**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**

It has become clear that nucleic acids have vast applications, spanning from therapeutics to nanotechnologies (21258382, 33900805), which originate from their ability to form base pairs with high fidelity. One specific interaction, which seems pervasive in both natural and synthetic systems, and unique to nucleic acids, is strand invasion. This phenomenon is observed in contexts ranging from riboswitches to homologous recombination to CRISPR/Cas9 gene editing (37095744, 36535150).

Strand invasion involves at least 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). The process of invading or exchanging happens when the invader strand fully displaces the incumbent, creating a novel duplex (35659535, 24402831). While the process where both strands compete for pairing with the substrate is random, the formation of a toehold can favor displacement by the invader (36535150). In this case, a toehold is a set of base pairs between the substrate and invading, which is allowed when there is a single stranded overhang on the substrate strand (36535150, 21258382, MORE).

Within stand invasion, there are multiple stages that must be understood, including toehold formation, branch migration, and displacement (**Figure 1A**). Branch migration is the stochastic process in which the invading and incumbent strands compete for base pairing with the substrate strand (34752075, 10949296). 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, MORE). Based on previous studies, factors that influence the rate of invasion include the length of the toehold, the number of mismatches (non-Watson Crick or non-wobble pairing), the location of the toehold on the substrate strand, and the type of nucleic acid present ([25762335](https://pubmed.ncbi.nlm.nih.gov/25762335), 34752075, 21258382).

Riboswitches serve as an example of highly folded, functional RNAs that often employ strand invasion to regulate associated genes (15919195, MORE). These elements have two principal components: a receptor domain, and a regulatory domain. The receptor domain, also called the aptamer domain, is a highly structured and conserved region that binds a ligand with high specificity (21925376, 14523920, 15173824). The regulatory domain, or expression platform, is a dynamic region that creates a genetic output by folding co-transcriptionally based on the ligand input in the receptor domain (20943759).

Riboswitches that control at the level of transcription must fold co-transcriptionally to exert control on the RNA polymerase before it escapes the element. The adenine responsive pbuE riboswitch from B. subtilis is an ON switch that folds co-transcriptionally. The interconversion between the two mutually exclusive states proceeds through strand invasion of the 3’ end into the ligand binding pocket (15919195**).**

* **Hole in understanding?**
* **This study**

This 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, optimized to increased dynamic range and highlight key features of RNA folding (33259551). Canonically, it is a transcriptional switch that regulates expression of a purine efflux pump in *B. subtilis*, turning ON expression when excess adenine is present in the local environment (16931335, 25550163). Transcriptional riboswitches are under immense temporal constraints as they must survey the environment and either form the ON or OFF conformation before the RNA polymerase escapes the switch. The formation of the stable OFF conformation is only accomplished when the 3’ end of the riboswitch successfully invades into P1, displacing the initial duplex. This strand invasion is toehold-mediated by the duplex region P4, which stabilizes the 3’ end to invade into the intact P1 helix (). 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 in a living system.

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. Modifications were made under the assumption that Watson-Crick (WC) base pairing in both the receptor and regulatory domains would create the best switch (33259551). Therefore, the switch became highly symmetrical, save the aptamer domain, and regions thought to be key to nucleation of the terminator helix (such as P4) were enriched with GC pairs. In addition to changing the primary sequence, the pre-aptamer sequence was shortened by 27 nucleotides to reduce misfolding. The prior mutagenic analysis, while yielding a high preforming variant did not aim elucidate the design principles behind this improved switching ability.

This study, through multiple genetic screens, 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 regulatory domain of the riboswitch was evaluated through multiple discrete libraries, each randomizing a different region of interest. Those libraries were analyzed through qualitative and quantitative methods to assess the riboswitch functionality, and functional variants were extracted and sequenced. Through the analysis of regions in the terminator helix, we find that strand invasion nucleation residues are the most essential and highly conserved. When looking at paired regions or regions in competition between the aptamer and terminator helix, balance between the two paired regions must be tuned to optimize functionality. 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.

**Results and Discussion**

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

Before investigating novel riboswitches, previously tested riboswitches were interrogated through cell-based assay alongside the parental riboswitch for this genetic screen, *NH5*. This included the wildtype pbuE responsive purine riboswitch (WT) and the wildtype with a deletion of the 27 nucleotides in the pre-aptamer domain (del27\_WT), each of which were expected to underperform compared to *NH5*. Consistent with the work by Drogalis and Batey, the del27\_WT variant had improved turn ON, slightly leakier turn OFF, and increased fold induction in comparison to the full WT variant (Figure XA). NH5, as suspected, displayed robust repression in absence of ligand, and expression in presence, with a fold induction 6x larger than the WT constructs.

Through screening of multiple regions of interest within the expression platform, 550 unique sequences were extracted and characterized through a cell-based activity assay. While *NH5* prevailed as the best solution for many of the libraries, a broad range of repression and fold induction was surveyed, ranging from those that outperformed *NH5*, to variants unable to function as well as the WT riboswitch (Figure 2B). After confirming that the control riboswitches behaved as expected, libraries were constructed looking at 3 overarching themes. Theme one was toehold formation, so two libraries were created to characterize the toehold domain which nucleates the formation of the terminator element. These libraries included one focusing on the paired region that is the toehold itself, P4, and the terminal loop adjacent to that element L4. Theme two was the exchange between the mutually exclusive ON and OFF states evaluating how the strength of one helix compares to the strength of the competitor helix through libraries that randomized regions in both helices. These includes 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).

**Toehold formation**

Inclusion of a toehold in a strand invasion system is imperative to ensure high fidelity displacement. P4 in the NH5 riboswitch serves as a toehold that includes 3 base pairs which are present in both conformations of the riboswitch (**Figure 1B**). Two libraries were designed to interrogate this region, the P4 library and the L4 library. L4, the terminal loop proximal to P4, was of interest because of the co-transcriptional nature of this riboswitch. To form a productive toehold, it was hypothesized that there would be sequence requirements that would either slow the processivity of the RNA polymerase or impact the flexibility of the region (Figure X). The P4 library spanned the two base pairs in P4 distal to L4 and an additional adjacent base pair that is in competition with P1. In the parental NH5, L4 is a polyuridine tract and P4 is a GC rich stem. The L4 library was of interest because a the hexauridine tract in the NH5 parental mutant is hypothesized to be a “programmed pause” site for the RNA polymerase in *E. coli* and showed marked improvements for both expression and repression compared to mutants without this feature (33259551). But, when tested against a variant with 5’UUCCUU rather than 5’UUUUUU, activity was very similar, so the sequence requirements of this region were unclear following mutagenesis (33259551).

To assess the sequence requirements of the two regions, two libraries were designed, each with 6 nucleotides randomized. Each library was screened with a different strategy based on performance. The L4 library produced numerous highly functional switches which rivaled NH5, so screening was halted due to the lack of insight and large burden of screening that region. Only ~200 sequences were observed of this library, which yielded 77 functional sequences, emphasizing the abundance of robust switchers within that library (table S#). Though only 4.8% of observing any potential sequence were screened in the L4 library, the data produced indicated that continued screening via this approach was unlikely to be informative. The requirements of the P4 library, which includes the three base pairs in P4 distal to L4, 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#).

**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).

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) 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 *NH5* 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 (). 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).



Figure 1: STRAND INVASION AND NH5. A schematic of strand invasion with the NH5 adenine responsive riboswitch (walk through steps). B secondary structure of ON and OFF states of NH5.

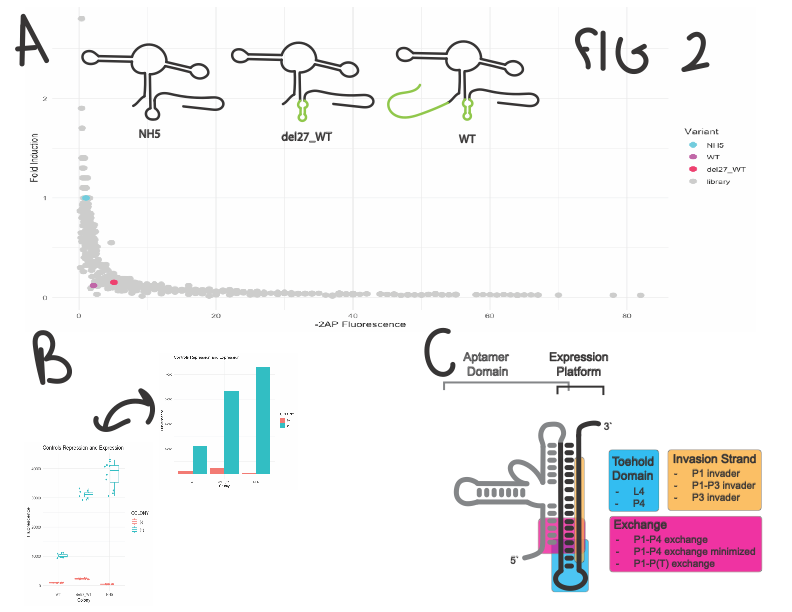


Figure 2: controls and libraries. A: controls highlighted in front of all libraries (P1-P4 exchange minimized not included). Cartoons of three controls shown in top corner, large structural changes in P4 and the pre-aptamer sequence are highlighted in green. B. activity of NH5, WT, and del27\_WT in presence and absence of 2AP. C. schematic of NH5 and regions surveyed.

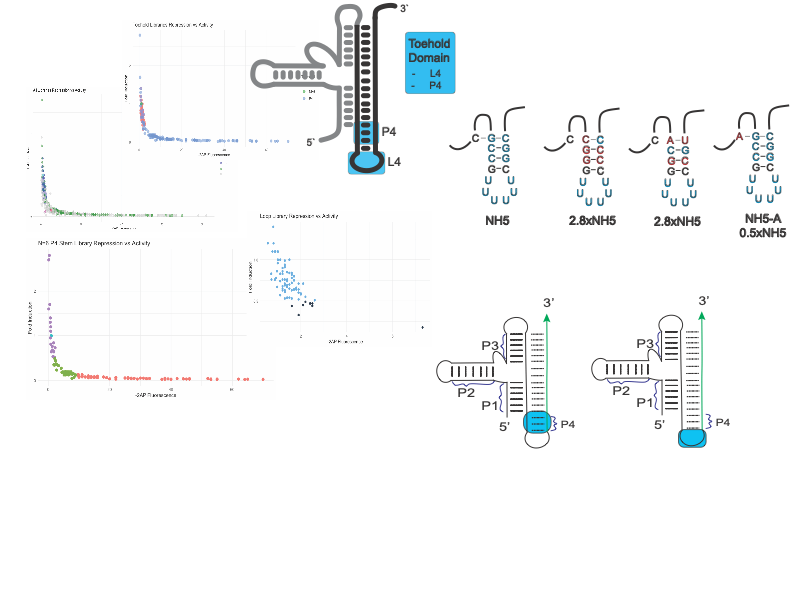


Figure 3: Toehold Graph, schematic, specific variants

Supplemental Table: Library completion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Library | nt varied | observed1 | functional2 | prob.3 |
| Anchor | 5 | 5,118 | 17 | 99.3 |
| Anchor\_exchange (P1\_gateway) | 5 | 5,069 | 97 | 99.3 |
| Gateway | 6 | 22,799 | 141 | 99.6 |
| L4 | 6 | 202 | 77 | 4.81 |
| P4 | 6 | 14,145 | 127 | 96.8 |
| P1\_exchange | 6 | 13,573 | 102 | 96.4 |
| P3 | 5 | 9,029 | 15 | >99.9 |