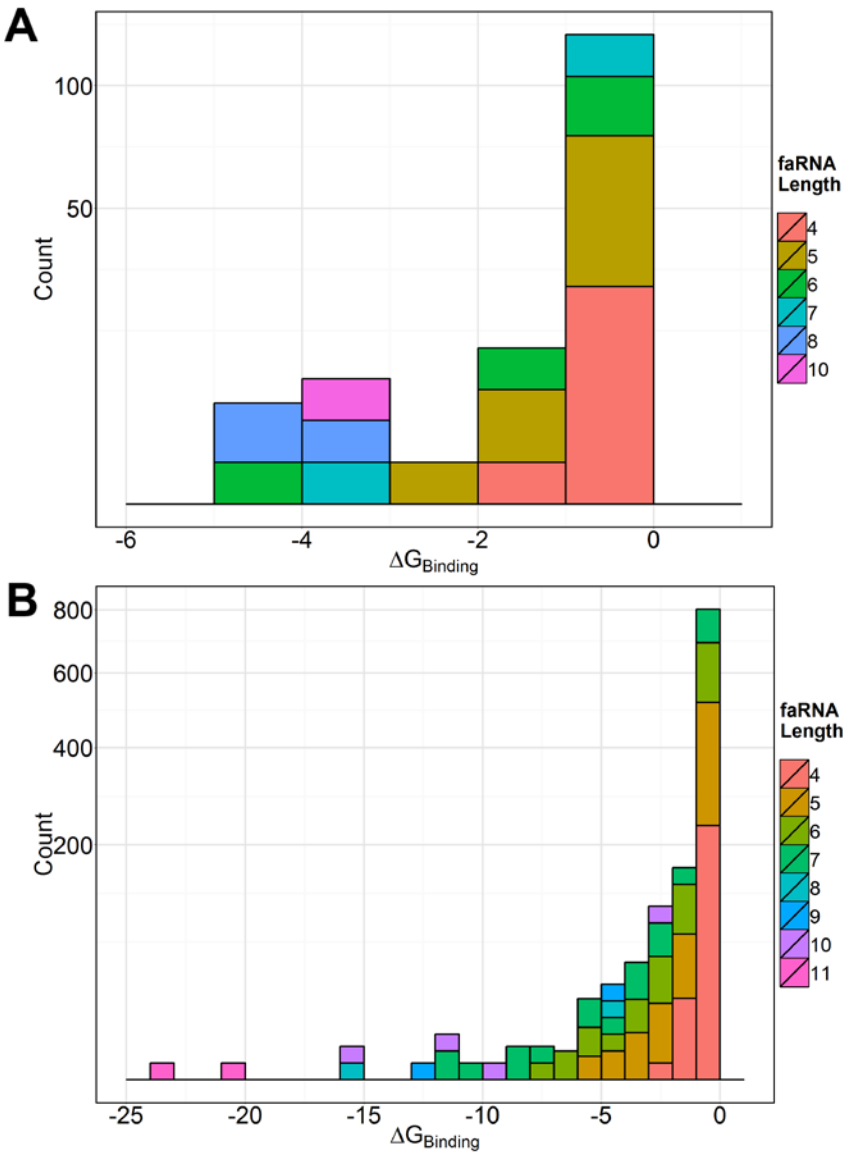

Fig 2. Distribution of computed $\Delta G°_{binding}$ values for predicted complexes between abortive initiation fragments and mRNAs occurring within functionally relevant genetic regions. We only report complexes which occur within the same transcriptional unit from which the abortive initiation fragment 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 0-100ºC, 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) $\Delta G°_{binding}$ distribution for complexes with footprints overlapping ribosomal binding sites (RBSs). 55 predicted complexes were found to overlap RBSs, with $\Delta G°_{binding} \geq$ -4.95 kcal/mol.

(B) $\Delta G°_{binding}$ distribution for complexes with footprints overlapping rho-independent terminators.

Sequence information on the three loci discussed below are in S1 Fig and S1 and S2 Tables (Supporting information online).

The existence of a minimum free energy requirement to achieve effective antitermination or anti-translation by antisense hybridization has not been shown. But sRNA-dependent regulation even with the aid of Hfq has been shown to be weak when hybridization energies are >-10 kcal/mol [17]. However, this energetic requirement may not apply to faRNAs as they can be present at extremely high concentrations (since each productive transcript may result in >10-100 abortives) in close proximity to their target. Therefore, as subjects for thorough investigation we chose the top matches from the free energy rankings for terminators and RBSs within the same transcript under both binding paradigms. We propose faRNA-mediated gene regulation mechanisms for these loci.

Proposed Mechanisms of faRNAs at Different Loci

***adhP*.** Expression of the alcohol dehydrogenase encoded by the gene *adhP* is known to be induced in the presence of ethanol [18]. However, little is known about the exact induction mechanisms. Immediately downstream of the transcription start site (TSS) for the TU lies a putative strong intrinsic terminator. 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 transcriptional unit are an exact reverse-complement match to part of the leading stem of the terminator. In the presence of sufficiently high concentrations of these *adhP* abortive fragments, the terminating hairpin may not be able to form. Conceivably, there could be

some as-yet-unidentified transcription factor that activates *adhP* transcription in the presence of ethanol and increases binding of RNAP to this promoter. The resultant increase in transcription rates (and effective affinity of RNAP complex for promoter) would favor abortive cycling 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 3).

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 [19]. *rpsA* lies directly upstream of and shares a transcriptional unit 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 [20]. *ihfB* transcription patterns have been shown to vary predictably based on the observed rate of growth of the cell [21]. 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 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 [21]. 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, allowing for anti-termination and co-transcription of *rpsA* and *ihfB* (Fig 4). 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.

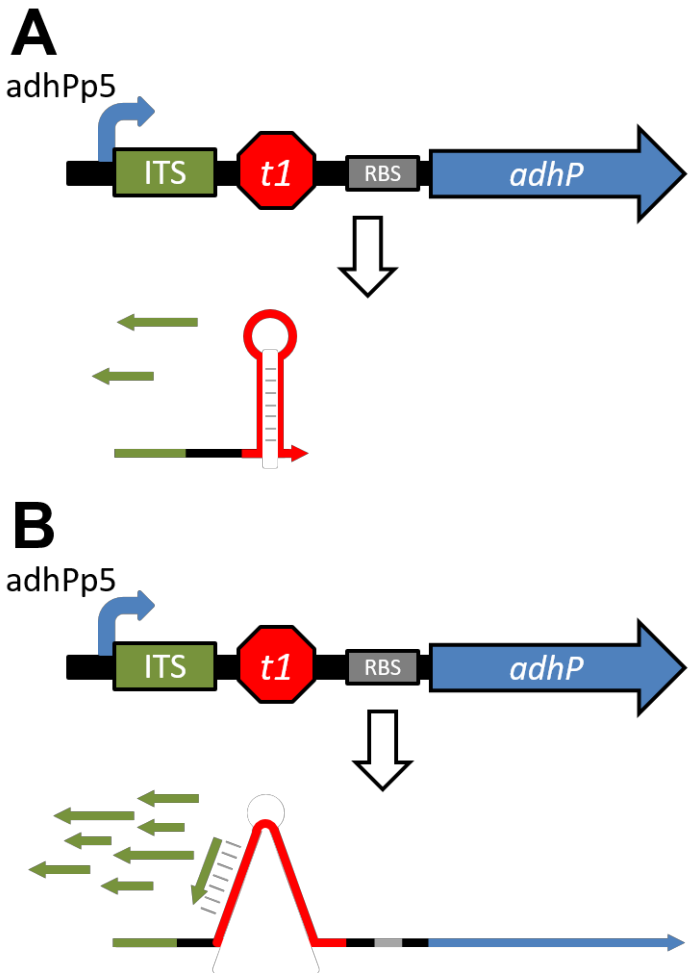

Fig 3. Proposed mechanism of faRNA action in the *adhP* transcriptional unit. Initial transcribed sequence (ITS) in green, *adhP* in blue, adhPt1 intrinsic terminator in red, and *adhP* ribosomal binding site (RBS) in grey. Top figures are DNA; bottom figures are RNA transcripts and bound/unbound abortive 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 affinity of transcription for adhp5 promoter results in enhanced abortive cycling. Excess abortive transcripts exert antitermination, enabling transcription of the *adhP* coding sequence.

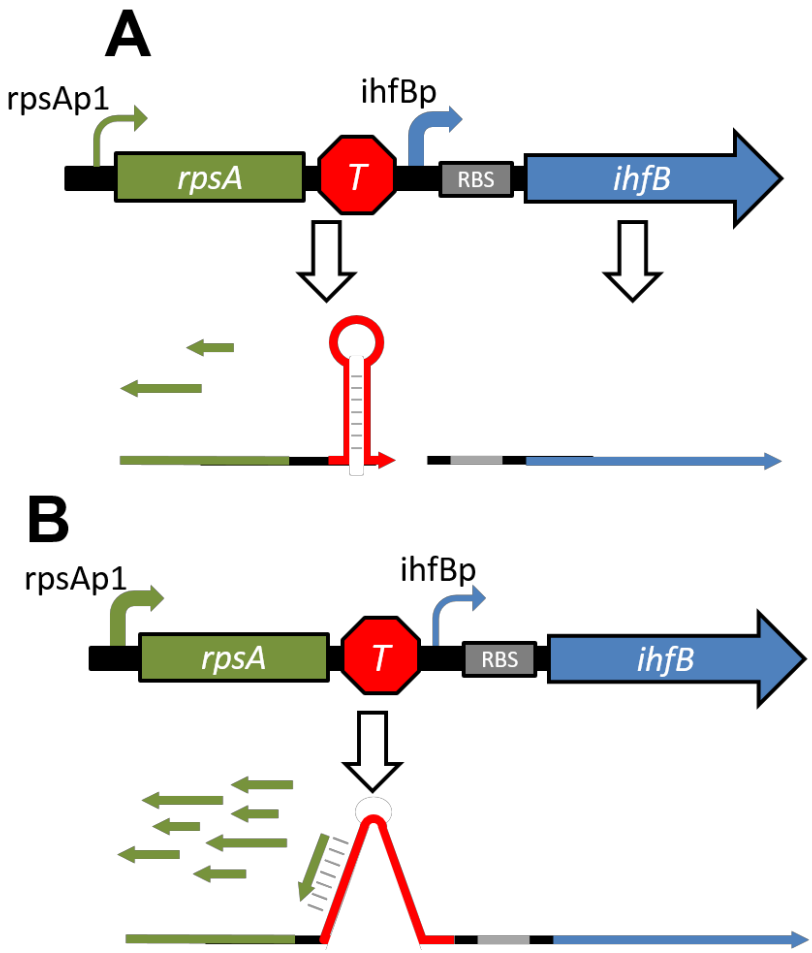

Fig 4. Proposed mechanism of faRNA action in the *rpsA-ihfB* transcriptional unit. *rpsA* in green, *ihfB* in blue, *rpsA* ribosomal binding site (RBS) in grey, and *rpsA* terminator in red. Top figures are DNA; bottom figures are RNA transcripts and bound/unbound abortive fragments. Curved arrows represent 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. As a result, 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.

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 [22]. 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. 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 [23]. 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 transcripts 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 found that abortive transcripts produced from the fecAp and fecIp promoters may be able to bind to the RBSs of mRNA of *fecA* and *fecR* 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 a 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.

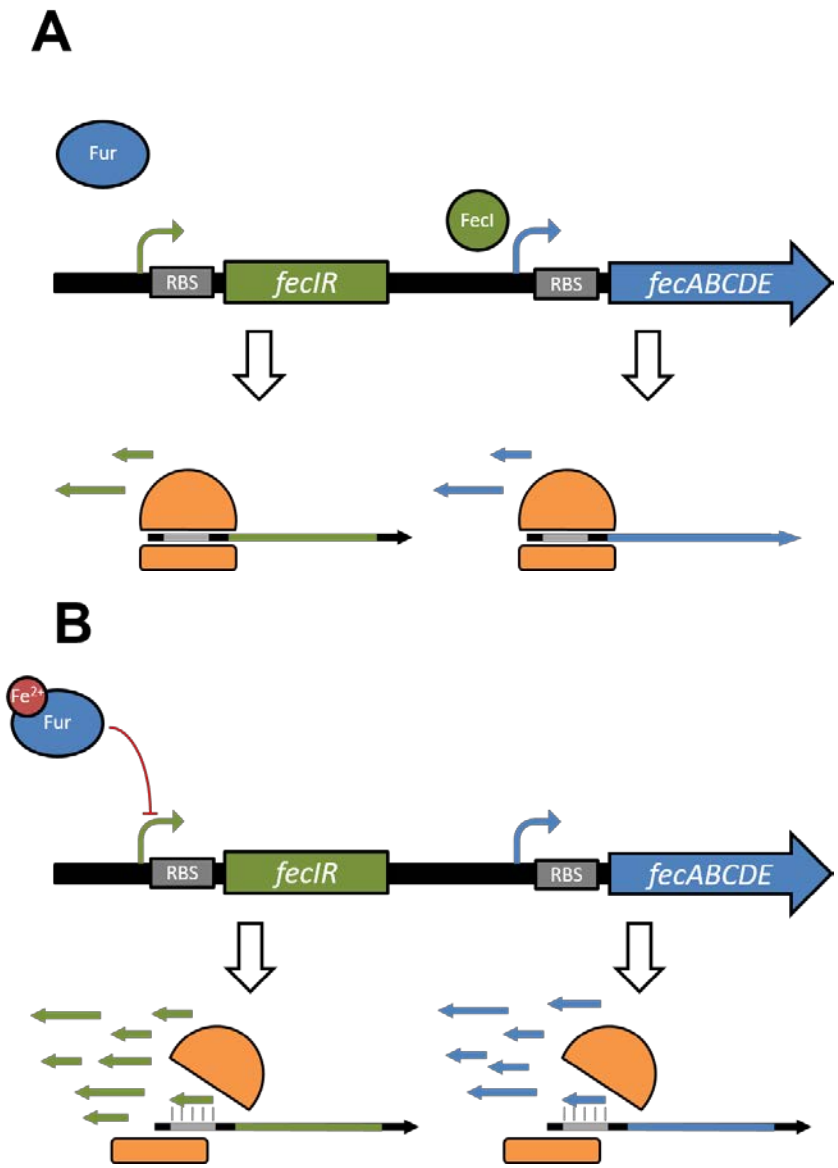

Fig 5. Proposed mechanism of faRNA action in the *fecIR* and *fecABCDE* transcriptional units. *fecIR* in green, *fecABCD* in blue, and ribosomal binding site (RBS) in grey. Top figures are DNA; bottom figures are RNA transcripts and bound/unbound abortive fragments. Curved arrows represent promoters.

(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 are able to 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.

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. We also chose to find hybridization targets only within the same transcription unit that produced the abortive. This was done because abortives are short unstructured RNA that are likely degraded quickly and 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 [24]. 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 ($\Delta G^{\circ}_{binding}$). We further hypothesized mechanisms of action for a number of our putative binding sites, three of which are described here. 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.

The unique attributes of faRNAs make them a complementary addition to the corpus of known varieties of ncRNAs (Table 3). 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 [25]. 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.

Table 3. Comparison of features of putative functional abortive RNAs (faRNA) with other bacterial noncoding RNA.

faRNA (functional abortive)	sRNA (small)	asRNA (anti-sense)
Small	Large	Large
Localized	Global	Global
Largely unstructured	Complex structure	Complex structure
Produced during regular transcription	Produced from own promoter	Produced from own promoter
Functions independently of other molecules	Function may be mediated by Hfq and other interactions	Functions independently of other molecules

ncRNAs of various types have been shown to be highly prevalent across species [26,27]. 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 fairly 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 were able to make a number of 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.

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. Understanding how differences between promoters affect abortive transcription might allow us to better model faRNA concentrations and predict binding affinity. 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 Tϕ 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.

Materials and Methods

Determining Transcriptional Unit Locations and Sequences

E. coli genome file 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 (http://regulondb.ccg.unam.mx/menu/download/datasets/files/UTR_5_3_sequence.txt). Based on these locations, forward and reverse strand 5'-3' nucleotide 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.

Determining Abortive Fragment Sequences

For each iteration of the abortive fragment matching algorithm, a new value *k* is chosen from a specified range of abortive fragment lengths (4-15 nt). Given the current *k* value, *k*-length abortive fragments are 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, … , 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, … , n-k+1}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).

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 (http://pallab.serc.iisc.ernet.in/gester/89318805/index.html, http://regulondb.ccg.unam.mx/menu/download/datasets/files/RBSSet.txt). Our program reports all abortive fragments of a given length that coincide with a regulatory feature and the number of nucleotides of overlap.

Free Energy Calculations

We calculated Gibbs free energy of hybridization ($\Delta G°_{binding}$) values for abortive-target pairs to rank the binding strength of putative faRNAs. $\Delta G°_{binding}$ values were obtained using the *hybrid2.pl* program included in the UNAFold software suite [28]. We used the RNA sequence and energy-only options, a temperature range of 0-100 ºC, 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 $\Delta G°_{binding}$ 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.

Manual Curation

We queried the EcoCyc database with all transcriptional units containing faRNA matches [29]. The information obtained from the database was used to investigate possible roles of faRNA binding in regulation of downstream gene products within each transcriptional unit.

Abbreviations

faRNA, functional abortive RNA; ncRNA, noncoding RNA; sRNA, small RNA; asRNA, anti-sense RNA; RNAP, RNA polymerase, RBS, ribosomal binding site.

References

1. Vazquez-anderson J, Contreras LM. Charming gene management styles for synthetic biology applications. 2013;10: 1778–1797.
2. Repoila F, Darfeuille F. Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. Biol Cell. 2009;101: 117–131. doi:10.1042/BC20070137
3. Thomason MK, Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? Annu Rev Genet. 2010;44: 167–188. doi:10.1146/annurev-genet-102209-163523
4. Nam SC, Kang CW. Transcription initiation site selection and abortive initiation cycling of phage SP6 RNA polymerase. J Biol Chem. 1988;263: 18123–18127.
5. Carpousis AJ, Gralla JD. Cycling of ribonucleic acid polymerase to produce oligonucleotides during initiation in vitro at the lac UV5 promoter. Biochemistry. 1980;19: 3245–3253. doi:10.1021/bi00555a023
6. Lescure B, Williamson V, Sentenac A. Efficient and selective initiation by yeast RNA polymerase B in a dinucleotide-primed reaction. Nucleic Acids Res. 1981;9: 31–45. doi:10.1093/nar/9.1.31
7. Goldman SR, Ebright RH, Nickels BE. Direct detection of abortive RNA transcripts in vivo. Science. 2009;324: 927–928. doi:10.1126/science.1169237
8. Luse DS, Jacob GA. Abortive initiation by RNA polymerase II in vitro at the adenovirus 2

major late promoter. J Biol Chem. 1987;262: 14990–14997.

9. Hsu LM. Promoter clearance and escape in prokaryotes. Biochimica et Biophysica Acta - Gene Structure and Expression. 2002. pp. 191–207. doi:10.1016/S0167-4781(02)00452-9

10. Lopez PJ, Guillerez J, Sousa R, Dreyfus M. The low processivity of T7 RNA polymerase over the initially transcribed sequence can limit productive initiation in vivo. J Mol Biol. 1997;269: 41–51. doi:10.1006/jmbi.1997.1039

11. Vo N V, Hsu LM, Kane CM, Chamberlin MJ. 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. 2003;42: 3798–3811. doi:10.1021/bi026962v

12. Houseley J, Tollervey D. The Many Pathways of RNA Degradation. Cell. Elsevier Inc.; 2009;136: 763–776. doi:10.1016/j.cell.2009.01.019

13. Lee S, Nguyen HM, Kang C. Tiny abortive initiation transcripts exert antitermination activity on an RNA hairpin-dependent intrinsic terminator. Nucleic Acids Res. 2010;38: 6045–6053. doi:10.1093/nar/gkq450

14. Keene RG, Luse DS. Initially transcribed sequences strongly affect the extent of abortive initiation by RNA polymerase II. J Biol Chem. 1999;274: 11526–11534. doi:10.1074/jbc.274.17.11526

15. Kapanidis AN, Margeat E, Ho SO, Kortkhonjia E, Weiss S, Ebright RH. Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. Science. 2006;314: 1144–1147. doi:10.1126/science.1131399

16. Goldman SR, Sharp JS, Vvedenskaya IO, Livny J, Dove SL, Nickels BE. NanoRNAs Prime Transcription Initiation In Vivo. Mol Cell. Elsevier Inc.; 2011;42: 817–825.

doi:10.1016/j.molcel.2011.06.005

17. Na D, Yoo SM, Chung H, Park H, Park JH, Lee SY. Metabolic engineering of Escherichia coli using synthetic small regulatory RNAs. Nat Biotechnol. Nature Publishing Group; 2013;31: 170–174. doi:10.1038/nbt.2461

18. Shafqat J, Höög JO, Hjelmqvist L, Oppermann UCT, Ibáñez C, Jörnvall H. An ethanol-inducible MDR ethanol dehydrogenase/acetaldehyde reductase in Escherichia coli: Structural and enzymatic relationships to the eukaryotic protein forms. Eur J Biochem. 1999;263: 305–311. doi:10.1046/j.1432-1327.1999.00323.x

19. Sørensen MA, Fricke J, Pedersen S. Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo. J Mol Biol. 1998;280: 561–569. doi:10.1006/jmbi.1998.1909

20. Freundlich M, Ramani N, Mathew E, Sirko A, Tsui P. The role of integration host factor in gene expression in Escherichia coli. Mol Microbiol. 1992;6: 2557–2563.

21. Węgleńska A, Jacob B, Sirko A. Transcriptional pattern of Escherichia coli ihfB (himD) gene expression. Gene. 1996;181: 85–88. doi:10.1016/S0378-1119(96)00468-4

22. Andrews SC, Robinson AK, Rodríguez-Quiñones F. Bacterial iron homeostasis. FEMS Microbiology Reviews. 2003. pp. 215–237. doi:10.1016/S0168-6445(03)00055-X

23. Braun V, Mahren S, Sauter A. Gene regulation by transmembrane signaling. Biometals. 2006;19: 103–113. doi:10.1007/s10534-005-8253-y

24. Nakashima N, Goh S, Good L, Tamura T. Multiple-Gene Silencing Using Antisense RNAs in Escherichia coli. In: Kaufmann M, Klinger C, editors. Methods. New York, NY: Springer New York; 2012. pp. 307–319. doi:10.1007/978-1-61779-424-7

25. Qi LS, Arkin AP. A versatile framework for microbial engineering using synthetic non-

coding RNAs. Nat Rev Microbiol. Nature Publishing Group; 2014;12: 341–354. doi:10.1038/nrmicro3244

26. Qu Z, Adelson DL. Evolutionary conservation and functional roles of ncRNA. Front Genet. 2012;3. doi:10.3389/fgene.2012.00205

27. Cech TR, Steitz JA. The noncoding RNA revolution - Trashing old rules to forge new ones. Cell. 2014. pp. 77–94. doi:10.1016/j.cell.2014.03.008

28. Markham NR, Zuker M. UNAFold: Software for nucleic acid folding and hybridization. Methods Mol Biol. 2008;453: 3–31. doi:10.1007/978-1-60327-429-6-1

29. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, et al. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 2013;41: D605–D612. Available: http://nar.oxfordjournals.org/content/41/D1/D605.abstractN2

Supporting Information

```
>adhP fRNA - Proposed binding to adhPt1/maeA terminator
actaccg

>rpsA-ihfB - Proposed binding to rpsA terminator
cgccttt

>fecIR fRNA - Proposed binding to fecR RBS
ctcatatt

>fecABCDE fRNA - Proposed binding to fecA RBS
ttctcgtt
```

S1 Fig. RNA sequences of abortive initiation fragments for top candidate faRNA matches.

Transcriptional Unit	RBS	Strand	Binding Site Sequence	faRNA Sequence	ΔG°$_{binding}$ of faRNA	TSS	Match Start	Match End
fecIR	fecR	reverse	agtatggg	ctcatatt	-4.95	4516305	4515743	4515750
fecABCDE	fecA	reverse	gatgggga	ttctcgtt	-4.11	4514750	4514705	4514712

S1 Table. Sequence and additional data describing predicted faRNA-RBS interactions in *E. coli*. All free energy (*ΔG*°) values are in kcal/mol.

Transcriptional Unit	Terminator	Strand	Binding Site Sequence	faRNA Sequence	ΔG°$_{formation}$ of Terminator	ΔG°$_{binding}$ of faRNA	TSS	Match Start	Match End
adhP	adhPt1/maeA	reverse	cggtagt	actaccg	-17.28	-8.15	1552018	1551980	1551986
rpsA-ihfB	rpsA	forward	aagggcg	cgccttt	-18.84	-8.22	960940	962975	962981

S2 Table. Sequence and additional data describing predicted faRNA-terminator interactions in *E. coli*. All free energy (*ΔG*°) values are in kcal/mol.