**Sequence constraints for efficient strand invasion for co-transcriptional gene regulation by an adenine sensing riboswitch**

**OR**

**Elucidating crucial elements of co-transcriptional strand invasion using an adenine sensing riboswitch**

**OR**

**Gleaning design principles of co-transcriptional strand invasion using an adenine sensing riboswitch**

**OR**

**Optimizing the functional core of an adenine sensing riboswitch informed by strand invasion**

**Lisa Hansen, Jesus Alicea, Ava Altenbern, Alexandra Brown, Parker Juels, Mia Kim, Aishah Hassan Scott Odierno, Matthew Payne, Peyton Roeder, and Robert T. Batey\***

Department of Biochemistry, University of Colorado, Boulder, CO, 80309-0596, USA

\* To whom correspondence should be addressed. Tel: 1(303) 492-2159; Fax: 1(303) 492-5894; Email: [Robert.Batey@colorado.edu](mailto:Robert.Batey@colorado.edu)

Keywords:

*Key points:*

**Abstract**

**Introduction**

Strand invasion generally involves three nucleic acid strands, two of which (substrate and incumbent strands) are in a duplex, and a third that is single stranded (invader strand) with similar or identical sequence composition to the incumbent (36535150, 34752075, 21258382) (**Figure 1A**). The process of strand invasion involves multiple steps including toehold formation, branch migration, and displacement (**Figure 1B**). Toehold formation begins the interaction between the three strands when the invader pairs with a region of ssRNA located on the end of the substrate (**Figure 1B,** “toehold formation”). Branch migration, which follows, is the stochastic process in which the invading and incumbent strands compete for base pairing with the substrate strand (34752075, 10949296) (**Figure 1B,** “branch migration”). This process can be biased towards the invading strand in the presence of a toehold because the invading strand will necessarily form a longer duplex (34752075, 37095744). Factors that influence the rate of invasion include the length of the toehold, the number of mismatches (non-Watson Crick and non-wobble pairing), the location of the toehold on the substrate strand, and the type of nucleic acid present (25762335, 34752075, 21258382). Finally, displacement occurs when the duplex is formed between the substrate and invader, fully displacing the incumbent (**Figure 1B,** “displacement”).



**Figure 1A: Strand invasion within the purine riboswitch.** (A)schematic of secondary structure of the minimized pbuE riboswitch P4-A. Notable regions for stand invasion highlighted: incumbent strand (orange), substrate strand (magenta), and invader (green). (B) Strand invasion schematic with purine riboswitch as system. Top row shows processes in absence of ligand, bottom row shows with ligand. Initiation begins with three strands in proximity, followed by toehold formation involving pairing of P4 with or without ligand. Branch migration proceeds when the invader strand begins pairing in P1 and proceeds into P3 until full displacement occurs. All steps are in reversible except for displacement because no toehold is available for the incumbent.

Riboswitches serve as an example of highly folded, functional RNAs that often employ strand invasion to regulate associated genes (15919195, 35348734). These regulatory elements have two principal components: an aptamer domain, and an expression platform. The aptamer domain is a highly structured and conserved region that binds a ligand with high specificity (21925376, 14523920, 15173824). The expression platform is a dynamic region that creates a genetic output by folding based on the ligand input in the aptamer domain (20943759). The adenine responsive *pbuE* riboswitch from *Bacillus subtilis* is an ON switch that folds co-transcriptionally. It is an ideal model system for studies of co-transcriptional folding and strand invasion due to its well characterized folding mechanism (23087247). This riboswitch, like most, toggles between two folded states, an ON-state where the ligand is bound and the anti-terminator is formed, and an OFF-state where ligand is unbound, and the terminator helix has formed. Based on a co-transcriptional study of this riboswitch, the OFF conformation is more stable and is the ultimate folding fate of all switches (23087247). The interconversion between the two mutually exclusive states proceeds through strand invasion of the 3’ end into the ligand binding pocket following toehold formation in P4 (15919195**).** As invasion proceeds further into the aptamer domain the invader strand encounters a two-nucleotide ligand gated toehold (J3/1, **Figure S1**) where invasion is stalled in the case of the ligand being bound, or further propagates through the formation of a second toehold (**Figure 1B,** “displacement”). This RNA is therefore a key model for understanding strand displacement *in vivo* as it coordinates the co-transcriptional folding of RNA and ligand-gated strand displacement to make genetic decisions.

The current study probes crucial elements of strand invasion within an RNA based system *in vivo* through a genetic screen. The RNA employed is a modified variant of the *pbuE B. subtilis* Adenine responsive, that was optimized to increased dynamic range and improve key features of RNA folding (33259551). This study offers insights into primary and secondary structural requirements for the processes of strand-invasion and nucleation that allow for the dynamic properties of this unique regulatory element. The modified parental switch that the following genetic screen is based on greatly outperformed the 5.6-fold switching of *pbuE* with a near 100-fold induction in cells (33259551). The switch was minimized, with all base pairs in the ON and OFF state being WC and regions thought to be key to nucleation of the terminator helix (such as P4) were enriched with GC pairs (33259551) (**Figure S1**). The prior mutagenic analysis, while yielding a high preforming variant did not aim elucidate the design principles behind this improved switching ability.

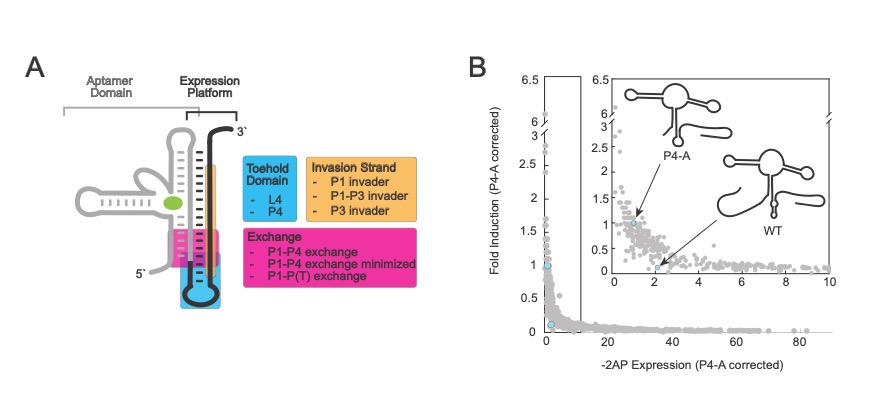
The expression platform of the riboswitch was evaluated by a structure-guided genetic screening approach using small libraries that interrogate a secondary structural feature of the switch. These libraries, which span the entire expression platform, were screened for activity in E. coli to identify functional variants. Analysis of over 500 functional variants revealed that residues directly involved in nucleation of strand invasion are the most essential and highly conserved. From these data it seems that the development of novel biosensors that use this scaffold can be easily designed i*n vivo* with limited manipulation to key areas near the initial 3-way branch point of the switch based on the parameters of the system. FINISH THIS ONCE PAPER IS DONE!

**Results and Discussion**

**Creation of artificial phylogeny through parallel genetic screens**

The “parental” riboswitch used for this study is the P4-A adenine-responsive riboswitch that was optimized in several important respects, as discussed above. To serve as benchmarks for this study, we determined the ligand-induced expression by the adenine analog 2-aminopurine (2-AP) of *wild type pbuE* responsive purine riboswitch (WT) and the parental P4-A. Consistent with the work by Drogalis and Batey (33259551), P4-A far outperformed WT in ability to repress expression without ligand and express in the presence of ligand, resulting in an increase of fold induction.

A genetic screen was preformed to look at specific regions of the riboswitch in an unbiased fashion. To do so, libraries were created that randomized 3-6 nucleotides to ask specific questions about the sequence and structural requirements in each area. The three questions we sought to answer were: what are the requirements for toehold formation, what are the sequence requirements for the invader strand, and what are the requirements for areas where the three strands must exchange during branch migration? To answer the first question there were two libraries spanning the ligand-independent P4/L4 stem loop. One focusing on the paired region that is the toehold itself, P4, and one with the terminal loop adjacent to that element L4 (**Figure 2A,** “toehold domain”). Three overlapping libraries were designed to understand the possible sequence space of the invading strand when the incumbent and substrate helix remains constant. The three libraries (P1 invader, P1-P3 invader, P3 invader) each include different obstacles including paired and unpaired regions of the aptamer domain (**Figure 2A,** “invasion strand”). The second theme was characterizing the invading strand. The third question, probing the exchange between the mutually exclusive ON and OFF states and evaluating how the strength of one helix compares to the strength of the competitor helix through libraries that randomized regions in both helices. These include at the site of initial invasion (P1-P4 exchange and P1-P4 exchange minimized), and a region further upstream where P1 is competing with the terminator helix (P1-P(T) exchange) (**Figure 2A,** “exchange”). Libraries were cloned and transformed into E. coli cells before undergoing phenotypic screening to find functional riboswitches. Those that appeared to function were then sequenced and assessed with a cell-based fluorescence assay to have a collection of sequences with associated activity. Once libraries were completed, completion being determined based on number of colonies observed regardless of their ability to function, data were analyzed, and variants were grouped based on performance. The two variables that were used to assess function were fold induction, the ratio of expression +2AP to expression -2AP, and the ability to repress (-2AP expression). Most variants had comparable turn on in the presence of ligand because the aptamer domain was not randomized, so we sought to find switches able to robustly repress expression in the absence of 2AP which should be those switches most able to efficiently strand invade (**Figure 2B**). Variants were grouped based on performance into top performers (P4-A corrected -2AP value of =<10), and all other variants (P4-A corrected -2AP value of >10).



**Figure 2: A genetic screen of the expression platform through multiple libraries.** (A) Schematic of riboswitch highlighting areas of interest for genetic screen. Aptamer domain in gray, expression platform in black and ligand shown in green. The toehold domain libraries (blue), exchange libraries (magenta) and invasion strand libraries (yellow) are listed (right) and highlighted on RNA (left). RNA shows all possible competing base pairs in ON or OFF state. (B) All variants from all libraries displayed on one axis (gray) with the wildtype *pbuE* riboswitch (WT) and parental P4-A riboswitch in cyan. Inset shows variants with -2AP expression >10-times that of P4-A.

Through screening of multiple regions of interest within the expression platform, over 500 unique sequences were extracted and characterized with a cell-based activity assay. Variants with a broad range of repression and fold induction were surveyed, including those that outperformed P4-A as well as variants unable to function even as well as the WT riboswitch (**Figure 2B**). In subsequent sections we will break down the findings from each library and discuss answers to the three questions above.

**P4 is essential for efficient ligand dependent gene regulation**

Two libraries were designed to interrogate the toehold, the P4 library and the L4 library (**Figure 3A**)**.** P4 in the P4-A riboswitch serves as a toehold that includes three base pairs which are present in both conformations of the riboswitch (**Figure 1B**). L4, the terminal loop proximal to P4, was of interest because of the co-transcriptional nature of this riboswitch.

*Loop sequence is not important for ligand dependent regulation.*

In P4-A, the loop sequence was engineered as a hexauridine tract to mimic the polyuridine tract in the WT pbuE riboswitch that has been proposed to be a programmed pause site. This pause site would temporarily stall RNAP to enable the aptamer domain to fold and to interrogate the cellular environment for the presence of adenine. Indeed, variants showed marked improvements for both expression and repression compared to mutants without this feature (33259551), and it has been shown that that motif pauses RNAP even in a heterologous system with a *B. subtilis* riboswitch expressed in *E. coli* (28541183). But, when tested against a variant with 5’UUCCUU rather than 5’UUUUUU, activity was very similar, so it was unclear if the hexauridine tract was important for folding and function of the riboswitch (33259551).



**Figure 3: A genetic screen of the toehold libraries.** (A) Schematic of P4-A with the P4 (magenta) and L4 (yellow) libraries highlighted. To right on top is schematic of secondary structure of stem loop 4 with the sequence. Residue implicated in elemental pausing, position -11, -10, -1 and +1, highlighted in cyan. On bottom is schematic of P4 helix showing naming of the 3 base pairs in P4 in white to right of paired region. One base pair of P1 included in gray to show that BP1 participates in P1 and P4. (B) L4 variants plotted in yellow on same axis as Figure 2B. P4-A and WT in cyan. Inset includes sequence of P4-A (top with arrow) and the sequence logo of all library variants. Logo includes regions of the non-randomized P4 region with elemental pausing numbered on x-axis. (C) TBD (D) P4 variants plotted in magenta, P4-A and WT shown in cyan. Inset shows variants with -2AP expression >10-times that of P4-A. (E) Pie charts illustrating base pair identity for 3 positions. BP1-3 labeled on left, first column includes all variant, second column super performers, third is top performers and fourth includes all remaining sequences. (F) Activity of specific variants with schematic of sequence above each variant. Error bars show SEM.

To assess the sequence requirements of L4, a 6-nucleotide randomized library was designed including the entire loop region (**Figure 3A**). The initial transformation of this library was grown on plates without ligand due to large number of false positives resulting from randomizing L4. Dim colonies were then selected from the unliganded primary plate. The L4 library produced numerous highly functional switches which rivaled P4-A, so screening was halted due to the lack of insight and large burden of screening that region. 200 colonies were screened for this library, which yielded 77 functional sequences, emphasizing the abundance of robust switchers within that library (**Supplementary Table 1**). Across libraries those variants with expression <10-times that of P4-A in the absence of ligand were grouped to compare against variants with diminished abilities to repress gfpUV. All loop variants fell in that category, emphasizing the functionality of the loop variants observed (**Figure 3B**).

We observed that functional variants in the L4 library lacked any identifiable sequence preference (**Figure 3B**, inset). Every position was observed as any nucleotide, a large portion of observed riboswitches (~39%) maintained the ability to function, and many variants functioned comparably to P4-A. While pausing has been established to be important for regulation by riboswitches and the polyuridine tract has been seen to slow the movement of the RNAP in *E. coli* (28541183), and variants containing the hexauridine tract in previous studies by our group significantly outperformed their counterparts without the motif and repressed more in the absence of ligand (33259551), the polyuridine tract was not necessary to have a highly functional riboswitch. These findings are consistent with work done by the Lucks group on the ZTP riboswitch who also saw a lack of functional impact from disruption of a predicted pause site (36864761). Similarly, pausing was not observed at poly(U) stretches of the *pbuE* switch in the co-transcriptional investigation performed by Frieda and Block, when working with physiological dNTP concentrations (23087247). While the poly(U) tract may not be crucial, pausing may still occur due to the sequence of stem loop 4. All variants in this library possess an elemental pausing sequence with G-11/C-1,G+1 motif in P4 which may be responsible for RNAP pausing to facilitate folding (25976475, 24789973) (**Figure 3B**). Considering this, the region may not be important for pausing, but rather may serve to facilitate formation of the toehold so that P(T) can form efficiently within the exit channel of RNAP to stop transcription. Based on the activity of these variants, a 6-nucleotide loop appears to allow for efficient helix formation with no clear requirement for any specific sequence.

*The length of L4 is critical for regulatory activity.*

While the above data indicated little preference for the sequence, we also wanted to define the length of the loop that could promote regulatory activity. To do this we created a series of loop sizes ranging from two nucleotides to 200 nt (sequences in Supplemenatary Table X). These loop sequences were specifically designed to be devoid of secondary structure, which was accomplished by designing sequences that are devoid of guanosine residues and checking the ability of the sequence to form secondary structure using mFold (ref).

RESULTS.

*The toehold formed by P4 is essential for riboswitch activity.*

The requirements of the P4 library, which includes the three base pairs in P4 distal to L4 (**Figure 3A**), were much more stringent based on colonies screened, ~14,000, and functional sequences extracted, 127 (**table S#**). The P4 library was screened until ~97% of potential sequences had been observed (**table S#**). While P4 formation is ligand-independent, this library was initially screened on +2AP plates with hopes of finding RNAs still able to form the anti-terminator in the presence of ligand.

In the P4 library base pairing was used to interpret the requirements of the region. The 127 sequences had variable performance in the liquid culture activity assay, with 62 of them able to repress less than 10-times as much as the parental P4-A (**Figure 3C**). Those performers were then broken into super performers, those with repression less than 4-times P4-A, and top performers, which fell in the 4- to 10-times P4-A window. While the base pairing seems stable across the 3 base pairs when looking at the entire library in aggregate, trends seem to fall out when looking at the different performance group. For BP1, super performers are enriched with wobble pairs compared to the other whereas those with 4-10 times repression of P4-A are largely composed of non-canonical pairing interactions (**Figure 3D**, “BP1”). This breaks the trend seen in BP2 and BP3 of increasing proportion of NC pairing as ability to repress is eliminated. BP2 seems to change with the groups. Super performers are primarily forming productive base pairs (not NC pairing) and half of them are GC/CG pairs (**Figure 3D**, “BP2”). While BP3 varies slightly with most GC/CG pairing amongst top and super performers, the distribution remains constant across the groups (**Figure 3D**, “BP3”).

While the elemental pause was constant in the L4 library, it was not present in every variant in the P4 library. Those sequences that had C-1,G+1 and a G at -11, -10, and or -9 were determined to have the elemental pause motif (25976475, 24789973). ~30% of the super performers, ~20% top performers, and ~20% bottom performers had the sequence motif. While the difference between bins could be because of increased function amongst switches with elemental pause motif, the difference could also be from more GC rich sequences amongst the highly functional switches. Therefore, no conclusions can be made with confidence about the relationship between performance and the elemental pause motif from this screen.

While the overall pool of sequences collected was unbiased, specific questions and insights can be gathered from looking at sequence variants pulled out from the screen. One sequence, with a fold induction less than a tenth of P4-A, still maintains 3 GC base pairs (**Figure 3F**, “variant 1”). The only difference is that BP3 has been changed from CG to GC. The deleterious impact of this change may be related to stacking energies or interactions with the pre-aptamer sequence, but overall suggests that analyzing the strength of each base pair will not directly translate to the in vivo activity of the system. Two more interesting sequences have exclusively AU pairing in the P4 randomized region. One sequence which switched each of the 3 base pairs to AU rather than the parental GC expressed 9x more FP in the absence of ligand than P4-A (**Figure 3F**, “variant 2”). But, in the presence of 2AP it expressed slightly more than P4-A, resulting in a marked change in fold induction likely due to reduced stability of the toehold. Many colonies in the bottom bin also formed 3 productive base pairs, so the formation of any WC or wobble pair is not sufficient for robust switching. For example, another variant that forms 3 AU pairs and differs only in the directionality of BP1 (**Figure 3F**, “variant 3”). That sequence lacks the ability to express or repress near the parental or the other sequence with 3 AU pairs.