**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.

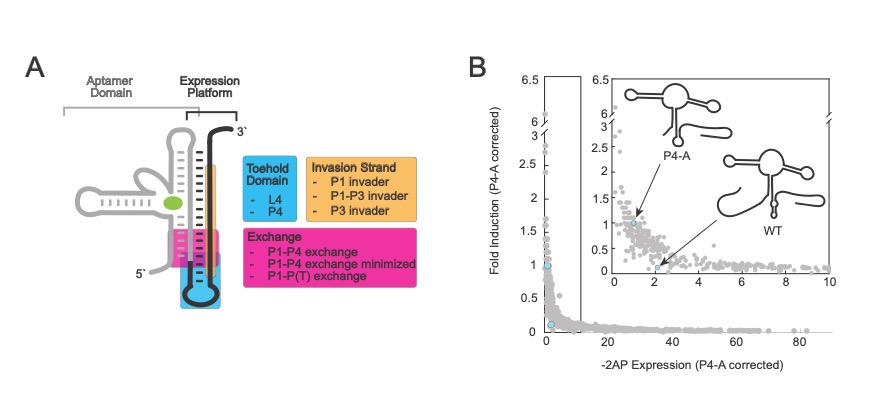
The expression platform of the riboswitch was evaluated by a structure-guided genetic screening approach using small libraries that interrogate a secondary structural feature of the switch. These libraries, which span the entire expression platform, were screened for activity in E. coli to identify functional 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 Supplementary 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.

**The invading strand favors WC base pairing but is not essential.**

The next family of libraries seeks to address sequences that can successfully invade a P1 and P3 of the aptamer domain (**Figure 4A**). The 3’ end of this riboswitch must invade into P1 and P3 to create P(T), a process facilitated by toehold formation in P4 and J3/1, P4-Awas designed to maximize the number of WC base pairs, both in the substrate-incumbent helix (P1) and the substrate-invader helix (P(T)). Three libraries were deigned to determine if that is the optimal design by varying only the invader strand. The P1 invader (n=5) includes the 3’ end of P4 and the invasion of P1, P1-P3 invader (n=6) spans P1 to J3/1 to P3, and the P3 invader (n=5) spans P3 (**Figure 4A**).

Each of the three libraries were screened until there more than 99% probability than any one sequence would be observed (18567049). Of the two 5-nucleotide libraries, P1 invader yielded 17 unique and functional switches after observing ~5,100 colonies and the P3 invader yielded 15 unique and functional sequences after observing ~9,000 colonies (Table S#). For the 6-nucleotide library, P1-P3 invader, 141 unique and functional sequences were extracted after observing ~23,000 variants (Table S#).

It appears that P4-A was the optimal solution for the invader strand across the three libraries (**Figure 4B**). The P1-invader library, encompassing the transition into P1, had one variant with a higher fold-induction than P4-A by changing the GC pairing formed at the third base pair to a GU wobble, meaning the n position was a U rather than a C. But that variant failed to repress as well as P4-A, meaning the decrease in stability likely limits the ability of the 3’ end to invading. The most conserved positions, likely the workhorses of the invasion process, were the first (n-2) and second (n-1) positions across all bins (**Figure 4B**). These positions are part of the toehold and do not compete with P1, reinforcing the importance of the toehold for successful invasion. The n-1 position was invariant, even amongst low preforming switches and likely creates a stabilizing WC GC pair to initiate invasion into P1. The last three positions, n, n+1 and n+2 were much more variable, suggesting that the actual invasion of P1 is more tolerant to mismatches than the toehold region preceding it.

The P1-P3 invader failed to produce a variant with a fold-induction as high as P4-A, with only 3 out of the 141 sequences falling in the top bin. However, one sequence was able to repress better than P4-A, due to a single point mutation at position 5 which again changed a GC to a GU. Like the P1 invader, this library spans a toehold region, J3/1, but this library invades two paired regions, P1 and P3 (**Figure 4C**). Uniquely, this library encounters the ligand gated potion of the invasion process. When ligand is bound, it forms a WC pair with U74, a nucleotide in J3/1, and invasion cannot proceed. But, when the nucleotides in J3/1 pair with the 3’ end of the expression platform, invasion is further stabilized and can proceed into P3. Across bins, there is high conservation in nucleotide identity at positions 4 and 5, corresponding to the n-1 and n positions of invasion into P3. (NEED TO USE/REFER to figures more!)

The final invasion library, P3 invader, had no sequences that preformed near P4-A and had 0 sequences falling in the top bin. All sequences found in the middle bin were single point mutants at position 1, 2 or 5, but all of them had a dramatic impact on repression reflected in a far lower fold-induction. While WC pairing clearly leads to the best functional performance in this area, there was high nucleotide identity at positions 3 and 4 in both bins (more so in the middle bin) which indicates again that not all positions are created equal. (Do we use Shannon entropy plots for this section??? No—something else…) This library aligns directly with P3 and has no ssDNA binding partners, but these two positions, n-2 and n-1 of the P(T) terminus, may be important for overall P(T) stability to actually effect change in RNAP activity. One variant in the mid bin with a point mutation at position 5 converting a UA pair to a UG pair had 7.4x more FP expression without ligand and 0.76x expression with ligand, again indicating the importance of a WC GC pair at the n-1 position. No point mutations were found for the n-2 position after extensive screening which may indicate that variation at that position does not allow for P(T) formation without a compensatory mutation. This is supported by the fact that all mutations of the n-2 position (position 4) are accompanied by at least 3 other mutations which may change the entire base pairing network in the region. (Huh?)

**Strand Exchange**

The final theme sought to examine the more complicated interplay between the three strands. Three libraries were constructed including nucleotides in each of the strands predicted to actively be competing during the initiation or branch migration. The P1-P(T) exchange (n=6) is a slightly larger library looking at sequence space for two base pairs that are P1 for the ON state and P(T) for the OFF state. The other two libraries, P1-P4 exchange (n=5) and the minimized P1-P4 exchange (n=3), specifically zoom in on the gate between P4 and P1 to focus on the initiation of the strand exchange process.

The larger P1-P(T) exchange library was screened until there was a 96.4% of observing any one sequence through observation of ~14,000 colonies which gave 102 unique and functional sequences (Table S#). The P1-P4 exchange library gave 97 unique functional sequences after observing ~5,000 colonies, with a probability of observing any given sequence of 99.3% (Table S#). MINIMIZED LIBRARY HERE.

Based on the P1-P(T) exchange library sequences, pairing in P1 seems to be more important than pairing in P(T). The best performer had 3 GC pairs and a GU pair in P4 proximal to L4 (again, should we show?). Productive pairing is important in this region but variation in pair strength between P1 and P4 may improve function. Of those 15 sequences in the top bin, 100% formed a WC or GU wobble pair for both P1 pairs while 60% (approximate? X of Y sequences) formed the P4 pair proximal to L4 and 80% formed a productive pair in P4 distal to L4. That trend continues through the bins, and it seems that the P1 pair proximal to the aptamer domain is the most important, with conservation above 90% in all bins, and the P4 pair proximal to L4 is least important with less than half of the 42 variants in the middle bin and 0% of the bottom variants forming a WC or wobble pair at that position. In previous work investigating this riboswitch *in vivo* it was established that the minimum size of P1 while still maintaining the ability to switch in response to 2AP was 3 base pairs (25550163). The highly conserved base pair in this library is the third base pair from the aptamer domain, adjacent to the two pairs that have ligand dependent interactions, and while it is a conserved RY in phylogeny 4/15 of the top bin switches has YR at that pair (25550163).

Unlike the P1-P(T) exchange library, P1 pairing seems to be unnecessary in the P1-P4 exchange library. (huh?) The highest functioning switch in terms of both repression and fold induction had 3 AU base pairs and none of the sequences that functioned better than P4-A had a GC pair in P1. All of those sequences formed at least 1 production (WC or wobble) base pair in P4 and none of them had two GC pairs in P4 like P4-Adoes. This indicates that P4-A could be improved by weakening the pairing in this region or removing the ability to form the terminal base pair in P1. (Is this the fortuitious one that was a result of P4-A?) the au was changed to a GC which ended up extending P1…should talk about this in terms of the design that was done…maybe break up into those that have longer P1 and those that have shorter P1 and see if there is a difference

In response to those findings, a set of directed mutants were created that were not observed in the screen. Since there was such a marked increase in function by substituting the GC pairs for AU pairs, two mutants were designed with one AU pair and one GC pair to determine if one position was more favorable and if it was due to the change at both pairs or instead if it was driven by a single pair. A mutant was also designed that maintained the P4-A sequence at all positions except for the first base at the 5’ end of P1. That base was changed from a C to an A, henceforth called P4-A*-A*, and served as an example of a switch with a longer P4 and shorter P1 due to elimination of that GC pair. These three mutants created a mutagenic cycle to elucidate the differences between the best functioning variant created in the screen and the parental P4-A*.*

The mutagenic cycle revealed that no single designed mutant outperforms R3AE6C4 or the parental P4-A. Changing either the L4 proximal or L4 distal base pair in this library from a GC pair to an AU pair resulted in 0.8x fold-induction of P4-A*.* While the distal AU pair mutant was able to repress better than P4-A, it was unable to express as much FP in the presence of 2AP resulting in a reduced fold-induction. The P4-A-A mutant interestingly had the same activity as P4-A in the absence of 2AP, but only expressed half as much FP in the presence of 2AP, again impacting the overall fold induction and indicating that there are benefits to having the first base pair of P1. This mutagenic cycle indicated that these mutations in concert allow for such a robust repression of FP (1/5th that of P4-A) while maintaining robust gene turn on in the presence of ligand.

MINIMIZED LIBRARY

**Plug and Play**

**Conclusions**

**Material and Methods**

Library construction

Each library was constructed via PCR with 3 outer primers and one library specific Ultramer (Integrated DNA Technologies) (sequences in Table S#). The libraries were created via PCR with Q5 DNA polymerase (New England Biolabs) and amplified with 12 or 20 rounds of amplification with an annealing temperature of 55 °C. Constructs were confirmed on 2% agarose gel and cleaned with E.Z.N.A® Cycle Pure Kit (Omega Bio-Tek). Libraries were then digested with the high-fidelity restriction digest enzymes NsiI and HindIII according to manufacturer’s protocol (New England Biolabs) and again cleaned with the E.Z.N.A® Cycle Pure Kit (Omega Bio-Tek). To improve ligation efficiency, digested libraries were then treated with PNK (New Englad Biolabs). The libraries were ligated into a low-copy plasmid containing an ampicillin resistance cassette and *gfpUV* as a reporter gene downstream of the site of HindIII cut site. The plasmid was ligated into the vector using T4 DNA ligase by the manufacturers protocol (New England Biolabs).

Directed mutant construction

Loop series preparation

Mix and match

Phenotypic screen of potential switches

Libraries were transformed into chemically competent *E. coli* K-12 BW25113 cells and plated onto chemically defined salt broth (CSB) plates with 1.2 % agar and ampicillin following ligation. The exchange libraries that destabilized P1 helices (P1) and L4 were grown on plates with 0.5 mM 2-Aminopurine and the invasion stand and toehold libraries that destabilized P4 or P(T) helices were grown on plates without ligand. Colonies were grown overnight at 37 °C and *gfpUV* expression was visualized with a 366 nm light. Plates were photographed and counting statistics were preformed using a software package from Opencfu (3.9.0). On liganded plates, bright colonies were selected and on unliganded plates dim colonies were selected. Colonies that were selected were transferred onto a set of gridded plates, one liganded and one unliganded, and those plates were grown overnight at 37 °C. Gridded plates were then visualized with the 366 nm light and variants that were dim on unliganded plates and bright on liganded plates were selected due to ligand inducible expression and moved forward to further characterization.

Sequence and quantification of ability to attenuate gene expression

Selected colonies were plasmid prepped with the E.Z.N.A® Plasmid DNA Mini Kit I, (Omega Bio-Tek) and sent for Sanger sequencing (QuintaraBio). Variants were also subjected to a cell-based activity assay to quantify ligand induced gene expression. Variants were inoculated overnight in 2xYT media supplemented with ampicillin spinning at 37 °C. Overnight growths were then diluted 1:1000 into 2 subcultures of defined CSB media supplemented with ampicillin. One of the two subcultures contained 0.5 mM 2-Aminopurine which the other did not contain ligand. Subcultures were grown spinning at 37 °C for about 6 hours until an OD600 range of 0.4-0.5 (mid-log phase growth) was achieved (). Subcultures were transferred into a Costar® 96-well half area microplate with 3 wells containing 200 µL for each subculture. Each plate included 200 µL of a 5.0 µg/mL fluorescein standard to set gain for fluorescent measurements. Plates were read with the Tecan Infinite M200® PRO plate reader and the OD 600 and fluorescence in arbitrary units were measured. The excitation and emission wavelengths for the measurements were 395 nm and 510 nm respectively. Each measurement was done in technical and biological triplicate. Each biological replicate of variants was accompanied by 3 controls, pBR327 to correct for the background fluorescence of cells, PRR5\_*gfpUV* which constitutively expresses the reporter gene, and the P4-A parental riboswitch.

Analysis of activity assay and library screen

To determine the probability of observing any one sequence for each library given the colonies counted from initial plating and the size of the randomized library was calculated based on the second equation proposed by *M. Reetz et al*. relating these factors (18567049). Libraries were screened until a threshold of 95% was passed unless otherwise stated.

The fluorescence values measured were processed in excel by dividing each fluorescence value by its OD600 and then subtracting the median OD600-corrected pbr327 fluorescence value (Microsoft). Those OD600- and pBR327-corrected values were then compiled into a .csv file which was then analyzed by a library specific R markdown, the R markdowns are open source and freely available (). There were two colonies in the P4 library that had negative repression values due to their ability to repress as much or more than the pBR327 control. Those variants were given an artificial repression value of 50 for comparison with other colonies, a value lower than the parental.

Processing of data included removal of outliers for OD600-correctly fluorescent values for measurements with and without ligand. The liganded and unliganded measurements were separated and outliers were determined to be those data points found more than 1.5-times the interquartile range away from the outer quartiles and were removed, a decision made assuming a Gaussian distribution (https://doi.org/10.1007/978-981-10-7563-6\_53). Data from the two conditions were then combined and the standard error of the mean and ratio of fluorescence with 2AP to fluorescence without 2AP (fold induction) were calculated. FASTA files were created by combining the sequences for each colony with values from the activity assay to create an activity associated sequence alignment (library sequence alignments found in XXX). To facilitate discussion of large data sets, each library was also divided into 3 groups based on performance. Those with a fold induction >=0.5-times that of P4-A and a repression value <1.5-times that of P4-Awere in the top bin. The middle bin was variants with <0.5 P4-A-correctedfold induction with a repression value of <=9-times that of P4-A*.* Bottom bin included all other variants.

Figure 1: STRAND INVASION AND P4-A. A schematic of strand invasion with the P4-A adenine responsive riboswitch (walk through steps). B secondary structure of ON and OFF states of P4-A.

Figure 2: controls and libraries. A: controls highlighted in front of all libraries (P1-P4 exchange minimized not included). Cartoons of three controls shown in top corner, large structural changes in P4 and the pre-aptamer sequence are highlighted in green. B. activity of P4-A, WT, and del27\_WT in presence and absence of 2AP. C. schematic of P4-A and regions surveyed. Don’t need B…want C to be main (should be A)

For C…extend the gray of AP to the bottom of P1 on 3’ end

Make the base pairs less bold and put 2AP in the binding pocket

Don’t display del27 WT…just do P4-A and WT (the FI super improved based on Lea’s work)

Figure 3: Toehold Graph, schematic, specific variants

A should be the schematic where we highlight P4 and L4 (SL4 is the combined, this is formed constitutively in both states!!!)

B L4 graph…make new figures in with the gray box as a bin (maybe better than WT)

C P4 graph

D P4 pie charts from Ava

If sequence variants are talked about they should be in the figure!!!!

Supplemental Table: Library completion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Library | nt varied | observed | functional | prob. |
| P1 invader  (Anchor) | 5 | 5,118 | 17 | 99.3 |
| P1-P4 exchange  (Anchor\_exchange P1\_gateway) | 5 | 5,069 | 97 | 99.3 |
| P3 invader  (Gateway) | 6 | 22,799 | 141 | 99.6 |
| L4  (loop) | 6 | 202 | 77 | 4.81 |
| P4  (nucleato/P4 stem) | 6 | 14,145 | 127 | 96.8 |
| P1-P(T) exchange  (P1\_exchange) | 6 | 13,573 | 102 | 96.4 |
| P3 invader  (P3) | 5 | 9,029 | 15 | >99.9 |