ABSTRACT

Title of Dissertation	Discovery and Characterization of Antiterminator Proteins in Bacteria
	Jonathan R. Goodson, Doctor of Philosophy, 2018
Dissertation directed by:	Professor Wade C. Winkler Department of Cell Biology and Molecular Genetics

Transcription is a discontinuous process, where each nucleotide incorporation cycle offers a decision between elongation, pausing, halting, or termination. In bacteria, many regulators—including protein antiterminators or cis-acting regulatory RNAs, such as riboswitches—exert their influence over transcription elongation. Through such mechanisms, these regulators can couple physiological or environmental signals to transcription attenuation, a process where RNA structure directly influences formation of transcription termination signals. However, through another regulatory mechanism called processive antitermination (PA), RNA polymerase can become induced to bypass termination sites over much greater distances than transcription attenuation can offer. These mechanisms are widespread in bacteria, although only a few mechanistic classes have been discovered overall. The aim of the research in this dissertation is two-fold: to identify novel genetic regulatory mechanisms targeting transcription termination and to systematically study the diversity and breadth distribution of these mechanisms among bacteria. This research focuses on two distinct mechanisms, each representing one of these mechanisms of antitermination. First, I detail discovery of LoaP, a specialized paralog of the universally conserved NusG transcription elongation factor. Our data demonstrate that *Bacillus velezensis* LoaP controls gene expression of antibiotic biosynthesis gene clusters by promoting readthrough of transcription termination sites. Additionally, we show that, unlike other bacterial NusG proteins, LoaP binds RNA with high affinity, and with apparent specificity for a sequence in the 5′ leader regions of its target operons. Second, we describe the interaction between a family of antitermination proteins containing the ANTAR RNA-binding domain with its target RNA. We show that ANTAR-containing proteins bind a tandem stem-loop RNA motif to prevent formation of terminator structures. Using a combination of mutagenesis strategies, we elucidate some of the RNA-binding requirements of a representative ANTAR protein. Finally, employed bioinformatic and phylogenetic approaches to place these regulators in the context of their entire protein families, learning about the distribution of these mechanisms, their association with particular potential regulons, and sequence composition of different protein subfamilies.

DISCOVERY AND CHARACTERIZATION OF ANTITERMINATOR PROTEINS IN BACTERIA

by

Jonathan R. Goodson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:

Professor Wade C. Winkler, Chair
Professor Kevin McIver
Professor Najib El-Sayed
Professor Carlos Machado
Asst. Professor Daniel Dwyer

Table of Contents

List of Tables

Tables S1 and S2 not present in the main text are electronically attached and available at: https://github.com/jgoodson/Dissertation-Tables

List of Figures

Chapter 1: Termination of Transcription Elongation is Often a Target of Genetic Regulation in Bacteria

Post-Initiation Transcriptional Regulation in Bacteria

An extraordinarily diverse range of genetic regulatory mechanisms has been discovered in the half century since Francois Jacob and Jacques Monod first proposed the operon model of gene regulation (1). Studies based on this model identified a soluble regulator, located distally from the targeted operon, that acts to repress transcription initiation of the lac operon. This discovery led to the identification and characterization of many more repressor proteins, each acting in modestly different ways to reduce the efficiency of transcription initiation. Soon followed discoveries of other types of transcriptional regulators, including those that activate gene expression by enhancing transcription initiation. And now, in an era where bacterial genome sequences can be acquired and draft-annotated in mere days and at low cost, it is clear that all bacteria encode for dozens or hundreds of proteins that regulate transcription initiation and that this 'layer' of genetic regulation is both ubiquitous and profoundly important. However, perhaps because transcription initiation is so universally recognized as a key point of regulatory influence (2), later stages of transcription elongation have not yet been sufficiently analyzed for genetic regulation. While the molecular mechanisms of transcription have been, and continue to be, intensively investigated, the biological extent of post-initiation regulatory mechanisms has been incompletely analyzed. Transcription initiation is only the first stage of gene expression. The stages that follow include transcription elongation, transcription termination, translation and mRNA degradation; each of these stages can be subjected to genetic regulatory control (3).

In this dissertation, I will focus on the discovery or characterization of two mechanistically distinct post-initiation regulatory systems that affect termination of transcription. One of these systems offers signal-responsive control of termination in a process called transcription attenuation, while the other acts to generally depress termination efficiencies at multiple locations along an operon in a process referred to as processive antitermination (PA). Termination signals normally induce rapid dissociation of the transcription elongation complex (TEC) and are most often located at the ends of operons (4). However, when placed within operons, they can serve as key points of regulatory control (5). In bacteria, there are two known classes of termination signals: intrinsic and Rho-dependent terminators (4). In many bacteria intrinsic terminators consist of a GC-rich RNA hairpin followed by a poly-uridine tract. Alone (6), or enhanced by a factor such as NusA (7, 8), these RNA elements promote pausing of the TEC, followed by release of the nascent transcript and dissociation of polymerase (9). In contrast, Rho-dependent termination depends upon the adenosine triphosphate (ATP)-dependent translocase Rho associating with Rho-utilization (*rut*) sites on a nascent mRNA and translocating the RNA to eventually promote TEC dissociation (10, 11). Both classes of termination sites may be specifically regulated by signal-responsive riboswitches (12, 13) or trans-encoded small RNAs (14, 15)

Termination of transcription at any given location is rarely 100% complete, with some proportion of elongation complexes proceeding past the point of termination. In general, two types of mechanisms can control transcription elongation to affect the efficiency of termination: transcription attenuation and processive antitermination (PA). For the former, regulatory mechanisms determine the formation of either Rho-dependent or Rho-independent termination sites (5). Importantly, transcription attenuation-based regulatory mechanisms exert their influence on only a single, defined terminator region by promoting formation of alternate structures incompatible with the formation of terminator elements. In other words, a regulatory RNA that promotes transcription attenuation by definition evolved in concert with the terminator region that it targets—it does not affect other terminator regions. Riboswitches, which are signal-responsive, *cis*-acting regulatory RNAs, oftentimes affect gene expression via transcription attenuation-based mechanisms (16). Riboswitches are widespread in bacteria and offer localized control of transcription termination sites throughout bacterial genomes. In many instances, these transcription attenuation-based regulatory elements can be considered modular, with a signal-responsive portion followed by a portion responsible for premature transcription termination (17). However, transcription attenuation is not limited to riboswitches, and is utilized broadly in bacteria in a variety of mechanisms where protein antiterminators will bind nascent RNA to promote antiterminator formation (5), and is the mechanism underlying the ANTAR family of transcription antiterminators discussed in Chapter 3.

Processive Antitermination

In contrast to transcription attenuation mechanisms, PA mechanisms do not necessarily target a specific terminator region, but instead manipulate elongating RNAP complexes to avoid termination signals throughout an individual transcript (18). In PA mechanisms, antitermination factors associate with a bacterial RNA polymerase (RNAP) elongation complex, leading to read-through of termination sites (18). However, whereas riboswitches, or protein-binding attenuators, exert control over a single intrinsic terminator site, or a particular entry point for Rho, PA systems differ in that they modify TECs to render them generally resistant to downstream termination sites (19). PA systems, therefore, are capable of causing read-through of multiple termination sites, even over long genomic distances. These PA strategies do not take a single form and may reduce transcript termination through a variety of direct and indirect effects. For example, some PA strategies rely on direct interference with factor-mediated termination (20). Alternatively, they can modify recruitment of transcription elongation factors, such as NusA, to affect nascent RNA behavior (21, 22). Additionally, they may alter recruitment of ribosomes in a manner that affects termination within coding regions (23). Furthermore, some PA systems have evolved to utilize multiple strategies simultaneously (21, 24).While only a few classes of PA mechanisms have been discovered in the past four decades, they vary widely in the molecular mechanisms they utilize and in their biological applications. Several new examples of PA mechanisms have been discovered more recently, which appear to be broadly used by bacteria for regulation of diverse sets of genes. We extrapolate from these discoveries that many new PA mechanisms still await discovery.

Phage Lambda Antitermination

During lytic growth, phage λ transcription temporally progresses from one large set of genes to another (26). In order to switch from intermediate-early gene expression to delayed-early gene expression, the phage utilizes a unique protein, λN, to promote antitermination, which enables expression of downstream genes (Figure I-1A) (27). λN is a small protein that is intrinsically disordered alone (28) but is stabilized by protein and RNA contacts in the final, λN antitermination complex (Figure I-2A) (21). Formation of the λN antitermination complex is triggered by synthesis of a *nut* sequence, composed of two RNA elements. The first, *boxB*, is a 15-nucleotide motif that resembles a GNRA tetraloop structure (Figure I-3) (29, 30) and serves as the substrate for λN binding (21, 22). In addition to binding λN, *boxB* also interacts with NusA. Formation of the antitermination complex occurs in steps, with initial association of λN to *boxB* followed by binding of NusA to the λN:*boxB* complex (31). This minimal λN:*boxB*:NusA complex is sufficient for antitermination of *nut*-proximal terminator sequences (18), although it is generally believed that the full antitermination complex *in vivo* relies on additional elongation proteins loaded at the second

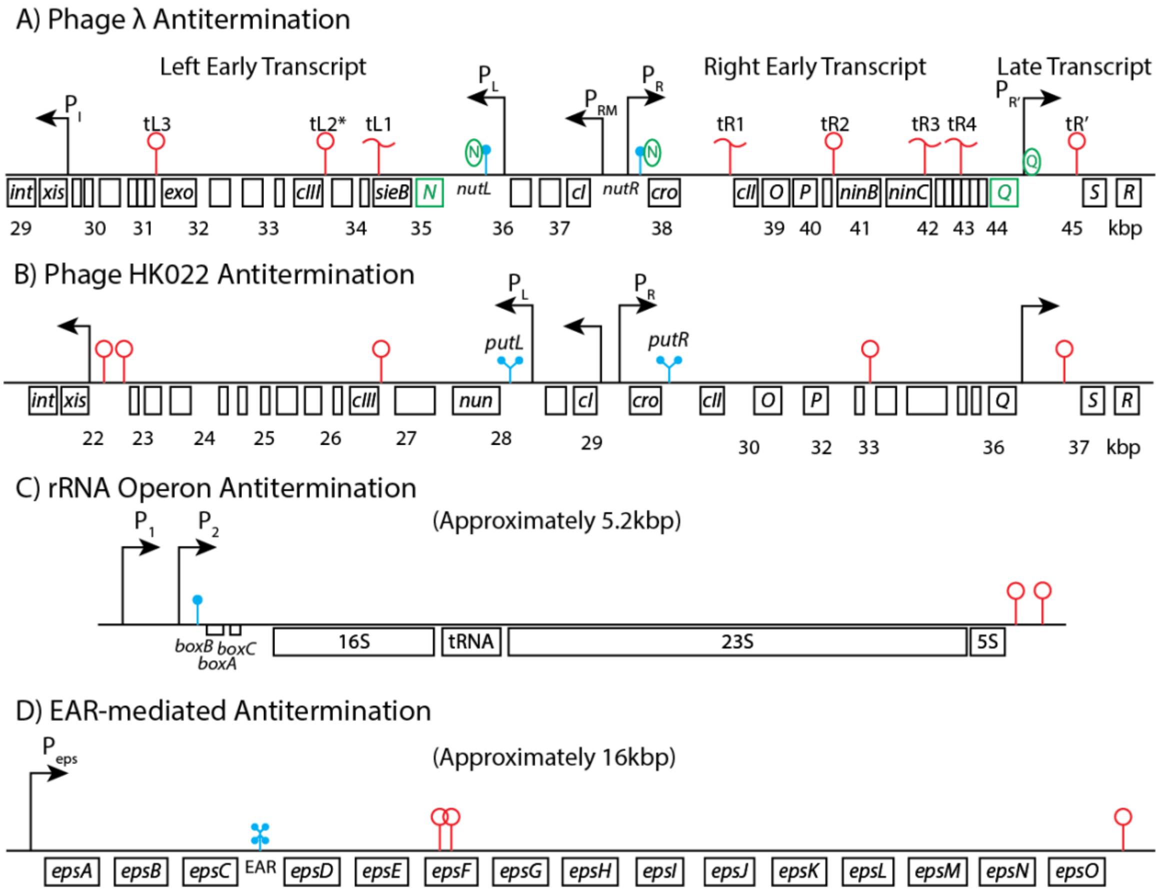


Figure I-1: Genomic Context of Processive Antitermination Systems.
This figure schematically illustrates the transcripts regulated by the λN, *put*, rRNA, and EAR RNA-based antitermination systems. (A) Phage λ early transcripts are initiated from two divergently facing promoters with *boxA/B nut* elements found early in the transcripts. The λN protein is encoded by the first gene in the left early transcript. RNA polymerase (RNAP) complexes associated with λN bypass multiple terminators in both transcripts. Using a different mechanism, the λQ protein promotes antitermination of the late transcript by binding to DNA near the late promoter and promoting a terminator-resistant configuration of RNAP. (B) Phage HK022 early transcripts are similar to phage λ, although they include *put* elements early in each transcript, which trigger λN-independent antitermination. Additional Rho-dependent terminators are likely present in these transcripts, although they have not been specifically characterized and are therefore not indicated on this particular diagram. (C) A representative *E. coli* rRNA operon is shown, containing *boxA/B/C* elements immediately downstream of the P_2 promoter. These elements promote read-through of Rho-dependent termination in the non-coding rRNA genes. (D) Several intrinsic terminators have been demonstrated in the *Bacillus subtilis eps* operon, which encodes for biosynthesis of biofilm exopolysaccharides. The *eps*-associated RNA (EAR) is found within the *epsBC* intergenic region and promotes read-through of the terminators within the operon. Intrinsic terminators are shown as sticks with empty circles, and Rho termination regions are shown as sticks with wavy lines, both in red. RNA elements involved in antitermination are show in blue, and proteins and protein-coding genes involved in antitermination are shown in green. Elements are not shown to scale. *Reproduced from* (25).

RNA element. This second RNA element, *boxA*, acts as a loading site for the NusB:NusE (S10) complex (32). Binding of the NusB:NusE (S10) complex to *boxA* promotes additional contacts between λN and NusA. This resulted in a unique complex of factors that are associated with RNAP near the RNA exit channel and remain together as a ribonucleoprotein complex (Figure I-2B) (21).

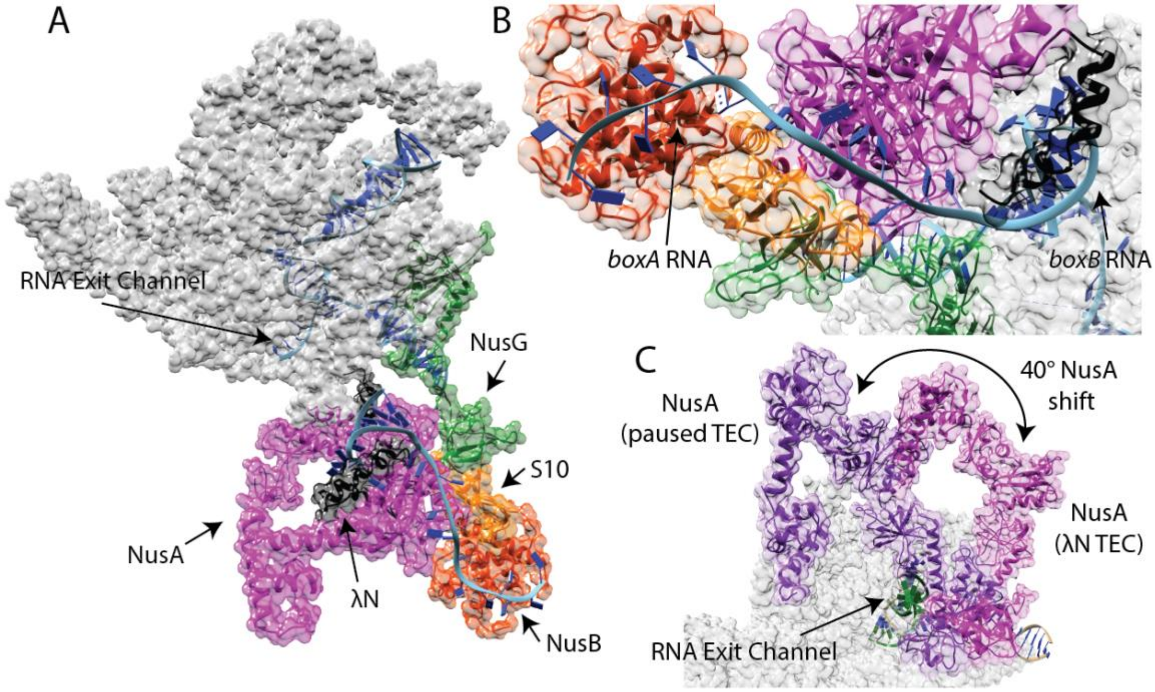


Figure I-2: Cryo-electron Microscopy Reveals Details of Antitermination Mechanism. This figure contains structural models generated from cryo-EM data of transcription elongation complexes (PDB 5MS0, PDB 6FLQ). (A) This panel shows the λN antitermination complex (PDB 5MS0) comprising λN (black), NusA (magenta), NusB (red), S10 (orange), NusG (green), and *boxA/B* RNA (blue), in addition to RNA polymerase (gray). (B) A zoom-in on the *boxA/B* and λN complex shows an extended binding of the *nut* RNA sequence with multiple protein components, with *boxA* bound to the NusB/S10 dimer and the *boxB* hairpin bound to λN and NusA. (C) Formation of the λN antitermination complex shifts the position of NusA (magenta) by 40 degrees away from the RNA exit channel, as compared to NusA (purple) in a transcription elongation complex constructed with the *E. coli his* hairpin-mediated pause sequence (PDB 6FLQ). Nascent RNA is shown in green. *Reproduced from* (25).

Binding of λN alone to RNAP modifies transcription elongation both *in vitro* and *in vivo*, promoting antitermination by modulating RNA exit channel elements and by suppressing melting of the RNA:DNA hybrid after terminator hairpin formation or in response to Rho activity (21, 22, 33, 34). However, formation of the complex with the full complement of transcription elongation factors is thought to further stabilize the interaction of λN with RNAP and increase its duration of occupancy—and, therefore, overall processivity—of λN antitermination (35). In “standard” transcription elongation complexes, NusA binds RNA polymerase near the RNA exit channel where it can enhance intrinsic termination (36). Indeed, NusA affects transcription termination at many locations across the genome and is even required for formation of some NusA-dependent termination sites (8). However, λN is thought to counteract the direct effects of NusA on terminator hairpin folding (22). A recent high-resolution structural model of the λN antitermination complex revealed that the C-terminal RNA-binding domains of NusA are repositioned such that they redirect nascent RNA away from the RNA exit channel (Figure I-2C). This is predicted to reduce formation of terminator hairpins, thereby essentially reprogramming NusA into a transcription antitermination factor (21). Formation of the λN complex also inhibits Rho-dependent termination. In “standard” elongation complexes, NusG helps recruit Rho to nascent RNA and thereby aids in Rho termination (37, 38). In contrast, the λN antitermination complex is likely to restrict NusG-mediated recruitment of Rho by instead promoting association of factors that compete for binding to NusG (e.g., S10:NusB), and also because of restricted access to the nascent RNA as it is looped out of the antitermination complex (21). Therefore, the λN complex acts as a physical roadblock to prevent Rho translocation and helps occlude access to Rho utilization (*rut*) sites.

Phage λ also contains a second antitermination system, which relies upon another unique protein (λQ) to promote antitermination of late-expressed genes (18, 39). However, unlike the N-antitermination system, λQ protein is a DNA-binding protein that associates with RNA polymerase within the promoter region during transcription initiation and triggers formation of an antitermination complex that is different from the N complex (40).

Ribosomal RNA Operon Antitermination

Dissociation of transcription elongation complexes by Rho helicase underlies the polarity which occurs when nonsense mutations reduce transcript abundance of downstream genes (41). Rho is capable of loading onto RNA molecules via C-rich binding sequences (*rut* sites), but the presence of ribosomes during coupled transcription-translation generally reduces Rho loading and translocation (42). Given that ribosomal RNA operons are not translated and are thereby not protected by ribosomes, their transcripts must be protected from Rho termination by other means. This protection may be partially explained by the extensive secondary structure of ribosomal RNAs, which acts to reduce loading of Rho at potential *rut* sites (43, 44). However, in Escherichia coli and many other bacteria, these operons are also subjected to an antitermination system that resembles closely the λN-antitermination mechanism (43–45). For example, the 5′ leader regions of E. coli rRNA operons contain *boxA* as well as a *boxB*-like hairpin, although only boxA appears to be essential for antitermination activity (Figure I-1C) (32, 46). Binding of the NusB:NusE (S10) complex to *boxA* RNA occurs in a manner similar to N-mediated antitermination and promoting a conformational state that strongly disfavors association of Rho (32).

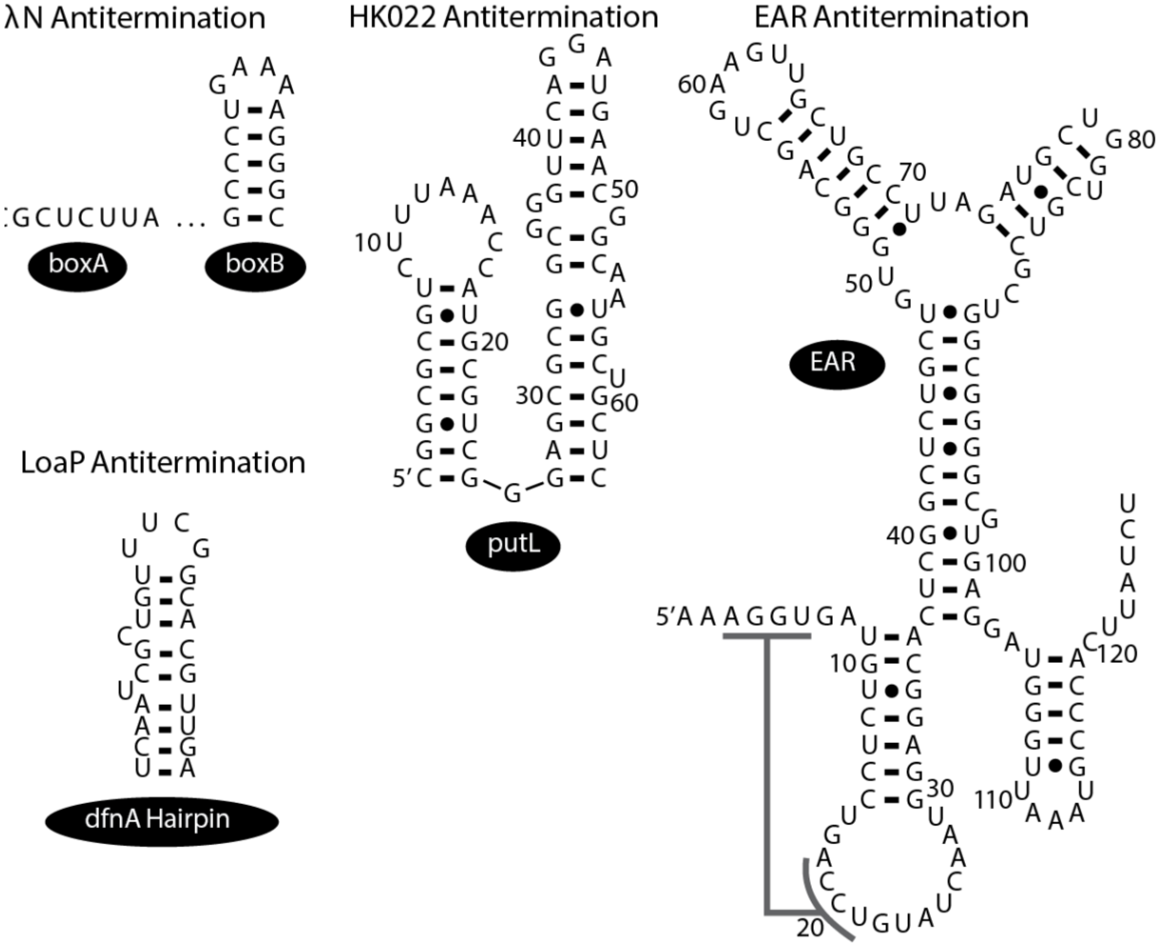


Figure I-3: RNA Elements Involved in Processive Antitermination.
This figure shows the sequence and secondary structure of RNA elements known or predicted to be utilized in processive antitermination mechanisms. Shown are the *boxA* and *boxB* elements forming the λN *nut* sequence as well as rRNA antitermination signal, the *put* RNA element from phage HK022, EAR from the *B. subtilis* exopolysaccharide pathway, and a UNCG-type hairpin implicated in LoaP antitermination. *Reproduced from* (25).

In contrast to λN antitermination, which requires N protein in addition to host Nus proteins, rRNA antitermination requires an additional host factor, SuhB (47). The complete elongation complex containing NusB:NusE, NusA, NusG, and SuhB is required not only for full rRNA antitermination activity *in vitro* but for correct rRNA maturation *in vivo* (47). In addition to regulation of rRNA transcription, *boxA* and Nus factors directly repress *suhB* translation in enterobacteria in a manner reminiscent of λN autoregulation and have been implicated in regulation of additional genes (48). Therefore, the rRNA antitermination system relies exclusively on general transcription elongation factors and their recruitment to the *boxA* RNA element. This system serves a dual purpose in rRNA operons, promoting both antitermination and RNA folding, and may regulate yet additional transcripts. Together, these observations suggest that N-antitermination may have arisen as a modification of the host Nus protein antitermination system, where λN protein evolved to reconfigure and further manipulate host transcription elongation factors.

RNA Elements that Promote Processive Antitermination

In addition to the role that RNA elements (*boxA* and *boxB*) play in antitermination of phage λ and rRNA operons, a few PA systems have been discovered that involve larger and more complicated RNA elements. Many if not most lambdoid phages utilize PA systems related to both N- or Q-antitermination (18). However, phage HK022 differs in that it encodes for λQ yet lacks λN, despite the fact it still requires antitermination of early-expressed genes (49). Moreover, HK022 does not utilize *nut* sites for antitermination. Instead, early gene antitermination is mediated directly by a larger RNA motif called *put*, found in regions analogous to λ *nut* sites (Figure I-1B) (50). HK022 *put* forms a two-hairpin RNA element of approximately 65 nucleotides in length that is critical for antitermination activity (Figure I-3) (50, 51). This element appears to directly affect RNAP elongation activity through pause suppression, potentially requiring no additional elements to promote antitermination (49). Evolution of this mechanism is likely interrelated with the evolution of a λN-like protein, Nun, which is also produced by HK022 (52, 53). Nun, found in the same relative genomic position as λN in phage λ, instead promotes Nun-termination at *nut* elements by binding to *boxB* and inhibiting RNAP translocation (54, 55). HK022 *put* promotes antitermination of both Rho-dependent termination and Nun-dependent transcription arrest in the HK022 early transcripts (54) as well as intrinsic terminators (56). While some mechanistic details of *put*-mediated antitermination are still lacking, its discovery was significant as it demonstrated proof-in-principle that PA could be driven primarily by RNA elements.

More recently, an even larger and more structurally complicated RNA element was discovered to trigger PA in bacteria. This RNA element, which is at least ~125 nucleotides in length and is constructed from an array of at least five helical elements and a characteristic pseudoknot, was discovered to be broadly conserved in Bacillales (Figure I-3) (57). Coined the EAR element, for *eps*-associated RNA, it is almost always associated with operons that encode for biosynthesis of biofilm or capsule exopolysaccharides (Figure I-1D). Either mutagenesis of conserved residues or deletion of EAR resulted in incomplete transcription of the *Bacillus subtilis eps* operon. Instead, transcripts were found to be prematurely truncated at the site of intrinsic terminators, located in the middle region of the *eps* operon. Indeed, placement of EAR directly upstream of this terminator site resulted in nearly complete read-through of the terminators *in vivo*, whereas, conversely, mutagenesis of conserved EAR residues resulted in termination. Moreover, placement of EAR upstream of unrelated intrinsic terminators, originating from sources other than the *eps* operon, still resulted in their read-through, demonstrating that EAR promotes general PA of intrinsic terminators. That EAR promoted read-through of intrinsic terminators is strikingly different than the biological utilization of the λN and rRNA PA systems, which are believed to function primarily for read-through of Rho termination. However, EAR PA has not yet been recapitulated *in vitro* or in a heterologous host, indicating that at least one additional factor may be required for its antitermination activity, in contrast to HK022 *put*. Regardless, discovery of EAR demonstrated that structurally complicated RNAs, with the size and apparent complexity resembling that of riboswitches, are sometimes used to promote PA. Moreover, the distribution of EAR PA determinants further showcases how PA mechanisms can be broadly important for biologically important functions such as biofilm formation.

Specialized NusG Paralogs

RfaH

Although most known PA systems are found in phage genomes or are reliant on general transcription elongation factors, some Gammaproteobacteria encode for the specialized PA and translation factor RfaH (24). RfaH is a paralog of NusG. NusG is an elongation factor generally associated with transcription elongation complexes and is an integral component of the λ and rRNA PA systems (58). RfaH, encoded by an essential gene in *E. coli*, is required for the expression of a regulon of virulence-related pathways—including synthesis of haemolysin, lipopolysaccharide, and the F-factor sex pilus (58, 59)—as well as additional targets involved in the production of membrane or extracellular components (60).

As a paralog of NusG, RfaH is a small protein containing two conserved domains. In general, the core domains of NusG homolog proteins exhibit strongly conserved structure (61, 62) and interface with RNAP in a similar fashion (62–64). The first domain is an N-terminal domain (NTD) unique to the NusG/Spt5 family of proteins (65). This domain is responsible for binding of RfaH to RNAP at the same site normally occupied by NusG. The C-terminal domain (CTD) contains a KOW (Kyprides, Ouzounis, Woese) motif found in several ribosomal proteins in addition to NusG (66). This characteristic CTD is shared among nearly all NusG homologs as well as several ribosomal proteins (66), and is believed to function as a tether that can interact with additional proteins (67).

While RfaH and NusG have distinct regulatory consequences, they rely on similar mechanisms to improve transcriptional processivity (64). The NTD of both proteins share highly similar sequences and structures (60, 62) and suppress pausing at many sites when added to purified transcription complexes *in vitro* (20, 68, 69). Both proteins are believed to suppress pausing by binding to the β' clamp and β pincer and stabilizing the active closed conformation of RNAP (62, 70). Recently, single molecule cryo-EM studies have clarified how stabilization of RNAP structure can promote processive elongation. Certain types of transcriptional pauses are affected by a swiveling of the RNAP β' pincer elements, resulting in an increase in pause lifetimes (71). However, binding of NusG or RfaH to RNAP disfavors this "swiveled" conformation, thereby suppressing pausing (64). Additional mechanisms for anti-pausing activity of NusG proteins have been proposed, including stabilization of the elongation complex by direct binding to non-template DNA (20, 69, 72) as well as upstream DNA (73–76). Indeed, both NusG and RfaH interact with the upstream DNA fork and promote re-annealing of the upstream DNA, although the specific effects of this activity on RNAP activity are unclear (64, 73). These mechanisms are conserved between NusG and RfaH, and are likely to be shared to varying degrees with other NusG paralogs.

RfaH is specifically recruited to the operons that comprise its regulon by a DNA element called the operon polarity suppressor, or *ops* (Figure I-4A). Deletion of this 8-bp conserved element reduces downstream gene expression (59); correspondingly, introduction of *ops* to other transcripts increases their expression (58). Depletion of RfaH mirrors these results, indicating that RfaH and *ops* are both required for expression of target operons (77). RfaH is specifically recruited to transcription elongation complexes by binding to the non-template DNA strand of the *ops*-element; this occurs during the lifetime of a programmed transcriptional pause (20). The *ops*-element forms both a consensus pause sequence as well as a DNA hairpin loop that makes specific, direct contacts to the RfaH NTD (78). RfaH and NusG are mutually exclusive, as both homologs share the same binding site on RNAP (79, 80). Moreover, once recruited, RfaH exhibits increased affinity for RNAP relative to NusG, allowing for extended association of RfaH with TECs (81). This increased affinity may also be responsible for the more pronounced effects of RfaH NTD on RNAP as compared to NusG (64). In this way, RfaH exerts its regulatory effects specifically on those operons that include the *ops* element.

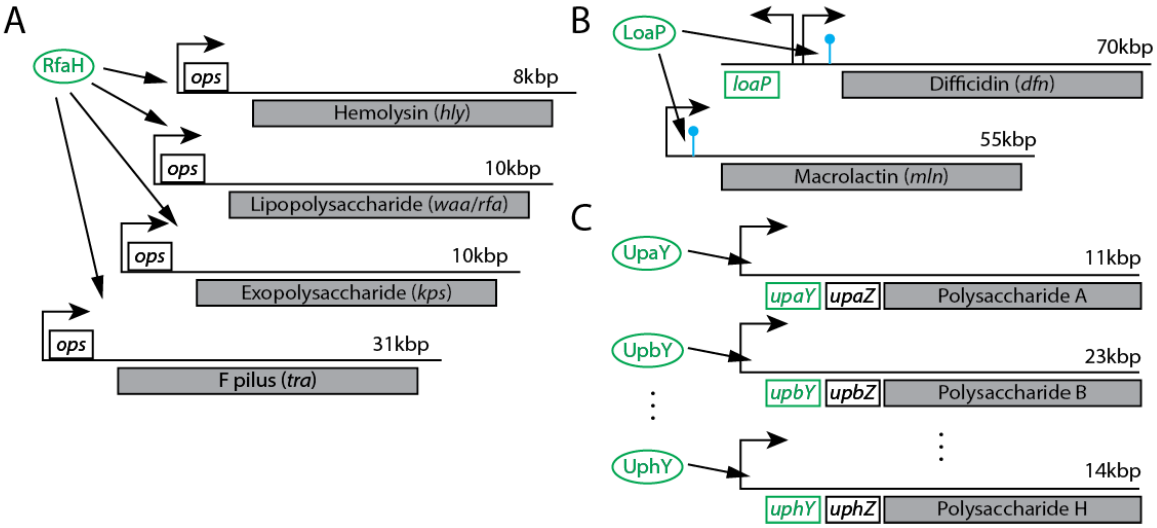


Figure I-4: Genomic Context of NusG Paralog Antitermination Systems.
This figure illustrates the transcripts regulated by the RfaH, LoaP, and UpxY antitermination systems. (A) RfaH regulates multiple pathways in *E. coli* including the hemolysin, F pilus, and lipo- and exo-polysaccharide operons. Each regulated transcript includes the DNA *ops* element for RfaH recruitment. RfaH promotes antitermination of Rho-dependent promoters. (B) LoaP regulates two polyketide antibiotic operons in *B. velezensis*: the *dfn* difficidin operon and the *mln* macrolactin operon. LoaP is found divergently oriented upstream of the *dfn* operon. Each transcript includes a required sequence region in the 5' leader region, which might include a functionally important hairpin followed by an intrinsic terminator. Additional intrinsic terminator sites have been implicated within the *dfn* and *mln* operons, although they are not shown in this figure. (C) UpxY proteins regulate multiple capsular polysaccharide pathways in *B. fragilis*. Each polysaccharide operon includes both a UpxY and UpxZ protein involved in targeted regulation, with 5' leader sequence required for antitermination. *B. fragilis* has eight distinct polysaccharide operons containing UpxY proteins. Grey rectangles represent multi-gene operons. RNA elements potentially involved in antitermination are shown in blue, and proteins and protein-coding genes involved in antitermination are shown in green. *Reproduced from* (25).

RfaH in solution differs from RNAP-associated RfaH. Instead of the common β-barrel fold found in most high-resolution structures of KOW domains, the CTD of free RfaH forms a dramatically different α-helical structure (79). This α-helical CTD interacts with the NTD, partially masking the RNAP-binding portion and thereby resulting in an autoinhibited form of the protein (Figure I-5) (81). After a conformational change is triggered, the NTD can associate fully with the transcription complex, which in turn promotes re-folding of the CTD to the β-barrel structure found in NusG (Figure I-5) (24). Because of this structural mechanism, RfaH adopts the classical NusG KOW domain structure only after the NTD has fully associated with RNA polymerase.

Though NusG and RfaH display nearly identical anti-pausing effects on transcription complexes *in vitro*, their overall regulatory outcomes are different. In some instances, NusG may promote pausing *in vivo* (82), perhaps as a result of increased affinity for certain non-template DNA strand sequences (69). More importantly, NusG is known to directly bind Rho (67). This interaction is likely to broadly promote Rho-dependent termination activity, possibly by increasing the rate at which Rho successfully binds RNA and forms a closed translocation-capable conformation (38). Association of NusG results in Rho-dependent termination and suppression of transcription, particularly in genomic regions that feature foreign DNA (23). This activity is essential in most *E. coli* strains primarily due to suppression of toxic genes in prophage DNA (23). However, in addition to its interaction with Rho, the NusG CTD can associate with NusE (S10), as well as NusA (83, 84). Similar to NusG, RfaH can associate with NusE (S10); however, in contrast to NusG, RfaH is incapable of binding Rho (24, 83). Because of this, RfaH strongly discourages Rho termination within its targeted operons (85).

Finally, the remaining mechanism by which RfaH may promote antitermination is through recruitment of ribosomes to nascent transcripts. NusG proteins are thought to couple transcription and translation by facilitating macromolecular interactions between both of these machines (83). RfaH in particular has been shown to exhibit much stronger polarity effects *in vivo* than its effects on transcription *in vitro* (81). Also, genes that are known to be regulated by RfaH display particularly poor ribosome binding sites, suggesting that translational enhancement is likely to be a key feature of RfaH regulation (85). It is possible that binding of NusG or RfaH to ribosomal S10 (NusE) may assist ribosome recruitment, thereby increasing local concentration of ribosomes and promoting translation initiation on nascent RNA (24, 86). This functional interaction might also affect transcription processivity. Indeed, recent data suggest that the leading ribosome—which conducts translation immediately upstream of RNAP, and that may participate in the RNAP-ribosome “expressome” (87, 88)—improves transcription processivity by directly blocking RNAP backtracking (89) and by obstructing Rho access (83, 90).

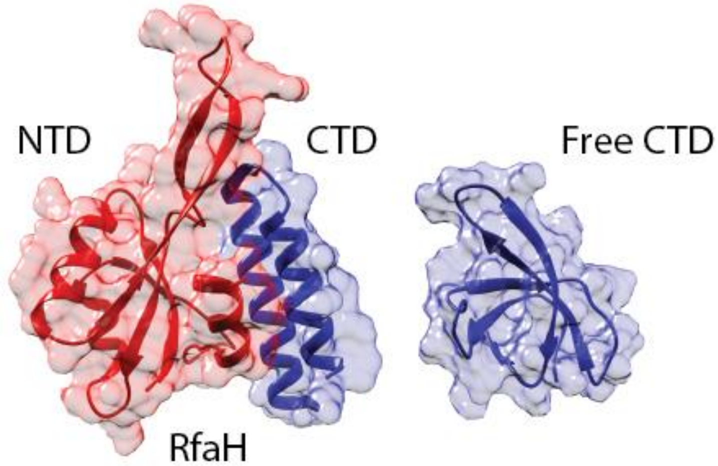


Figure I-5: RfaH-CTD Undergoes a Large Conformational Shift from an α helix to β barrel. Full length RfaH (left) exists as an autoinhibited structure with the CTD (blue) in an α helix conformation bound to the NTD (red) (PDB: 2OUG). Upon binding to RNAP and the *ops* DNA, the CTD (right) is released and forms the β barrel conformation characteristic of NusG KOW domains (PDB: 2LCL). *Reproduced from* (25).

Through these aggregate mechanisms, RfaH acts as a specialized elongation factor that exhibits anti-pausing activity, prevents NusG-mediated Rho termination, and encourages ribosome recruitment, for each of the operons that display *ops* elements.

Other NusG Paralogs (ActX, TaA, UpxY)

Although RfaH is the most prominent and best studied NusG paralog, other examples have been identified, several of which have been predicted to function in transcription antitermination (91–94). All of these homologs share significant sequence similarity to NusG and RfaH and undoubtedly share conserved structural features. Moreover, for those NusG paralogs where a functional role has been demonstrated, they have inevitably been found to affect transcription of certain targeted transcripts, suggesting that NusG paralogs are broadly used by bacteria as specialized transcription regulators (92–94).

ActX and TraB proteins are most phylogenetically similar to RfaH (91, 95) and are found in a variety of conjugative plasmids conferring antibiotic resistance in Gammaproteobacteria (91, 96). Though a function has not been demonstrated for these proteins, they are often transcribed as the first open reading frame in long pilus biosynthesis operons and are suspected to be involved in the transcription of conjugation genes (97).

Myxococcus xanthus, a Gram-negative soil bacterium, produces the well-studied polyketide antibiotic TA (also called myxovirescin) (98, 99). The first open reading frame of the TA-producing gene cluster is *taA*, which encodes for a NusG paralog (92). Disruption of the *taA* gene eliminated antibiotic production, suggesting a regulatory relationship. However, the specific role of TaA in expression of the TA gene cluster is unknown, although as a NusG paralog and relative of other known NusG specialized paralogs, it has been proposed to regulate transcription elongation, perhaps through PA.

More recently, a NusG paralog called UpxY has been proposed to function as a family of regulators for complex polysaccharide pathways in Bacteroidetes (93, 100). They are widely used by these microorganisms. Indeed, many Bacteroides encode between six and nine copies of the UpxY proteins. The genes encoding these proteins, initially described in *Bacteroides fragilis*, are each associated with a different capsular polysaccharide gene cluster (Figure I-4C) (100). These proteins have been shown to affect transcription of their associated gene cluster and it has been proposed that they participate in antitermination-based regulatory mechanisms that involve unique sequence features located within the 5′ leader regions of their respective operons. Additionally, while these regulators might be co-transcribed with the operons they affect, they can also affect gene expression when moved to a distal location, supporting the claim that they are recruited to their targeted operons, perhaps via sequence elements within the transcript leader regions. Yet, despite these observations, little is known regarding the molecular mechanisms of UpxY proteins. Adding a new wrinkle to the overall family of NusG paralogous proteins, *Bacteroides fragilis* also encodes a set of unique proteins (UpxZ) alongside genes encoding UpxY proteins. The UpxZ proteins can act as *trans*-inhibitors of UpxY proteins, and have been hypothesized to hierarchically regulate the expression of different sets of capsular polysaccharides, although the underlying mechanism of this inhibition is also unknown (93).

Outlook

It is in this context that we began investigating the NusG family of proteins with the goal of discovering new processive antitermination systems and expanding our understanding of the importance and utilization of these PA systems in bacteria. In particular, we initially focused on a more specific question: do processive antitermination elements often regulate large, complex, gene clusters? Most of the PA elements previously discovered regulate larger multi-cistronic transcripts, including large polysaccharide and phage pathways. A single example of a putative processive antiterminator, TaA from *M. xanthus* is predicted to regulate a polyketide antibiotic pathway. Therefore, we investigated whether NusG paralogs could affect polyketide pathways, leading to the discovery of LoaP, described in Chapter 2. The discovery of LoaP, along with the description of TaA, suggests that transcription elongation may be a broad point of regulatory control for secondary metabolite gene clusters in bacteria.

Chapter 2: Discovery of the LoaP Family of Processive Antiterminators

Introduction

After nearly a century of searching for bioactive natural products from microbes, bacteria still constitute a major target of modern drug discovery (101, 102). Most bacteria include in their genomes at least a few gene clusters encoding for biosynthesis of complex specialized metabolites, of which a subset exhibit biomedically relevant activities. In recent years, it has become increasingly productive to screen for bioactive molecules from culture supernatants of random environmental bacterial isolates. However, the characterization of the biochemical pathways of these molecules remains a critical bottleneck to their analyses and industrial application. One of the key problems in characterizing these pathways is a shortage in knowledge on the range of genetic mechanisms that can affect them. Indeed, an incomplete understanding of the genetic regulatory mechanisms that affect specialized metabolite pathways limits the overall field in at least two ways. First, it limits discovery of new natural products, as many metabolite production pathways may be transcriptionally silent in culture conditions. Second, it stifles attempts at expressing specialized metabolites from within heterologous hosts. Therefore, discovery of new classes of genetic regulatory mechanisms will greatly impact development of new natural products.

Characterization of genetic mechanisms affecting specialized metabolite synthesis pathways has been generally restricted to analysis of transcription factors (2). These DNA-binding proteins largely affect the efficiency of transcription initiation by binding within the promoter region and altering its interaction with RNA polymerase holoenzyme. However, initiation is only the first stage of the information processing pathway. In general, the post-initiation stages that follow initiation include transcription elongation, transcription termination, translation, and mRNA degradation. Notably, each can be subjected to genetic regulatory control, oftentimes involving different types of regulatory RNAs (3). Yet, while many post-initiation regulatory mechanisms have been discovered, the extent of their influence on specialized metabolite pathways has been incompletely studied. It is possible that the unusually long length of many specialized metabolite gene clusters may present molecular challenges that can be resolved by post-initiation regulatory mechanisms. For some specialized metabolite synthesis gene clusters, there may be an important role for regulatory mechanisms that improve efficiency of transcription elongation and that ensure transcript completion.

Bacterial RNA polymerase (RNAP) is a highly processive enzyme that can rapidly catalyze addition of numerous individual nucleotides, extending the nascent transcript for tens of thousands of residues. Despite its ability to transcribe RNA molecules over long distances, RNAP largely lacks an ability to restart transcription from a truncated RNA molecule that is greater than ~2-4 nucleotides in length. As a consequence, termination of transcription is irreversible and an RNAP elongation complex must transcribe the entire targeted operon in one transcription cycle. For some operons, processive antitermination (PA) mechanisms are used to assist this goal.

As discussed in Chapter 1, a few examples of bacterial and phage PA mechanisms have been discovered (18, 19, 103). In these mechanisms, antitermination factors associate with an RNAP elongation complex to bypass terminator sites. The best-characterized PA examples are mediated by phage lambda proteins N and Q, which primarily prevent Rho termination for early and late gene transcripts, respectively (18, 19, 103). For the related phage, HK022, a ~65-nucleotide, *cis*-acting RNA sequence ('PUT') can supplant the role of N or Q antitermination proteins by protecting against Rho termination sites in their absence (50). Although featuring a completely different sequence and structure, another *cis*-acting RNA sequence capable of promoting PA is the ~125-nucleotide 'EAR' element, which promotes readthrough of termination sites located within operons encoding exopolysaccharide biosynthesis for *Bacillus subtilis* and other bacteria (57). In certain gram-negative bacteria, a NusG paralog, RfaH, spurs readthrough of Rho termination sites within horizontally acquired operons (20, 85, 97). RfaH adjoins with RNA elongation complexes through recognition of a characteristic nontemplate DNA sequence, and does not require additional factors, making it a dedicated antitermination factor. However, RfaH is not the only NusG paralog to be discovered for genetic regulation. Indeed, a different NusG paralog, UpxY, is widespread in Bacteroidetes, and is also presumed to trigger PA (93). Interestingly, the genomes for these bacteria encode several different UpxY proteins, and each is associated with a distinct polysaccharide pathway. While the molecular mechanism of UpxY regulation has yet to be revealed, it is already clear there is a broadly important role for antiterminator proteins in regulation of Bacteroidetes capsular polysaccharides.

Previously, a *Myxococcus xanthus* NusG paralog, TaA, was identified as a putative genetic regulatory protein; mutational disruption of the gene resulted in decreased expression of a nearby polyketide synthase gene cluster responsible for myxovirescin (92). While antitermination activity has yet to be actually observed for TaA, this discovery suggested proof-in-principle that transcription elongation factors may participate in regulation of some specialized metabolite pathways. Indeed, overexpression of a NusG homolog led to discovery of a new natural product synthesized by *Clostridium cellulolyticum*, although antitermination activity has not yet been examined for this factor either (104). These observations further support the idea that an improved understanding of transcription elongation-based regulatory mechanisms will aid discovery, characterization and production of microbial specialized metabolites. Inspired in part by these observations, we hypothesized that there could be a other NusG-related proteins that specifically promote PA within specialized metabolite biosynthesis gene clusters.

In this chapter I present the discovery of a NusG specialized paralog in *Bacillus velezensis*, to which we assign the name LoaP. LoaP promotes readthrough of intrinsic termination sites located within the polyketide synthase (PKS) gene cluster encoding for the antibiotic difficidin (*dfn*). We find that LoaP can also promote readthrough of heterologous intrinsic terminators through a mechanism that requires a portion of the *dfn* 5′ leader region. Also, RNA-seq analyses of LoaP-lacking strains revealed that LoaP promotes PA of a second PKS gene cluster, encoding for a different antibiotic, macrolactin (*mln*). Yet, strikingly, it has no effect on a third PKS pathway encoding for production of bacillaene. Therefore, our data show that *B. velezensis* LoaP promotes PA of a selective regulon of antibiotic gene clusters. Additionally, we show that this antitermination activity is dependent on a specific sequence found in the leader region of both regulated transcripts, and ultimately that LoaP is capable of direct, specific binding to this RNA sequence. In a later chapter, a broad scale phylogenetic analysis reveals that the LoaP protein is widespread in Firmicutes, Actinobacteria and Spirochaetes, and is oftentimes associated with specialized metabolite gene clusters or polysaccharide biosynthesis operons. These data together demonstrate that transcription elongation is subject to regulatory control by PA for some specialized metabolite gene clusters, and implicate a cohesive subgroup of NusG paralogs, with distinct mechanisms of action or recruitment, for this purpose.

Identification of potential antiterminator proteins associated with specialized metabolite pathways

Hmmer and AntiSMASH

Beginning after discovery of the EAR antitermination system regulating complex polysaccharide pathways in *Bacillus* species, we initially hypothesized the existence of processive antitermination systems regulating large specialized secondary metabolite clusters in bacteria. The NusG paralog RfaH was the only known specific processive antitermination factor in Bacteria (19), so we initially began an extensive survey of all NusG homologs found in all sequenced bacterial genomes available in the NCBI genomes database. We identified NusG homologs using HMM searches for the NusG N-terminal domain unique to NusG/Spt5 proteins utilizing the Pfam NusG N-terminal domain (PF02357) model (105, 106). Regulators of specialized metabolite synthesis pathways in bacteria are often located near or in the pathways that they regulate (107), so we hypothesized that specialized NusG paralogs might be found in the genomic region of their regulated gene cluster. We utilized the antiSMASH pipeline to predict and identify potential secondary metabolite gene clusters in the same set of bacterial genomes (108). We then filtered these sequences for those found within 5,000 bp of a predicted metabolite pathway in a genome with at least two NusG homologs. This additional restriction avoids core NusG protein coding sequences that happen to be present near an antiSMASH gene cluster and enriches for likely paralogs. Manual examination of these results revealed a class of NusG sequences found immediately adjacent to predicted polyketide biosynthesis gene clusters in several species of Bacillales. We then chose a representative sequence for further study which was found in a bacterial species which has been previously studied and found to be naturally competent (109). Specifically, we chose to investigate RBAM_022090, a *B. velezensis* FZB42 NusG homolog found adjacent to the *dfn* gene cluster, which is responsible for synthesis of the polyketide antibiotic difficidin (109).

Development of genetic tools for *B. velezensis*

We chose *B. velezensis* FZB42 as a target because integration of DNA into the chromosome had been previously established, transforming cells with genomic DNA from other *Bacillus (110)* or plasmid DNA (111). Despite this use of the strain in the literature, no plasmids were available specifically for integration into its genome; therefore, I constructed a plasmid for this purpose. We initially began with construction of a plasmid that could be used for marker-replacement of the *loaP* gene by introducing a two kilobase DNA fragment containing *loaP* into the standard pUC18 cloning plasmid using Gibson assembly (112), followed by interruption and replacement of the *loaP* gene in this plasmid with an erythromycin resistance cassette using similar methods. Initial attempts to transform *B. velezensis* using the classic Spizizen methods for *Bacillus* described in the literature (109) were unsuccessful, however growth in nitrogen-limited conditions as described for induction of natural competence in *B. subtilis (113)* were successful in inducing competence and integration of the marker-replacement plasmid into the genome.

After initial experiments indicated a possible role for LoaP, we constructed an additional plasmid system for integration of ectopic DNA into the *amyE* region of the *B. velezensis* genome. To do this, we modified the *B. subtilis* integration vector pDG1662 to replace the *B. subtilis amyE* sequence with homologous sequence from *B. velezensis amyE*. Additionally, we cloned the *loaP* gene sequence into a xylose-inducible vector pSWEET-III and transferred the full expression cassette into this *B. velezensis* integration vector, which was integrated in wild-type *B. velezensis* as well as the Δ*loaP* (‘knockout’) strain. This permitted inducible complementation of LoaP in the Δ*loaP* background strain. Later, we additionally integrated a constitutive promoter-YFP reporter cassette from our plasmid pJG019 adjacent to the xylose-inducible promoter. This created a dual-function integration vector to allow for simultaneous control of LoaP expression and measurement of LoaP’s effects on the fluorescent reporter in a genetic context that resembled the diverging promoters of *loaP* and *dfnA*.

LoaP is required for transcription and production of polyketide antibiotic pathways in *B. velezensis*

mRNA abundance of the *dfn* gene cluster is dependent on LoaP

Bacillus velezensis contains multiple specialized metabolite gene clusters, which encode for production of several polyketides, non-ribosomally produced lipopeptides, and bacteriocins (114). These gene clusters range from five to ten kilobases for bacteriocins, twelve to forty kilobases for non-ribosomal lipopeptides, and from fifty to seventy-five kilobases for polyketide-producing gene clusters (114, 115). The second longest of these gene clusters, *dfn*, encodes for the production of the polyketide antibiotic difficidin (115). Immediately upstream of this *dfn* gene cluster, which is arranged as a single very long operon, is the gene *loaP*, encoding for a protein distantly related to the transcription elongation factor NusG (Figure II-1A).

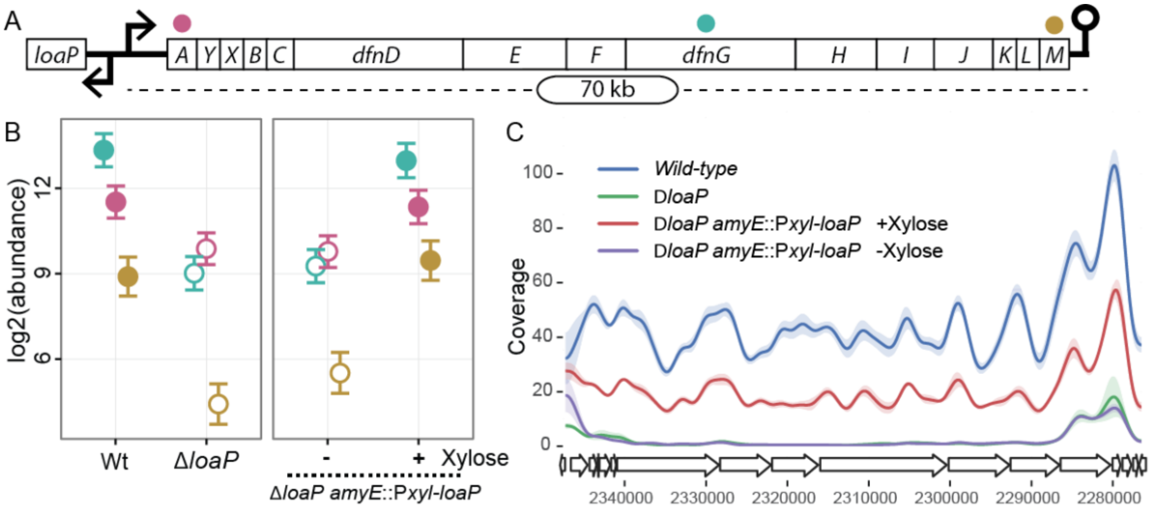


Figure II-1: LoaP is required for expression of the *B. velezensis* difficidin gene cluster.
(a) Schematic depiction of the dfn gene cluster, including the general location of *dfnA*, *dfnG* and *dfnM* amplicons used for quantification by qRT–PCR. (b) Normalized transcript abundance at the beginning, middle and end of the dfn operon (*dfnA*, *dfnG*, *dfnM*) as measured by qRT–PCR. Filled symbols represent samples with loaP expression and open symbols represent samples with no or minimal loaP expression. Colours correspond to amplicon locations in a, with *dfnA* in pink, *dfnG* in teal and *dfnM* in gold. Error bars represent Bayesian 95% highest posterior density estimates of mean expression. Data resulted from biological triplicate cultures with qPCR technical duplicates. WT, wild type.(c) RNA-seq coverage across the dfn gene cluster normalized with DESeq2 normalization factors. Traces represent coverage data smoothed with Gaussian smoothing with a bandwidth of 500 nt. Shading represents standard deviation from libraries from three independent cultures for each condition. *Reproduced from* (94).

To test its potential role in genetic regulation, *loaP* was replaced with a gene responsible for erythromycin resistance, resulting in strain JG091. Levels of *dfn* transcript were monitored by real-time qRT-PCR in early stationary phase. Specifically, *dfn* abundance was monitored at the beginning, middle, and end of the operon, for amplicons located within *dfnA*, *dfnG*, and *dfnM* respectively. Deletion of *loaP* resulted in a 3-fold drop in expression at *dfnA*, and a 20-fold reduction at *dfnG* and *dfnM* (Figure II-1B), suggesting that *loaP* affects *dfn* transcript abundance. To investigate the possibility that marker replacement of the *loaP* gene could have altered the local transcriptional landscape and deleteriously affected the *dfnA* promoter (P*dfnA*), a xylose-inducible copy of *loaP* (P*xyl-loaP*) was ectopically integrated into the genome of *B. velezensis FZB42* at the nonessential *amyE* locus. We integrated this plasmid into both wild-type *B. velezensis* and JG091 (*loaP*::*erm*) to produce strains JG098 and JG099 respectively. The transcript abundance of the *dfn* transcript was measured by qRT-PCR for these strains, in the absence and presence of 0.5% xylose. This revealed that the mRNA abundance of the *dfn* gene cluster measured at *dfnA*, *dfnG*, and *dfnM* is unchanged in the uninduced complementation strain (JG099) but is restored to wild-type levels when *loaP* is induced (Figure II-1B). Therefore, *loaP* strongly affects expression of the *dfn* transcript.

Induction of loaP elevates mRNA abundance across the full *dfn* operon

To examine *dfn* transcript abundance further, we performed whole-transcriptome RNA-seq on wild-type (FZB42), *loaP* deletion (JG091), and LoaP complementation (JG099) strains. We constructed transcriptome libraries from three independent cultures of each strain and subjected them to Illumina sequencing. Analysis of differential expression between the wild-type and Δ*loaP* showed a 14-fold decrease in *dfnA* transcript levels, which increased to a 30-fold drop in abundance towards the middle and end of the *dfn* operon (Fig II-1C). The majority of the observable decrease in transcript level occurs in the first 8-10 kilobases of the *dfn* transcript. Specifically, there is a rapid drop in transcript level at the very beginning of the operon to approximately 10% of wild-type for the first few kilobases, with a more gradual drop over the next few kilobases to 3-4% of wild-type for the majority of the operon. Interestingly, the amount of *dfn* transcript in the deletion strain increases slightly beginning at *dfnJ* and continuing to the end of the operon at *dfnM* implying the possible presence of an internal promoter near the end of the gene cluster.

Induction of *loaP* expression significantly increased expression across the entire length of the *dfn* gene cluster, restoring transcript levels to 35-50% of wild-type levels. The gradual drop in transcript levels within the first ten kilobases of the transcript was also eliminated upon complementation of *loaP* (Figure II-1C). Together, these data demonstrate that *loaP* had a dramatic effect on *dfn* mRNA abundance. Moreover, the pattern of changes in *dfn* transcript abundance hinted that *loaP* is likely to utilize a genetic mechanism distinct from traditional transcription factors, which, by affecting transcription initiation, traditionally affect mRNA evenly after the transcription start site.

LoaP promotes processive antitermination within the *dfn* mRNA leader region

Previously, investigators used high-throughput sequencing approaches to map transcription start sites across the *B. velezensis* FZB42 genome (116). Our inspection of these data identified a single transcription start site (TSS) for *dfnA* located significantly upstream of the *dfnA* coding sequence. 5′ leader regions that are longer than 100 nucleotides in length are unusual; most mRNA leader regions in *Bacillus* species are only long enough to permit translation initiation (~35 nts) (116, 117). Correspondingly, unusually long leader regions are typically involved in post-initiation regulation of their downstream genes. At 417 nucleotides, the *dfnA* leader region is exceptionally long (Figure II-2A), and we therefore hypothesized that it is involved in genetic regulation. No additional TSSs were annotated within the intergenic region upstream of *dfnA (116)*. Additionally, the *dfnA* TSS is positioned just downstream of a near-consensus SigA promoter sequence and is consistent with the pattern of cDNA reads as recorded by our RNA-seq analysis. Together, these data demonstrated that the *dfn* transcript is preceded by an unusually long leader.

RNA-seq coverage data revealed that abundance of the 5′ portion of the *dfnA* leader region is not significantly affected by expression of *loaP*, suggesting that LoaP altered *dfn* expression at a point after transcription initiation. In general, abundance within the *dfnA* leader decreased in the 5′→3′ direction, with a particularly acute and steep decrease near the midway point (94) (Figure S1). This drop in transcript abundance occurred within a portion of the leader that exhibited the ability to form a large stem-loop structure, followed by a poly-uridine tract (Figure II-2A). This suggested that the stem-loop element comprised a candidate for an intrinsic terminator site. Therefore, while interpretation of coverage profiles at high resolution can be complicated by the sequence-specific bias of the Illumina technology, particularly in regions containing inverted repeats (118), we hypothesized that the decrease in transcript abundance resulted from premature transcription termination. Indeed, this putative terminator corresponded to a site of permanent drop in coverage for *loaP*-deficient samples. In contrast, the moderate drop in coverage at this site for the wild-type sample was temporary and recovered shortly thereafter, which we hypothesize resulted from bias in Illumina library construction and sequencing due to the significant local secondary structure. To further test the transcript coverage patterns in this region, we performed qRT-PCR using amplicons targeted to regions immediately before and after the putative intrinsic terminator in the *dfnA* leader region. Using the geometric mean of three reference genes (*rpoB*, *gyrB*, and *dnaG*) as the normalization control, we calculated the proportion of cDNA fragments before and after the putative terminator for four independent cultures. Upon induction of *loaP*, cDNA abundance was significantly increased after the terminator, with a minimal change in upstream abundance (Figure II-2B). Moreover, examination of the RNA-seq data for the *loaP*-complemented strain revealed that the number of Illumina sequencing paired cDNA reads (average length 125 nts) spanning the putative terminator site was significantly increased upon induction of *loaP* (Figure II-2C). Together, these data suggest that LoaP promoted readthrough of a *dfnA* leader intrinsic terminator.

To examine whether LoaP antitermination determinants are derived from within the *dfnA* leader RNA, or are instead incorporated within the promoter region, we constructed promoter-replacement strains. Specifically, the native *dfnA* promoter sequence was replaced by a constitutively active promoter (P*const*) for strains that either contained or lacked the *dfnA* leader region. We quantified the transcript levels at *dfnA*, *dfnG*, and *dfnM* for these strains, in the presence and absence of xylose-mediated induction of *loaP* (Figure II-2D). With the leader region under control of P*const*, transcript level changes appeared to closely resemble wild-type, albeit with slightly higher basal transcript levels. In contrast, removal of the leader region resulted in complete loss of *loaP*-induced transcript levels. Therefore, determinants for LoaP regulation are contained within the *dfnA* leader region.

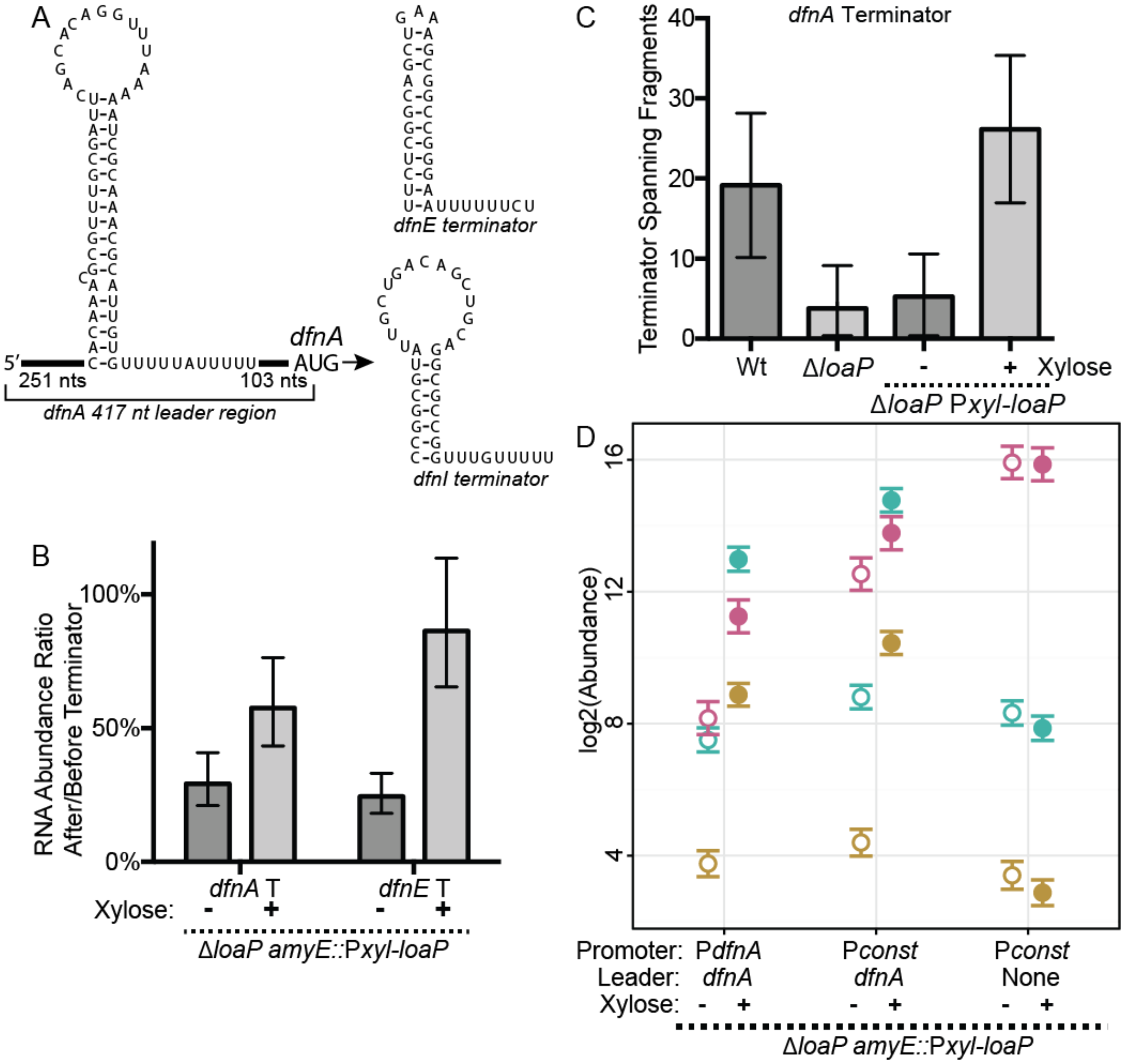


Figure II-2: LoaP promotes readthrough of intrinsic terminator sites.
The *dfnA* leader region contains determinants for LoaP-mediated processive antitermination. (a) Schematics of intrinsic terminator candidates identified within the dfnA leader region, or within coding sequences of *dfnE* or *dfnI*. (b) Estimated transcription termination efficiencies for putative intrinsic terminators (T) in the *dfnA* leader region and *dfnE* coding sequence. Efficiencies are calculated as the ratio of transcript abundance immediately before and after the terminator sequences measured by qRT–PCR. Error bars represent 95% highest posterior density of ratios calculated directly from posterior estimates of normalized transcript abundance. Experiments were performed with four independent cultures for each condition and qPCR technical duplicates. (c) Normalized count of RNA-seq read pairs spanning the termination site of the *dfnA* leader intrinsic terminator. Error bars represent standard deviation (n = 3). (d) Normalized transcript abundance at the beginning, middle and end of the dfn operon in strains where the *dfnA* promoter has been replaced by the constitutive promoter Pconst, with or without the *dfnA* leader region. Colours correspond to amplicon locations in Figure II-1a, with *dfnA* in pink, *dfnG* in teal and dfnM in gold. All strains contain a marker-replacement of *loaP* and an ectopic xylose-inducible *loaP*. Error bars represent Bayesian 95% highest posterior density estimates of mean expression. Data represents six biological replicates with technical duplicates. Open symbols represent uninduced cultures and filled symbols represent cultures with *loaP* induction. *Reproduced from* (94).

LoaP regulates a second polyketide synthase gene cluster

For a global analysis of gene expression changes, we performed differential expression analysis using DESeq2 between the wild-type and *loaP* deletion strains. Deletion of *loaP* resulted in significant (adjusted p-value < 0.01) differential expression of only 30 genes (Figure II-3A). Unexpectedly, with very few exceptions, every differentially expressed gene (p < 0.01) belonged to either the difficidin biosynthesis operon or a second polyketide biosynthesis gene cluster (Figure II-3B-C), located elsewhere in the genome. The latter operon (*mln*) encodes for biosynthesis of the polyketide antibiotic macrolactin (119). Every gene in the nine-gene macrolactin gene cluster was reduced on average between 4-fold and 12-fold in the *loaP*-deficient strain. The decreased expression of the *mln* gene cluster under *loaP*-deficient conditions was independently supported by qRT-PCR measurement of *mlnA, mlnD,* and *mlnI*, which were decreased between 5- and 13-fold, comparable to the transcriptomic data (Table S1).

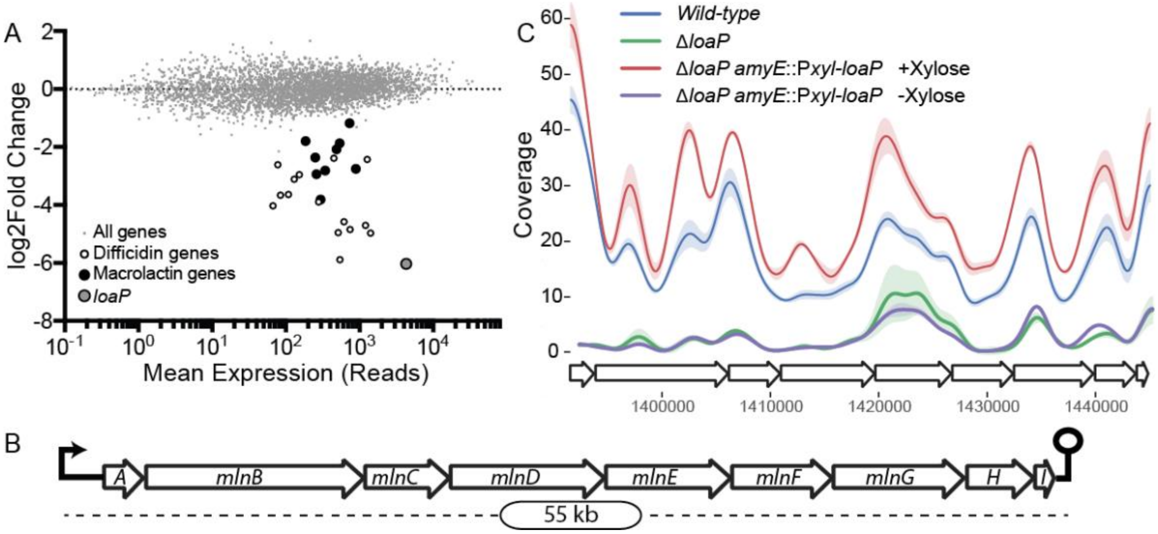


Figure II-3: LoaP expression affects transcription of both difficidin and macrolactin operons. (a) MA-plot showing mean expression and log-fold-changes for all genes between wild-type and ΔloaP strains from RNA-seq analysis. Large open points represent dfn synthesis genes and large filled points represent mln synthesis genes. Data represents the average expression of three (wild-type) and two (ΔloaP) replicates. (b) Schematic of the mln macrolactin synthesis operon. (c) RNA-seq coverage across the mln gene cluster normalized with DESeq2 normalization factors. Traces represent coverage data smoothed with Gaussian smoothing with a bandwidth of 500 nt. Shading represents standard deviation from libraries from three independent cultures for each condition. *Reproduced from* (94).

Differential expression analysis on the *loaP* complementation strain resulted in many differentially expressed genes (373), although most were annotated either specifically as xylose metabolism genes or with other carbohydrate metabolism functions (Table S1). Therefore, we conclude that most differentially expressed genes were altered from the use of xylose as an inducer molecule for controllable expression of *loaP*. However, the remaining differentially expressed genes agreed well with analysis of the *loaP* deletion strain. Of the 33 genes differentially expressed in the *loaP* deletion strain, 28 were also differentially expressed for the complementation strain upon induction of *loaP*, including the *mln* pathway. From these transcriptomic data, we conclude that *loaP* specifically affects transcript abundance of a regulon of *B. velezensis* antibiotic biosynthesis genes.

To confirm the specificity of the *loaP* regulon for the difficidin and macrolactin biosynthesis gene clusters, we performed HPLC analysis on extracts of *B. velezensis* culture supernatants and measured production of polyketides (difficidin, macrolactin, and bacillaene). The ∆*loaP* strain showed a specific and near-complete elimination of both difficidin and macrolactin, and as expected from the transcriptional data, had no effect on bacillaene production (Figure II-4A). Thus, the production of difficidin and macrolactin requires the LoaP function in a manner consistent with regulation of transcriptional elongation of the respective biosynthetic operons. Because *loaP* inserted at the *amyE* locus restored *dfn* and *mln* gene expression, we asked whether this complementing strain also restored metabolite production to *B. velezensis*. We measured quantitatively the output of both difficidin and macrolactin in the ∆*loaP*, *amyE::*Pxyl-*loaP* strain with and without xylose induction. Xylose induction of *loaP* restored metabolite production, confirming that the *loaP*-dependent changes in gene expression indeed correlate with antibiotic production (Figure II-4B).

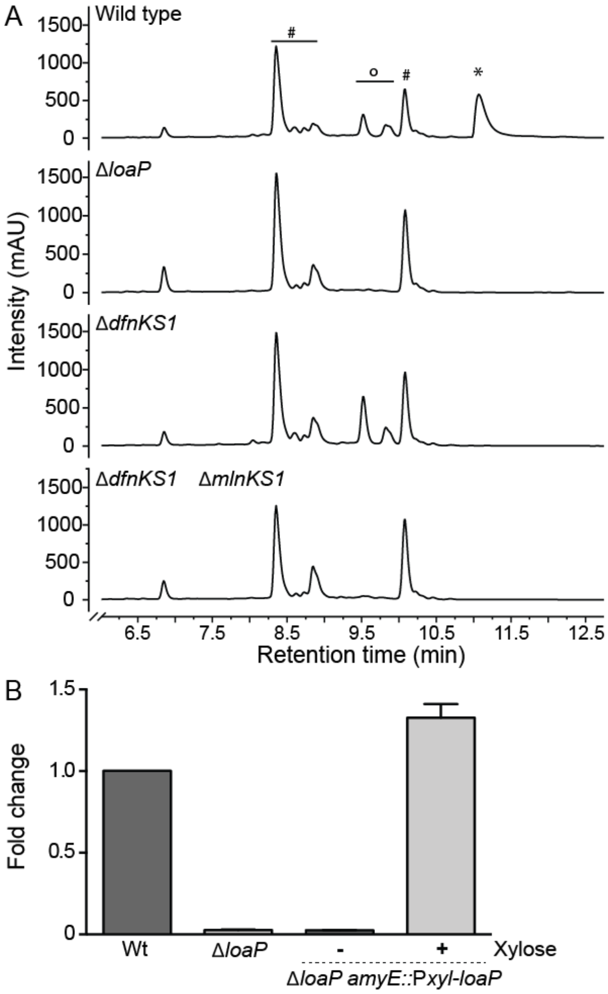


Figure II-4: LoaP-dependent production of difficidin and macrolactin.
(a) A comparison of ΔloaP to wild-type *B. velezensis* FZB42 production of difficidin, macrolactin and bacillaene by HPLC. Deletion of *loaP* specifically disrupts production of difficidin and macrolactin, while bacillaene production is maintained. HPLC peaks corresponding to difficidin (*), macrolactin (°) and bacillaene (#) are labelled for reference on the chromatographs. The *Δpks3KS1* strain is deficient in difficidin production and the *Δpks2KS1*, *Δpks3KS1* double mutant strain is deficient in both difficidin and macrolactin. The mutant strains serve as reference controls for specificity of HPLC peaks. Metabolites were detected at λ = 280 nm. Representative traces for each genotype are shown. mAU, milli-absorbance units. (b) Quantitative comparison of difficidin production by *B. velezensis* FZB42 strains, wild-type (WT), ΔloaP and ΔloaP, amyE::Pxyl-loaP (+ and −1% xylose). Relative production of difficidin was compared between the wild-type and mutant strains. Peak values were compared to bacillaene as a reference. Data and error bars represent the average and standard deviation of two biological replicates. *Experiments performed by Chengxi Zhang. Reproduced from* (94).

Demonstration of activity in a fluorescent reporter system and in *B. subtilis*.

Together, our data strongly suggested that *loaP* is responsible for readthrough of an intrinsic terminator site located within the *dfnA* leader region, and that this readthrough activity does not require the native promoter. These data are consistent with two scenarios: (1) *loaP* has evolved to exert regulatory control over a single transcription terminator, as a new example of a transcription attenuation mechanism, or, (2) *loaP* indiscriminately promotes readthrough of multiple termination sites, as a new example of a PA mechanism. To investigate the latter, we searched for intrinsic terminator sites in the *dfn* operon and identified several additional candidates, suggesting that LoaP might indeed assist readthrough of terminator sites downstream of the leader region. As a preliminary test of this theory, an analysis of a putative terminator site in *dfnE* (Figure II-2A-B) suggested that, indeed, readthrough of the *dfnE* terminator site is also dependent on expression of LoaP.

LoaP promotes antitermination of intrinsic terminators in a *yfp* reporter

To further differentiate regulatory scenarios for LoaP, and to divorce our analysis of LoaP activity from other, possibly complex effects on the full *dfn* transcript, a reporter construct was created with the *dfnA* leader region (including its intrinsic termination site) placed upstream of a *yfp* gene. Due to low fluorescence intensity of the reporter-containing strains, we initially quantified reporter activity using qRT-PCR on the *yfp* transcript. Analysis of *yfp* transcript by qRT-PCR showed that induction of *loaP* resulted in approximately 8-fold increase in *yfp* transcript abundance for this strain (Figure II-5A), providing key support for *loaP*-dependent readthrough of the *dfnA* leader terminator site.

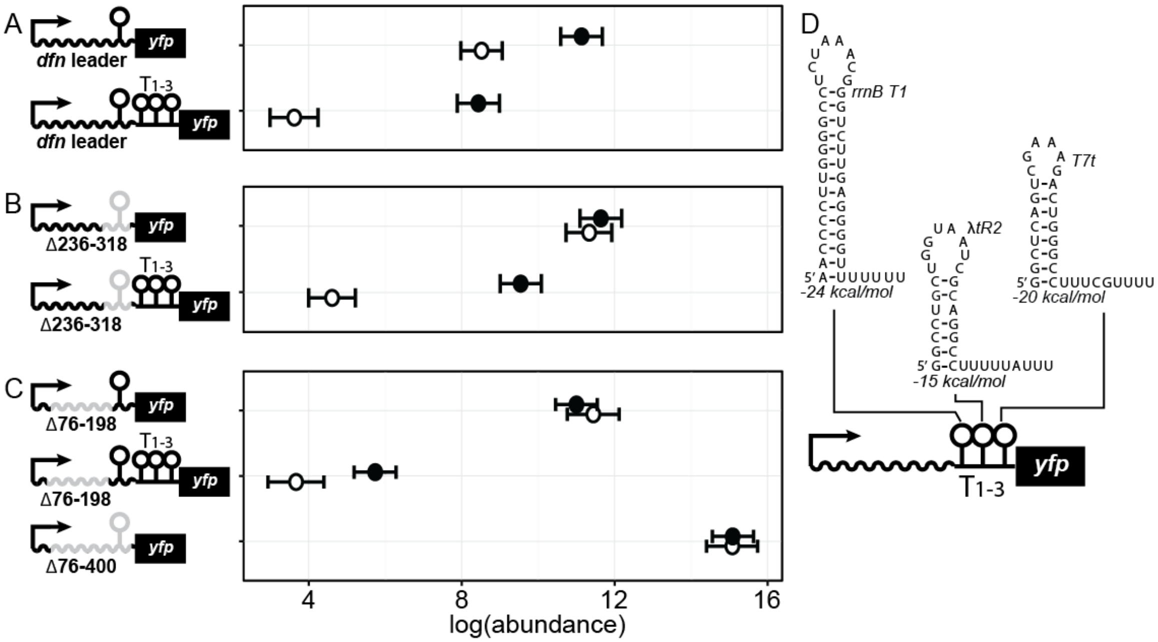


Figure II-5: LoaP mediates transcription antitermination in reporter constructs. Normalized transcript abundance of *yfp* mRNA measured by qRT–PCR. All strains contain a marker-replacement of loaP and an ectopic xylose-inducible *loaP* integrated into *amyE*. In addition, all strains contain a single copy of the P*dfnA* promoter transcriptionally fused to *yfp* with different modified *dfnA* leader regions. (a) In these constructs, a wild-type *dfnA* leader region was included upstream of *yfp*. A variant of this construct contained a *dfnA* leader region followed by an array of three tandem intrinsic terminators. (b) In these constructs, the region of the *dfnA* leader containing an intrinsic terminator was deleted, but they were otherwise identical to constructs in (a). (c) Deletions were introduced into the *dfnA* leader region of the *yfp* reporter fusions for constructs with and without the terminator array. In all rows, open symbols represent conditions without xylose induction and filled symbols represent conditions with 1% xylose induction of *loaP*. All conditions were measured with duplicate independent cultures and qPCR technical duplicates. Error bars in (a–c) represent Bayesian 95% highest posterior density estimates of mean expression. Data resulted from biological duplicate cultures with qPCR technical duplicates (d) Sequences of the tandem intrinsic terminators that were incorporated into some *yfp* reporter constructs, as indicated by schematics. *Reproduced from* (94).

In a separate construct, we then added an array of three validated but completely unrelated intrinsic termination sites (57), located downstream of the *dfnA* leader region but upstream of *yfp* (Figure II-5D). Again, induction of *loaP* resulted in significantly increased expression of the *yfp* reporter gene, despite the presence of the heterologous terminator sites. As a negative control for antitermination activity, a separate reporter construct lacked the *dfnA* intrinsic termination site (Figure II-5B). Without any intrinsic termination activity in the leader region, the transcript levels with and without *loaP* induction remained correspondingly high. Addition of the three heterologous termination sites to this terminator-less construct restored dependency on *loaP* for transcription of *yfp*. Finally, we tested a few large truncations within the *dfnA* leader region, which essentially eliminated terminator readthrough, indicating that determinants for LoaP-mediated PA are likely present within the 5′ portion of the *dfnA* leader region (Figure II-5C).

LoaP promotes antitermination activity on reporter transcripts in *B. subtilis*

While we established genetic tools for working in *B. velezensis* and utilized these in a variety of ways to demonstrate antitermination activity by LoaP, we wished to demonstrate activity in a different bacterium for two reasons. First, we wished to determine whether *B. velezensis* contained any other unique factors required for LoaP antitermination. Second, the *B. velezensis* reporter system we established has several limitations. While the nitrogen limitation competence-inducing protocol works with *B. velezensis*, the efficiency of transformation is several orders of magnitude lower than with laboratory strains of *B. subtilis*. Also, the fluorescence of our single-copy *dfnA*-YFP reporters is quite low relative to the background fluorescence of *B. velezensis*. Finally, *B. velezensis* is very proficient at forming bacterial biofilms and forms aggregates in liquid culture at later stages of growth, which interfere with single cell fluorescence quantification approaches such as flow cytometry or fluorescence microscopy.

Given the increased availability of integration vectors for *B. subtilis*, we decided to integrate the LoaP expression cassette and the fluorescent reporter at different loci. We transferred the P*xylA-loaP* expression cassette into pDG1664 for integration into the *thrC* locus. For the fluorescent reporter, we transferred the cassette back into the original pJG019 *B. subtilis* integration vector and integrated this into *B. subtilis* 168 at the *amyE* locus. With these two modifications, we generated a complete strain for testing the hypothesis that LoaP can induce antitermination in *B. subtilis*. Indeed, when this strain was grown with xylose we observed an increase of reporter activity by fluorescence microscopy relative to the uninduced strain (Figure II-6). This led us to believe that, much like lambda-N protein or RfaH, LoaP does not require any other specific factors to promote antitermination, beyond the standard bacterial transcription elongation factors shared between these bacteria. Additionally, observing activity on these reporters in *B. subtilis* allows us to use this more genetically accessible system for additional experiments.

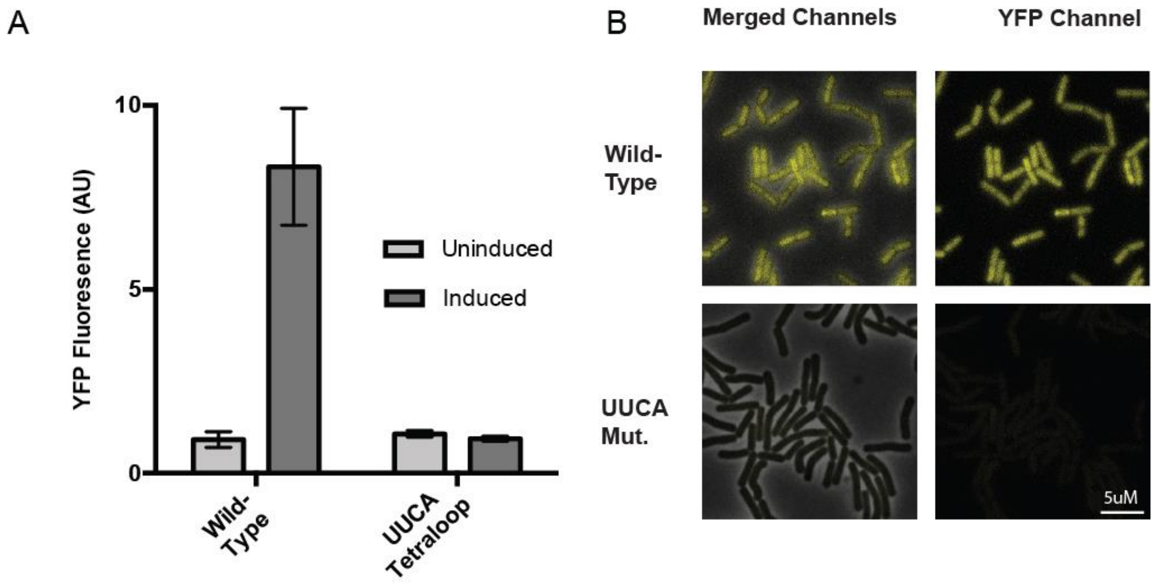


Figure II-6: LoaP antiterminations reporter constructs in *B. subtilis*.
(A) Quantified fluorescence microscopy data of *B. subtilis* strains carrying a xylose-inducible *loaP* and a *dfnA* leader-*yfp* reporter cassette. The *yfp* reporter contained a constitutive promoter upstream of the *dfnA* leader region, which was transcriptionally fused to a downstream *yfp* gene. Reporter constructs were created with or without a mutation converted the UUCG tetraloop sequence to UUCA, a mutation that is predicted to abolish proper RNA hairpin formation. These data demonstrate that *B. velezensis* LoaP antitermination can be recapitulated in the heterologous *B. subtilis* host. Data and error bars represent the mean fluorescence and 95% confidence interval (CI over means of each replicate) for all cells in three fields of view for each of three biological replicate cultures of each strain with and without induction. (B) Representative images of induced reporter strains quantified in (A) showing lack of reporter expression when the tetraloop sequence is mutated. *Reproduced from* (94).

LoaP is an RNA-binding protein with specificity for an RNA stem-loop in the leader regions of its regulon.

Antitermination activity in reporters requires a RNA stem-loop sequence

The loss of apparent antitermination activity in the *yfp* reporters that contained deletions of segments of the *dfnA* leader region indicated that there may be specific sequences required by LoaP. To more rigorously determine which region of the *dfnA* leader region is important for LoaP activity, we created additional reporter variants containing deletions of approximately 30-nt non-overlapping segments. These deletions spanned the entire *dfn* leader region up to the intrinsic terminator (Figure II-7A). Analysis of reporter activity by flow cytometry revealed that only deletion of segments 2 and 3 completely abolished apparent antitermination activity, while deletion of neighboring segments exhibited minor reductions in activity (Figure II-7B). This suggested that the primary determinants for LoaP activity are likely present in this region.

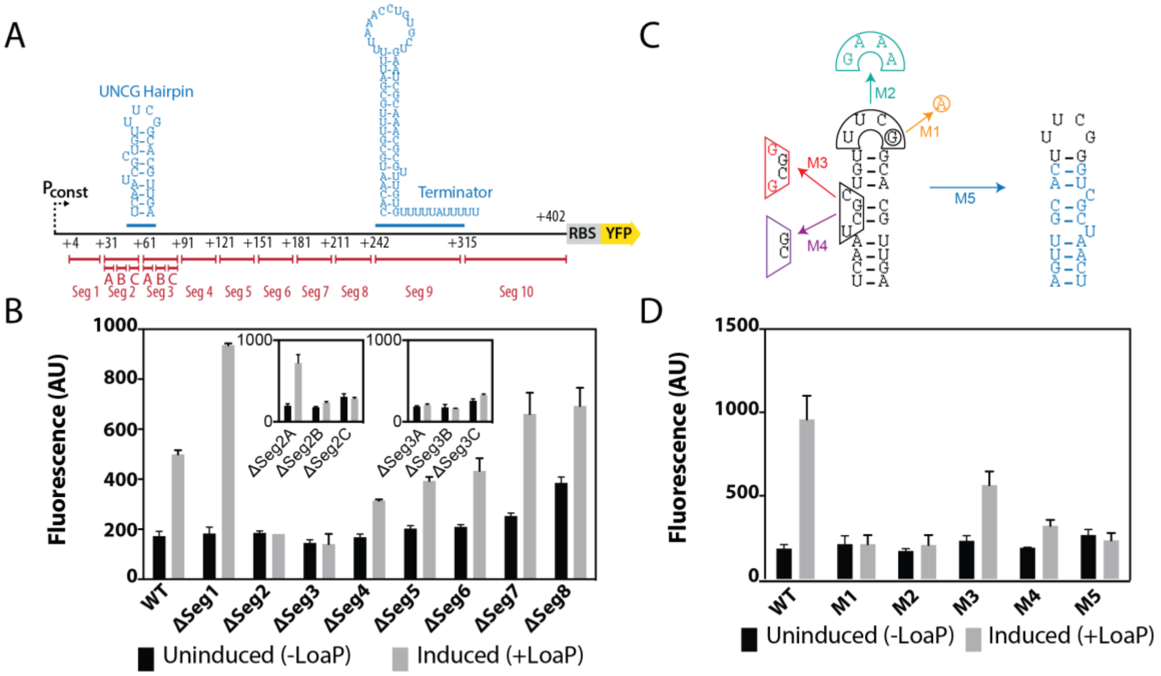


Figure II-7: A UNCG-type Hairpin in the *dfnA* leader region is required for antitermination in reporter assays.
(A) Schematic showing the promoter, 5' leader region, and beginning of the *yfp* gene for the *B.subtilis* LoaP antitermination reporters. Red markers illustrate regions deleted in individual mutant reporters. (B) Flow cytometry analysis of YFP fluorescence for strains containing mutant reporters with deletions of each 30 bp segment 1-8 and 10bp sub-segments of segments 2 and 3 (inset) as illustrated in (A). (C) Targeted mutations of the putative UUCG stem-loop sequence used in reporter assays. (D) Flow cytometry analysis of YFP fluorescence for strains containing mutant reporters with targeted hairpin mutations as illustrated in (D). *Assistance from Steven Klupt.*

The *dfn* and *mln* leader regions share a few common sequence features that may participate in LoaP regulation. One of these is an inverted repeat sequence (Figure I-3), located proximal to the 5′ terminus, which may correspond to a

UNCG-type tetraloop (30). This sequence is contained entirely within the segment 2 and 3 deletions. While the core NusG and KOW domains of bacterial NusG proteins are not known to bind RNA with any particular affinity (120), this led us to hypothesize that this RNA stem-loop may be somehow involved in LoaP recruitment to the nascent *dfn* or *mln* transcripts.

To test this hypothesis, we targeted the putative RNA stem-loop directly with mutations designed to test the requirement for this RNA structure and some of its particular sequence features (Figure II-7C). Some of these mutations are predicted to eliminate the potential for a UNCG-type loop structure (M1 and M2), while others substitute or remove the potentially bulged residues in the helix (M3 and M4). Yet other mutations swap the upstream and downstream halves of the helix (M5). We made targeted mutations to the *dfn-yfp* reporter in *B. subtilis* and quantified the apparent LoaP antitermination activity by flow cytometry (Figure II-7D). Both mutations that targeted the terminal loop abolished LoaP antitermination. Of these mutations, M1 is predicted to destabilize the structure by substituting a nucleotide essential for UNCG stem-loop folding and M2 substitutes the loop for a thermodynamically stable alternative, the GNRA loop sequence. This indicates to us that if this RNA stem-loop is directly involved in the LoaP mechanism, the loop structure itself is likely to be involved, rather than simply providing favorable folding energy. In contrast, the two mutations affecting the bulged helix residues have different effects. While M3, a mutation affecting only the identity of the bulged residues, retained antitermination activity, the M4 mutant, which removes these altogether, significantly reduced antitermination.

Given that the identity of the bulged residues could be changed but their deletion was not tolerated, we speculate that either the bulged residues are involved in direct interactions with the antitermination complex or, alternatively, they promote a specific change in helix geometry. M5, which mirrored the two strands of the helix, including moving the position of the bulged residues, eliminated antitermination activity. The combined results of these mutations suggest to us that this RNA stem-loop is likely to be fundamentally involved in the LoaP antitermination mechanism.

LoaP directly binds *dfn* leader region RNA *in vitro*

Upon observing the importance of the putative stem-loop RNA for LoaP activity *in vivo*, we decided to further investigate the relationship between LoaP the RNA motif using biochemical assays *in vitro*. Specifically, we chose to determine whether LoaP could associate directly to the RNA element. For this test, we employed *in vitro* equilibrium binding assays between purified LoaP protein and purified stem-loop RNA.

Initial attempts to purify LoaP were largely unsuccessful. Purification of hexahistidine-tagged LoaP from *E. coli* overexpression strains in various conditions resulted in severely low yields of LoaP from cell lysate soluble fractions (Figure II-8B). Attempts to re-nature LoaP after purification from inclusion bodies in denaturing conditions resulted in protein that precipitated quickly. Purification of MBP-tagged LoaP was more successful; however, cleavage of these fusion proteins with Factor Xa protease resulted in rapid precipitation of free LoaP protein (Figure II-8A). The best results were obtained after purification of a His_6-MBP-LoaP fusion protein by IMAC chromatography under high salt conditions followed by anion-exchange chromatography; these conditions significantly reduced the co-purification of LoaP with nucleic acids, as measured by the ratio of A_{260} and A_{280}. Eventually, extensive optimization of conditions revealed that cleavage by bdSENP1 (SUMO-like) protease(121) could yield sufficient quantities of full-length, non-fusion LoaP, when the protein sequence included the appropriate cleavable domain. Most binding data, however, was performed using MBP-LoaP fusion protein, which showed very similar binding activity but was available in higher quantity and with more favorable handling characteristics.

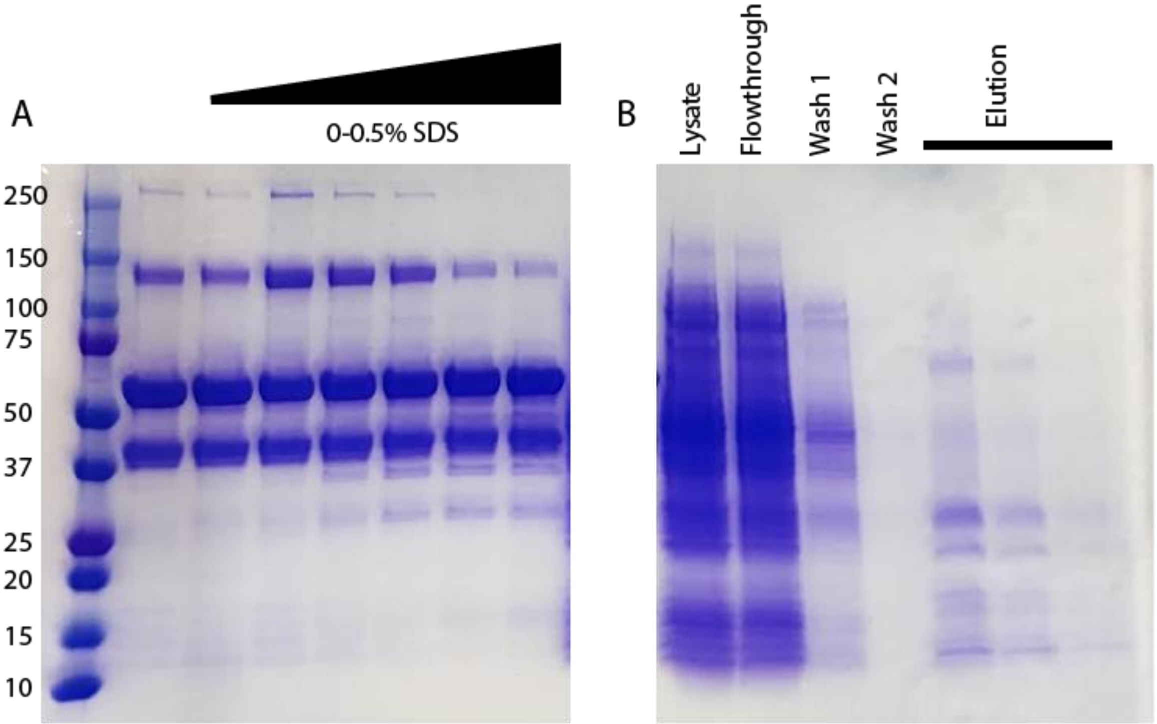


Figure II-8. LoaP has poor purification and solubility properties.
(A) Cleavage of purified MBP-LoaP fusion protein using Factor Xa protease results in both poor cleavage as well as insolubility of the cleaved LoaP peptide. 65 kDa band represents MBP-LoaP, 40 kDa band represents MBP, and 25 kDa band (faint) represents LoaP. (B) 6xHis-Loap (25 kDa)

expresses and purifies poorly. Only a small amount of LoaP protein is eluted and no strong band is visible in soluble cell lysate.

To test the potential interaction between MBP-LoaP and the *dfn* leader stem-loop, we chose to develop a fluorescence anisotropy equilibrium binding assay. We purchased synthetic *dfn* leader stem-loop RNA consisting of only the stem-loop sequence itself (Figure II-9A) with a covalently linked Cy3 fluorophore at the 3′ terminus. Additionally, we used a second, unrelated Cy3-labeled RNA of similar size (‘P1P2’), which contained two RNA stem-loop structures (discussed more in Chapter 3). Therefore, the unrelated RNA was used to estimate non-specific binding activity. Using a SpectraMax M5 plate reader, our initial tests of these RNA molecules showed that they exhibited sufficient anisotropy and acceptable variability at concentrations above 10 nM under our desired conditions (Figure II-9B). Equilibrium binding assays revealed a strong, specific interaction between MBP-LoaP and the *dfn* leader stem-loop with an affinity of approximately 50 nM, while the unrelated P1P2 RNA bound with a poorer affinity of approximately 500 nM (Figure II-9C). Binding of the *dfn* RNA to LoaP is not due to the presence of the MBP fusion protein, which alone did not exhibit detectable RNA binding activity (data not shown). As a second test of the putative LoaP-RNA interaction, MBP-LoaP and 5′-radiolabeled *dfn* leader RNA were mixed and assayed by differential radial capillary action of ligand assay (DRaCALA). This revealed an apparent binding affinity similar to the anisotropy measurements (Figure II-9D).

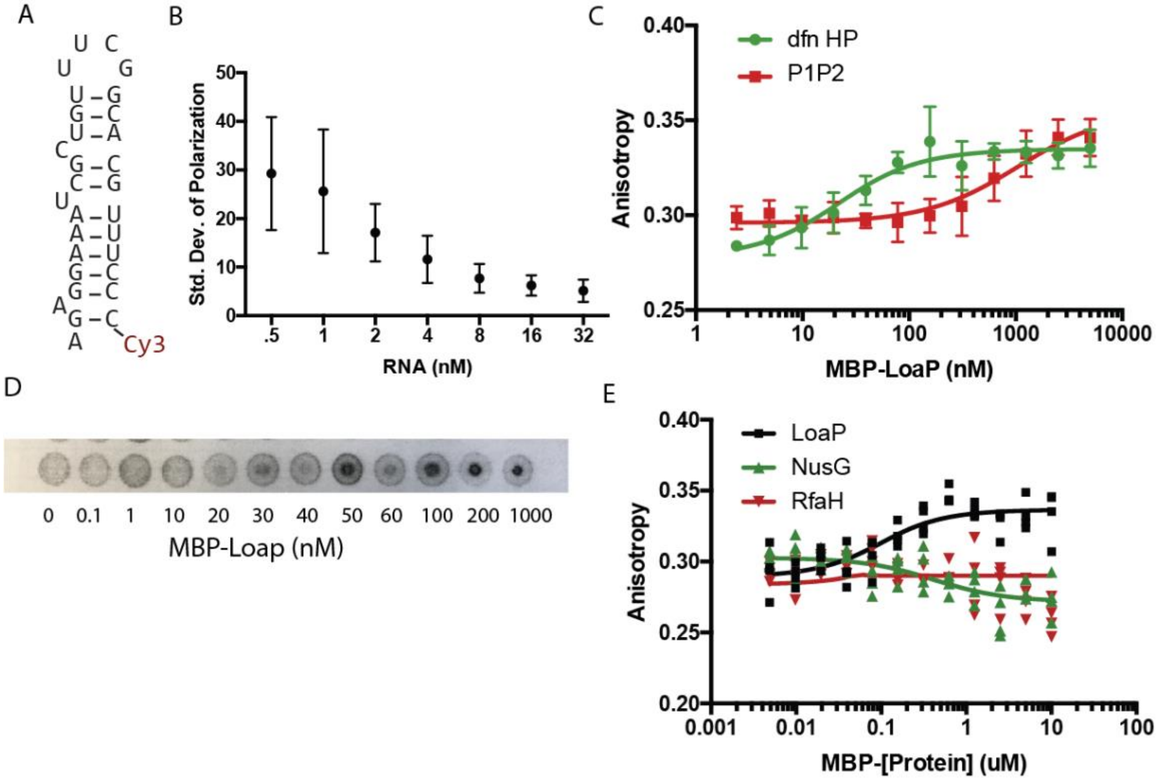


Figure II-9: Fluorescence anisotropy binding analysis reveals LoaP is a strong RNA-binding protein.

(A) Schematic of the synthetic RNA JRG646 labeled with a Cy3 fluorophore at the 3' terminus used for fluorescence anisotropy. (B) Titration of RNA JRG646 in binding buffer showing the standard deviation of fluorescence polarization across six replicates at different concentrations. (C) MBP-Loap binds the *dfn* hairpin RNA JRG646 with high affinity (approximately 40-50 nM apparent affinity), while binding to a similarly labeled RNA from the *nasF* leader region of *K. oxytoca* binds LoaP with a significantly lower apparent affinity. (D) Representative DRaCALA titration showing MBP-LoaP binding to the *dfnA* leader RNA. (E) Fluorescence anisotropy saturing binding curves for MBP-LoaP, MBP-NusG (*B. velezensis*), and MBP-RfaH (*E. coli*) with the Cy3-labeled JRG646 RNA. *DRaCALA assistance from Amy Elghondakly.*

Other investigators have tested whether NusG or RfaH can bind RNA and found no measurable interaction (20, 120). Correspondingly, it has been generally assumed that NusG family proteins do not associate with RNA and the observation that LoaP is an RNA-binding protein is therefore unexpected. However, these prior experiments were performed on different RNA sequences. To investigate whether the *dfn* hairpin RNA sequence might associate with NusG proteins in general, we performed binding assays between the Cy3-labeled RNA and either *B. velezensis* MBP-NusG or *E. coli* MBP-RfaH. Neither of these proteins demonstrated detectable binding activity with the *dfn* RNA (Figure II-9E). From the combined information from our *in vivo dfn-yfp* reporter assays and the *in vitro* equilibrium binding assays, we speculate that the RNA stem-loop shared between the *dfn* and *mln* leader regions might be integrally involved in the LoaP antitermination mechanism via direct binding with the LoaP protein.

RNA determinants for binding LoaP are similar to RNA determinants for antitermination

At this point, our experiments had revealed that specific sequence elements of the *dfn* hairpin were required for *in vivo* antitermination activity and that LoaP could bind this RNA element with moderately high affinity. Therefore, we sought to explore whether the mutations that affected antitermination activity *in vivo* also affected *in vitro* binding to LoaP. We hypothesized that changes to the RNA sequence that decreased antitermination activity *in vivo* would be also likely to exhibit lower binding affinity *in vitro*. To that end, we synthesized a panel of RNA sequences containing desired mutations (Figure II-10), including most of those mutations that were tested by the fluorescent *in vivo* reporters.

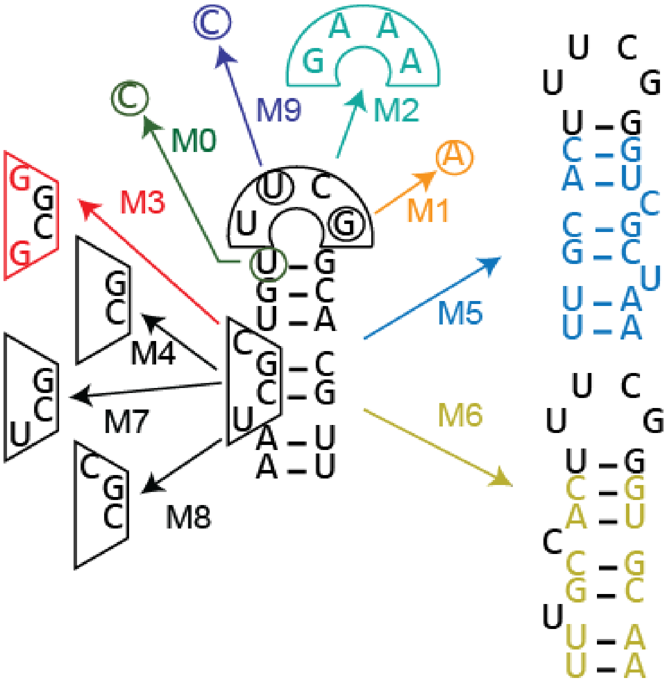


Figure II-10: Suite of mutants targeting aspects of the *dfnA* leader hairpin. Each mutation is illustrated, where any nucleotide with a different identity is shown in an altered color. Some mutants only contain deletions, and are show in the same box as the native sequence.

In experiments performed in our lab by Amr Elghondakly, binding of LoaP to the different RNA sequences was assessed by equilibrium competition assays, using a fixed concentration of radiolabeled wild-type *dfnA* hairpin RNA and LoaP protein, while titrating an increasing amount of unlabeled mutant RNA, and quantifying the LoaP-WT RNA complex by DRaCALA. The behavior of the mutant RNAs in this competition assay varied widely, with a few exhibiting competitive binding affinity comparable to wild-type RNA (Figure II-11A), some exhibiting reduced affinity (Figure II-11B), and several with no detectable competition (Figure II-11C). When fit to either a two-ligand one-site binding equation or a traditional one-site competitive inhibition equation, those mutants with strong binding affinity contained a much steeper slope than predicted from a one-site model, indicating that there may be additional sites of interaction that should be investigated. The relatively poor fit to the model for these titrations suggest that we should interpret this data qualitatively, comparing the ability to compete wild-type RNA binding against the titration curve for WT RNA itself.

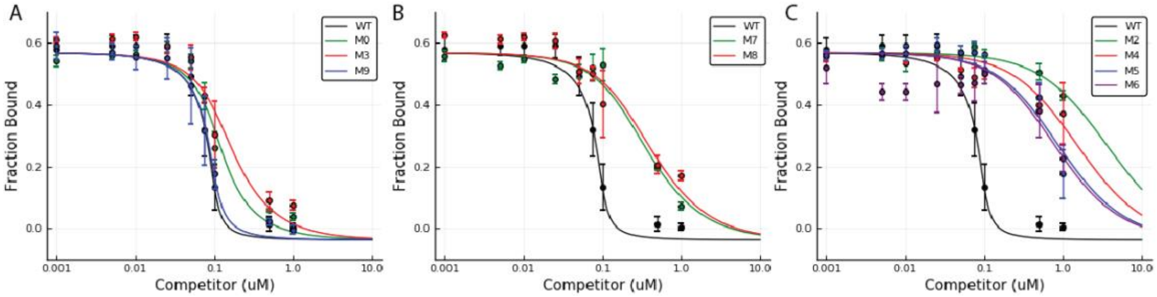


Figure II-11: Equilibrium competition assays show varied impacts of *dfnA* stem-loop mutations.

For each mutant, purified RNA was used to compete for binding to native *dfnA* stem-loop RNA (0.7 nM) binding to MBP-LoaP (100 nM) in DRaCALA assays. (A) Binding curves for native *dfnA* stem-loop RNA as well as mutants M0, M3, and M9 which showed the least reduction in competition. (B) Mutants M7 and M8 showed reduced competition, but reduced binding to less than 50% of initial at 1 uM. (C) The remaining mutants, M2, M4, M5, and M6 show reduced competition comparable to unrelated RNA (data not shown). All curves contain data from three or four replicate titrations. *Experiment performed by Amr Elghondakly*.

The three mutations that retained strong binding affinity, M0, M3, and M9, either introduce only minor changes to the loop region, either changing a U-G base pair into the canonical C-G (M0), changing the often non-conserved second position of the UNCG tetraloop (M9), or changing the identity of the bulged helix residues (M3). Of the four sequences shared between the *in vivo* and *in vitro* mutant experiments, only the M3 mutant retained significant antitermination activity, and is also the only mutant to exhibit near-WT binding affinity. Two mutations exhibited moderately reduced binding affinity, M7 and M8, both of which eliminate one of the two bulged helix residues. Four mutations that demonstrate greatly reduced binding affinity target a variety of sequence features: M2, M4, M5, and M6, three of which also eliminate antitermination activity *in vivo*. M2 changes the UNCG tetraloop to a GNRA, which would be predicted to result in a similarly-stable RNA stem-loop with a different structural conformation. M5 and M6 both make major changes to the stem of the RNA, although M6, which would contains reversed base pairs but retains the bulged residues in the same position, also greatly reduced binding affinity. Finally, M4, which simply removes the bulged nucleotides on the helix and would be predicted to stability the stem-loop, displayed negligible competition. In total, these mutations suggest that the bulged helix and the UNCG tetraloop are both important for binding LoaP and for antitermination activity *in vivo*.

Discussion

Bacteria appear to have evolved distinct PA mechanisms for different circumstances. Several types of PA systems appear to promote full synthesis of long transcripts (*e.g.,* phage transcripts), or for transcripts that might otherwise be susceptible to Rho termination (*e.g., rrn* operons). However, the functional range of PA mechanisms has not been fully examined. More antitermination systems undoubtedly await discovery, and, furthermore, these genetic regulatory mechanisms may prove to be essential for transcription of important pathways, such as those encoding for biosynthesis of specialized metabolites.

An abbreviated search for NusG paralogs proximal to specialized metabolite gene clusters led us to the discovery of *B. velezensis RBAM_022090*, which we renamed *loaP* (long operon-associated protein). Deletion of *B. velezensis loaP* dramatically affected transcript abundance of its neighboring gene cluster (*dfn*), which could be reestablished with ectopic expression of *loaP*. Surprisingly, with very few exceptions, every gene that was differentially expressed in response to the presence or absence of LoaP belonged to either the difficidin biosynthesis operon or the macrolactin PKS gene cluster (*mln*). In contrast, a third PKS cluster, encoding bacillaene production, was unaffected by changes in *loaP* expression, suggesting that *dfn* and *mln* share discrete determinants for LoaP regulation.

To investigate whether LoaP antitermination determinants were located within promoter regions, the *dfnA* promoter was swapped for a known constitutive variant. This resulted in no loss of *dfn* dependency for LoaP, demonstrating that LoaP determinants are positioned downstream of the promoter and transcription start site. This is consistent with a regulatory mechanism that targets transcription elongation. Relevant to this observation, both *dfn* and *mln* transcripts share a particularly striking feature in common—an unusually long 5′ leader region. Within each of the *dfn* and *mln* leader regions is a moderately strong, putative intrinsic termination site, which is bypassed upon induction of LoaP. Indeed, we find that LoaP promotes general readthrough of intrinsic terminators, including other intrinsic termination sites located within the *dfn* and *mln* operons, as well as unrelated intrinsic terminators introduced into reporter fusions.

It originally seemed unlikely that the direct mechanism of LoaP-mediated antitermination depends on an RNA-level binding interaction of LoaP. NusG family proteins are comprised of two domains separated by an unstructured linker (82, 85). The N-terminal portion is responsible for binding to RNAP, although this binding activity competes with sigma subunit for access to polymerase (85). When associated with RNAP, the NusG/Spt5 family proteins RfaH (20), core bacterial NusG (69), and Spt5 (72) have all been shown to interact with nontemplate DNA strand. This recognition of DNA sequence is particularly significant for RfaH, which is recruited to RNAP elongation complexes by a characteristic stretch of nontemplate DNA sequence called *ops*. Only operons that contain the *ops* sequence are targeted for genetic regulation by RfaH. Therefore, by extension, we might anticipate that there are likely to be DNA determinants for LoaP antitermination, conceptually similar to *ops* sequence, located somewhere within the *dfn* and *mln* 5′ leader regions.

The interaction between LoaP proteins and a conserved RNA stem-loop, at least in Bacillales species containing LoaP, introduces a new wrinkle into the understanding of NusG specialized paralog recruitment. It is very unlikely that this direct RNA interaction could replace the conserved non-template strand interaction. Although we have not been able to identify an additional conserved DNA element, the *ops* element that recruits RfaH is quite minimal, consisting of only a few specific base interactions in a defined position downstream of a pause-inducing dinucleotide (122, 123). Most intriguing, however, is the recent revelation that the *ops* element actually forms a small, 4-nt, non-template DNA loop inside of the open transcription bubble (78). The *ops* loop does have significantly different characteristics, forming only an abbreviated one base pair "stem” with a GGTA loop (78). This results in a thermodynamically unstable structure that is instead allowed to form during transcriptional pausing at the very strong *ops* pause, stabilized by interactions with RNAP and RfaH. However, while intriguing, the parallels between the LoaP-associated RNA stem-loops and the *ops* non-template DNA loop are likely coincidental.

Interestingly, the lambdoid bacteriophage antitermination factor, N, associates with a related but structural distinct stem-loop (*boxB*) that features a terminal GNRA tetraloop motif (124), or, for some lambdoid phages, an alternate sequence that mimics the GNRA fold (125). The *boxB* RNA element, in combination with an adjacent unstructured sequence (*boxA*), acts as loading region for N protein and bacterial elongation factors, respectively, and is important for antitermination of the lambda early gene transcript. By extension of this model, we speculate that the UNCG tetraloop found in *dfn* and *mln* leader regions may also serve as a loading site for elongation factors other than LoaP. In the case of lambda N antitermination, the elongation proteins bound to *boxB*/*boxA* RNA elements remain in a complex, associated with one another and with the transcription elongation machine, while the emerging nascent transcript continues to loop from the exit channel (18). It is possible that the LoaP antitermination complex exhibits a similar feature during transcription elongation of *dfn* and *mln* transcripts. We would expect, however, that a transcriptional pause site would still be required to allow the LoaP N-terminal domain to load onto RNAP. *B. subtilis* NusG is known to have increased interactions with degenerate T-rich sequence in the non-template DNA, increasing the lifetime of pauses displaying these sequences in the transcription bubble. One hypothesis is that LoaP may have two-step recognition requirements: first, to be physically recruited by interactions with the RNA hairpin, and second, to be loaded onto RNAP, perhaps requiring a less-conserved non-template strand sequence than RfaH.

After binding to RNAP via its N-terminal portion, the NusG C-terminal Kyprides-Onzonis-Woese domain (KOW) is responsible for association with other regulatory factors. One of these KOW-binding factors is ribosomal protein S10 (NusE). Association of RNAP-bound NusG (or RfaH) to ribosome-bound S10 results in improved coupling of transcription and translation machinery (85). A second KOW-binding factor is Rho, whose transcription termination activity is either enhanced or reduced upon interaction with NusG (82, 85). In order to inhibit Rho-mediated termination, RfaH outcompetes, and thereby excludes, the NusG:Rho complex from RNAP. However, Rho is dispensable in *B. subtilis* (82, 85). Moreover, LoaP is the first NusG/Spt5 family member shown to promote readthrough of intrinsic terminators. Therefore, it is unclear what relationship—if any—LoaP might have with Rho termination factor, or whether the KOW domain plays an important role in LoaP antitermination.

Bacillus velezensis loaP is shown to affect readthrough of intrinsic terminators located in an adjacent polyketide synthesis gene cluster. These data suggest that LoaP antitermination may function primarily on intrinsic termination, in contrast to the suppression of Rho-dependent termination known for RfaH (20, 94). However, while this might be due to the preference of Gram-positive bacteria for Rho-independent antitermination, it is also true that Rho termination has been insufficiently characterized in *B. velezensis* and other Firmicutes in general (126).

Therefore, the relationship between LoaP proteins and Rho termination is still unknown. While LoaP affects transcript abundance across the length of the targeted operon, it appears to require sequence elements located somewhere within the 5′ leader region. Interestingly, we found a small RNA stem with a UNCG tetraloop in the leader regions of both the difficidin and macrolactin operons in a sequence region required for antitermination. Moreover, this element is both required for LoaP antitermination activity in *in vivo* reporter assays and is capable of direct, high-affinity binding to LoaP itself. In total, LoaP exhibits significant differences from both NusG and RfaH, suggesting that subfamilies of NusG paralogous proteins may exhibit fundamental differences in the molecular mechanisms they employ.

Finally, the discovery of LoaP suggests that transcription elongation may be a broad point of regulatory control for secondary metabolite gene clusters in bacteria. As screening through bacterial secondary metabolites for those that display desired biomedical properties constitutes a primary pathway for the discovery of new drugs, it is of profound importance to study the PA mechanisms that might govern their expression patterns. Understanding the mechanisms and requirements of these PA mechanisms will have a significant impact on improving the discovery and production of new natural products from bacteria.

Methods

Construction of *Bacillus velezensis* FZB42 plasmids and strains

To construct a marker-replacement of *loaP*, a backbone plasmid derived from pUC19 was digested with BamHI/EcoRI into which a PCR product for the *B. velezensis loaP* region was subcloned via Gibson assembly (112). A PCR product containing an erythromycin resistance cassette was then inserted into the above plasmid using restriction-free cloning to construct plasmid pJG030. This plasmid was transformed into *B. velezensis FZB42* (obtained from Bacillus Genetic Stock Center, Columbus, Ohio) using a one-step transformation protocol modestly adapted from a *B. subtilis* transformation protocol, which resulted in construction of strain JG091 (113). Plasmid sequences were verified by Sanger sequencing and the genome sequence of JG091 was verified by analysis of Illumina RNA-seq data.

To construct a plasmid for inducible expression of *loaP*, the *B. subtilis* integration vector pDG1662 (obtained from Bacillus Genetic Stock Center, Columbus, Ohio) was modified for integration into *B. velezensis.* Specifically, the *B. subtilis amyE* homology arms were replaced with the corresponding *amyE* homology arms from *B. velezensis* by Gibson assembly of *amyE* homology PCR products with pDG1662 backbone PCRs, resulting in plasmid pJG031. pJG031 will be submitted to the Bacillus Genetic Stock Center for distribution upon request. A *B. velezensis* region encompassing *loaP* was then PCR-amplified and subcloned into a NheI/BamHI digestion of plasmid pSWEET-III via Gibson assembly. The resulting plasmid was digested with BamHI/HindIII, thereby releasing a restriction fragment containing the xylose-inducible promoter region followed by *loaP*, which was subcloned into pJG031, resulting in construction of pJG032 (127). This plasmid was integrated into wild-type *B. velezensis* and JG091 for construction of overexpression and complementation strains.

Promoter replacement strains were constructed using variants of plasmid pJG030, which was designed for marker replacement of *loaP* (*loaP*::*erm*). Specifically, in plasmid pJG105, the promoter region for *dfnA* was replaced by a semisynthetic constitutive promoter, P*const*. Similarly, in plasmid pJG102, the *dfnA* leader region was included downstream of the P*const* promoter. These plasmids were integrated into strain JG098 (containing *amy*::P*xyl*-*loaP*). All plasmid sequences are available as described in (94).

For construction of *yfp* reporter plasmids, we modified a base *B. subtilis* plasmid, pJG019, which contains a semisynthetic promoter that drives constitutive expression of *yfp*. To construct *yfp* reporter strains the *dfnA* promoter and leader region were subcloned upstream of *yfp* sequence to give an intermediate plasmid. The *dfnA*-*yfp* reporter construct was then varied by mutagenesis. For example, some of the modifications included removal or addition of terminator sequences, which was accomplished by Q5 Site-Directed Mutagenesis (NEB). All *yfp* reporter cassettes were subcloned into plasmid pJG031 to create combination *yfp* reporter/inducible *loaP* plasmids. These plasmids were transformed into strain JG091 to generate reporter strains with inducible *loaP* complementation.

For experiments in *B. subtilis*, the region of the reporter plasmid containing the *loaP* expression and the *dfn-yfp* reporter cassettes was subcloned into plasmid pDG1662 and transformed into *B. subtilis* 168 for integration into *amyE*.

Full sequences of all plasmids are available on GenBank and the plasmids used to construct each *B. velezensis* strain are detailed in (94). All plasmids were verified by Sanger sequencing of the inserted region.

Extraction of total RNA for qRT-PCR and Illumina sequencing

Total RNA was extracted from *B. velezensis* cultures by bead-beating using 300 micron acid-washed glass beads, followed by extraction with TRI Reagent RT (MRC, Inc.) according to manufacturer protocol. Total RNA was treated with RQ1 RNase-Free DNase I (Promega), or PerfeCTa DNase I (Quanta Bio), re-purified using Zymo Research Direct-Zol columns and the overall integrity of the extracted RNA was assessed by agarose gel electrophoresis and concentration by Nanodrop (Thermo Scientific).

Quantification of transcript abundance by qRT-PCR

DNA-depleted total RNA was converted to cDNA with Quanta Bio qScript cDNA Supermix. The cDNA was diluted in TE buffer and added to qPCR reactions made with Quanta Bio PerfeCTa SYBR Green FastMix. All qPCR reactions were prepared for 18 µL volumes in 96- or 384-well plates, and analyzed using a Roche Lightcycler 480 with the recommended three-step fast cycle. All oligonucleotides used for each amplicon are listed in (94). All experiments utilized negative RNA-only controls for DNA contamination using reference gene amplicons. C_q values and amplification efficiencies were determined using LinRegPCR (128). Relative quantification analysis and statistics were performed using the MCMC.qpcr R library using efficiency and C_q values from LinRegPCR (129). Relative quantification was performed relative to three reference genes using the soft-normalization approach in MCMC.qpcr and significance testing was performed using two-sided MCMC posterior sampling in MCMC.qpcr (Bayesian p-values). Mean expression values are reported with per-condition 95% posterior credible intervals representing all sources of between-sample variability. All experiments used at minimum three biological replicates to achieve power sufficient to control false-negative rate to 5% at an effect size of less than 2-fold between samples, given an estimate of the average between-sample variance, except for the *yfp* reporter samples which targeted an effect size of 4-fold. Diagnostic plots from MCMC.qpcr were used to assess model assumptions of normality, homoscedasticity, and linearity. R scripts used for analysis of data are available online (https://github.com/jgoodson/LoaP-2016).

Construction, sequencing and analysis of Illumina libraries

We constructed Illumina transcriptome libraries from extracted RNA of three independent cultures of each strain using Illumina ScriptSeq. We subjected these to paired-end 75-bp sequencing on a NextSeq 500 and obtained high-quality sequence for all libraries, with the exception that one library of the JG091 marker-replacement strain did not provide sufficient reads for analysis and was excluded from further analysis. FASTQ reads were quality-filtered using fastq-mcf and aligned to the FZB42 reference genome using bwa-mem algorithm in BWA (130, 131). Normalization and differential expression analysis and statistics were performed using the DEseq2 R library (132). Significance testing was performed using multiple-testing adjusted Wald test p-values in DEseq 2. Coverage plots with standard deviations were constructed using custom Python scripts using the normalization factors calculated by DEseq2. Scripts used are available online (https://github.com/jgoodson/LoaP-2016). Sequencing data is available in NCBI SRA (BioProject PRJNA327241).

Extraction and detection of difficidin, macrolactin, and bacillaene

To investigate the effects of *loaP* mutant, overnight cultures of *B. velezensis* FZB42 strains were diluted to an optical density at 600 nm (OD_{600}) of 0.08, growing in 25 mL Landy medium (133) for 8 hours as described (134). The supernatant from each culture was harvested. A 25 ml volume of each culture was centrifuged for 30 min at 11,000 rpm and loaded on an SPE column (3M, Empore, C18-SD). After loading, the columns were washed once with dH_2O and once with 20% methanol. Samples were eluted using 2 ml 100% methanol followed by 1 ml 100% ethanol. The eluates were dried in a rotary evaporator and samples re-dissolved in 100 µl 90% methanol.

Metabolite production was performed using High Performance Liquid Chromatography (HPLC) with Agilent 1200 device. A 10 µl volume of each sample was injected onto a ZORBAX Eclipse Plus C18 column (4.6×100mm, 5µm) (Agilent). The run was performed with a flow rate of 1.0 ml/min at 30°C. Samples were eluted with a gradient of 20% CH_3CN and 80% of 0.1%, v/v HCOOH, which reached 95% CH_3CN and 5% of 0.1%, v/v HCOOH after 12 min. The 95% CH_3CN – 5%HCOOH step was held for further 5 min. The column was equilibrated with 20% CH_3CN – 80%HCOOH for 2 min. Difficidin and macrolactin peaks were detected at l 280 nm as previous reported (134). We confirmed the specificity of difficidin, macrolactin and bacillaene peaks in the HPLC chromatographs by comparison to samples from *B. velezensis* ∆*pks3KS1* (abolished difficidin biosynthesis) and ∆*pks2KS1*, ∆*pks3KS1* (abolished difficidin and macrolactin biosynthesis) strain and by LC-MS/MS.

(Metabolite extractions and analyses performed by collaborators in Paul Straight's laboratory at Texas A&M)

Quantification of fluorescence reporters by flow cytometry

For each strain used in each experiment, two 5 mL cultures were inoculated from 100 µL of overnight culture of an approximate OD_{600} of 2.5 into LB broth containing the appropriate antibiotic. To each pair of independent cultures, 100 µL of 25% w/v xylose (for a final concentration of 0.5%) was added to one culture to induce expression of LoaP, while 100 µL of sterile water was added to the other. The cultures were incubated shaking at 37 C at 250 rpm for 3 hours, reaching mid-exponential phase, and were then pelleted by centrifugation at 5700 x G and washed once in 5 mL phosphate-buffered saline (PBS). The pellets were then resuspended and diluted in PBS to an OD_{600} of 0.05 and quantified by flow cytometry on a BD FACSCanto instrument with excitation at 488 nm and detection at 535 nm.

(Experimental approach partially performed by Steven Klupt)

Protein expression and purification

Hexa-histidine- and MBP-tagged LoaP (from plasmid pAmr003) was cultured in 2xYT and expression induced at A_{260}= 0.5 with 1mM IPTG at room temperature overnight. The cell pellet was resuspended in Resuspension buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 5% glycerol, 1 mM PMSF, 2 U DNase). Cells were lysed by incubation with 0.5 mg/mL lysozyme on ice for 30 minutes followed by sonication. Crude lysate was clarified by centrifugation at 12,000 x G. Protein was purified by loading the lysate on a Ni-NTA (HisTrap HP) column with an AKTA FPLC. Bound protein was washed with 10 column volumes of wash buffer (50 mM Tris-HCl pH 8.0, 150 mM - 1M NaCl, 5% glycerol, 40 mM imidazole) and eluted with 50 mM Tris-HCl pH 8.0, 150 mM, 5% glycerol, 250 mM imidazole. Sample was buffer exchanged into 50 mM Tris-HCl pH 8.0, 150 mM, 5% glycerol either by repeated dialysis or with a HiPrep 26/10 Desalting column. Protein was assayed for purity by SDS-PAGE and staining with Coomassie solution.

Equilibrium binding assays

Fluorescence anisotropy

Purified protein was diluted to 40 mM with Binding Buffer (50 mM HEPES pH 7.5, 100 mM KCl, 1 mM $MgCl_2$). Generally, 2-fold serial dilutions were performed on the purified protein in binding buffer. Each of the diluted protein samples were added to 40-100 nM Cy3-labeled RNA in a 1:1 (v:v) ratio and mixed. Each mixture was aliquoted into 4 wells of a 384-well opaque-bottom microplate and incubated in the dark at room temperature for 30 minutes to reach equilibrium. The fluorescence polarization was then read in a Molecular Devices Spectramax M5 plate reader at 535 nM excitation and 580 nM emission. Data was fit using GraphPad Prism 6 using the quadratic steady-state equilibrium binding equation.

Differential radial capillary action of ligand assay (DRaCALA)

RNA transcripts were transcribed and radiolabeled with γ-^{32}P ATP. Radiolabeled RNA (~300 fmol) was incubated with increasing concentrations of protein in a 20 μl reaction containing 50 mM HEPES pH 7.5, 75 mM sodium chloride, 1 mM MgCl2 for 30 minutes at 25°C. Approximately 2 μL of each sample was spotted on a nitrocellulose membrane using a slotted pin-tool and allowed to air dry for 30 min. Storage phosphor screens were exposed with the nitrocellulose membrane and visualized using a FLA5000 phosphorimager.

(Experimental approach performed by Amr Elghondakly and Christopher Zhang)

Chapter 3: Characterization of the ANTAR Family of RNA-binding Antitermination Factors

Introduction

A prominent example of post-initiation control of gene expression that is widely used by bacteria is that of riboswitches, which oftentimes control transcription attenuation in a signal-dependent manner. Their initial discoveries, however, have been significantly aided by the extensive conservation of their sequences and secondary structures(12, 135). This level of sequence conservation is not observed for many other types of transcription elongation regulatory strategies, a limitation that may have slowed discovery of the latter. How, then, may other transcription elongation-based regulatory strategies be systematically discovered if experimentalists cannot rely primarily on bioinformatics searches of highly conserved regulatory RNAs? And what kinds of transcription elongation regulatory mechanisms have not yet been found? One mechanism by which this can occur, transcription attenuation, is through the signal-responsive control of an antiterminator, which is a structural element that is mutually exclusive with respect to formation of a terminator hairpin. The signals that influence the interconversion of terminator and antiterminator structures can vary, depending on the different regulatory RNAs. Multiple classes of RNA-binding proteins have been discovered to control transcription attenuation. For these systems, association of the appropriate RNA-binding protein influences whether terminator or antiterminator elements are formed. For example, certain members of the

BglG/SacY protein family contain the PTS regulation domain, which is an RNA-binding domain that associates with a characteristic antiterminator element that overlaps a mutually exclusive, adjacent terminator site. Phosphorylation of the PTS domain by the appropriate carbohydrate transport system controls the RNA-binding activity, thereby coupling signal responsiveness to direct stabilization of the antiterminator structure. In contrast, the *trp* RNA-binding attenuation protein (TRAP) associates with a tandem series of triplet sequences in order to prevent formation of a default antiterminator element, thereby permitting formation of an alternate intrinsic terminator structure(136–138).

Another important family of proteins with RNA-binding activity contains the “AmiR and NasR Transcriptional Antiterminator Regulator” domain (ANTAR) (139). The ANTAR domain is composed of three helices with five highly conserved residues (three alanines, one alanine/serine and one aromatic residue) that are exposed in the three-helical structure (140). Sequence homology based searches have predicted more than 7,000 occurrences of the ANTAR domain, widely distributed across at least 3,300 bacterial species (Figure III-1; http://pfam.sanger. ac.uk/; Pfam: PF03861). ANTAR-containing proteins typically occur as multi-domain proteins. There are several prominent subclasses, including those containing an N-terminal pseudo-receiver domain, characterized by the AmiR regulator of *Pseudomonas aeruginosa (140)*, those containing a two-component system response-regulator domain, represented by the EutV response regulator of *Enterococcus faecalis (141)*, and those containing the NIT nitrate-responsive sensing domain, typified by the *Klebsiella oxytoca* nitrate-responsive regulator

NasR (142). Additionally, many bacteria possess ANTAR-domain containing proteins with a N-terminal GAF or PAS domain, although no examples of this class have yet been characterized.

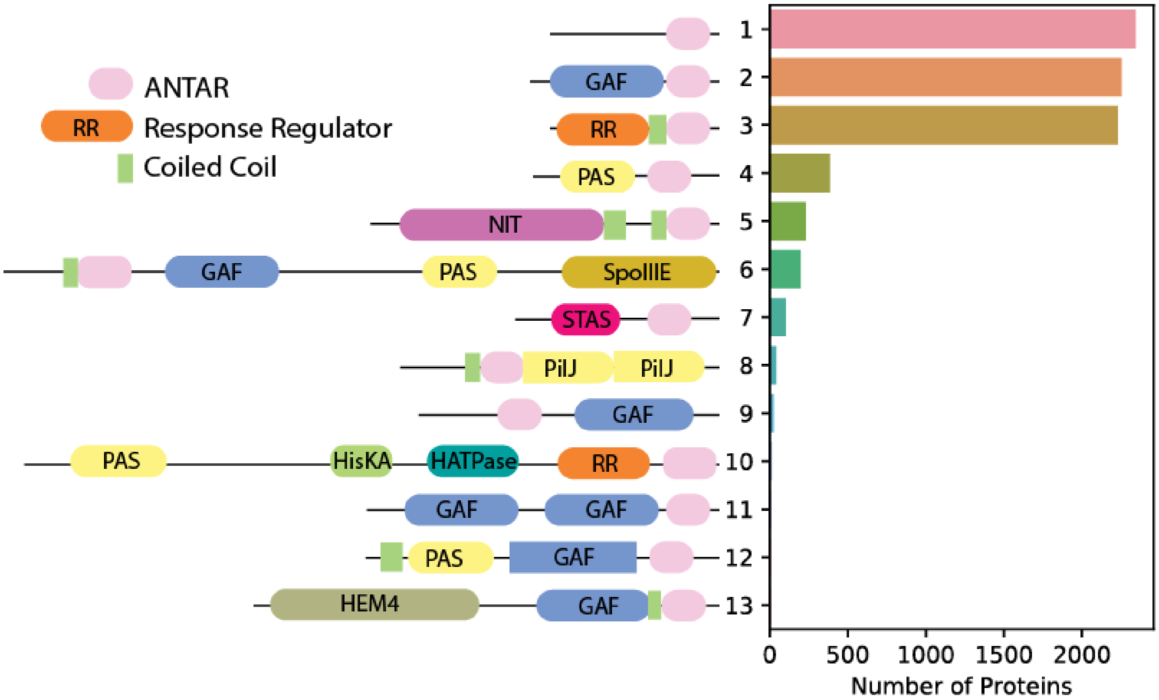


Figure III-1: Distribution of ANTAR-containing proteins according to their domain organization (Pfam: PF03861).

The bar graph shows the number of ANTAR proteins for each of XX domain architectures that have been identified. The latter are schematically represented to the right of the bar graph. Only domains patterns containing 3 or more unique sequences were included. Additionally, domain patterns containing only minor variations such as a different Pfam version of a similar domain or containing non-significant matches for individual domains in larger patterns were merged into a single pattern.

One class of ANTAR regulators, typified by AmiR, may regulate gene-expression via interactions with a modulator protein, which itself may possess signal-sensing function. *Pseudomonas aeruginosa* AmiR dimerizes upon binding two molecules of its negative regulator AmiC. Under inducing conditions AmiC binds one of a variety of small amide compounds, allowing association of AmiR with the 5′ leader of the appropriate target mRNA. This had been hypothesized to prevent formation of an intrinsic terminator, however, the molecular mechanism of antitermination had yet to be revealed (143, 144). The ANTAR domain also occurs in combination with a diverse set of signal-sensing domains (Figure III-1), with the most prevalent being the NIT domain. NasR, a protein with a nitrate and nitrite sensing NIT domain fused to an ANTAR domain, regulates the *nasFEDCBA* operon in *Klebsiella* species, which is required for nitrogen assimilation(142). In the presence of nitrate, NasR is activated and binds to the 5′ leader region of the nascent nasF transcript (145). Association of NasR inhibits formation of a transcription terminator within the 5′ leader region, thereby allowing synthesis of the downstream *nas* operon. Like the AmiR system, the molecular mechanism of antitermination had not been identified. In fact, it had been speculated that the mechanism might not even involve formation of a specific antiterminator structure, in contrast to the BglG/SacY family of antiterminators (146).

The largest individual class (nearly 50%) of ANTAR-containing proteins is comprised of proteins that are part of bacterial two-component regulatory systems (TCS). TCS consist of two proteins, the first of which is a sensor histidine kinase where a signal triggers autophosphorylation. The histidine kinase subsequently transfers the phosphoryl group to a receiver domain of the partner response regulator protein, usually triggering the regulatory activity of the effector domain of the response regulator (147, 148). These effector domains may affect gene regulation through a variety of mechanisms, whether transcriptional regulation through DNA-binding, promotion of protein-protein interactions or allosteric effects on enzyme activity (149). In contrast, ANTAR-containing response regulator proteins, such as *Enterococcus faecalis* EutV, regulate gene expression via post-initiation RNA-binding mechanisms. Until recently, this class of response regulators was the least understood, despite the frequent presence in a wide variety of bacterial genomes.

EutV, a representative of ANTAR-containing response regulators, was discovered to regulate the ethanolamine utilization operon (*eut*) in *E. faecalis* and this mode of regulation appears to be conserved in many Firmicutes that contain *eut* operons (141, 150, 151). For *E. faecalis*, the corresponding sensor kinase, EutW, undergoes autophosphorylation in response to ethanolamine and the phosphoryl group is transferred to EutV (141, 151). Phosphorylated EutV disrupts terminator sites located just upstream of each of the genes *eutP*, *eutG*, *eutS* and *eutA and* its association activates downstream gene expression (141, 150–152). These locations within the *eut* operon were previously found to share a common 13-nucleotide sequence (AGCAANGRRGCUY) overlapping the upstream strand of their corresponding intrinsic terminator elements. These sites were previously proposed to serve as part of the recognition sequence for ANTAR-based regulators in order to promote antitermination and allow production of the downstream transcript (151).

Ultimately, we presented evidence that a novel RNA motif comprises a specific antiterminator structure containing the full determinants for recognition by the

EutV ANTAR domain (152). This structure consists of a pair of small stem-loops, one of which contains the previously identified 13-nucleotide sequence (Figure III-2). Importantly, the same RNA motif could be identified for the other ANTAR-based regulatory systems that have been studied (AmiR and NasR), suggesting that it is likely to constitute the general recognition element of ANTAR-based regulatory proteins. Recognition of RNA by EutV relies on a combination of structure and primary sequence determinants (152). Specifically, certain residues within the hexanucleotide terminal loops share primary sequence conservation, particularly at the first and fourth positions, and are important for binding.

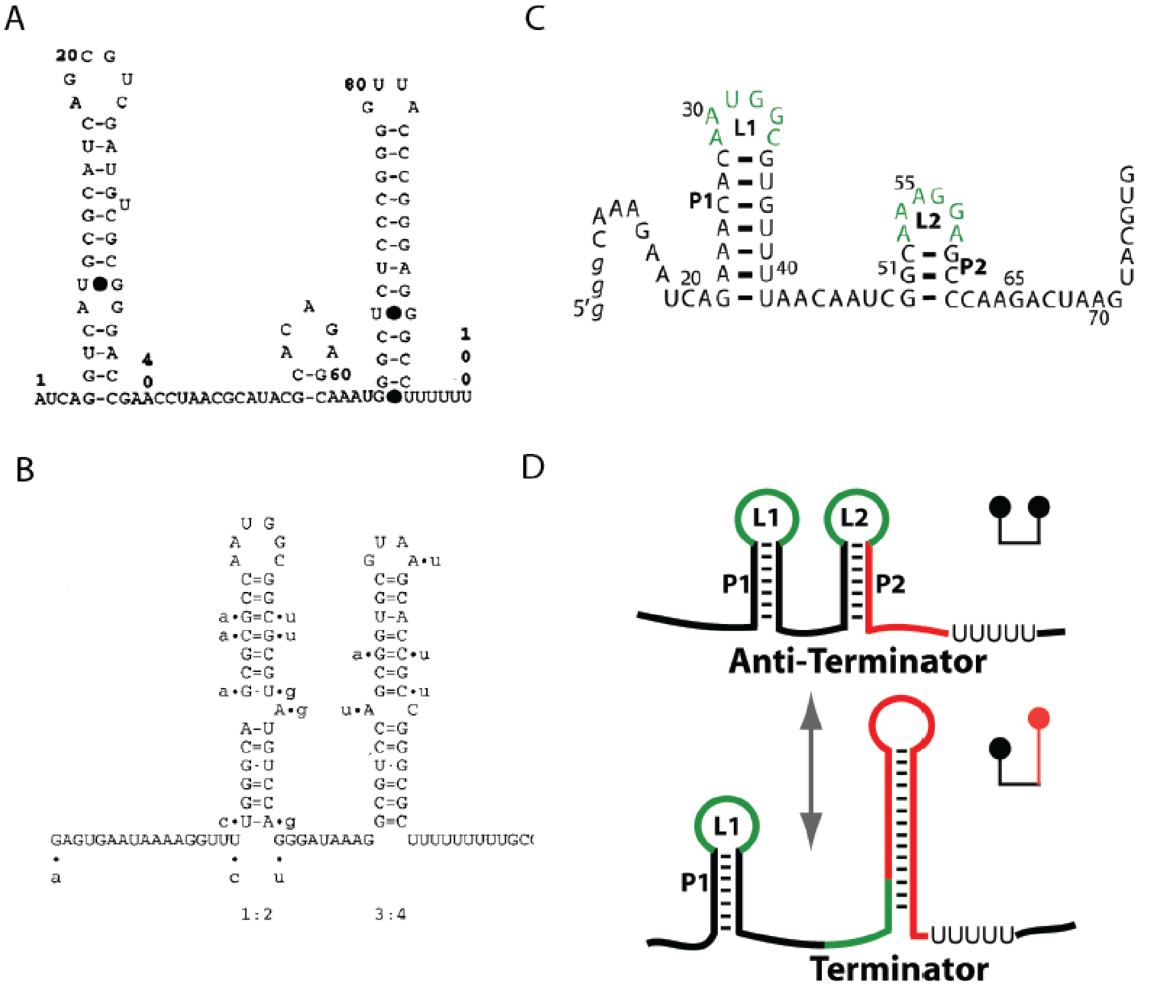

A
AUCAG-CGAACCUAACGCAUACG-CAAAUG●UUUUUU
B
GAGUGAAUAAAAGGUUU GGGAUAAAG UUUUUUUUUGC
1:2
3:4
C
L1
P1
L2
P2
5′g
UCAG-UAACAAUCG-CCAAGACUAAG
D
L1
L2
P1
P2
UUUUU
Anti-Terminator
L1
P1
UUUUU
Terminator

Figure III-2: ANTAR Substrate RNA motifs as described in different works.
(A) *Pseudomonas aeruginosa amiE* leader region containing the motif as described by Wilson et. al. (143). (B) *Klebsiella oxytoca nasF* leader region containing the motif as described by Chai et. al. (146) (C) *Enterococcus faecalis eutP* leader region showing the P1-P2 motif tandem stem-loop structure. (D) Schematic showing the transition between the P1-P2 tandem loop and the mutually incompatible terminator hairpin structure. *Panels C and D reproduced from (152).*

In this chapter, to assess whether the dual stem-loop RNA structure might be present in other bacteria, we searched for this element across many bacterial genomes. These searches led to the discovery of many new regulons that are likely to be coordinated by ANTAR recognition elements. These searches also revealed that the ANTAR recognition elements described herein are generally involved in coordinating expression of nitrogen metabolism genes. Intriguingly, all three characterized examples of ANTAR regulators ultimately regulate nitrogen metabolism in response to nitrogenous compounds, although these compounds are quite distinct, amides such as lactamide or butyramide by AmiR, nitrate and nitrite ions by NasR, and ethanolamine by EutV. This raises the interesting hypothesis that ANTAR domains may be generally used by bacteria to regulate nitrogen metabolism, although the modular RNA-binding activity conferred by the ANTAR domain could also be repurposed for other RNA-binding regulatory mechanisms.

From this point, we continued our investigation into the mechanism of ANTAR recognition of RNA elements in two directions. The first investigation begins from the protein side. We asked what elements of the short ANTAR domain are generally required for RNA binding and antitermination activity. Our lab has investigated both EutV antitermination activity *in vivo* as well as NasR RNA binding *in vitro*. In this chapter we focus on the NasR RNA binding and additionally compare the results with those from EutV experiments. The second investigation into ANTAR activity begins from the RNA side. While we generated an alignment of conserved RNA elements found in a wide variety of bacteria, we are limited to experimental results from only the three systems described above, and our bioinformatic expansion of this alignment could completely miss ANTAR substrate sequences with significant sequences differences not represented in our seed alignment. To address this limitation, we generated an unbiased alignment of ANTAR-binding RNAs using an *in vitro* evolution approach. This approach extends the traditional SELEX process (153), which focuses on identifying a small number of strongly binding sequences, by high-throughput analysis of the output sequence pools (154), enabling more thorough analysis of diversity of selected sequences (155). HT-SELEX has been used to identify sequences binding a wide variety of targets, including aptamers to small molecules or proteins (156, 157), identifying transcription-factor binding sites (158, 159), RNA-binding protein ligands (160), and fitting quantitative models of binding (161, 162). Together, the elaboration on our understanding of both the ANTAR protein and RNA substrates expand our understanding of the role these regulators play in transcription regulation.

Initial identification of two-stem-loop motifs in selected operons

Previous works had identified a variety of sites of ANTAR-mediated transcription attenuation, and a candidate recognition sequence was identified as a small conserved sequence overlapping the intrinsic terminator sequences. Manual inspection of the sequence context of these candidate sequences in the *E. faecalis eut* gene cluster indicated an upstream sequence feature shared between each. This sequence feature is a short potential stem-loop sequence which is predicted to form a hexaloop sharing similar sequence to the previously identified candidate motif (Figure III-2A-B). Indeed, the conserved sequence overlapping the terminator sequence is capable of forming the same short stem-loop structure (Figure III-2C). This dual stem-loop structure appeared in other RNAs known to be regulated by ANTAR proteins, including the *nasF* leader region regulated by NasR and the *amiE* leader region regulated by AmiR. The formation of the second stem-loop, which significantly overlaps the upstream strand of the intrinsic terminator, is incompatible with formation of the terminator (Figure III-2C). These observations served as the basis for the hypothesis that this dual stem-loop structure forms an antiterminator structure bound by ANTAR regulators responsible for the observed transcription attenuation (Figure III-2D).

The EutV ANTAR domain is sufficient for recognition of the dual stem-loop

Several of these dual-stem loop motifs can be identified in the EutVW-containing *E. faecalis eut* gene cluster and *Listeria monocytogenes eut* gene cluster, in addition to the single examples found in the *K. oxytoca nasF* and the *P. aeruginosa amiE* leader regions. All of the ANTAR-containing proteins associated with these elements contain two protein domains. EutV itself is capable of binding the dual-stem loop motifs, in particular the 5′ leader region of *eutP*, in an *in vitro* electrophoretic mobility shift assay (EMSA), albeit with a poor apparent affinity (or probable avidity due to separate interaction of each stem-loop) of approximately 10 µM (Figure III-3A).

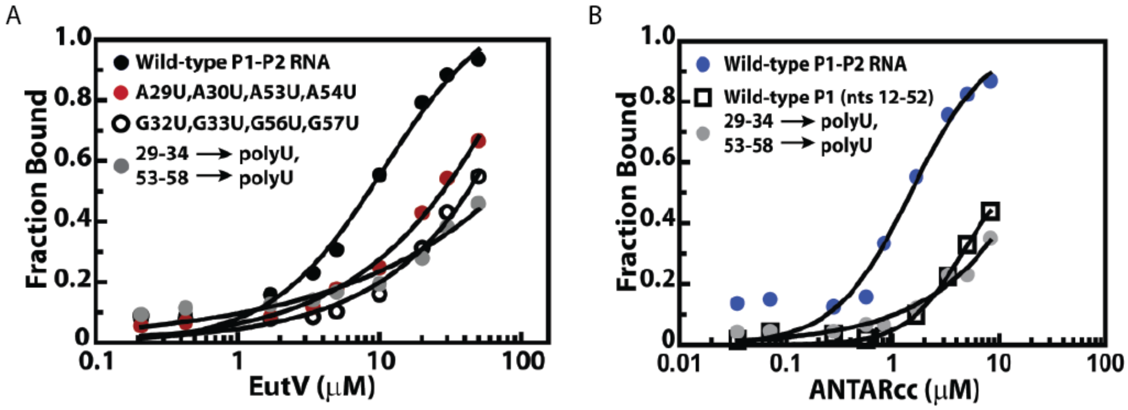


Figure III-3: EutV proteins bind P1-P2 tandem stem-loops *in vitro*.
Saturation binding curves from electrophoretic mobility shift assays (EMSA) are shown. Fractional binding is plotted against protein concentration. (A) EutV (unphosphorylated) bound the eutP 5′ leader region with an apparent KD of 10 μM (black). Binding was significantly deceased in an RNA mutant where the hexanucleotide terminal loops were mutated to uridines (grey). Binding was significantly weaker with RNAs mutated in the first (red) and fourth (open circle) positions of the terminal loops. (B) ANTARcc, a truncation mutant lacking the response-regulator domain, binds different RNA constructs with variable affinities. While RNA that included the dual hairpin motif (blue) was bound with micromolar affinity, mutation of the terminal loop sequences (grey) as well as deletion of the second stem loop (squares) abolished binding. *Figures reproduced from* (152).

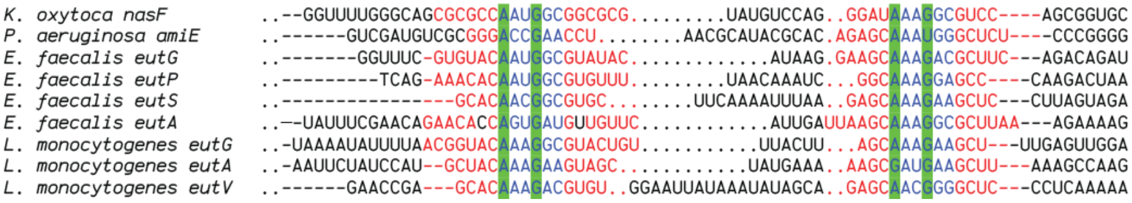
In the case of EutV, there is an N-terminal response receiver domain separated from the C-terminal ANTAR domain by a coiled-coil linker. Similar coiled-coil structures are known to exist in the structures of other ANTAR-containing proteins (140, 145, 163). Response regulator proteins innately require a change in phosphorylation status to create the response. As the avidity of full-length, unphosphorylated EutV, as tested *in vitro*, is quite high, the unphosphorylated state may inhibit ANTAR RNA binding activity. To test the hypothesis that the ANTAR domain itself, with or without the associated coiled-coil region, is sufficient for binding to the putative substrate RNA, we expressed and purified two mutant forms of *E. faecalis* EutV, lacking either the response receiver domain (ANTARcc) or both the receiver and coiled-coil regions (ANTAR). In similar EMSA experiments, the ANTAR mutant displays very little binding activity, with an apparent affinity much higher than full-length unphosphorylated EutV (152). In contrast, the ANTARcc mutant retaining the coiled-coil region binds *eutP* leader RNA with a much higher apparent affinity (Figure III-3B), approximately 700 nM, a comparable but stronger interaction than that previously reporter for the NasR-*nasF* RNA interaction (146). This suggests that, indeed, the unphosphorylated state of the receiver domain of EutV may inhibit RNA binding activity of the ANTAR domain. Phosphorylation of the receiver domain is likely to be accompanied by structural reorganization, perhaps allowing the ANTARcc domain to adopt a conformation better suited for RNA-binding.

This interaction between EutV protein and the two stem-loop motif in the *eutP* leader region is both specific and dependent on several conserved characteristics of this motif. Mutation of either the conserved adenine residues at the beginning of each loop, the guanine residues in the middle of the loop, or the entire loop sequences all result in significantly reduced binding affinity (Figure III-3A). The ANTARcc mutant protein exhibits similar behavior, with reduced apparent affinity for a single P1 stem-loop, or RNA containing poly-U loop sequences (Figure III-3B). These data support the premise that ANTAR domains, potentially requiring the coiled coil region, promote antitermination by recognizing and binding to the terminal loop residues of the dual stem-loop motif, stabilizing that conformation and preventing formation of the mutually-exclusive terminator stem-loop conformation.

Development of search strategies for two-stem-loop motifs

Covariance model and explicit minimal conserved element search.

Given the close sequence and structural similarity between the dual stem-loop RNA motif in the three different characterized ANTAR systems (AmiR, NasR, EutV), we hypothesized that that RNA motif might be generally representative of ANTAR substrates in other organisms. Also, the three previously characterized ANTAR regulatory systems each affected a single locus, albeit with multiple RNA substrates, in their respective host organisms. Therefore, we reasoned that a subset of bacteria might instead incorporate multiple ANTAR-responsive RNA elements at disparate genomic locations for coordination of ANTAR-based regulons. To this end, we searched for additional occurrences of the putative ANTAR RNA substrate using a bioinformatics-based approach. Specifically, we used a covariance model-based approach (164) wherein a basic sequence alignment of a target RNA element, including certain secondary structure and primary sequence determinants, is used as input information for discovery of additional representatives from fully sequenced bacterial genomes.



```
K. oxytoca nasF            ..--GGUUUUGGGCAGCGCGCCAAUGGCGGCGCG.........UAUGUCCAG..GGAUAAAGGCGUCC----AGCGGUGC
P. aeruginosa amiE         ..------GUCGAUGUCGCGGGACCGAACCU........AACGCAUACGCAC.AGAGCAAAUGGGCUCU----CCCGGGG
E. faecalis eutG           ..-------GGUUUC-GUGUACAAUGGCGUAUAC.............AUAAG.GAAGCAAAGACGCUUC---AGACAGAU
E. faecalis eutP           ..---------TCAG-AAACACAAUGGCGUGUUU.........UAACAAAUC...GGCAAAGGAGCC----CAAGACUAA
E. faecalis eutS           ..----------------GCACAACGGCGUGC........UUCAAAAUUUAA..GAGCAAAGAAGCUC---CUUAGUAGA
E. faecalis eutA           ..--UAUUUCGAACAGAACACCAGUGAUGUUGUUC............AUUGAUUAAGCAAAGGCGCUUAA---AGAAAAG
L. monocytogenes eutG      ..-UAAAAUAUUUUAACGGUACAAAGGCGUACUGU...........UUACUU...AGCAAAGAAGCU---UUGAGUUGGA
L. monocytogenes eutA      ..-AAUUCUAUCCAU--GCUACAAAGAAGUAGC............UAUGAAA..AAGCGAUGAAGCUU---AAAGCCAAG
L. monocytogenes eutV      ..------GAACCGA---GCACAAAGACGUGU..GGAAUUAUAAAUAUAGCA..GAGCAACGGGGCUC---CCUCAAAAA
```

Figure III-4: Seed alignment of ANTAR substrate hits.
Seed alignment of the tandem stem-loop motifs from the ANTAR-regulated leader regions of *K. oxytoca*, *P. aeruginosa*, *E. faecalis*, and *L. monocytogenes*. Putative stem residues are shown in red, while terminal loop residues are shown in blue. The highly conserved A and G residues at positions 1 and 4 of the terminal loop are highlighted in green.
Reproduced from (152).

This method has been successfully employed for larger, structured RNAs such as riboswitches, and is also the underlying algorithm currently used by the Rfam database to curate bacterial noncoding RNAs (165, 166). Therefore, a seed alignment was created based on the putative ANTAR RNA substrates (the dual hairpin element) from the *eut* loci of *E. faecalis*, *Clostridium* and *Listeria* species, as well as the corresponding RNA sequences for *K. oxytoca nasF* and *P. aeruginosa amiE* (Figure III-4). This RNA element was defined as a dual stem-loop motif with a minimum of three base-pairs in each stem and a variable linker region connecting the two stems. Sequence conservation in the loops, with an adenine at position 1 and a guanine at position 4 of each loop was maintained, although these sequence and structural constraints are not effectively enforced in the resulting covariance model. Given the relatively small size of the motif and the small number of residues conserved at the primary sequence level, the first search was targeted against a narrowly defined subset of genomic sequence. We reasoned that this would allow us to limit the initial high false-positive rate and fully examine the quality of our individual RNA hits. For this targeted analysis we searched against 83 bacterial genomes that were previously predicted (167) to specifically encode for a putative *eut* locus. Some *eut* loci are regulated by a

DNA-binding regulator called EutR (e.g., *Salmonella*, *Escherichia*) whereas others, especially the Firmicutes, are regulated by a ANTAR-containing homolog of EutV, as in *E. faecalis* (167). Therefore, a subset of these genomes contains putative *eut* pathway homologues but lack any ANTAR-encoding genes, while other genomes contain both. As predicted, we recovered less RNA hits in genomes that lack ANTAR-encoding genes (Figure III-5A).

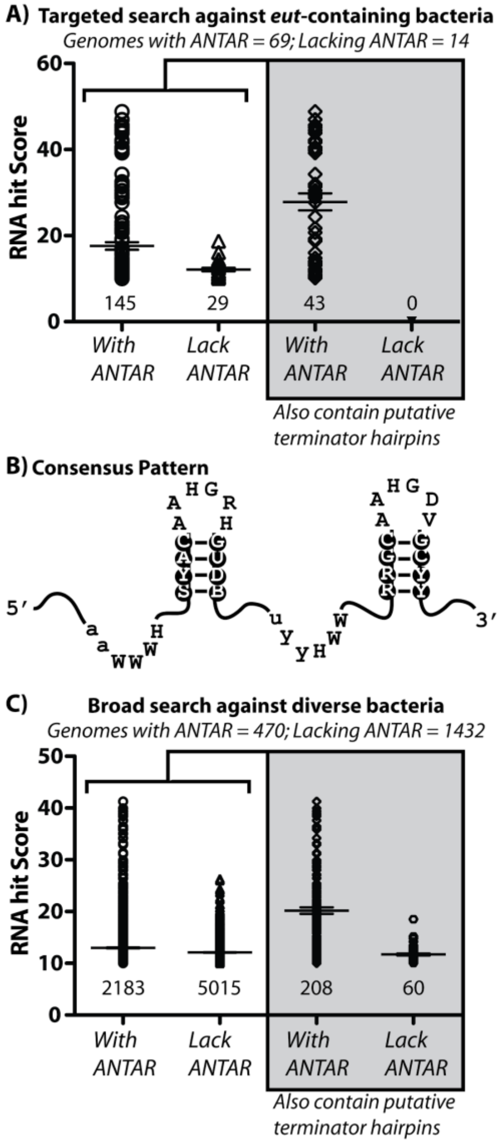


Figure III-5: Bioinformatic analysis of the ANTAR domain and its two stem-loop RNA substrate.
A) Using a covariance-based search approach (Infernal (168)), we identified additional occurrences of the ANTAR RNA substrate in bacteria that contained eut pathways. A scatter plot is shown for the resulting RNA hits, where each data point represents a different RNA hit. The hit scores for these sequences were plotted for two classes of microorganisms used in this search. Specifically, organisms that are predicted to encode for ANTAR domain proteins (see Table S1) have more RNA hits with higher scores than a control set of organisms that appear to lack any ANTAR domain proteins. Also, these hits were screened using TransTerm for the presence of an intrinsic terminator hairpin located immediately downstream of the P2 helix. Only a subset of the hits satisfied this important criterion. B) A consensus secondary structure was derived from this sequence alignment and is shown herein. C) The covariance-based search approach was then employed against 1902 bacterial genomes to search more broadly for putative ANTAR-based regulatory pathways. Again, a scatter plot is shown for the resulting hits, and for the subsequent screening of these hits for the presence of an overlapping downstream intrinsic terminator hairpin. *Reproduced from* (152).

This covariance model-based search revealed the presence of many putative ANTAR RNA targets (152). Our approach was validated in part by the identification of all 17 input sequences that were used to derive the seed alignment. Most hits (>83%) originated from bacteria that encoded for at least one ANTAR-encoding gene (Figure III-5A). Moreover, the average model score was higher and E-values lower for RNA hits from organisms that encoded for at least one ANTAR gene (Figure III-5A), suggesting that matching RNA elements are is at least partially correlated with the presence of ANTAR-containing genes. These newly identified putative ANTAR substrates originated from diverse bacteria, including Gram-positive bacteria (e.g., *Mycobacterium*, *Streptococcus*, *Fusobacterium*, *Alkaliphilus*, etc.) and Gram-negative bacteria (e.g., *Pseudomonas*, *Burkholderia*, etc.), and resulted in a consensus pattern that resembled the input consensus pattern (Figure III-5B).

A common approach to identification of families of sequences, whether coding sequences, non-coding genes, or just conserved sequences motifs, is to begin from a seed sequence or alignment and iteratively search and incorporate new hits into the alignment at each step (169–171). We initially attempted to expand the search for ANTAR RNA substrates by incorporating additional hits into the alignment for future rounds. As mentioned above, the covariance model searches resulted in a high apparent false-positive rate, and poor separation in bit-score between convincing two stem-loop motifs in non-coding regions and highly similar sequences lacking key features. For this reason, automatic incorporation of high-scoring hits into the alignment either resulted in limited improvement when the score threshold was high, or a dramatic increase in false-positive rates when the threshold was lowered. Manual curation of additional sequences and improved criteria for screening potential hit sequences was eventually required. We utilized one potential screen, effective for Gram-positive bacteria, detailed in the next section. Additionally, we also will use an *in vitro* selection approach to catalog RNA sequences capable of binding ANTAR proteins, generating many high-confidence hits.

Identification of two-stem-loop candidates overlapping intrinsic terminator sequences.

The ANTAR systems that have been previously characterized are each used to regulate transcription attenuation. To examine whether some or all of the hits acquired in this analysis are also likely to mediate transcription attenuation, we screened them using TransTermHP for candidate intrinsic transcription terminator stem-loops that overlapped with the P2 helix. Approximately 30% of the hits satisfied this criterion for organisms that encoded for ANTAR genes, whereas none of the RNA hits satisfied this criterion from organisms lacking an ANTAR gene (Figure III-5A). Moreover, the average hit score increased further for the hits that contained terminator stem-loops. Therefore, these putative hits represent the best possible candidates for new ANTAR-based regulatory systems. However, it is important to note that many of the remaining hits (lacking terminator stem-loops) may still function as actual ANTAR regulatory elements, but via regulatory strategies other than transcription attenuation, such as control of translation initiation. Indeed, manual inspection of some of these hits revealed instances where they were arranged near to or overlapping with the ribosome binding site of the downstream gene (152).

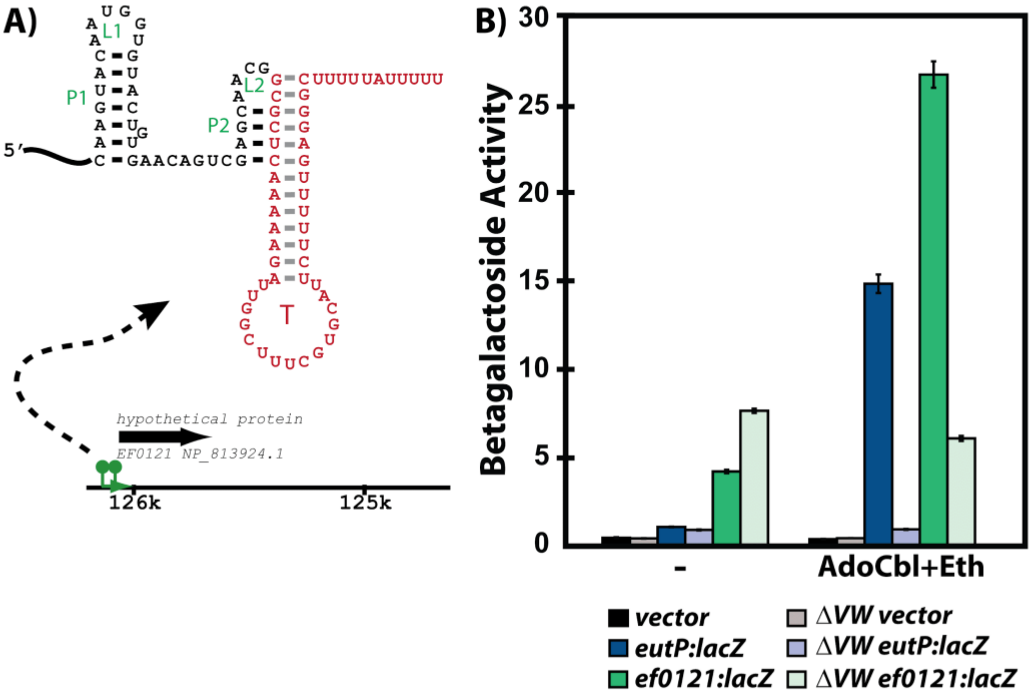


Figure III-6: Responsiveness of the *ef0120* leader region to regulation by EutVW.
(A) One outcome of the bioinformatic search for new ANTAR RNA candidates was identification of a new putative hit within the *E. faecalis* genome. This hit contained an overlapping intrinsic terminator hairpin, as identified by TransTermHP. B) To test whether this hit was functionally responsive to EutVW *in vivo*, the leader region of this gene was translationally fused to a *lacZ* reporter and monitored with and without AdoCbl and ethanolamine. *Reproduced from* (152).

Interestingly, while most of the new hits in this search were associated with *eut* genes, many were not. For example, a new ANTAR substrate was unexpectedly identified in the *E. faecalis* genome outside of the *eut* locus and within the 5′ leader region of *ef0120*, suggesting that EutVW might indeed control a regulon rather than a single locus. To investigate this observation further, we fused the *ef0120* 5′ leader region to *lacZ* and monitored expression in the presence and absence of adenosylcobalamin (AdoCbl), ethanolamine, and the EutVW genes (Figure III-6). Indeed, expression was activated by AdoCbl and ethanolamine in a EutVW-dependent manner. Therefore, our covariance model search for putative ANTAR substrates is likely to have revealed ANTAR-based coordination of *E. faecalis* genes both inside and outside of the *eut* locus. In fact, this search revealed many examples where putative ANTAR substrates were associated with multiple functionally-related operons, as one might expect for regulons. For example, new ANTAR substrates appeared to be co-transcriptionally linked to different glutamate synthase genes in the *Desulfotomaculum reducens* genome (Figure III-7A). Similarly, in *Mycobacterium vanbaalenii*, a putative ANTAR RNA substrate is positioned upstream of multiple uncharacterized gene clusters unrelated to *eut* genes (152). In *Pelobacter carbinolicus,* new ANTAR substrates were located within three separate transcriptional units, which are each predicted to be contain nitrogenase function, suggesting an ANTAR-based regulon for nitrogenase regulation in this microorganism (Figure III-7B).

To broaden this search outside of organisms containing *eut* pathways, we repeated the covariance search with moderately more restrictive criteria against 1902 bacterial genomes, of which 470 included homologues for ANTAR-containing genes (Figure III-5C). After screening these hits for the presence of an intrinsic transcription terminator overlapping the P2 stem, the average hit score was higher for organisms encoding an ANTAR gene as compared to organisms lacking ANTAR genes. This search revealed many more examples of excellent candidates for ANTAR-mediated regulons in diverse bacterial species, including Gram-negative and Gram-positive bacteria (Table S2). Remarkably, the majority of these hits were consistently located upstream of nitrogen metabolism genes. Indeed, both searches revealed a close association between nitrogen metabolism genes and ANTAR-based regulation. For example, RNA hits were discovered upstream of genes annotated as coding for genes, including but not limited to: nitrogen regulatory protein P-II, ammonium transporters, urea transporters, nitrate and nitrite transporters, nitrite reductase, nitrogenase subunits, and synthase enzymes for glutamate, glutamine, and arginine. This point is particularly illustrated by *Desulfotomaculum acetoxidans*, which contains at least 13 new regulatory RNA hits within six putative transcriptional units that encompass many of these metabolic functions (Figure III-7C). It is worth noting that the previously studied *eut*, *nas* and *ami* operons encode for ethanolamine catabolism, nitrate assimilation and ammonia-releasing amidases, respectively, which are all tied to nitrogen metabolism. Therefore, this aggregate data reveal clearly that ANTAR-based regulons are widely used in bacteria for control of nitrogen metabolism genes.

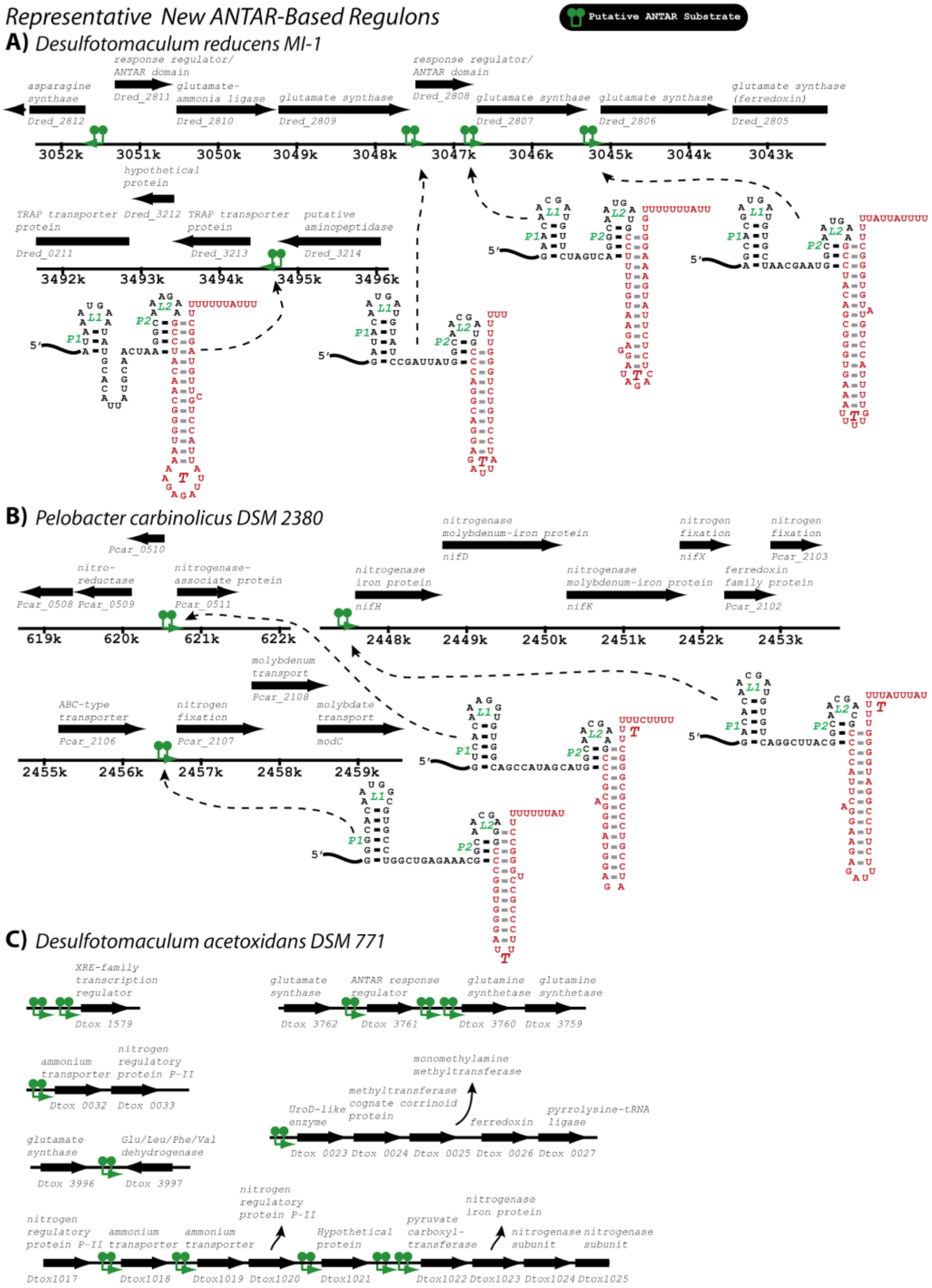
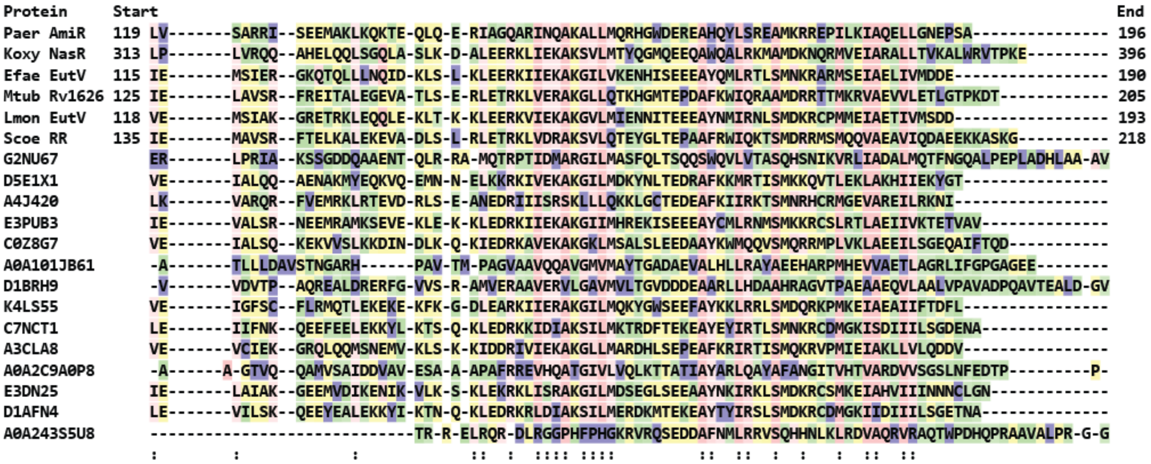


Figure III-7: Representative ANTAR-based regulons identified through bioinformatics. RNA hits (green) from the covariance searches are shown within their genomic contexts for two representative organisms. Genes are shown along with their annotations (black). The putative ANTAR substrate RNAs appear to be present in multiple operons in the same bacterium, and are thereby likely to participate in control of ANTAR-based regulons. In these examples, the regulons are predicted to be functionally related to control of glutamate metabolism and nitrogenase expression, respectively. *Reproduced from* (152).

Requirement for conserved residues in the ANTAR domain for RNA-binding and antitermination activity.

The EutV ANTAR domain is sufficient for binding ANTAR substrate RNA, although inclusion of an adjacent, extended α-helical region dramatically increased RNA binding affinity (Figure III-3B)(172). However, little is known about which residues of ANTAR domains are required for binding RNA substrates and for antitermination activity. To begin investigating these requirements, we aligned sequences of ANTAR domains for proteins related to those known to bind the canonical ANTAR RNA substrate. This comparative sequence alignment revealed 18 strongly conserved and 14 moderately conserved amino acids spread between the ANTAR-domain and the adjacent α-helical segment (Figure III-8). To investigate where these residues are likely to be positioned, we located these in the high-resolution structural model of NasR derived by X-ray crystallography (145). Using this three-dimensional model, it appeared that of the 25 conserved amino acids, 12 hydrophobic residues are likely to be involved in intramolecular interactions within the protein core, such as the structural interactions formed between alpha helices. The remaining residues were therefore determined to be candidate sites for protein-RNA recognition.



```
Protein     Start                                                                                                           End
Paer AmiR   119 LV-------SARRI--SEEMAKLKQKTE-QLQ-E-RIAGQARINQAKALLMQRHGWDEREAHQYLSREAMKRREPILKIAQELLGNEPSA--------------- 196
Koxy NasR   313 LP-------LVRQQ--AHELQQLSGQLA-SLK-D-ALEERKLIEKAKSVLMTYQGMQEEQAWQALRKMAMDKNQRMVEIARALLTVKALWRVTPKE--------- 396
Efae EutV   115 IE-------MSIER--GKQTQLLLNQID-KLS-L-KLEERKIIEKAKGILVKENHISEEEAYQMLRTLSMNKRARMSEIAELIVMDDE----------------- 190
Mtub Rv1626 125 IE-------LAVSR--FREITALEGEVA-TLS-E-RLETRKLVERAKGLLQTKHGMTEPDAFKWIQRAAMDRRTTMKRVAEVVLETLGTPKDT------------ 205
Lmon EutV   118 VE-------MSIAK--GRETRKLEQQLE-KLT-K-KLEERKVIEKAKGVLMIENNITEEEAYNMIRNLSMDKRCPMMEIAETIVMSDD----------------- 193
Scoe RR     135 IE-------MAVSR--FTELKALEKEVA-DLS-L-RLETRKLVDRAKSVLQTEYGLTEPAAFRWIQKTSMDRRMSMQQVAEAVIQDAEEKKASKG---------- 218
G2NU67          ER-------LPRIA--KSSGDDQAAENT-QLR-RA-MQTRPTIDMARGILMASFQLTSQQSWQVLVTASQHSNIKVRLIADALMQTFNGQALPEPLADHLAA-AV
D5E1X1          VE-------IALQQ--AENAKMYEQKVQ-EMN-N-ELKKRKIVEKAKGILMDKYNLTEDRAFKKMRTISMKKQVTLEKLAKHIIEKYGT----------------
A4J420          LK-------VARQR--FVEMRKLRTEVD-RLS-E-ANEDRIIISRSKLLLQKKLGCTEDEAFKIIRKTSMNRHCRMGEVAREILRKNI-----------------
E3PUB3          IE-------VALSR--NEEMRAMKSEVE-KLE-K-KLEDRKIIEKAKGIIMHREKISEEEAYCMLRNMSMKKRCSLRTLAEIIVKTETVAV--------------
C0Z8G7          VE-------IALSQ--KEKVVSLKKDIN-DLK-Q-KIEDRKAVEKAKGKLMSALSLEEDAAYKWMQQVSMQRRMPLVKLAEEILSGEQAIFTQD-----------
A0A101JB61      -A-------TLLLDAVSTNGARH------PAV-TM-PAGVAAVQQAVGMVMAYTGADAEVALHLLRAYAEEHARPMHEVVAETLAGRLIFGPGAGEE--------
D1BRH9          -V-------VDVTP--AQREALDRERFG-VVS-R-AMVERAAVERVLGAVMVLTGVDDDEAARLLHDAAHRAGVTPAEAAEQVLAALVPAVADPQAVTEALD-GV
K4LS55          VE-------IGFSC--FLRMQTLEKEKE-KFK-G-DLEARKIIERAKGILMQKYGWSEEFAYKKLRRLSMDQRKPMKEIAEAIIFTDFL----------------
C7NCT1          LE-------IIFNK--QEEFEELEKKYL-KTS-Q-KLEDRKKIDIAKSILMKTRDFTEKEAYEYIRTLSMNKRCDMGKISDIIILSGDENA--------------
A3CLA8          VE-------VCIEK--GRQLQQMSNEMV-KLS-K-KIDDRIVIEKAKGLLMARDHLSEPEAFKRIRTISMQKRVPMIEIAKLLVLQDDV----------------
A0A2C9A0P8      -A------A-GTVQ--QAMVSAIDDVAV-ESA-A-APAFRREVHQATGIVLVQLKTTATIAYARLQAYAFANGITVHTVARDVVSGSLNFEDTP---------P-
E3DN25          IE-------LAIAK--GEEMVDIKENIK-VLK-S-KLEKRKLISRAKGILMDSEGLSEEAAYNKIRKLSMDKRCSMKEIAHVIIINNNCLGN-------------
D1AFN4          LE-------VILSK--QEEYEALEKKYI-KTN-Q-KLEDRKRLDIAKSILMERDKMTEKEAYTYIRSLSMDKRCDMGKIIDIIILSGETNA--------------
A0A243S5U8      -----------------------------TR-R-ELRQR-DLRGGPHFPHGKRVRQSEDDAFNMLRRVSQHHNLKLRDVAQRVRAQTWPDHQPRAAVALPR-G-G
                :        :           :           ::  :  :::: ::::         ::  ::  :  :   :  ::  ::
```

Figure III-8: Alignment of ANTAR and coiled-coil domains of ANTAR-containing proteins. T-Coffee multiple sequence alignment of 20 selected ANTAR domain sequences with 20 residues toward the N-terminus from the start of the ANTAR Pfam model. Six proteins, comprising the proteins studied in this manuscript or studied in the literature (142, 152, 163, 173, 174). Three of these (AmiR, NasR, Rv1626) have had structure solved crystallographically and those structures were included to guide this alignment in 3DCoffee mode. The alignment was performed using 120 additional sequences chose from 7700 total ANTAR domains clustered by sequence identity to 120 clusters. This alignment represents the original six proteins along with fourteen additional sequences extracted from the larger alignment shown as Uniprot accession numbers. Residues are colored by conservation in the larger alignment, and conserved columns are highlighted with colons (:).

Previous attempts to address *in vitro* binding of the ANTAR domain to its RNA substrate have been complicated by poor solubility of EutV, the high proportion of inactive protein after purification, and the difficulty in creating or mimicking phosphorylated protein *in vitro*. For these reasons, we chose to target the *K. oxytoca* NasR protein for additional characterization, as it is a one-component regulator directly responsive to nitrate or nitrite, and can actively bind RNA and antiterminate *in vitro (146)*. We purified a His_{10}-MBP tagged NasR to test whether it could bind the *nasF* two-hairpin RNA motif. We tested equilibrium binding using fluorescence anisotropy by titrating RNA labeled at the 3′ terminus with a Cy3 fluorophore with increasing amounts of the His_{10}-MBP-NasR dimer, similar to the method described in Chapter 2 for analysis of LoaP-RNA complexes. While changes in fluorescence anisotropy were detectable with the SpectraMax M5 plate reader with as little as 5 nM RNA, higher concentrations reduced assay noise, with concentrations above 10 nM preferable (Figure II-9B). In the presence of 1 mM nitrate, His_{10}-MBP-NasR bound *nasF* P1P2 with an affinity of approximately 500 nM (Figure III-9A). This is comparable to the previously reported dissociation constant (146). Under the same conditions, His_{10}-MBP-NasR displayed no detectable binding to an unrelated Cy3-labeled RNA hairpin (from the 5′ leader region of *B. velezensis dfnA*) (Figure III-9A). RNA binding activity was only observed in conditions containing nitrate or nitrate, with maximum binding occurring above 1 mM KNO_3 or 500 nM KNO_2 (Figure III-9B) again comparable to previously reported data (146). Finally, no detectable binding occurred with additional of urea or guanidine (Figure III-9B).

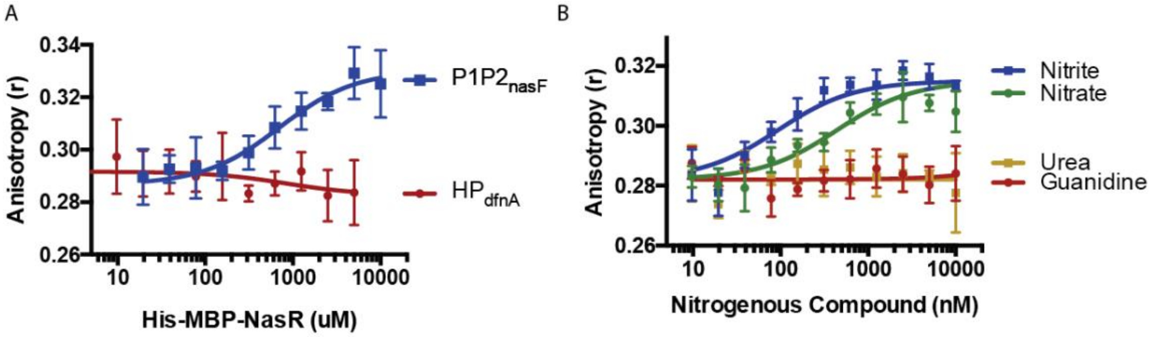


Figure III-9: Fluorescence anisotropy saturation binding curves for His-MBP-NasR. Equilibrium binding assays using fluorescence polarization of Cy3-labeled synthetic RNA. (A) Saturation binding of His-MBP-NasR using either RNA derived from the *K. oxytoca nasF* leader region (P1P2$_{nasF}$) or from the *B. velezensis dfnA* leader region (HP$_{dfnA}$, Figure II-9A) in the presence of 1mM nitrate. (B) Equilibrium binding of of P1P2$_{nasF}$ RNA in the presence of varying amounts of sodium nitrite, sodium nitrate, urea, or guanidine. **Assistance from Daniel Trettel*

Targeted substitution mutants reveal protein elements required for RNA binding

We constructed and purified single alanine point mutants of nineteen conserved residues in the ANTAR domain of His$_{10}$-MBP-NasR. Each mutant protein was successfully purified and appeared predominantly as a single consistently-sized band when analyzed using SDS- and native-PAGE (Figure III-10A and III -10B). We generated saturation binding curves by fluorescence anisotropy for each mutant in the presence of 1 mM nitrate and 50 nM Cy3-labeled *nasF* RNA. Of the nineteen mutants, eight bound with comparable or better affinity as compared to wild-type NasR sequence (Figure III-11A), five had quantifiable binding with lower affinity than wild-type sequence (Figure III-11B), and RNA binding activity was completely abolished for six mutants (Figure III-11C). All mutants were analyzed by native PAGE to identify potential severe structural differences from WT protein. Three of the mutants lacking RNA binding activity appear to be predominately shifted to a much faster-migrating band (W362A, N373A, M376A), potentially indicating significant defects in folding or dimerization. Two mutants with no binding activity (R340A, K347A) and two displaying near-WT activity (K345A, K372A) display moderately anomalous patterns in native PAGE, although it is not clear why (Figure III-10B).

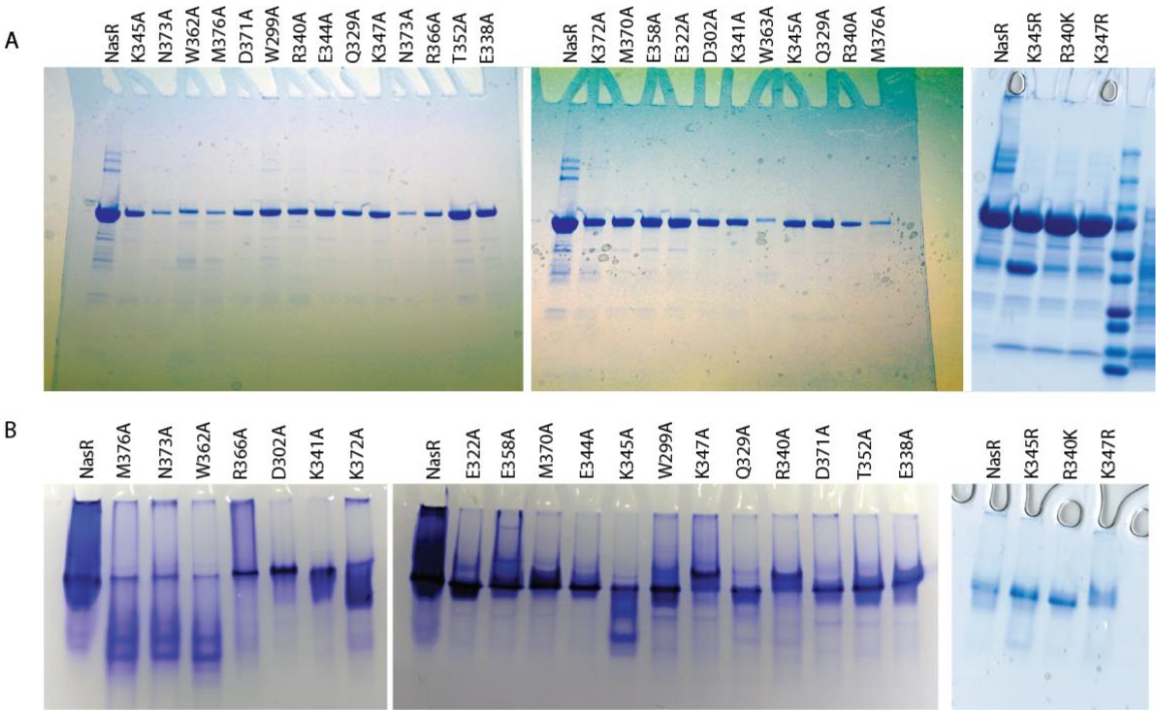


Figure III-10: SDS- and Native-PAGE of purified His-MBP-NasR mutant proteins. SDS-PAGE (A) or Native-PAGE gel images stained with Coomassie dye showing the integrity and purity of His-MBP-NasR mutant protein purifications. **Assistance from Christopher Zhang*

In addition to alanine point mutants, we mutated three charged residues (R340, K345, K347) in the putative positive match, swapping lysine for arginine, or, instead, arginine for lysine (*i.e.,* R340K, K345R, K347R). All three proteins migrate similar to wild-type His$_{10}$-MBP-NasR on SDS-PAGE gels, although K347R exhibits slightly anomalous migration comparable to K347A (Figure III-10). These three mutants exhibit differing RNA binding activity (Figure III-11D), with K345R showing similar activity to both wild-type NasR and K345A. K347R and R340K each have decreased binding activity, exhibiting an apparent affinity around ten-fold higher than wild-type. Therefore, both K347 and R340 residues appear essential for RNA binding activity, as both alanine mutations and conservative charge-swap mutations dramatically decrease binding activity, although the alanine mutation may be more disruptive.

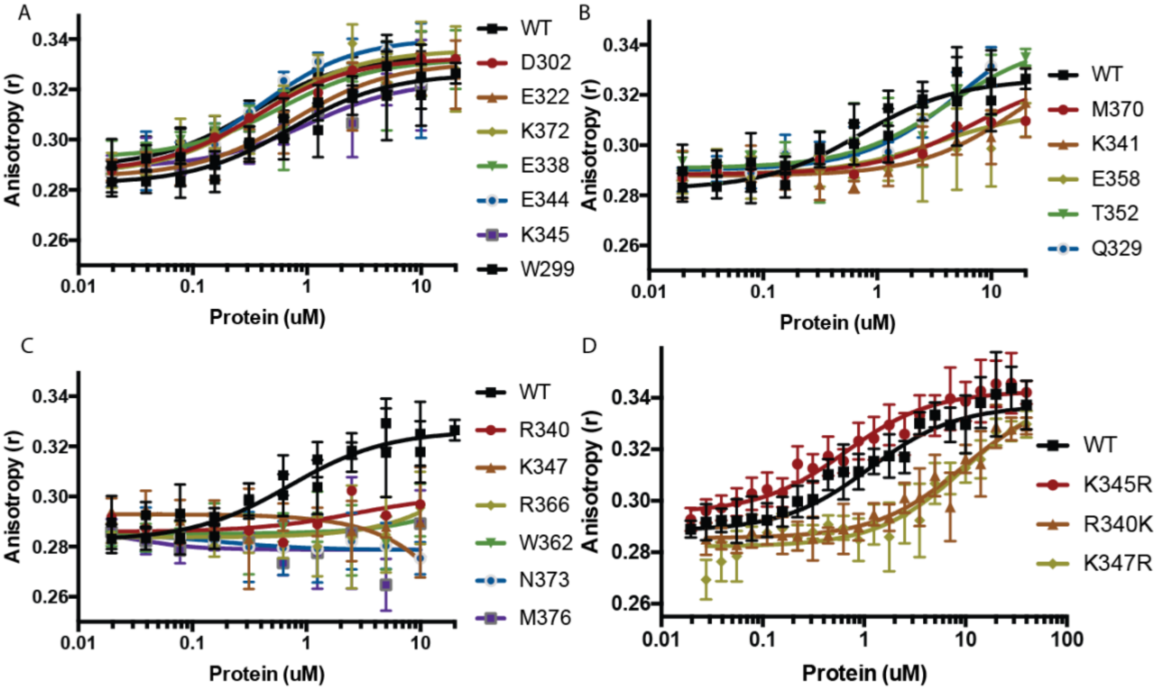


Figure III-11: Targeted mutations of NasR ANTAR and coiled-coil domains have a mix of effects.
Equilibrium binding assays using fluorescence polarization of Cy3-labeled synthetic RNA to a variety of His-MBP-NasR mutant proteins. Each subpanel represents saturation binding titrations of His-MBP-NasR-(mutant) proteins against P1P2*nasF* Cy3-labeled RNA. Mutants in subpanels (A), (B), and (C) represent alanine substitutions. Mutations in (D) contain charge-conserved mutations from one positively-charged amino acid to another. **Assistance from Christopher Zhang*

Comparison with results of targeted reporter mutations

Although EutV has proven difficult to work with *in vitro*, our lab has subsequently developed *in vivo* models of EutV antitermination using fluorescent reporters in *B. subtilis*. In short, our lab recapitulated the EutVW genetic circuit in *B. subtilis* by an inducible cassette for the expression of *eutVW* as a two-gene operon and a *eut* P1/P2 two-hairpin attenuator upstream of a *yfp* reporter gene. In the presence of inducer for production of EutVW, addition of 5 mM ethanolamine triggers antitermination of the P1/P2 attenuator and results in a 10-fold increase in YFP fluorescence as quantified by flow cytometry (175).

This genetic tool allowed another member of the lab to analyze the activity of targeted amino acid substitution mutants in *in vivo* antitermination assays. In that experiment, alanines were individually substituted at fifteen amino acids conserved among ANTAR domains and the resulting EutV variants were expressed in *B. subtilis* alongside the other components of the EutVW genetic circuit. These cells were subjected to flow cytometric analysis with and without ethanolamine. The alanine substitutions resulted in a variety of activities, with two alanine substitutions (N173A, R175A) exhibiting a moderate effect on antitermination, nine of the mutants (Y101A, K104A, M117A, R142A, K147A, E160A, Y164A, M172A, and M178A) appearing to fully abolish antitermination activity, and four of the alanine substitutions (Q124A, Q131A, E140A, and S171A) exhibiting wild-type antitermination activity (Figure III-12). Two of the mutants, however, appeared to exhibit poor expression by Western blotting (M117A, K147A) (175).

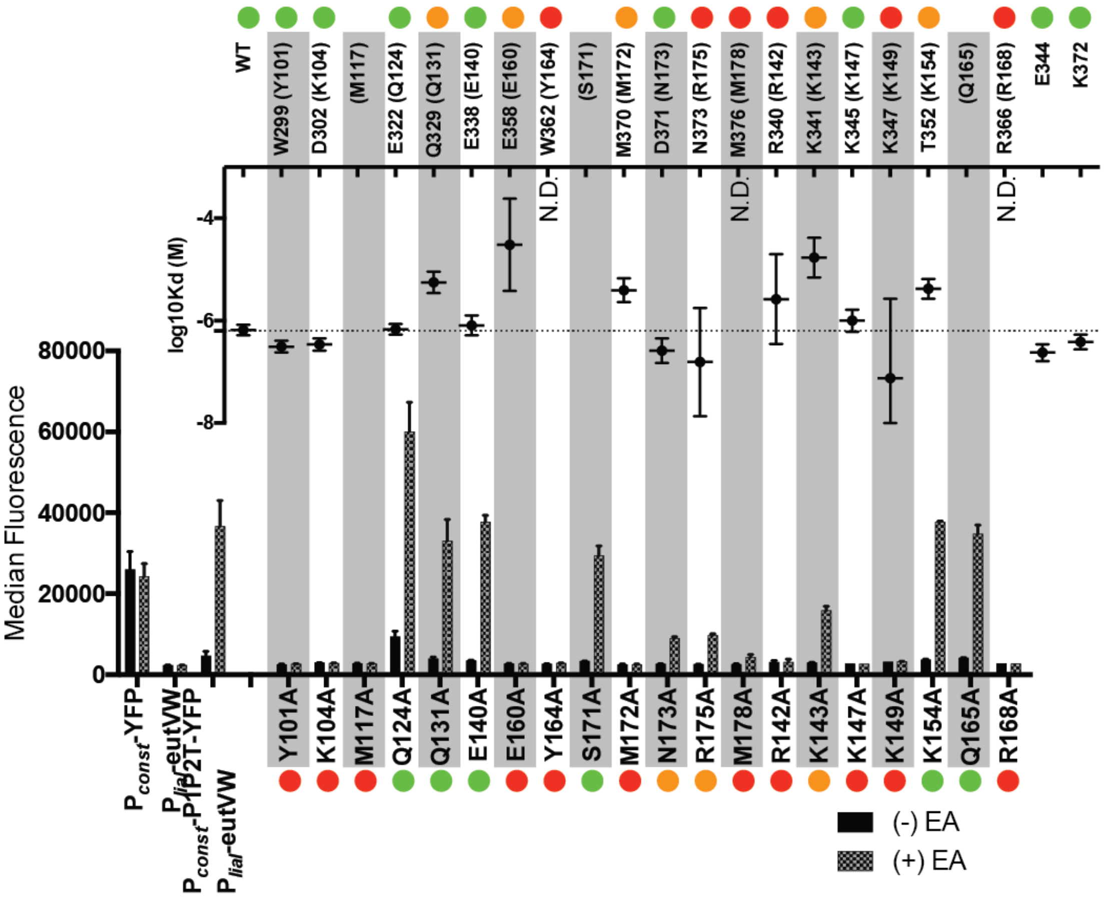


Figure III-12: Comparison of results from *in vitro* binding with NasR and reporter assays with EutV mutants.

Comparison of results from His-MBP-NasR RNA-binding assays with results from flow cytometry quantification of EutV-induced antitermination in a fluorescent reporter construct. The top of the figure shows the estimated log10-binding affinity from the fluorescence anisotropy saturation binding titrations from Figure III-11. The bottom half of the figure shows the median fluorescence as measured by flow cytometry with and without the induction of EutV antitermination by addition of ethanolamine (EA) in *B. subtilis* reporter constructs. Each column contains equivalent residues from the two proteins, and missing data represents mutations which were not tested in both assays. **Assistance from Christopher Zhang. EutV antitermination data from Margo Gebbie.*

In total, we estimated binding affinities for nineteen different alanine-mutant NasR proteins and have quantified antitermination values in *B. subtilis* reporter assays for twenty alanine-mutant EutV proteins. This allows us to compare between mutagenesis of different ANTAR proteins, and to correlate mutations that affect RNA-binding activity to those affecting antitermination activity *in vivo.* Seventeen mutants from each experiment corresponded to mutations of equivalent residues according to the alignment in (Figure III-8). While there is some correspondence between results from each experiment, there is not perfect overlap between the two datasets. We delineated the results of each into three categories: (1) either binding/no-binding, (2) termination/antitermination, or (3) by inclusion of an intermediate category for each experiment representing detectable binding or antitermination substantially lower than for the wild-type protein (Figure III-12).

Half of the equivalent alanine substitutions are highly consistent between experiments, with six mutant pairs neither binding nor promoting antitermination (EutV/NasR: E160/E358, Y164/W362, M178/M376, R142/R340, K149/K347, and R168/R366), two appearing indistinguishable from wild-type behavior (Q124/E322 and E140/E338), and one with intermediate results in each (K143/K341). Three mutant pairs exhibited both antitermination and RNA-binding activity (N173/D371, Q131/Q329, and K154/T352), although with one or the other impaired compared to wild-type. Two additional mutant pairs exhibit either completely impaired binding or antitermination in one assay and only partially impaired activity in the other (M172/M370 and R175/N373). While no mutants retained antitermination activity while completely losing binding activity, three EutV mutants appeared to be completely defective in antitermination while their corresponding NasR mutants retained near-native levels of RNA binding activity (Y101/W299, K104/D302, and K147/K345). These mutants may represent residues required for signal transduction from the EutV response receiver to the ANTAR RNA-binding domain which prevent proper activation of the ANTAR domain, but not the equivalent signal from the NasR NIT nitrate-sensing domain. However, these may also simply represent residues which are not essential for NasR structure or activity but are required for EutV, especially as, mentioned earlier, K147A exhibited poor expression by Western blotting. In total, these experiments provide a preliminary map for investigating the ANTAR residues that associate with RNA ligands.

Identification of potential RNA-binding targets by SELEX

SELEX approach

In addition to our interest in the protein sequences required for ANTAR-RNA binding, we also wished to reassess our knowledge of the RNA sequence requirements. While we have cataloged many examples of two stem-loop motifs in bacterial genomes, our approach may be blind to potential RNA substrates with very different composition that have not yet had seed sequence experimentally identified. To address this deficiency, we elected to study the RNA sequence landscape formed by RNAs capable of binding specifically to the ANTAR domain using a high-throughput sequencing *in vitro* selection approach (HT-SELEX). The dual stem-loop motif is capable of fitting into a 30-nucleotide span suitable for randomization and SELEX, theoretically allowing us to study a comprehensive landscape of small RNAs capable of strongly binding ANTAR domains.

After several rounds of *in vitro* selection, a pool of RNA aptamers emerged that appeared to associate with *K. oxytoca* His_{10}-MBP-NasR. For this experiment we utilized a commercial SELEX RNA library containing a T7 RNAP promoter and two specific primer binding sites for reverse transcription and amplification (TriLink Biotechnologies O-32003). These experiments were performed in parallel for several other proteins. In addition, two separate groups performed parallel selections using similar approaches with distinct methods for enrichment of bound RNA; we used nitrocellulose filter-binding (153), while the other group utilized Ni-NTA paramagnetic bead-binding (176). In both cases, we performed rounds of selection by binding RNA to protein, incubating and applying the mixture to the physical substrate, washing the beads or filter with binding buffer, and elution from the substrate using by either heating or incubation with imidazole for Ni-NTA. After each round, the RNA was purified, reverse transcribed, PCR-amplified to add a T7 promoter, and finally transcribed to generate an RNA pool for input to the next round. In total, we performed four rounds of selection on each protein, lowering the protein:RNA ratio after the second round. Results and parameters from each step are shown in Table III-1.

Selection	Round	Protein (pmol)	RNA (pmol)	RNA:Protein Ratio	PCR Cycles
NasR (Filter)	1	150	400	2.7:1	14
	2	150	400	2.7:1	14
	3	75	400	5.3:1	8
	4	18	400	22.2:1	10
LoaP (Filter)	1	44	400	9.1:1	13
	2	44	400	9.1:1	12
	3	22	400	18.2:1	8
	4	5.5	400	72.7:1	12
NasR (Ni-NTA)	1	500	500	1:1	20
	2	500	500	1:1	13
	3	39.6	396	10:1	16
	4	28.7	287	10:1	16
LoaP (Ni-NTA)	1	500	500	1:1	25
	2	500	500	1:1	13
	3	50	500	10:1	13
	4	8.1	81	10:1	16

Table III-1: Selection parameters for HT-SELEX.
Table shows the input quantities of protein and RNA for each step of the selection process as well as the number of PCR cycles required for amplification of cDNA for the next round.

After SELEX, we characterized the apparent binding affinity of each pool to target protein using DRaCALA (Figure III-13A). While the original pool of randomized RNA sequences bound His$_{10}$-MBP-NasR, it did so with a similar affinity to the unrelated *dfn* RNA tested previously. Each round resulted in an increase in apparent affinity, culminating in an apparent affinity equal or greater than *nasF* P1P2 sequence. This observation suggested to us that the selection had reached sufficient affinity to characterize a pool of sequences with affinities comparable to native substrates.

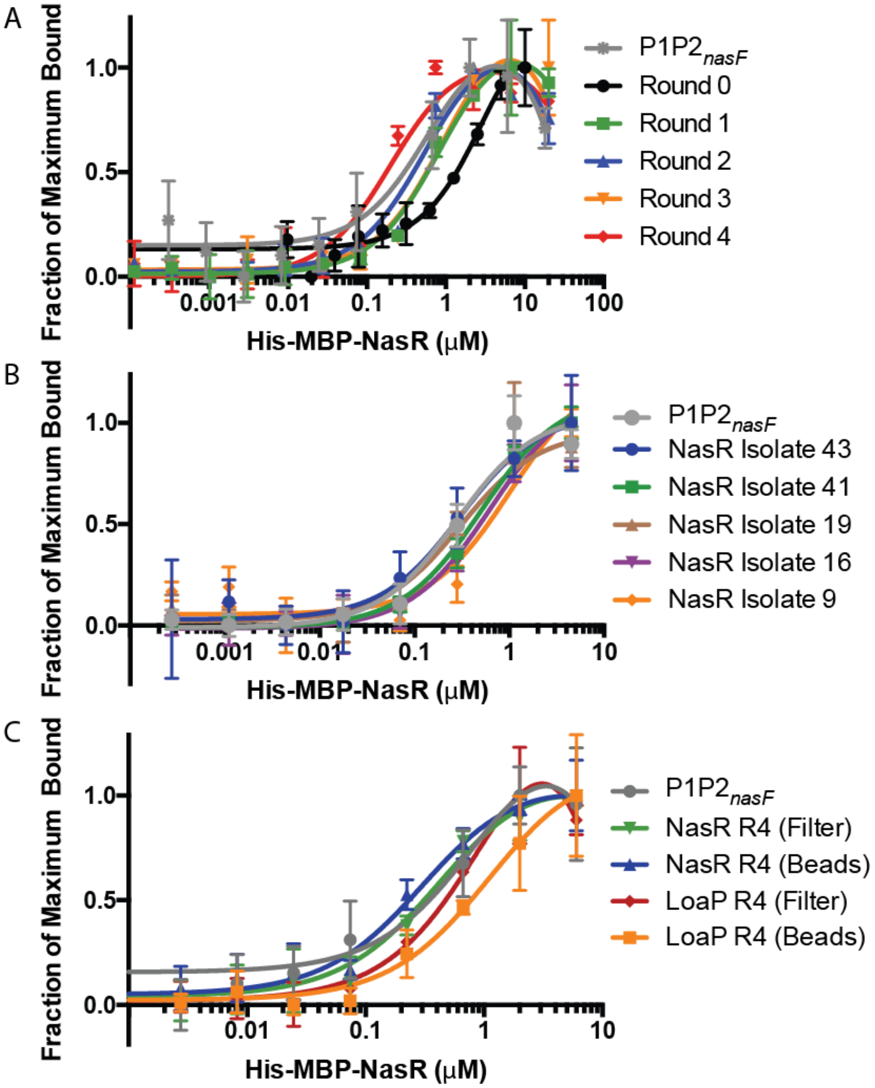


Figure III-13: Saturation binding curves for His-MBP-NasR to SELEX pools.
(A) DRaCALA binding curves for His-MBP-NasR to the P1P2$_{nasF}$ leader region RNA (grey), as well as to the starting SELEX pool (Round 0) and to the output of each of the four rounds of *in vitro* selection (Round 1-4). (B) DRaCALA binding curves for His-MBP-NasR to the P1P2$_{nasF}$ leader region RNA (grey) and to five sequence isolates from Round 4 of the filter-binding NasR selection process. (C) DRaCALA binding curves for His-MBP-NasR to the P1P2$_{nasF}$ leader region RNA (grey), as well as to the Round 4 output RNA from each selection against either NasR or LoaP. **Assistance from Christopher Zhang.*

Sequencing of final pools sequences yields candidate ANTAR-binding RNAs

Determination of NasR-binding sequences by Sanger sequencing

This SELEX approach is a multistep molecular biology process involving repeated steps of PCR, and as such may be vulnerable to problems such as contamination or takeover by unexpected sequences. We planned to thoroughly analyze the resulting sequence pool by high-throughput sequencing. That approach, however, requires investment of considerable resources and time. To get an initial overview of the sequences arising from the selection of random RNA sequences capable of binding NasR *in vitro*, we chose to clone and sequence molecules from the output of the fourth round of SELEX to determine whether the sequence pools contain reasonable sequence content. Additionally, these sequences may also allow for initial evaluation of the consensus patterns for NasR binding. Using the Taq PCR product of the output cDNA, we utilized TOPO TA cloning to generate 60 individual clonal isolate plasmids each from the filter-binding selection with wild-type NasR and LoaP target proteins. Fifty isolates were sequenced by Sanger sequencing, resulting in 33 and 32 valid sequences respectively (Table III-2).

NasR (Filter) Sequences	LoaP (Filter) Sequences
GTGCCGGCGGCTGGGTAAGCGCTCGGTCCA	GGGTGTTAGAGGGGTGTGCG
CACTTTGGCGTCTGGTTGTGGGACGTTT	TAGGGGGGGGGCGTCAGGA
TAGGTGGTTTGGGTGGCTCGCGTTGTTGGT	CAAATGGGAGGGCTGCATG
CACTTACGCCATGAGGTAGCACTTTATTTG	CAGAGGTTGCGGCCGGGCAC
GTCGGCCAGGAGGCGGTGACGGTGCGAGGG	GCGCGGCGGGAGGAGGATGG
GTGCTCGTGGGCAACGAGGCTGTTTGTCC	GGAACTGCGAATGTCGGCCT
TGGTGCGTGGGGCGTCCGGCAGATCGGGGC	ATGCATGAGGGCGGCGGGCC
GAGTGCACAGGAAGCGCTCTTTGGTGGGG	GTGGAGTAACAGCGGGCGTA
CGCGAGCCAAATAACCCTCTGCGCCTTAGG	GTGGAGTAACAGCGGGCGTA
TATAGAGCTGCAGGAGATTGGGGCTACTCC	GCGGGCTGGCAGGAAGGCCA
TAGCTTAGGCACAGGAGGTTGGGCGC	GGGACGTTATATGTTGGAAA
GGAGCCGCGGGCGTGAGTGGGTGCTTGCG	GGTGAGTTCGGGGAGAATGG
TTTGGAAGGGGGCATAGCAACGTGGGGGC	AATGCAATGGGGGGCGGGAA
CAGAGAGAGGGGGGGGGTGAGGGATCGACA	GGAGCTGCGTGCCCAACCGA
AAATTGGTTACTCAGCAGTTATTTCGTCAG	GTGCACGGAGCCGCTCGGAG
GCGCGAGACGAAGCGCAGGGGTAGGGCC	CGCGAACGGCAACTGAGGCTTT
TGGGAGCAGCGGGCGGATCATCGGGGGGGG	TGTACAGGGGGTGCTTCGCC
ATCAGCTGGGCAGCGTGGAGGGGGGGGT	GCGGGGTAGGGTAAGTAGGG
GGCCGTGTGAGCGCACGGATTGCAGCGCAT	TGGGCGCGCCCGTAAGCCGG
AGAGGGGAGGCGGGTAAGTATGGCGGACTG	AGCTCGTTATAAGGCGCGCG
ACGGGGCGGGATTTGCAGTAGTTGGTGTAT	GCGGGGGGTGAAGTCGTGGG
AATTAGCAGCTCACGTAGAACCTTGTTCA	TAGGGGGGGGGCGTCAGGA
TGGGAGGGCTTCGGGGAGAGAAAGGGTTTA	AACACGCCGCACAGCACACA
GTCATTGCAGAGGCTGCCTCAGTAGTCC	GTATAGGGGTTAGTTTCGGC
ATCCCATAGAGGGCGGCTCTGTAGCACGG	CATCTGGCAGGGTGGGTGATT
GGGCTTTTGGAGGAGGCAGGGAGCGCGT	CGTATGCTGAGGTGATTTGG
GGGTGGTCCTGCGCTTCGACATCCATGAGC	GATCATGCGGGAGGGCGCAG
TTATGAAGGCTATTGGGGTGAATCAGCGTT	GGTGTTGCGGCTCATTCACATT
AAGATGCAGTGCCCCTCGCACCCAGCTCGG	GGCGGAGGAGGGGGGCGTTA
AGACGTATAGTGGTTGGGCTGTGCCGCG	GTGTGGGTTTCGTTCAAGGT
TTAGGTACCAGGTATTGGCTAGAGCGCGGT	AGGTGGTGGCGCGGCGGTGC
GCAGCGGCGGCAGCAGTATCAGCGGGCTAA	GTGTTGTGTTACGTTGCGGC
GAACGGGAGGGTTACCGGCAGTGGCCCACC	

Table III-2: HT-SELEX sequenced isolates from filter-binding NasR and LoaP selections.

The first striking characteristic of these sequences is the strong bias toward guanine nucleotides, at 40% across the total dataset, with the median. This bias is at the cost of all other nucleotides (A: 18%, C: 20%, T: 20%), indicating a bias at the RNA level as opposed to a bias in the GC content of the DNA templates. With only this information, we cannot determine whether this bias is due to the true selection process or is instead a byproduct of the starting RNA or the molecular biology of the SELEX process. Of the 33 NasR sequences, none appeared to contain a canonical dual stem-loop motif matching the consensus we previously identified. Eight sequences matching a permissive version of the loop motif (CANNGNNG) can be identified in these sequences, compared to the expected value of 6.5 motifs given a zero-order model (based on single-base composition frequencies) of the same amount of random sequence, an insignificant enrichment of this motif. To us, this indicates one of two likely reasons for this lack of expected enrichment. The first possibility is that this selection did not adequately select for ANTAR binding RNAs, either due to a fault in the SELEX process or by selection for alternate parts of the target His_{10}-MBP-NasR protein. The second possibility is that the SELEX process did enrich for ANTAR binding RNAs, but these bind the NasR protein via sequence that does not fit the biological consensus, possibly because we are selecting only for equilibrium binding and not the full behavior required for antitermination *in vivo*.

As there is evidence that the final SELEX pool binds NasR with increased affinity, we decided to test the latter hypothesis, that these RNA sequence do bind this protein individually, just with a distinct sequence motif. We took five random clones from the Sanger sequence pool and transcribed these RNAs to perform saturation binding DRaCALA (Figure III-13B). In this assay, four of the five RNAs bound His_{10}-MBP-NasR with an affinity comparable to, but lower than, the *nasF* leader P1-P2 RNA sequence. As the starting pool (referred to as Round 0) exhibited dramatically lower apparent binding affinity overall, this indicates that we selected for better-binding RNAs. These RNA sequences may represent ANTAR-binding RNAs with a distinct mode of binding as compared with the biological sequences found in the ANTAR regulons. Complicating this, however, is that this DRaCALA shares many characteristics with the filter-binding SELEX approach, including requiring RNA binding to some component of the full His_{10}-MBP-NasR protein bound to a nitrocellulose membrane. We compared NasR-binding for the output RNA from the final round of each selection against both LoaP and NasR proteins (Figure III-13C). While both selections against NasR bind His_{10}-MBP-NasR with high affinity as expected, the filter-binding selection against the distinct His_{10}-MBP-LoaP protein also shows strong binding, while the Ni-NTA based selection does not. We suspect that the filter-binding selections may have resulted in RNAs which prefer binding to nitrocellulose-immobilized proteins and that the Ni-NTA selection, with more specificity for NasR, may result in more useful sequences.

Cataloging of NasR-binding sequences by Illumina amplicon sequencing

The preliminary Sanger sequencing of the final round of SELEX indicated very high remaining sequence diversity, as no duplicate sequences were identified. We used high-throughput sequencing to thoroughly catalog the sequences capable of binding NasR and to quantify the enrichment throughout the four-round selection process. As part of a larger effort including selections targeting unrelated projects, we started with two pools of randomized sequence, of either 20 or 30 nucleotides in length. We subjected these to four rounds of selection as described above, targeting four proteins in total. Two of these proteins, including NasR, were used for selection using both nitrocellulose filter-binding and Ni-NTA magnetic bead-binding for the enrichment steps. ultimately, we chose to sequence 24 samples: 5x 4 rounds for five proteins (NasR Bead/Filter, LoaP Bead/Filter, NasR R340K Filter), 1x 2 rounds for one protein (RdRP filter), and 2 for the starting RNA pools.

To prepare DNA libraries for high-throughput Illumina sequencing, we used a two-step PCR approach to add Illumina sequence primer, adapter, and barcode sequence to enable multiplex sequencing of all 24 samples. Using primers JRG655 and JRG656, we amplified the cDNA pools made from each RNA sample using the same adapter sequences utilized in SELEX, while also adding additional sequence equivalent to the NEBNext Universal Adapter, with a five base-pair randomized spacer (molecular barcode) separating the Illumina sequencing primer binding site from the start of the SELEX library sequence (177). This random segment is included for the purpose of identifying PCR duplicate sequence reads and to improve the sequencing quality by introducing sequence diversity in the first rounds of sequencing which is important for cluster registration in some Illumina platforms (178). These libraries have an expected read length of 76 and 86 nucleotides for N20- and N30-starting pools respectively. We then prepared sequencing-ready libraries from these amplicons by performing an additional step of limited-cycle PCR using full-length dual-index barcode NEBNext primers, adding a unique pair of i5 and i7 Illumina barcodes to each amplicon. We finally pooled and sequenced these libraries using a NovaSeq 6000 instrument with a 2x150bp run, resulting in approximately 1.5-2 million read pairs each spanning the entire insert for each of the 24 samples.

Reads from the NovaSeq platform have automatically eliminated adapter sequence for reads shorters than the run read length. Accordingly, most read pairs contained only the expected insert sequence. We merged the read pairs using FLASH to improve sequence accuracy and simplify processing (179). The merged libraries contained high-quality sequence, with all rounds prior to round 87 having a mean Phred quality score above 36.7. Read length distributions for most libraries predominately show a spike at the expected read length of 76/86, although some later round 3 or 4 libraries showed an additional peak at 56-57 bases, indicating an enrichment of empty libraries containing only the SELEX adapter sequences. Only three selections show this effect for round 3 or 4, with a maximum of 35% adapter-only sequence, still yielding at minimum 1.3 million quality reads, while most samples contain only a fraction of a percent of adapter-only reads.

Detection of ANTAR-binding RNA enrichment

Due to the high accuracy of the sequencing, we were able to extract the SELEX sequence from each read by identifying the exact forward and reverse adapters, and selecting the intervening sequence, keeping sequences with a minimum length of 15 bases. As our analysis of the 33 Sanger sequencing isolates from indicated a strong bias for high guanine content, we immediately wanted to know whether this was a general characteristic of our selection pools, or a particular quirk of the round 4 libraries we sequenced. In fact, the G-rich nature of the output sequences is common to most of the samples, in particular the starting N20 and N30 libraries are approximately 45% guanine, indicating that this bias may be a characteristic of the commercial RNA library and not a function of our SELEX approach (Figure III-14A). The libraries do vary in G-content, with some of the later round selections containing a minor reduction to as low as 39-40% guanine, perhaps indicating that these libraries are a result of weak selection against G-rich RNA sequences (Figure III-14B).

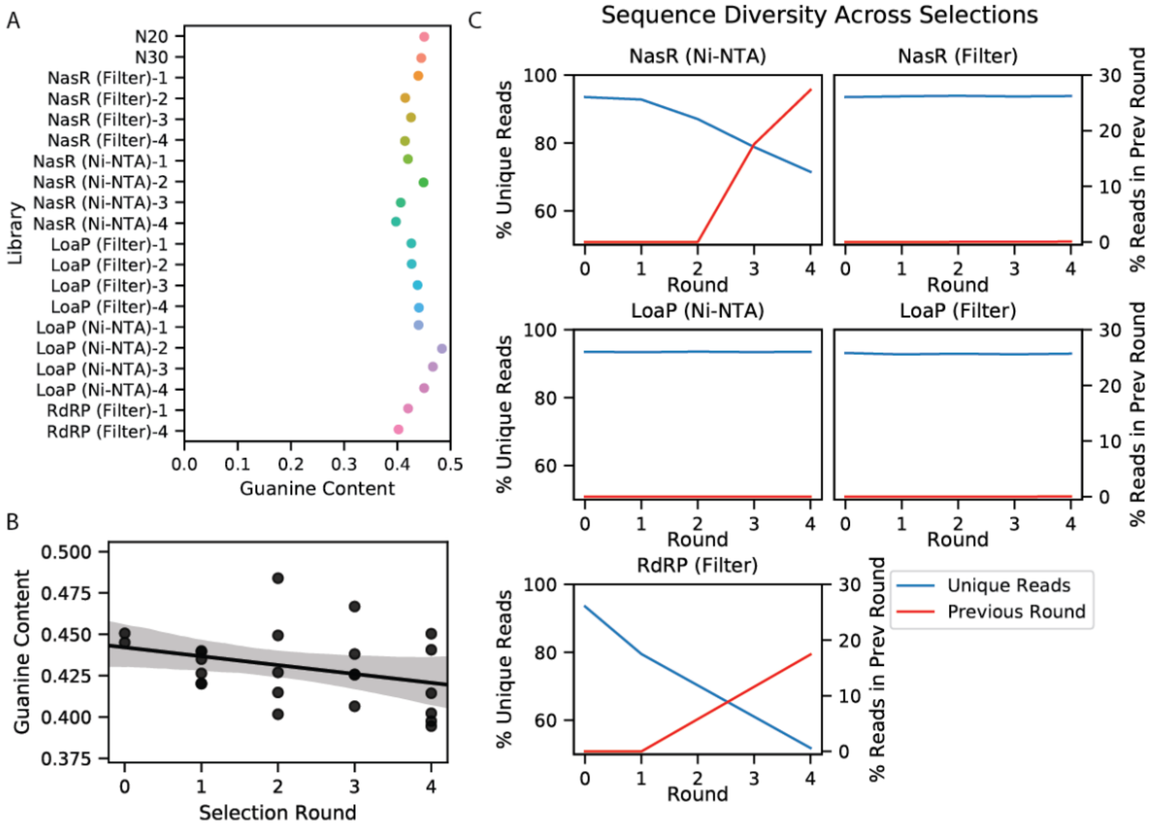


Figure III-14: HT-SELEX Sequencing reveals strong guanine-bias in the starting libraries and high residual sequence diversity in several selections.
(A) A strong bias toward guanine nucleotide content is present in all HT-SELEX libraries, including the starting N20 and N30 pools. (B) There is a slight overall decrease in guanine nucleotide content as selection proceeds through multiple rounds. (C) Individual sequence counts for each sample in individual selections were checked to identify the percent of each library which was composed of unique sequences, as well as the number of individual sequences which were present in the previous round for that selection.

Ultimately, the goal of any SELEX approach is to enrich for high-binding sequences. Although HT-SELEX experiments occasionally contain a greater depth of sequence diversity than sequencing depth, masking observable sequence enrichment, we quantified the number of sequences found in multiple, sequential, rounds of selection to determine if enrichment values could be calculated (Figure III-14C). Only two selection experiments display sufficient enrichment to observe a substantial proportion of sequences in multiple rounds, including one experiment targeting NasR. As our pool-binding experiments indicated that the later rounds of selection do have notably increased affinity for the target proteins, this lack of observable sequence enrichment may indicate that the remaining selections retain sequence diversity beyond the level our approximately one-million read sequencing can detect individual sequence-enrichment for.

To detect lower levels of sequence enrichment, we turned to K-mer enrichment analysis, which has been used for HT-SELEX in the past (160). As our samples already contain high levels of guanine, k-mers with very high guanine content are abundant but present at comparable levels to that expected from the single-nucleotide frequencies (Figure III-15A). We expect these extremely high-guanine K-mers would not contribute to high-affinity binding RNAs, and are likely selected against during SELEX. To this end, we tracked the ratio of the rate of individual 8-mers against that expected from the nucleotide content of that library. In the starting pools, these largely show the expected prevalence, with the exception of the extreme 8-mers with seven or eight guanines (Figure III-15B). In three of the five selections, including those two showing sequence enrichment in later rounds, there is a substantial decrease in prevalence for the 8-mers with more extreme guanine content, while in the other two there is, if anything, an increase in extreme guanine-content. This indicates that two selections may have experienced very little effective selection, while a third may have experienced selection, but simply to a degree not detectable in individual sequence enrichment, demonstrating the value of k-mer analysis for tracking HT-SELEX enrichment in early rounds.

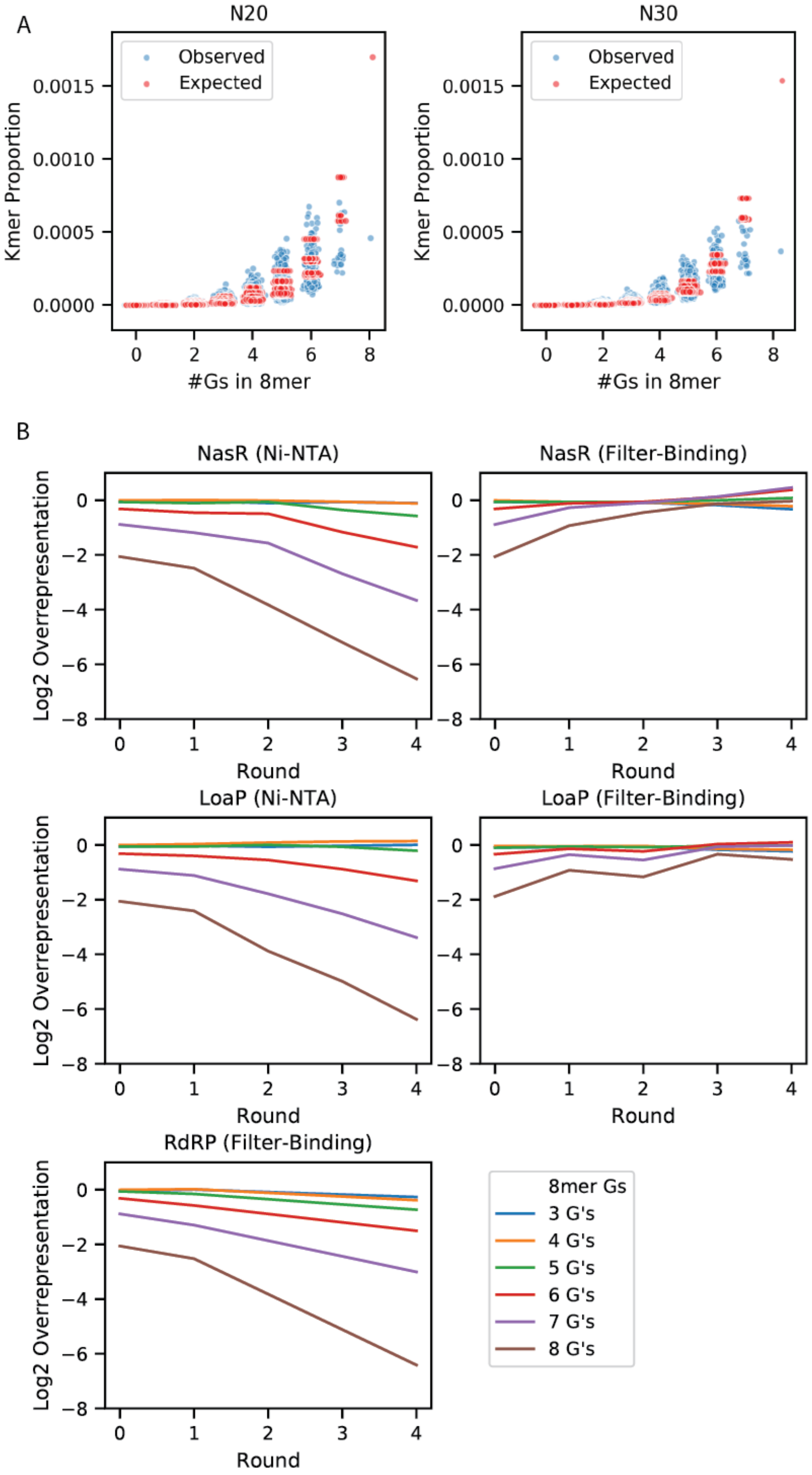


Figure III-15: HT-SELEX selections with low sequence diversity also exhibit no selection against guanine-bias. (A) The proportion of the total 8-mer pool of each starting library for each 8-mer containing a specific number of guanine residues (blue). In red is the predicted abundance of that 8-mer given the single base composition of the library. (B) The average log2 ratio of the abundance of each set of 8-mers containing three or more guanines relative to what would be expected from the single base composition of the library is shown for each round of selection.

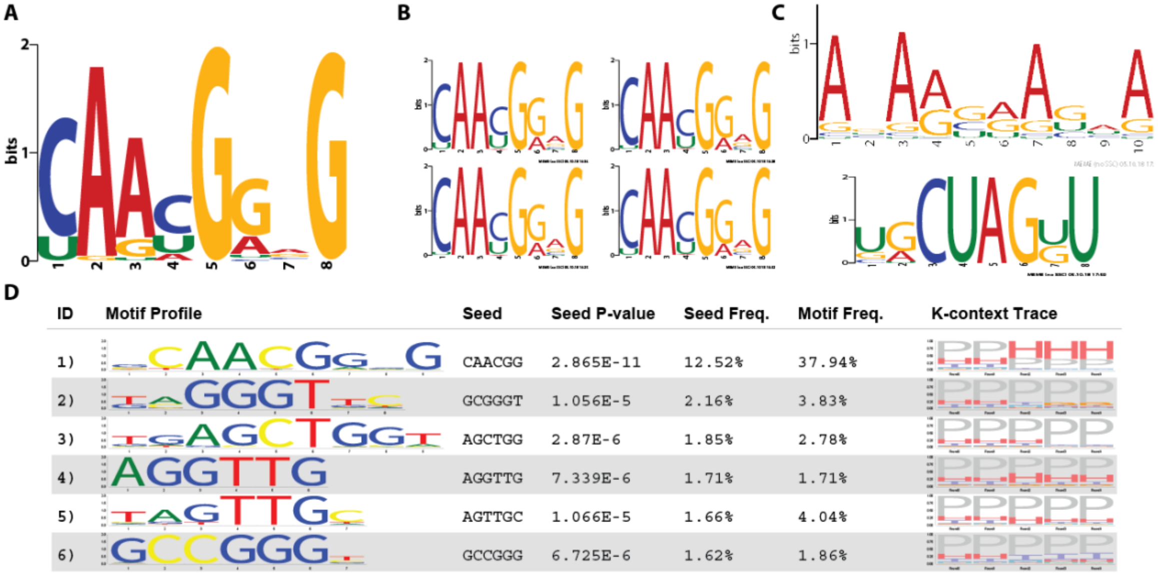
Detection of enriched ANTAR-binding candidate motifs

With perhaps the most effective selection experiment, at least by the metrics available, being one of the His_{10}-MBP-NasR target selections, we wanted to identify ANTAR-binding RNA to generate consensus patterns, mirroring that found in bacterial genomes, with the *in vitro* HT-SELEX results. Initially, we wanted to identify sequences corresponding to the known two stem-loop motif described earlier, and then expand this search to identify potential additional motifs, which may actually be present, undiscovered, in nature.

MEME

Perhaps one of the most-used tools for identify sequence motifs is the MEME algorithm, used to find recurring short, ungapped motifs in a set of sequences. In the HT-SELEX field, MEME is often described as attractive but not practical, due to the inefficient scaling to large sequence databases (154, 160, 180). The version 5 update of the MEME suite of algorithms introduces a number of updates to MEME, with two important changes relevant to HT-SELEX (181). First, MEME now exhibits much improved algorithmic complexity on large inputs by splitting the database and using a subset for the initial motif search step. Second, it includes a new objective function allowing for scoring of motifs based on enrichment compared to a baseline sequence database. These changes significantly improve the applicability of MEME to analysis of large HT-SELEX datasets.

We initially ran MEME version 5.0.2 on a random subset of 250,000 sequences from the fourth round of selection on His_{10}-MBP-NasR, along with a random subset of 100,000 sequences from the starting N30 pool in differential enrichment and any-number-of-motifs mode, looking for three motifs of width=8-10. This should maximize the ability of MEME to identify the native CAAHGRHG/CAAHGDVG loop motifs. This resulted in identification of a 9-mer motif, sCAAWGRnG with an E-value of 2.0e-7514 (Figure III-16A), quite similar to the native RNAs, at approximately 83,000 sites in the 250,000 sequences. Using FIMO to identify this motif in the complete 1,754,515 sequence set, the motifs is present in 758382, or 43% of the round 4 output at a p-value < 0.01. Two instances of this motif are present in approximately 16% of total sequences, increasing the likelihood that these sequences might represent true dual stem-loop RNAs. These results are highly reproducible, with similar motifs found in each instance when a different, disjoint, random set of 100,000 input sequences was used (Figure III-16B).



ID	Motif Profile	Seed	Seed P-value	Seed Freq.	Motif Freq.	K-context Trace
1)		CAACGG	2.865E-11	12.52%	37.94%	
2)		GCGGGT	1.056E-5	2.16%	3.83%	
3)		AGCTGG	2.87E-6	1.85%	2.78%	
4)		AGGTTG	7.339E-6	1.71%	1.71%	
5)		AGTTGC	1.066E-5	1.66%	4.04%	
6)		GCCGGG	6.725E-6	1.62%	1.86%	

Figure III-16: MEME and AptaTRACE detected motifs.
Illustration of the common motifs detected by MEME or AptaTRACE in sequence data for the final round of Ni-NTA based selection against NasR. (A) The most common motif detected by MEME in a random sample of 250,000 sequence reads. (B) The most common motif detected in four disjoint random samples of 100,000 sequence reads. (C) The additional next most abundant motifs detected MEME in the largest sample, found in 6% and 1.2% of reads respectively. (D) Motif profiles detected by AptaTRACE as enriched in the four rounds of the Ni-NTA based selection against NasR.

Slightly over half of the sequences in round 4 are not detected by the near-canonical MEME motif, and most do not have two sites, which are required for specific binding. To determine whether this is due to legitimately low-binding sequences and the low number of SELEX rounds, or whether there are suboptimal consensus sequences or alternate sequences, we removed the motif hits from the database and repeated motif identification. We did this in two ways, first, by removing all sequencing containing any significant hits to the motif, and second, by selectively removing the identified motif from sequences containing only one instance. When searching a random sample of 100,000 of the sequences which did not contain significant motif hits, a single common motif of vAnbGGYH was found with an E-value of 3e-381 in 68,526 sequences, with strong consensus at the 2 and 5 positions, corresponding to the most conserved A and G residues in the ANTAR terminal looks, perhaps indicative of a minor contribution to binding affinity of suboptimal terminal loops in most of the remaining sequences. Re-detecting this motif with FIMO as previously done, 26% of original no-hit sequences contain this more degenerate motif, with only 4% containing two instances. When searching the portions of sequences originally containing but not including the hit to the original motif, no motifs are found in more than 9% of the remaining sequence, suggesting that the original single-motif sequences in fact only contain the one conserved element. Together, these suggest to us that it is unlikely that a secondary consensus motif for ANTAR recognition exists.

AptaSUITE

In addition to general tools applicable to identifying motifs in sequences, the HT-SELEX field has developed several specialized tools (161, 182, 183). One very prominent set of tools is distributed as AptaSUITE (184). This suite contains tools for parsing (185) and clustering (154) HT-SELEX reads, as well as tools to predict structural features and identify conserved sequence-structure motifs (180). AptaTRACE, in particular, is optimized to analyze multiple rounds of HT-SELEX, combining specialized k-mer metrics with structure context in a manner similar to an extended RNAContext (186). We applied a standard pipeline of these tools to the 5 rounds of the His_{10}-MBP-NasR target selection. Only one motif profile was found with a frequency higher than 5% of the total pool sequences in the final round of selection (Figure III-16D), representing a highly similar motif to that found by MEME. There was no significant resemblance between additional motifs found by MEME and AptaTRACE (Figure III-16CD) although the purine-rich motif identified by MEME may bear some resemblance to motif profiles 2 and 4 detected by AptaTRACE.

Validation of NasR-binding of selected sequences

Initially, we took five plasmid isolates from the filter-binding NasR target selection TOPO cloning and used these to prepare transcribed and labeled RNA for saturation binding analysis with DRaCALA. Of these five isolates, none contained detectable dual stem-loop motifs, although four did contain sequences similar to the P1 or P2 terminal loop sequence (9, 16, 19, and 41), although with no potential base-pairing to anchor them. Despite the lack of either canonical stem-loop structures or significant conserved motifs, four of these isolates bound His_{10}-MBP-NasR with comparable apparent affinity to the native *nasF* P1P2 sequence (Figure III-13B), although as mentioned above, the filter-binding selection against LoaP resulted in a final pool with comparable apparent affinity (Figure III-13C). Additional testing with additional isolates from the more effective bead-binding selection which match consensus patterns will need to be tested, both with DRaCALA as well as alternate assays such as EMSA or fluorescence anisotropy to validate these results.

Discussion

The ANTAR protein domain had previously been known to promote antitermination in specific leader regions by binding specific RNA sequences, but the particular nature of the RNAs required were unknown. Through a combination of manual sequence gazing and biochemistry, our lab, along with our collaborators, identified the shared features common to the known ANTAR-regulated transcripts required for protein binding and antitermination. These consist of a two tandem stem-loop RNA motif with conserved loop and closing base-pair residues overlapping an intrinsic transcription terminator, a unique feature among RNA-binding antiterminators (138, 187, 188). In most cases, the second stem-loop of this motif overlapping the transcription terminator results in two mutually-incompatible RNA structures, one containing both P1 and P2, which we predict is dominant when bound to ANTAR domains, and the other containing P1 and the terminator stem-loop, forming and terminating transcription in the absence of ANTAR-binding. This domain, and the proteins containing it, are widely conserved in bacteria, and we found similar tandem RNA elements in many of those bacteria containing ANTAR-domain proteins. While our initial searches were hampered by low specificity, addition of a screen for overlap with intrinsic terminator sequences revealed the widespread distribution of potential ANTAR antiterminator elements in Gram-positive ANTAR-domain containing genomes. As many bacteria do not rely extensively on intrinsic transcription of termination, we believe alternate forms of ANTAR regulation may exist utilizing the same tandem stem-loop motif, instead coupling RNA-binding activity to control of Rho-termination or translation initiation.

As the conservation of the RNA elements associated with ANTAR regulation is quite strong, it is likely the interaction between this domain and the RNAs are as well. Three structures exist of proteins containing ANTAR domains, structures for *P. aeruginosa* AmiR in complex with AmiC (140), *K. oxytoca* NasR (145), and the *M. tuberculosis* response-regulator protein Rv1626 (163). All three of these structures are in inactive states, and none contain RNA. In addition, while both NasR and AmiR crystallized as dimers containing ANTAR-domain interactions, the interacting ANTAR domain surfaces are completely different. While the ANTAR domains of all three contain extremely similar structures, they contribute little to understanding how ANTAR might interact with RNA. We targeted a variety of conserved residues in the ANTAR domains of both NasR and *E. faecalis* EutV with the goal of determining which residues might be involved directly in RNA binding. Six of these residues were required for activity of both proteins. Three of these (E160/E358, K149/K347, and R168/R366) we believe are likely to be in charge-charge interactions with other ANTAR residues responsible for ANTAR folding. Two of these mutants (Y164/W362 and M178/M376) are both non-polar mutants with side chains within 3.5 Å and likely internal in the NasR structure. When the remaining mutant (R142/R340) is converted to an alanine in either EutV or NasR activity is lost, and conversion to a lysine residue does not restore activity as the other three charged residues do in this category. We believe that this residue may be involved in direct charge-

charge interactions with RNA, and, as the arginine-to-lysine mutations do not restore activity, may be important for ligand specificity.

Finally, *in vitro* selection of RNA sequences which bind to NasR resulting in RNA libraries with comparable binding to the *nasF* leader region RNA. Interestingly, the selection performed without nitrate in the binding reactions resulted in the greatest enrichment, yet did not result in sequences capable of binding without nitrate. We suspect this occurred as a result of NasR protein forming an RNA-binding form at a low frequency, and that this much lower effective protein concentration increased the stringency of selection allowing enriched sequences to be detected after only four rounds of selection. Sequence analysis of this library revealed only a single enriched motif perfectly matching the motif identified from our previous analysis. This result increases our confidence that this specific tandem stem-loop motif is the sole ligand for ANTAR domains similar to those found in AmiR, NasR, and EutV. These three proteins, however, have high sequence similarity in comparison to the whole ANTAR domain family (Chapter 4), and may not be representative of the entire family. We will extend the results of this RNA sequence identification, as well as the analysis of the sequence composition of the ANTAR domain in an extended analysis of this domain family in Chapter 4.

Methods

Protein expression, RNA preparation, EMSA

Data shown for biochemistry of EutV and EutV mutants is from (152).

Experiments largely performed by Arati Ramesh and Sruti DebRoy.

Identification and analysis of ANTAR target RNAs

Previous work by Tsoy et al. listed 84 bacterial genomes that contain putative ethanolamine utilization genes (167). Eighty-three of these reference genomic sequences were available at the NCBI RefSeq genome database and were downloaded and combined in a searchable database. An input RNA alignment consisting of 17 ANTAR substrate sequences was manually created. These sequences represented *eut* intergenic regions from *E. faecalis*, *Clostridium* and *Listeria sp.* where we manually identified the ANTAR-binding dual hairpin motif. This motif consisted of 3 bp stem regions with hexa-nucleotide terminal loops. The distance between the two stems was variable. The alignment was used as input for Infernal (168). Potential RNA hits identified by Infernal were scored according to the level of similarity to the consensus sequence alignment and sorted according to bit scores. An arbitrary bit score cut-off of 10 was applied in order to catalog putative ANTAR substrate hits; this value was high enough to include all of the ANTAR substrates from *Enterococcus*, *Clostridium*, and *Listeria* included in the seed alignment. This cut-off was retained for all further searches. Also, all hits located within coding regions were eliminated from any further analysis. To create a template alignment for the catalog of ANTAR substrate hits, the portions of the hits that included the stem 1 and stem 2 regions were separately subjected to the comparative sequence alignment software LocaRNA (189) and RNAalifold (190). This consensus pattern alignment was then used for an additional Infernal search. A catalog of NCBI annotated bacterial genomes was retrieved from the NCBI Genomes FTP and filtered to remove all but one genomic sequence from each available strain. Again, an arbitrary bit score cutoff of 10 was used to catalog putative ANTAR substrate hits. To further filter these hits, they were screened for the presence of putative intrinsic transcription terminator hairpins using TransTermHP. For this step, we utilized the default model with an adjustment to allow for a larger (up to 26 nucleotide) terminal terminator loop. The only hits that passed both filters were those that were arranged such that the second stem-loop (P2) overlapped with a putative intrinsic terminator that was also oriented in the correct direction.

Mutagenesis of NasR ANTAR domain

Beginning from expression vector pMG1130 (175) containing His_{10}-MBP-NasR, we created single-residue mutants using Q5 Site-Directed Mutagenesis (New England Biologicals). Using primer pairs from primers CZ001-CZ042, we amplified pMG1130 by PCR with Q5 DNA Polymerase, and circularized these products as described in the standard kit protocol. We transformed these reactions into XL10 Gold chemical competent cells. Using standard protocols, we miniprepped and sequenced isolates and screened for accurate mutations by

Sanger sequencing. Miniprepped plasmid DNA was transformed into *E. coli* strain T7 Express LysY/Iq (New England Biologicals) for protein expression.

Purification of His_{10}-MBP-NasR and mutants

BL21(DE3) cells containing an expression vector (pMG1130 (175)) for His_{10}-MBP-NasR were cultured in 2xYT and expression induced at A260= 0.5 with 1mM IPTG at room temperature for 18 h. The cells were pelleted and resuspended in Resuspension Buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM $MgCl_2$, 2 mM β-ME, 5% glycerol, 1 mM PMSF, 2 U DNase). Cells were lysed with 0.5 mg/mL lysozyme on ice for a total of 30 minutes. After cell disruption by bead-beating for small samples or sonication for large samples, the lysate was clarified by centrifugation at 12,000xG.

For small samples, the supernatant was passed over 200 µL Ni-NTA resin (Thermo scientific), followed by 6 column-volumes (CV) wash buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM MgCl2, 25 mM Imidazole, 5% glycerol). The protein was eluted in 3 fractions of 1 CV elution buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM MgCl2, 250 mM Imidazole, 5% glycerol). The protein was eluted in 3 fractions of 1 CV elution buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM MgCl2, 250 mM Imidazole, 5% glycerol). For large samples, the supernatant was passed over a 5 mL HisTrap column (GE Healthcare) on an FPLC at 5 mL/min, followed by 5 column-volumes (CV) wash buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM MgCl2, 25 mM Imidazole, 5% glycerol). The protein was eluted with 5 CV elution buffer (50 mM HEPES pH 8.0, 150 mM

NaCl, 1 mM MgCl2, 250 mM Imidazole, 5% glycerol) with fractions collected every 3 mL with concentration monitored by UV absorbance.

Eluted protein was dialyzed against dialysis buffer (50 mM HEPES pH 8.0, 150 mM NaCl, 1 mM MgCl2, 5% glycerol) for three 2 hour buffer exchanges. All steps were performed either on ice or at 4°C. The purity of NasR was judged by SDS/PAGE followed by Coomassie-staining. The mutant NasR proteins were additionally subjected to native PAGE and Coomassie-staining to identify potential protein-folding problems.

Fluorescence anisotropy binding analysis

For a description of the methods used for saturation binding analysis of NasR and mutant NasR proteins by fluorescence anisotropy, see the methods in Chapter 1.

Experiments performed with the assistance of Christopher Zhang.

Mutagenesis and *in vivo* assays for EutV antitermination

Data shown for antitermination by EutV and EutV mutants *in vivo* comprises material from an unpublished manuscript (152). Experiments are covered in more detail in the doctoral dissertation of Margo Gebbie (175).

Experiments performed by Margo Gebbie.

In vitro selection for RNA aptamers for RNA-binding proteins

Two separate *in vitro* selections for RNA aptamers to NasR were performed. One selection utilized filter-binding of the His-MBP-NasR protein to a nitrocellulose filter for separation (153), while the other used binding of the His-tagged protein to HisPur Ni-NTA magnetic beads (191). Both selections began from the TriLink N30 RNA library for *in vitro* selection.

For filter-binding, we performed an initial negative selection by passaging approximately 500 µL of 800 nM RNA library in binding buffer (50 mM HEPES, 100 mM KCl, 1mM $MgCl_2$) over a nitrocellulose filter. For positive selection, we then mixed this RNA with 500 µL of 600 nM His-MBP-NasR protein and incubated for 30 minutes. We conducted selection by passing this solution over a 25 mm nitrocellulose filter at approximately 1 drop/second. We then washed the filter with 3 mL of binding buffer in the same fashion. We eluted our RNA and protein by placing the filter in a microcentrifuge tube with 100 µL binding and incubating at 95C for 5 minutes, repeating this step once. (*Selection performed by Christopher Zhang)*

For Ni-NTA selections, we performed an initial negative selection by incubating 100 µL of 5 µM RNA library (500 pmol) and 5 µM His-MBP with 30 µL of HisPur Ni-NTA beads in binding buffer (50 mM HEPES, 100 mM KCl, 1mM $MgCl_2$) for 5 minutes at room temperature, then placing on a magnetic stand to separate the beads and retaining the supernatant. For positive selection, we then mixed the RNA with 10 µL of 50 µM His-MBP-NasR and 30 µL of Ni-NTA magnetic beads and incubated for 10 minutes at room temperature. We then placed the tube on a magnetic stand to separate the beads and remove the supernatant. We then resuspended the beads in 500 µL of binding buffer, and repeated the washing three times. Finally, we transferred the beads to a new tube and placed on a magnetic stand, removing the supernatant. We eluted the bound protein and RNA by incubating the beads with Zymo RNA Clean & Concentrator Binding Buffer with 250 mM imidazole for five minutes, placing the beads on the magnetic stand, and retaining the supernatant. (*I designed the selection protocol and selection was performed by students in Engineering Biosensors FIRE Stream under the supervision of Catherine Spirito)*

For each cycle, from this step onward both selections used a similar protocol. We cleaned the eluted RNA using the Zymo RNA Clean & Concentrator kit, eluting in 25 µL water. We then reverse-transcribed this RNA to make cDNA using the standard SuperScript IV (filter-binding) or ProtoScript II (Ni-NTA) protocols with 25 µL reactions and 5 µL of input RNA and the selection reverse primer (TriLink). We then performed cycle-course PCR using 2 µL of this cDNA in 40 µL reactions using Taq DNA polymerase and the selection forward and reverse primers (TriLink), removing 2 µL of PCR reaction every two cycles from rounds 10 to 30. We ran these aliquots on a 2% TAE agarose gel to determine the cycle at which DNA was visibly amplified and the number of cycles required for PCR amplification. We then performed a 400 µL Taq PCR reaction using the number of cycles determined by the cycle-course experiment. We cleaned up these reactions using the Zymo DNA Clean and Concentrator columns. Finally, we transcribed 450 ng of this PCR-amplified template DNA using T7 RNAP in a 25 µL reaction, and again cleaned up this RNA using the Zymo RNA cleanup kit. This RNA was then used to repeat the next round of selection. Subsequent rounds did not include a negative selection step, and reduced the protein-to-RNA ratio. In total, four selection rounds were performed.

High-throughput sequencing of SELEX RNA pools

For each round of selection, we began from the cDNA prepared during the selection process from each round. We amplified 5 ng of cDNA from each rouch of each selection in standard 50 mL reactions using Q5 DNA Polymerase (Initial denaturation: 98 ºC for 30s, Cycle, 15 rounds: 98 ºC for 10 seconds, 60 ºC for 15 seconds, 72 ºC for 20 seconds, Final extension: 72 ºC for 2 minutes). These PCR products were cleaned using DNA Clean and Concentrator-5 columns (Zymo Research). These reactions were quantified by UV absorbance and normalized. We performed a second PCR reaction with 100 ng total template DNA for each reaction and unique primer pairs from the NEBNext Dual-Index Barcode Set 1 (Initial denaturation: 98 ºC for 30s, Cycle, 4 rounds: 98 ºC for 10 seconds, 65 ºC for 75 seconds, Final extension: 65 ºC for 5 minutes). We again purified these PCR products using DNA Clean and Concentrator-5 columns. We then quantified each one of the 24 libraries using UV absorbance as well as fluorescence using the QuantiFluor dsDNA Dye (Promega) along with a standard curve in a Molecular Device SpectraMax M5 plate reader. We then mixed the individual libraries at an equimolar ratio using the geometric mean of the concentrations determined by the two methods. We then submitted these for sequencing utilizing 5% of a NovaSeq 6000 2x150bp run, resulting in 44,421,626 total read pairs.

Analysis of HT-SELEX sequencing data

Demultiplexed sequencing reads were checked for quality using AfterQC and FastQC (192, 193). No important issues were noted for any samples. For all analyses except AptaSUITE, read pairs were merged using FLASH (179) and the N20- or N30-derived sequences were extracted used custom Python scripts which identified the SELEX library adapters in each sequence, then filtered out sequences <16 nt or >35 nt and stored the intervening sequence as FASTA files.

For MEME, random samples of sequence reads were obtained using Python scripts. All MEME searches in this text were performed using the RNA alphabet, background frequencies using a Markov-order of 1 (dinucleotide frequencies), and the differential enrichment objective function with version 5.0.2.

For AptaSUITE, the raw sequence reads were input directly to AptaPLEX, which handled identification of the SELEX library adapters and extraction of sequence. AptaCLUSTER was run with LSHDimension=25, EditDistance=5, LSHIterations=5, KmerSize=3 and KmerCutoffIterations=10000. AptaTRACE was run with KmerLength=6, FilterClusters=True, and Alpha=10.

Chapter 4: Targeted Sequence Analysis of Protein Subfamilies Containing Domains Specialized for Antitermination

Introduction

In both Chapters 2 and 3 we identified novel transcription antitermination regulatory systems. In the former, we identified a new specialized NusG paralog, LoaP, capable of promoting antitermination in targeted RNA transcripts at intrinsic terminators. We initially identified this protein by analysis of bacterial genomes to identify potential antiterminator proteins genetically linked to antibiotic biosynthesis operons. In this analysis, we noticed a small cluster of similar genomes, *Bacillis*, *Brevibacillus*, and *Paenibacillus*, that contained NusG proteins immediately upstream of antibiotic polyketide synthase (PKS) pathways. Although these could all be highly similar pathways that were horizontally transferred while retaining the same arrangement, as is common with PKS systems (194–196), we wished to know how extensive this association might be and where else LoaP proteins might be found. This exploration ultimately revealed that the LoaP protein is widespread in Firmicutes, Actinobacteria and Spirochaetes, and is oftentimes associated with specialized metabolite gene clusters or polysaccharide biosynthesis operons. This analysis additionally evolved into a comprehensive analysis of all bacterial NusG proteins, to explore their relationship with specialized metabolite pathways and to identify clusters of potentially new and distinct NusG paralogs.

Both the core NusG and RfaH family of antiterminators have been thoroughly studied at many levels, with genetic (58, 69), biochemical (20, 33, 58, 197), and structural data available for both proteins alone and in combination with RNA polymerase (61, 63, 64, 78, 198) as well as with other transcription elongation factors (21, 83, 199). These studies have implicated many individual protein elements or residues in specific NusG-family activities, including RNAP-binding, non-template DNA strand interaction, upstream DNA interaction, and interactions with transcription control proteins, such as NusE (S10), NusA, and Rho. Other paralog groups have been studied to a much lower extent, with very few examples of direct characterization of any other paralog group. In this chapter, we attempt to determine whether any expectation for the presence of these activities in different paralog subgroups can be determined from a comparison of sequence conservation of different regions of the NusG family.

Our analysis of two ANTAR-domain containing antiterminator proteins in Chapter 3 resulted in an improved understanding of how ANTAR family regulators promote antitermination in their target operons. Several outstanding questions remain, including, but not limited to the following. How do the diverse sensory domains maintain proper interaction with the extremely small and highly conserved ANTAR domain, which must itself maintain very specific RNA-binding contacts? Second, although the AmiR, NasR, and EutV proteins have a very strong preference for the specific RNA motif described in Chapter 3, are these representative of all ANTAR-containing proteins? Do all ANTAR proteins have regulons containing the established RNA substrate? In this chapter, we will attempt to address parts of these questions through a large-scale comparison of sequence and domain conservation between members of the ANTAR family.

Comprehensive phylogenetic analysis of bacterial NusG proteins reveals additional specialized paralog groups associated with specialized metabolites.

LoaP is a member of the NusG-homolog subfamily and is broadly conserved

While *loaP* plays a specific role in the expression of two of the three important polyketide antibiotic gene clusters in *B. velezensis*, it may be present and performing similar functions in other organisms. To investigate this possibility, we initially searched for close homologs of LoaP protein sequence using phmmer from the HMMER3 software suite, manually checking the genomes of resulting hits for the presence of a core *nusG* gene in addition to a *loaP* homolog (200). The most highly homologous protein sequence hits—specifically, those found in genomes that also featured an additional core nusG gene—were generally distributed among other species of *Bacillus*, *Brevibacillus*, and *Paenibacillius*. Other closely related homologs were found among Clostridia, including *Clostridium* and *Thermoanaerobacter* species. Many high-scoring hits, however, appeared to be the sole *nusG* gene in the genome, likely to encode the core NusG protein.

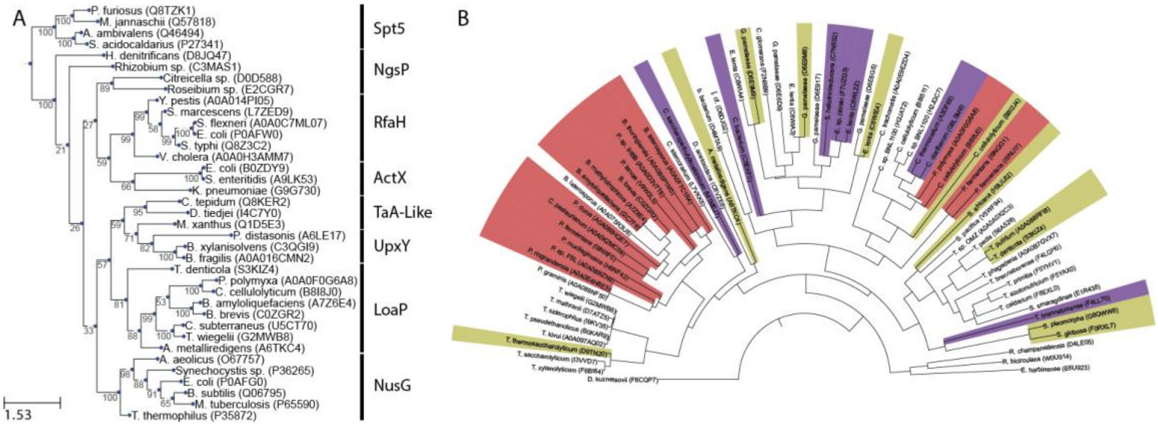


Figure IV-1: LoaP represents a distinct group of NusG specialized paralogs and is commonly associated with large biosynthetic gene clusters.

(A) Phylogenetic tree of curated NusG homolog sequences labeled with predicted subtype. (B) Subset of large-scale phylogenetic analysis showing LoaP homolog sequences. Background shading represents association of gene sequences with large gene clusters. Dark grey sequences are found near PKS or NRPS gene clusters. Medium grey sequences are found near polysaccharide gene clusters. Light grey sequences are found near other types of antiSMASH gene clusters. Unlabeled sequences were not found nearby an antiSMASH predicted gene cluster, although some appear to be next to long stretches of coding sequences in one direction.

To gain insight into the relationship between the LoaP proteins and the larger NusG family, we combined a few LoaP sequences with selected examples of other NusG family member proteins—Spt5, NusG, RfaH, ActX, UpxY, and TaA—and constructed a structure-assisted multiple sequence alignment and maximum-likelihood phylogenetic tree (85) (Figure IV-1A). The LoaP sequences formed a monophyletic clade separate from NusG and from the other specialized paralogs. We began to differentiate the top scoring hits between LoaP and other NusG subfamilies by sequentially adding the sequences to the reference alignment and reconstructing the phylogenetic tree with the additional sequence to determine where the protein fit in the reference tree. The position of the protein determined the preliminary subfamily assignment.

To obtain a very broad picture of the NusG family paralogs we extracted all sequences of proteins containing the NusG N-terminal domain from the Uniprot complete database (14,435 sequences in total). To reduce the number of nearly identical protein sequences while maintaining the majority of sequence diversity, we utilized sequence-similarity clustering to limit putative core NusG sequences (determined above) to a set of at most 60% identical sequences, and NusG paralogs to at most 95% sequence identical (1205 total sequences). We constructed a large-scale multiple alignment using the accurate multi-domain progressive alignment algorithm of the MAFFT software, and constructed a maximum-likelihood phylogenetic tree of the trimmed alignment using RAxML (Figure IV-2). The underlying topology of the large tree matched very closely the topology of the small reference tree. We rooted the tree by assigning the root to the internal node maximizing the amount of Archaeal (Spt5) sequences. Protein subfamilies were labeled by assigning a subtype to the monophyletic group formed by the most recent ancestor of all of the curated protein examples found in the small tree analysis. In general, additional protein sequences corresponding to putative LoaP homologs were identified in Bacilli and Clostridia classes, as well as the Coriobacteriia class of Actinobacteria and in a variety of Spirochaetes. Not all proteins were assigned a type. Some subtrees were found adjacent to known paralog groups and exhibit distinct characteristics, and may represent additional subtypes.

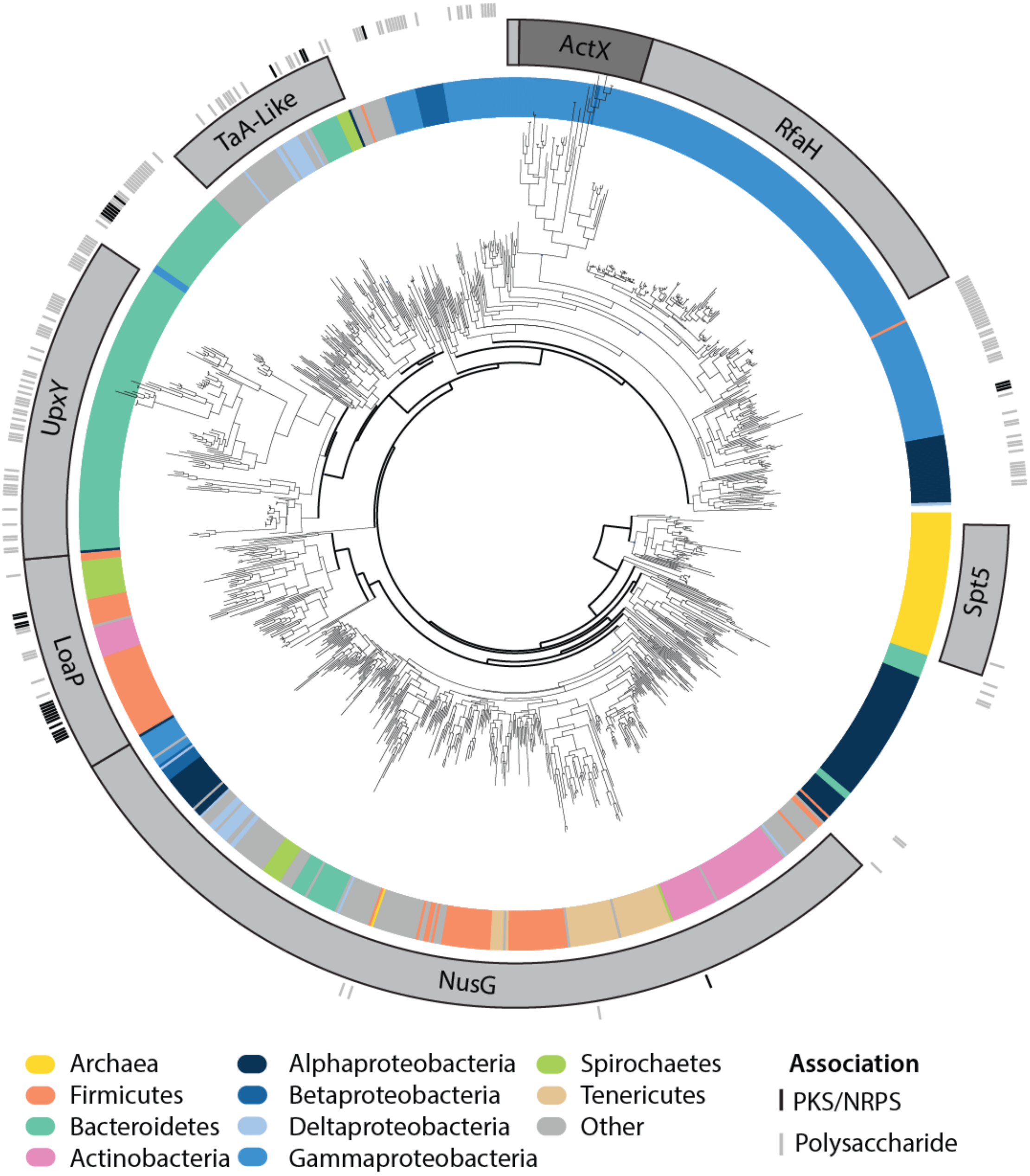


Figure IV-2: Large-scale phylogenetic analysis of NusG family proteins reveals several subclasses of specialized paralogs.

A large-scale phylogenetic tree composed on 1205 representatives of NusG homologs. The bacterial phylum (or class for Proteobacteria) of the organism containing each protein sequence is represented by color by the inner ring. Subtrees formed from the most recent ancestor of curated subgroups are labeled with grey boxes in the middle ring. Tick marks representing the association of particular sequences with PKS or NRPS gene clusters (black) or polysaccharide gene clusters (gray) are found in the outer ring.

LoaP is associated with polysaccharide and secondary metabolite biosynthesis gene clusters

Given that *B. velezensis loaP* is located adjacent to the difficidin gene cluster, we hypothesized that other examples of *loaP* subfamily genes may be present adjacent to gene clusters that they regulate. We initially manually surveyed a selection of close *loaP* homologs to look for large gene clusters in the genomic region. Almost all *loaP* homologs were located adjacent to large gene clusters containing either polyketide synthase genes or sugar-related enzymes consistent with polysaccharide synthesis operons. We then took a more rigorous approach and collected all 1,205 NusG family members from the large phylogenetic tree and, using the antiSMASH pipeline, searched their surrounding genomic sequence for putative specialized metabolite gene clusters within five kilobases of the NusG family gene sequence (Fig. 6B, 7). Most *loaP* sequences were found immediately adjacent to the large gene clusters encoding specialized metabolite synthesis or polysaccharide synthesis. Interestingly, the different NusG paralog groups showed distinct patterns of gene cluster association. UpxY, as previously shown (93) is very commonly associated with Bacteroidetes polysaccharide gene clusters. Yet there also appeared to be a separate UpxY/TaA-like cluster found in Bacteroidetes associated with gene clusters for either polysaccharides or specialized metabolites. RfaH and ActX appear to never be associated with antiSMASH-identifiable gene clusters, although there are some RfaH-like clusters found in both Alpha- and Gamma-proteobacteria that are associated with long gene clusters, as well as a subgroup found largely in Alphaproteobacteria independent of putative gene clusters.

Comparison of functionally important residues among NusG paralogs

There is significant conservation domain composition of NusG proteins among prokaryotes. Both bacteria and archaea possess NusG proteins with a single N-terminal NusG and C-terminal KOW domain, occasionally with an additional taxa-specific insert domain in the middle (62, 201, 202). The secondary and tertiary structure of NusG proteins is also strongly conserved, with core bacterial and archaeal NusG proteins, as well as Gammaproteobacterial RfaH, sharing highly similar structures and conserved interactions with RNA polymerase (63, 64, 75). For this reason, we believe it likely that all NusG paralogs share similar structures and interactions with RNA polymerase.

As illustrated by the important distinctions between *E. coli* NusG and RfaH, small differences in sequence can result in completely different regulatory activity *in vivo (81)*. We consolidated much of the available information concerning specific functional interactions with transcription elongation complex (TEC) partners to allow comparison of these important residues among NusG paralogs (Figure IV-3). These activities include direct interactions with RNAP (63, 64, 75) , interactions with non-template strand DNA in the transcription bubble (69, 78), interactions with elongation factors NusA (84), NusE (83), and termination factor Rho (199, 203), and finally interactions between the two domains of RfaH (198).

We also mapped these residues onto a model of *E. coli* NusG to allow visualization of the many interaction zones on NusG (Figure IV-4).

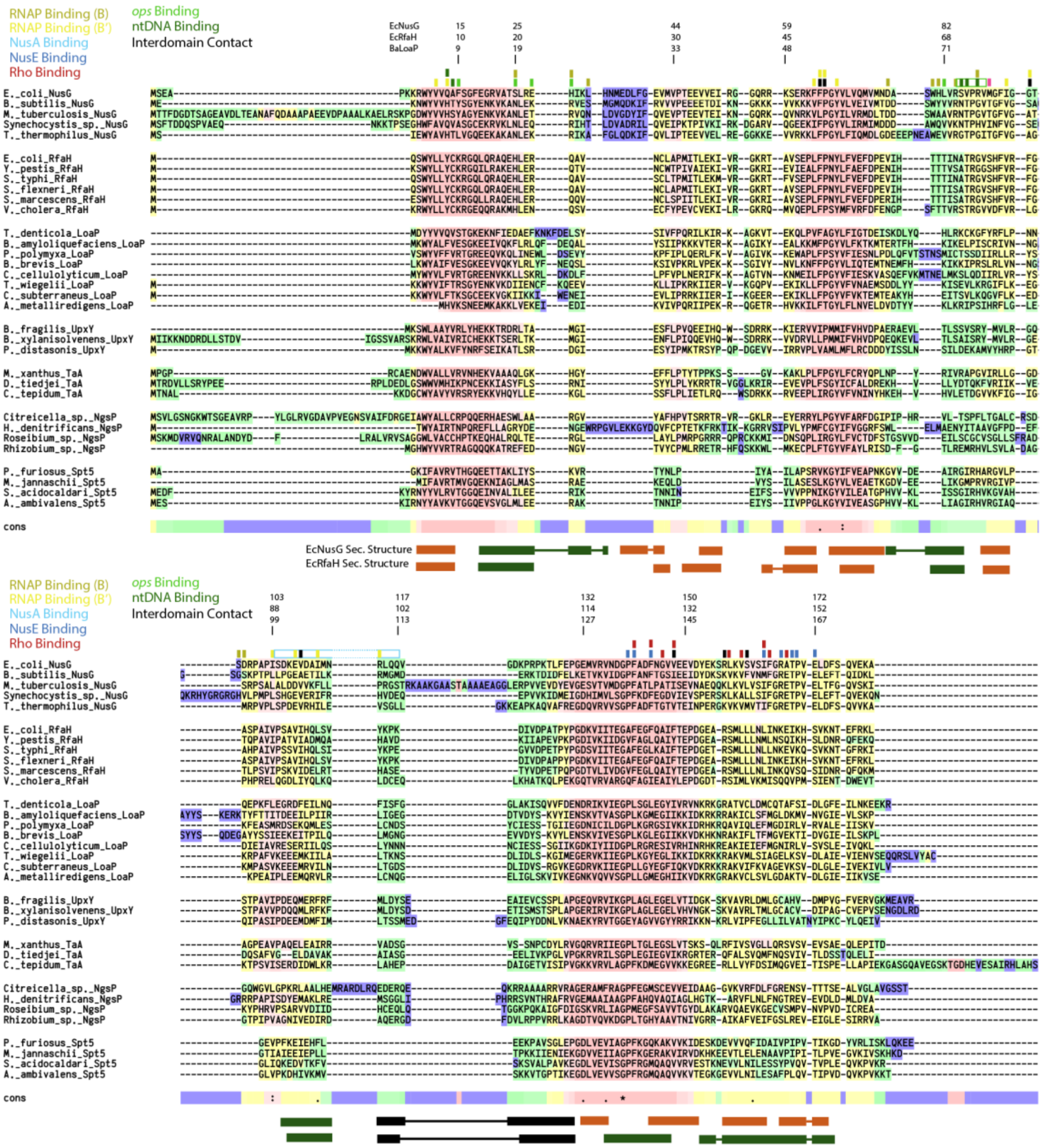


Figure IV-3: Multiple sequence alignment of NusG family protein sequences reveals conserved differences between subtypes of specialized paralogs.

Alignment contains 33 protein sequences containing representatives of multiple subtypes of NusG paralog aligned by T-COFFEE using the "accurate" alignment method utilizing available crystal structure data. Residues are colored according to T-COFFEE consistency score. Also shown are secondary structure diagrams representing the secondary structure from PDB structures of E. coli NusG and RfaH (Orange represents beta-strands and green represents alpha-helices, black represents relatively unstructured inter-domain linker). Colored tick marks above the alignment indicate residues implicated in specific interactions with other protein or nucleic acid components of the transcription elongation complex.

The interaction between NusG and RNAP is the most likely conserved interaction among all NusG paralogs. The residues involved in this interaction, however, are not strongly conserved, instead falling into general classes of polar (residues corresponding to EcNusG V11, F64, F65, P66, I93, I103, and I111) or non-polar (for example residues corresponding to EcNusG H29, R84, R88) residues moderately conserved among all paralog groups. Some residues involved in direct RNAP interactions are not conserved among different groups, often residues also involved in non-template DNA interaction. EcNusG Q13, for example, is a polar glutamine or histidine in core NusG also involved in non-template DNA binding, but an aromatic tyrosine or phenylalanine in RfaH and LoaP families respectively. RNAP-interacting residues also differ between NusG and specialized paralogs at ecNusG S25, H29, G95, and D97.

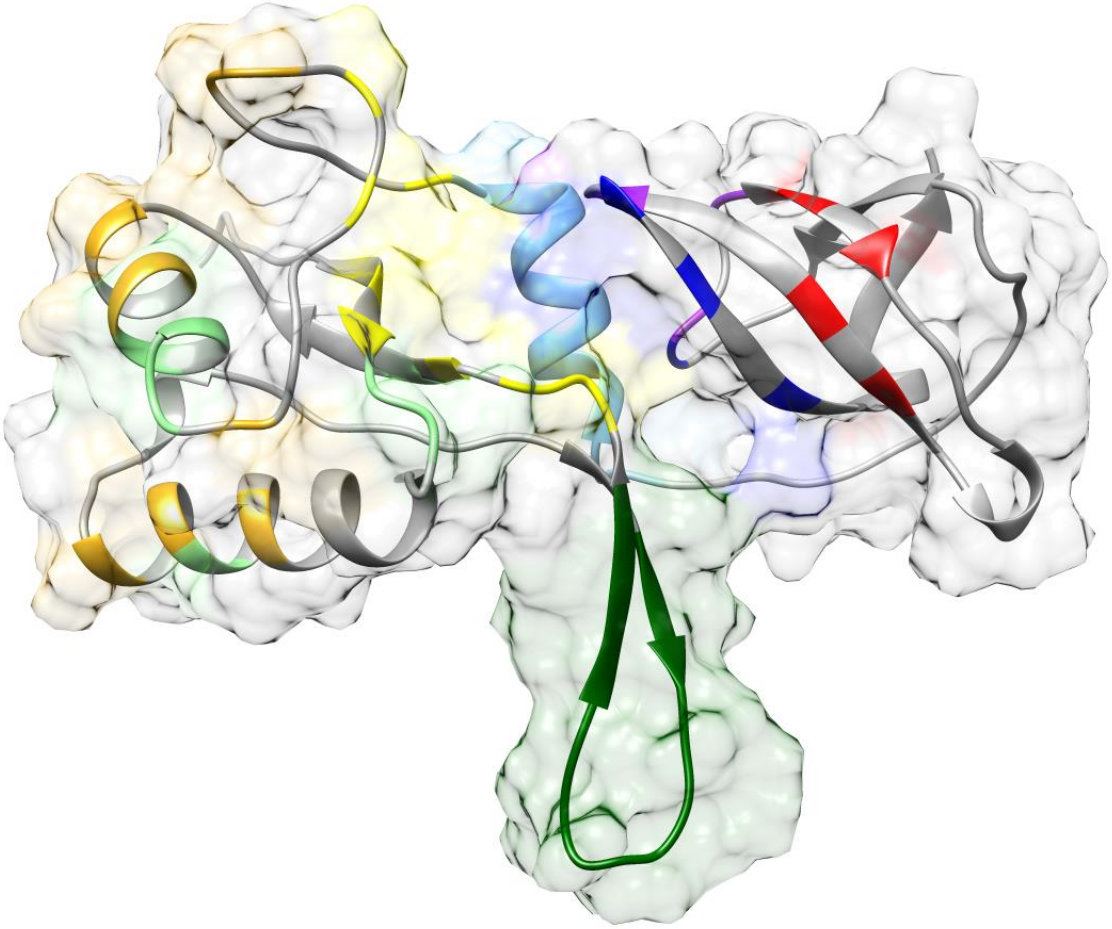


Figure IV-4: Model of *E. coli* NusG showing known interactions with different partners. Structure model generated by I-TASSER of full-length ecNusG. Individual residues or regions reported to interact or be required for functional interaction with protein or nucleic acid partners are colored. Interactions with RNA polymerase by *E. coli* NusG (ecNusG) are shown in gold (Beta subunit) or yellow (Beta' subunit). Interactions with non-template DNA by *B. subtilis* NusG or *ops* DNA by ecNusG are shown in light green. The beta-hairpin predicted to interact with upstream DNA is labeled in dark green. A helix required for ecNusG interaction with ecNusA is shown in light blue. Known interactions by ecNusG with NusE (S10) are shown in blue while interactions with Rho are shown in red, with overlapping interactions displayed in purple.

NusG paralog interactions with non-template DNA may also be likely, even unavoidable, given that the conserved RNAP-binding position of NusG directly spans the ntDNA interface. These interactions have been demonstrated for both *E. coli* RfaH and *B. subtilis* NusG (bsNusG), although DNA sequence preferences for both differ (20, 69). These residues often diverge considerably between RfaH and NusG. A three amino acid region corresponding to EcNusG

13-15 (Q, A, F) that has been shown in RfaH (Y, C, K) to interact with *ops* DNA and has been implicated in ntDNA recognition by bsNusG (H,T,Y). This region exhibits group-specific conservation in LoaP (F, V/T, E/R) and Spt5 (R, V, T), but is more variable in other subgroups, perhaps indicating group-specific ntDNA binding preferences. Not all *ops*-binding residues exhibit this pattern, EcRfaH H20, R23, Q24 and residues 70-74 are strongly conserved in RfaH but exhibit more variability in all other subgroups.

A few residues in each domain of RfaH have been determined to interact in the autoinhibited form of RfaH (198). In the NTD these often overlap with residues involved in binding to RNAP, which help prevent RfaH-RNAP binding (ecRfaH F51, P52, F81). The conservation of these residues in the NusG family largely fall in to two groups, those that are not strongly conserved outside of RfaH (ecRfaH F81, V93, L142) and those that are conserved among the whole family (V93, F130, and R139). EcRfaH F51 and P52 are strongly conserved among RfaH, NusG, and LoaP subgroups and are more variable in others. These observations do not significantly support a similar form of domain interaction in other NusG paralogs.

Interactions with other regulators of transcription elongation (NusA, NusE, and Rho) largely occur in the C-terminal KOW domain. Both NusG and NusA are generally associated with TECs and are also capable of forming a direction interaction between the NusG N-terminal domain (ecNusG 103-117) and the NusA AR2 domain (84). This interaction involves residues also required for RNAP binding, and was predicted to be mutually incompatible, implicating this interaction in the complicated combinatorial effects of these alternately antagonistic or synergistic elongation factors. This region does not exhibit significant conservation outside of two residues involved in RNAP binding, although the final 4-5 residues are not present in any Spt5 proteins, potentially reflecting the lack of this elongation factor in archaea.

The final set of interacting residues consist of those between NusG and NusE or Rho. NusG is important, but not required, for Rho termination. NusG may act to recruit Rho to the TEC (126) and help trigger the transition of Rho from an open, loading, state to the active, closed conformation capable of translocation (199). The interaction between NusG and Rho occurs at the L1 (EcNusG 139-144) and L2 loops (164-167) of NusG-CTD. This binding site is overlaps heavily with that observed in a NusG:NusE complex (83), and an interaction between NusG and NusE in the TEC is capable of preventing Rho termination (21). The L1 and L2 loops are strongly conserved in core NusG proteins, and extremely strongly conserved in Beta- and Gamma-proteobacteria (199). The L1 loop region also exhibits subgroup-specific conservation, with the L1 loop containing one conserved sequence in RfaH and a distinct, but shared, conserved sequence among most other specialized paralogs. The L2 region is not as widely conserved, although each subgroup appears to contain a distinct, individually conserved sequence. This may be reflective of the important of the KOW domain in functional specialization of the different NusG subgroups. Core NusG must maintain the NusA interaction, while RfaH acts to exclude Rho and also uses residues in this same region to interface the N- and C-terminal domains. LoaP, which largely appears in bacteria lacking significant Rho termination activity, may utilize this region for other functions, perhaps RNA binding. When considering that hypothesis, it is important to note that as in other contexts the KOW domain is nucleic-acid binding (75, 204, 205).

Analysis of ANTAR RNA-binding domain reveals subgroups of genetic regulators.

Phylogenetic analysis of the ANTAR domain and associate proteins

The ANTAR domain is present in a broad range of bacteria including Actinobacteria, Firmicutes, and Proteobacteria, although perhaps most prevalent in Actinobacteria. Only a few ANTAR domain-containing proteins have been characterized and we wished to obtain a broader understanding of the phylogenetic and domain context these domains are found in, in a manner similar to our analysis of the NusG family. These two protein families have distinct characteristics which necessitate a different approach. Where prokaryotic NusG proteins have a near-universal two-domain composition, ANTAR domains are found in proteins with multiple, distinct, domain compositions (Figure III-1). The NusG family also has several well-established subgroups with known functional differences and an established outgroup (Archaeal Spt5), whereas ANTAR proteins have only been distinguished by their sensory domain function.

	GAF	Resp. Reg.	AmiR -Like	PAS	NIT	Other	None	
Actinobacteria	535	22	0	88	0	57	204	906
Alphaproteobacteria	0	51	66	0	2	0	2	121
Betaproteobacteria	0	48	20	1	10	0	9	88
Deltaproteobacteria	2	10	0	0	0	0	5	17
Gammaproteobacteria	0	60	20	0	29	0	16	125
Firmicutes	0	189	0	0	0	0	45	234
Other	29	48	1	0	0	0	5	83
Total	566	428	107	89	41	57	286	

Table IV-1: ANTAR-domain containing proteins with specific addition domains in each bacterial phylum.
ANTAR-domain containing proteins found in the 60% identity clustered phylogenetic tree. The top scoring HMM for each sequence was identified as the associated domain.

	GAF	Resp. Reg.	AmiR-Like	PAS	NIT	Other	None	
Actinobacteria	2522	1103	0	404	0	339	637	5005
Alphaproteobacteria	0	554	190	0	6	0	10	760
Betaproteobacteria	0	286	71	1	86	0	37	481
Gammaproteobacteria	0	262	38	0	127	0	48	475
Deltaproteobacteria	15	31	0	0	0	0	6	52
Firmicutes	0	574	0	0	0	1	226	801
Other	98	135	2	0	0	0	12	247
Total	2635	2945	301	405	219	340	976	

Table IV-2: ANTAR-domain containing proteins with specific addition domains in each bacterial phylum.
ANTAR-domain containing proteins found in the entire UniProt dataset. The top scoring HMM for each sequence was identified as the associated domain.

To analyze this family, we initially took a similar approach, obtaining a comprehensive list of ANTAR-domain containing sequences (7,830 total) from UniProt. We then extracted the sequence corresponding to the ANTAR domain for each sequence, as well as for sequences matching the Pfam domains for Response_reg, PAS, GAF, and NIT domains as well as a custom HMM representing the AmiR-like pseudo-response regulator domain. We used MMSeqs2 (206) to cluster the ANTAR domain sequences to a level of 60% sequence identity, resulting in 1,549 cluster centroid sequences, to which we added five previously studied ANTAR sequences (NasR, AmiR, Rv1626, and *E. facaelis* and *L. monocytogenes* EutV). We aligned each extracted domain from these sequences individually, and built maximum likelihood phylogenetic trees with both the concatenated alignments as well as the ANTAR domain alone (Figure IV-5). For each sequence in the final tree, we indicate the bacterial phylum and the primary associated sensory domain in colored rings around the unrooted tree.

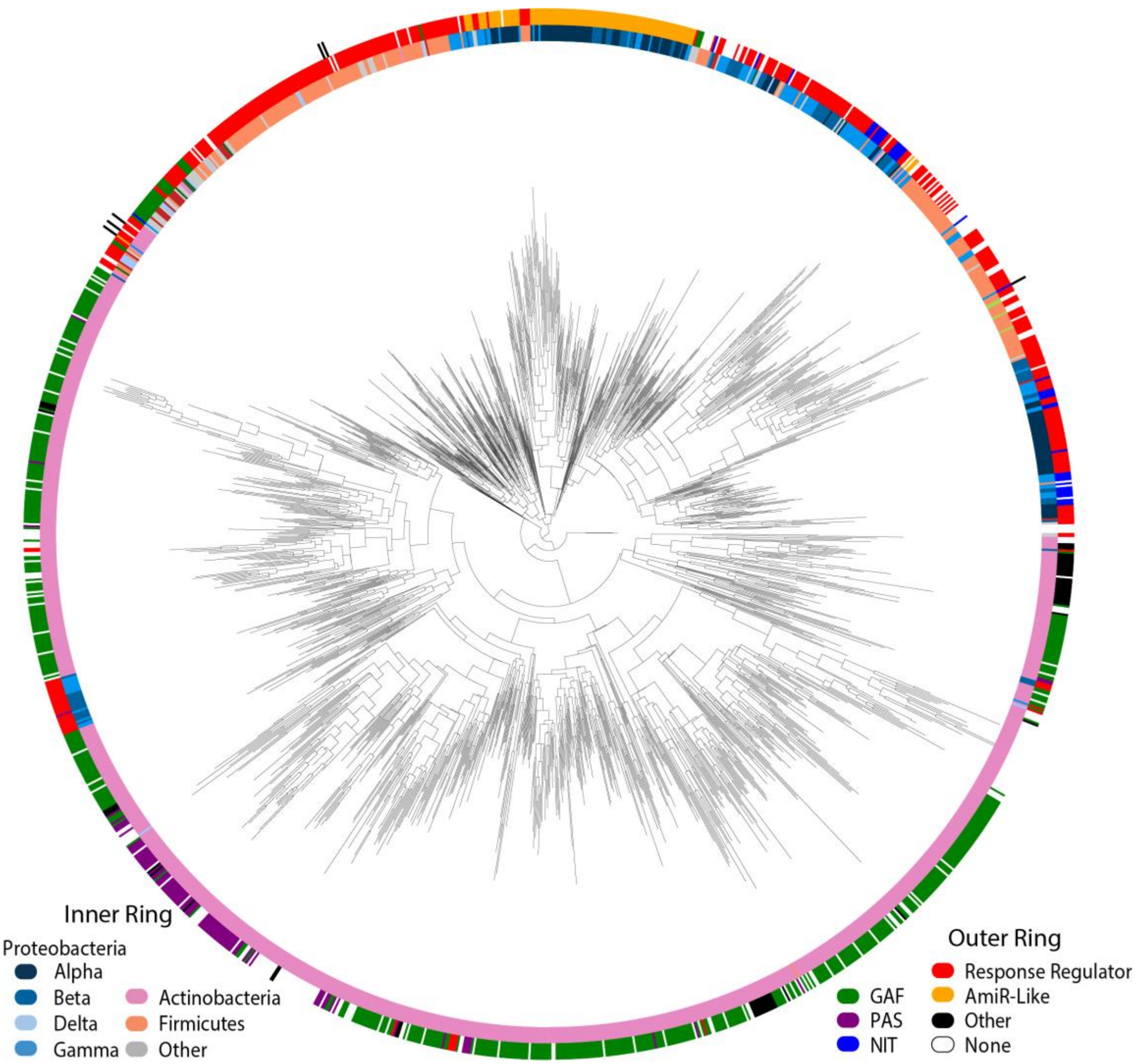


Figure IV-5: Phylogenetic tree of 1,549 ANTAR domains.
Unrooted phylogenetic tree illustrating subgroups of ANTAR domain-containing proteins. The inner ring represents the bacterial phyla the sequence is present in. The outer ring represents the associated sensory domain for that protein sequence, if any.

Even without including additional domain sequences, the ANTAR domain sequences themselves generally cluster according to taxonomic location and domain composition. Additionally, there appears to be a general separation between ANTAR sequences present in Actinobacteria and those present in Firmicutes and Proteobacteria, although one proteobacterial cluster of response regulators appears in this generally actinobacterial putative clade in both trees. Interestingly, the *M. tuberculosis* Rv1626 and *S. coelicolor* Sco2013 (174) response regulators appear near AmiR in a different section of the tree. The four most prevalent ANTAR-associated domains are response regulator and AmiR-like pseudo-response regulator domains, GAF domains, and PAS domains. In general, GAF and PAS domains appear associated with ANTAR domains in the broadly actinobacterial group, while response regulators are found in Firmicutes and Proteobacteria. GAF- and PAS-associated ANTAR domains are almost exclusive to Actinobacteria, while response-regulator domains are found in all ANTAR-containing phyla and are the only domains identified in Firmicutes. AmiR-like and NIT domains appear almost exclusively in Proteobacteria. Other domains found in this analysis include STAS and SpoIIE domains, representative of domain compositions 6 and 7 from Figure III-1, also largely restricted to Actinobacteria.

Discussion

The ANTAR RNA-binding domain is prevalent in bacteria, although largely restricted to three of the largest bacterial phyla, Actinobacteria, Proteobacteria, and Firmicutes. While this domain regulates transcription attenuation in the studied examples, phylogenetic analysis reveals that the majority of ANTAR domains are present in Actinobacteria associated with GAF and PAS domains, distinct from the two-component response-regulator and NIT domains which have demonstrated antitermination activity. GAF domains are widely prevalent and often associated with nucleotide signaling, sensing molecules like cAMP or cGMP (207). PAS domains, representing a related structural fold, are incredibly widespread ligand-binding domains sensing diverse ligands and coupled to many output domains (208, 209). The wide variety of ligands sensed by PAS domains, including heme, flavin cofactors, carboxlates, divalent metals, and fatty acids, makes predictions about the role of these PAS-ANTAR proteins difficult. Perhaps, the majority of ANTAR domains are found in Actinobacteria, in protein subfamilies distinct from those represented by AmiR, NasR, and EutV and may have entirely different regulons or even different modes of regulation beyond transcription antitermination.

NusG paralogs putatively involved in antitermination have been identified in a variety of bacteria, including but not limited to Alpha-, Gamma- and Delta-proteobacteria, Bacteroidetes, and most recently Firmicutes, Actinobacteria, and Spirochaetes (85, 94). Of the general transcription elongation factors, only NusG is found in all three domains of life, suggesting its function is important in all organisms. Therefore, essentially all bacteria encode for a core NusG protein, while archaea and eukaryotes encode for a similar protein, Spt5 (79). As a result, all NusG family proteins share core conserved sequence and structure features (79).

Although analysis of the paralogs supports grouping them within the overall NusG family, each sub-grouping displays significant sequence diversity, with some subgroups displaying very limited overall sequence identity despite sharing remarkably conserved structural elements (94). This large-scale phylogenetic analysis utilized structural modeling to efficiently align specialized NusG paralogs with limited sequence similarity, and focused on comprehensively covering the diversity of paralog sequences without restriction to the known subgroups. The resulting phylogenetic tree (Figure IV-1) confirmed that each set of NusG paralogs forms its own distinct group, separate from core bacterial NusG and archaeal Spt5, while also revealing a few new candidate subgroups (81, 94). It is likely that each subgroup will be defined by specific sequence differences. Indeed, a number of characteristic differences between sequences—such as between RfaH or UpxY and core NusG—have been identified as being important for the distinct activities of those specialized paralogs (81, 100, 198).

As NusG paralogs were found in a variety of distinct genetic contexts (81), it was important to systematically identify associations between these genes and potential target pathways. Overall, they were found in diverse genomic contexts, with some positioned alone, at the beginning of complex polysaccharide or secondary metabolite gene clusters, at the end of operons, or in unique contexts (81, 94). For example, NusG paralogous sequences from Betaproteobacteria and Bacteroides are located in or near large polysaccharide pathways. TaA and LoaP sequences are generally present in or near large polyketide biosynthesis pathways, which suggests they share a broad relationship to secondary metabolites (81, 94). Indeed, there appears to be a general association of NusG specialized paralogs with polysaccharide biosynthesis gene clusters, and to a lesser extent polyketide synthase gene clusters. In fact, of all the paralog groups, only the Gammaproteobacterial RfaH and its related ActX gene sequences were not frequently identified near or in these classes of gene clusters (94).

There also appear to be several subgroups of NusG paralogs with interesting genomic association and evolutionary distribution, but that have not been characterized or named. For example, a group of sequences closely related to RfaH and found in Alpha-, Beta-, and Gamma-proteobacteria is oftentimes associated with polysaccharide gene clusters. Similarly, at least two more uncharacterized and unnamed putative groups of sequences are consistently associated with polysaccharide and polyketide biosynthesis gene clusters. From this, it can be tentatively speculated that NusG specialized paralogs evolved as regulators of these long operons (polysaccharides and secondary metabolite biosynthesis genes) and became further specialized into RfaH in Gammaproteobacteria. Finally, an additional set of paralog sequences in Alphaproteobacteria was not found in a consistent genomic context, and remains unnamed. ultimately, the evolutionary relationship between all these different NusG paralogs remains unclear, as bootstrap support for early branches after divergence from core NusG is low, likely due to the extensive sequence divergence in this family. Elucidating the true history of this family may require different approaches, integrating more information about the structural changes and sequence insertions and deletions during evolution of the NusG paralogs. However, it is already clear that the NusG family of proteins is widely used in bacteria as specialized transcription elongation regulatory factors, and that they regulate expression of fundamentally important pathways, albeit through largely unexplored molecular mechanisms.

Methods

Bioinformatic analysis of NusG family protein sequences

A representative phylogenetic tree was constructed by obtaining, via manual curation of the literature, 3-8 sequences representing each NusG family subgroup. These sequences were aligned using T-COFFEE in Expresso mode using 3D structure information for alignment (210). A maximum-likelihood phylogenetic tree was constructed using RAxML with automatic evolution model selection and 1000 rapid bootstraps (211).

A comprehensive list of prokaryotic NusG protein sequences was obtained by searching the full UniprotKB protein sequence database with the NusG N-terminal domain PFAM HMM using HMMER 3.0 with the default PFAM score cutoff and filtering for proteins from prokaryotic organisms (170). Sequences were assigned a preliminary subgroup classification by addition to the small representative alignment and rapid tree construction using FastTree, and assigned to the nearest subgroup (212, 213). Near-duplicate sequences were removed by UCLUST clustering preliminary core-NusG sequences to 60% sequence identity, and non-core-NusG sequences to 95% sequence identity, resulting in 1205 sequences (214). These were aligned using MAFFT in E-INS-I local domain alignment mode, and a phylogenetic tree was constructed using RAxML with automatic evolution model selection and 1000 rapid bootstraps.

The representative genomic sequences for each protein were obtained from the UniProt database and the surrounding region for each sequence was searched for biosynthetic gene clusters using antiSMASH 3.0.5 with additional

ClusterFinder search (108). Each protein coding sequence was checked for proximity to detected gene clusters and this information mapped onto the large-scale phylogenetic tree. Labeling of NusG paralog subtypes was performed by identification of the sub-tree corresponding to the most recent common ancestor of curated representative sequences.

Bioinformatic analysis of ANTAR family protein sequences

A comprehensive list of ANTAR-containing protein sequences was obtained by searching the full UniprotKB protein sequence database with the ANTAR domain PFAM HMM using HMMER 3.1 with the default PFAM score cutoff (170). We identified associated domains by detecting the presence of all Pfam domains in Pfam-A in addition to a custom HMM created using jackhmmer beginning with the AmiR N-terminal domain (170). Near-duplicate sequences were removed by MMSeqs 2clustering preliminary core-NusG sequences to 60% sequence identity, and non-core-NusG sequences to 95% sequence identity, resulting in 1549 sequences (206). We obtained the ANTAR domains alone as well as several common associated domains by extracting the HMM hit envelope for each domain found. Each set of domain sequences was aligned using MAFFT in G-INS-I global domain alignment mode, and a maximum-likelihood phylogenetic tree was constructed using FastTree.

Chapter 5: Progress in Transcription Antitermination

NusG-family processive antitermination

In Chapters 2 and 4 of this dissertation, we describe the discovery of the LoaP processive antitermination protein. Although several other NusG paralog proteins have been previously identified, this was not simply addition of another paralog to the list. LoaP represents the first distinct NusG paralog group since RfaH to demonstrate processive antitermination activity. Additionally, while both RfaH and LoaP are paralogs of the NusG elongation factor, LoaP exhibits substantially distinct functions, acting on intrinsic termination instead of Rho-dependent termination. Similarly, while both proteins exhibit nucleic acid-binding activity, LoaP alone has been shown to exhibit strong RNA-binding activity. The widespread distribution of not just RfaH and LoaP, but additional groups characterized by UpxY and TaA proteins and yet more currently undescribed paralogs, indicates a broad role for these presumed antiterminators. The substantial differences in function between the two functionally characterized groups suggests the potential for more important differences in function between other groups.

This discovery of LoaP provides important context for other research on NusG proteins. In a unique experiment, investigators previously demonstrated the utility of NusG proteins in discovery of novel antibiotics by overexpressing a NusG protein (104). When overexpressing this protein, *Clostridium cellulolyticum* produces a variety of polythioamide compounds, antibiotics from a novel class of compounds typified by closthioamide. Although this experiment originally intended to use NusG to globally activate transcription antitermination, the gene chosen to overexpress is not NusG, as *C. cellulolyticum* contains three NusG family proteins. In fact, the sequence chosen represents a LoaP protein (B8I8J0, Ccel_0849), included in our phylogenetic analysis (Figure IV-1B). Overexpression of this LoaP protein presumably acts on a transcript containing genes required for production of these closthioamide compounds. Recently, the pathway required for production of closthioamide has been identified (215) and is found several megabases from Ccel_0849. Although introduction of this pathway into *E. coli* results in production of intermediate compounds, the final closthioamide compound is not produced (215). The authors suggest that inactivity of one of the final genes in the pathway in *E. coli*, *ctaJ*, would explain the pattern of compounds present. It is possible that the LoaP protein Ccel_0849 may be required for expression of the full-length *cta* transcript, and that without this activity the *ctaJK* genes are not properly expressed to produce the final compound. LoaP proteins in a variety of species are found adjacent to polyketide pathways, although many are not. The ability of *B. velezensis* LoaP to act on the macrolactin transcript and *C. cellulolyticum* LoaP to presumably affect the *cta* pathway for production of a non-polyketide, non-NRPS antibiotic suggest a potentially larger role for LoaP proteins in regulation of antibiotic production.

An open question following this work concerns the nature of these processive antiterminator proteins. Although they may be thought of as regulators, controlling termination by coupling downstream transcription to some input signal affecting their own expression and activity, processive antiterminators may have other uses. Ribosomal RNA antitermination in particular does not follow this pattern of regulation. Instead, it acts to ensure complete transcription and processing of the complex non-coding RNA transcripts. Secondary metabolite and polysaccharide pathways—such as those affected by the EAR RNA and LoaP or UpxY proteins—are often very long, metabolically expensive, and require expression of most or all genes to produce the intended product (57). Processive antitermination mechanisms may act as tuning mechanisms to ensure production of all gene products in these transcripts even when transcribed at low frequencies. The deficiency in closthioamide production during heterologous expression of the *cta* pathway in *E. coli* may indicate a role for the Ccel_0849 LoaP in ensuring sufficient expression of the distal *ctaJ* gene. Further work to determine whether LoaP is directly involved in signal-dependent regulation of difficidin and macrolactin or whether it is simply essential to bypass the intrinsic terminators present in multiple sites may yield insight into the regulatory nature of processive antiterminators.

RNA-binding transcription attenuation proteins

The single-compartment bacterial cytoplasm requires many cellular processes to take place in the same environment. As both transcription and translation happen in the same environment in which most metabolism takes place, genomic DNA and transcribing mRNAs are directly accessible by many potential regulatory signals or factors. DNA-binding transcription factors, which are often members of the nearly universal class of helix-turn-helix proteins, are highly prevalent and widely used to control much of gene expression. RNA-binding proteins, on the other hand, do not have such a consistently utilized motif and often contain unique domains responsible for RNA-binding (216). In particular, transcription attenuation mechanisms, whether solely dependent on RNA or on RNA-binding proteins, are often more difficult to identify. These mechanisms often contain unique nucleic acid or protein sequence motifs. Experimental approaches to identify these motifs, such as PAR-CLIP (217) or Term-seq (218) are often more complicated or depend on transcription initiation of regulated transcripts to ensure their presence, as opposed to DNA immunoprecipitation which requires only activity of the DNA-binding factor (219).

In some cases, individual examples of regulation may be identified, but generalization of these mechanisms, allowing for their identification elsewhere, depends on an understanding of their mechanism and definition of their requirements. To this end, in Chapter 3 we set out to understand how ANTAR regulators of transcription termination identified their target transcripts and how their RNA-binding activity led to regulation of transcription termination. Through this process we defined the sequence and structure requirements for ANTAR binding and antitermination. This understanding allowed us to catalog potential ANTAR regulons in hundreds of bacteria. Additionally, following up on studies of the original *E. faecalis* ethanolamine utilization gene cluster resulted in identification of a novel riboswitch-containing small RNA capable of sequestering the EutV ANTAR protein and reducing antitermination activity on other ANTAR substrates (14).

The requirements for ANTAR antitermination in *Klebsiella*, *Pseudomonas*, and *Enterococcus* are very similar, and can be extended to hundreds of other bacterial genomes. Despite this similarity, the representatives of the ANTAR clan studied up to this point are all representatives of just two or three subtypes and the ANTAR domains, at least, are relatively similar in comparison to all ANTAR domain-containing proteins. Even with these examples sharing nearly identical RNA substrates, the *Enterococcus* EutV and the *Klebsiella* NasR have completely distinct sensory inputs and appear to have distinct methods of activation: response regulator domain-induced dimerization or ligand-dependent conformational changes, respectively. Thousands of additional ANTAR domain-containing proteins exist, largely in actinobacterial species, which have been not directly studied nor do they contain input domains comparable to any of the studied proteins. A better understanding of how ANTAR RNA-binding domains interact with their input domains and how this allows for activation of RNA-binding activity of very similar small domains in different ways would improve our understanding of these domain-domain interactions. Unbiased identification of RNA motifs capable of binding NasR protein even in the absence of ligand led to the same motifs identified in genome-mining approaches. Even if the phylogenetically distinct actinobacterial ANTAR domains recognize a different RNA substrate in response to unknown input signals, similar HT-SELEX approaches may directly shed light on their RNA substrate preferences.

Final Thoughts

Throughout this document we sought to discover and better understand some perhaps understudied but fascinating mechanisms of transcriptional regulation in a variety of bacteria. Building on years of work in the fields of processive antitermination and transcription attenuation, we provided information about new and functionally distinct mechanisms, broadening our field's understanding of the range of forms these regulators may take and the variety of biochemical functions they may hold. We learned that not all NusG PA homologs act on the same types of termination or even have the same range of interactions. We learned that highly conserved regulatory domains with near-identical regulatory targets may be activated in very different ways. Studying individual systems in great detail is necessary to truly understand how biology accomplishes gene regulation, but investigation of other related systems can reveal the enormous diversity of solutions biology may utilize to evolve to solve similar problems.

Appendix A: High-Throughput Reporter Assays to Quantify Genetic Reporter Libraries

Introduction

This chapter sets out to detail a concrete method for simultaneous quantification of a large number of mutant genetic reporters influencing the expression of a fluorescent reporter by randomizing the first five base pairs of a bacterial transcript.Introduction

The advent of high-throughput DNA sequencing technologies has revolutionized how biology can be done and made available a variety of new tools. One way these technologies can allow scientists to investigate sequence-function relationships is by allowing for massive parallelization of assays that previously required large amounts of manual labor or expensive automation to approach (Figure A1-1). These approaches can generate quantitative data for thousands of unique cell genotypes simultaneously, whether by measuring the abundance of each genotype by sequence barcode tags (220) or directly sequencing mutant variants (221).

Either approach may rely on assays which differentially select for different genotypes in various conditions. One approach which this chapter focuses on, utilizes flow cytometry and fluorescence-activated cell sorting to physically separate individual cells on the basis of fluorescence intensity (222). When combined with high-throughput sequencing of DNA amplicons generated from the sorted cells, this approach, called Sort-Seq or FACS-Seq, allows for simultaneous measurement of fluorescence activity for thousands of distinct genotypes (221, 223). Similar approaches have been used to investigate transcription factor binding sites (221, 223), promoters and ribosome binding sites (224–

226) and small RNAs (227) as well as to engineer improved genetic control elements (228).

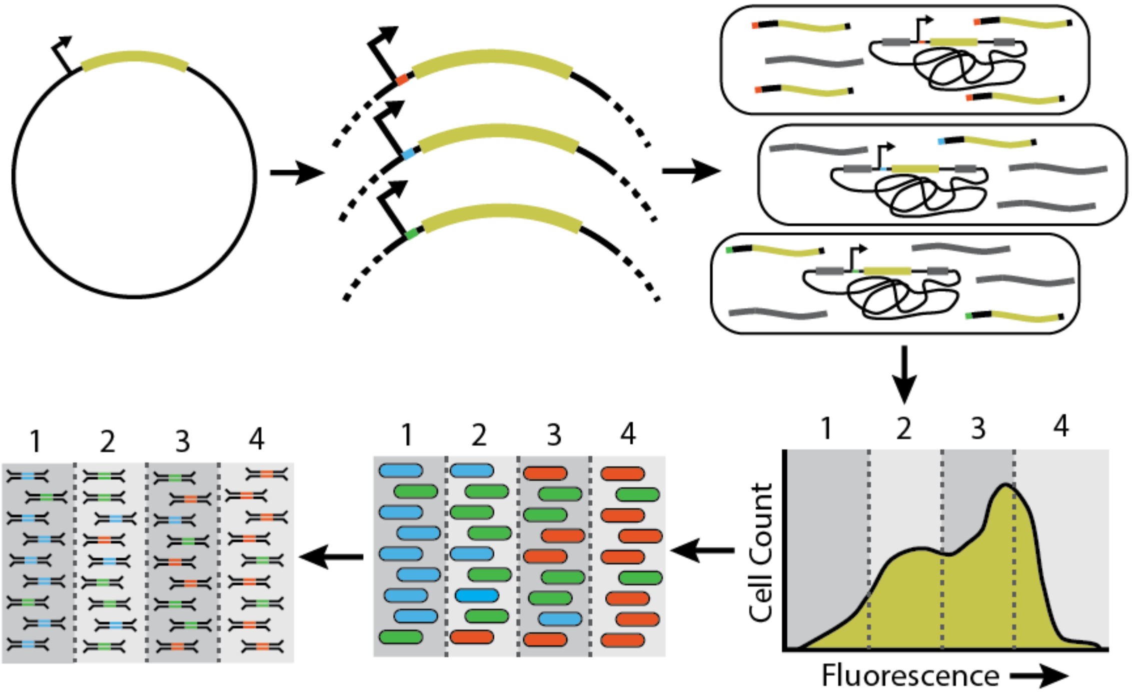


Figure A1-1: Overview of the Sort-Seq process.
The Sort-Seq process in this chapter consists of four major steps. First, a fluorescent reporter is mutagenized to create a library of mutant reporters. Then, these reporters are introduced into individual cells to create a mixed population of fluorescent cells each representing one mutation. Next, these pools are sorted into subpopulations by FACS. Finally, DNA from each subpopulations is extracted and sequenced to determine the relative abundance of each mutant sequence in the different subpopulations.

Primer	Sequence	Purpose
ACS012	ACAATATCAGCATCCTTGCAGGGTATG	Site-Directed Randomization of pJG019 Transcription Start
ACS013	TGTGTGT**NNNNN**CTTAAGGAGGAAAGTCACATTATGAG	Site-Directed Randomization of pJG019 Transcription Start
JRG629	**GTAAGCTG**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 1
JRG630	**GAACTGCT**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 1
JRG631	**GCTACTCA**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 2
JRG632	**TGGAGTAC**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 3
JRG633	**ACCTCATG**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 4
JRG634	**ATGTGGCA**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 5
JRG635	**CCAGTTAG**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 6
JRG636	**TCATACCC**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 7
JRG637	**TGAGCGTA**ACATCCAGAACAACCTCTGCT	Forward Barcoded Amplicon Primer 8
JRG638	**GGCTTCAA**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 2
JRG639	**GAGTATGG**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 3
JRG640	**TCGCTAGA**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 4
JRG641	**TACACCGT**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 5
JRG642	**GGTCAATC**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 6
JRG643	**TTCCAGAG**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 7
JRG644	**GTGGCAAT**TCACCATCCAGCTCAACCAG	Reverse Barcoded Amplicon Primer 8

Table A1-1: Primers used in the Sort-Seq experiment

Ultimately, these experiments seek to use FACS to generate subpopulations each containing cells representing a small range of fluorescent reporter activities. The relative abundance of a single genotype in the different subpopulations should be dependent on its characteristic reporter activity. Statistical approaches allow experimenters to estimate the underlying cell-to-cell distribution of fluorescence for every genotype in the experiment (223, 229). This rapid, parallel, quantification of very large numbers of distinct reporters can be used to fit quantitative biological models or to screen large numbers of sequences quickly. This chapter seeks to explain some of the considerations for design of a Sort-Seq experiment and to provide an example protocol for a Sort-Seq experiment investigating the total biological effects of sequence composition at the initial transcription start site for a bacterial promoter in *Bacillus subtilis*. This initial sequence may affect the efficiency of a variety of molecular processes including transcription initiation and 5′ end-dependent degradation. In this experiment, we will characterize both the effects of initial transcribed sequence on steady-state reporter levels as well as attempt to detect any sequence-specific effects from the 5′-end pyrophosphohydrolase RppH.

Construction of Reporter Libraries

The initial step of any Sort-Seq strategy is creation of the pool of variant reporters to be measured. How this is accomplished will depend on experimental considerations including several factors, the variety of reporter used, the amount of nucleic acid sequence that must be mutated, and the ultimate genetic context the reporter will be used it. Often, these reporters will take the form of a transcription unit in a self-replicating plasmid, which may either be used directly in the target organism or used to integrate the reporter construct into the genome. Depending on what aspects will be investigated, the mutations may be made in small, targeted reasons such as promoters or factor bindings sites, or potentially larger regions such as entire coding sequences. If mutations will only be targeted at smaller regions up to 100 nucleotides, site-specific mutagenesis techniques can quickly generate large numbers of variants using degenerate primers. Larger regions of interest may be targeted using methods such as chemical mutagenesis of mutagenic PCR. Variants libraries may also be created directly, perhaps using methods utilizing microarray-based oligo synthesis which may ensure the presence of all desired variants, with more effort and at greater cost.

However the reporters are generated, care must be taken to ensure that a sufficient number of unique sequences are present and in sufficient abundance to successfully measure the desired biological effects. For instance, experimenters may wish to develop a quantitative sequence-function model to describe the activity of their genetic variants. The final library must contain variants for every factor in the model (every position, nucleotide identity), but not necessarily every possible combinatorial variant (229). Some experiments may only desire quantification of a large number of variants to identity specific sequences with desired behavior, in which case experimental considerations should determine how many variants are required to ensure sequences with the desired behavior. Certain techniques may introduce bias to randomly generated variants. For instance, mutagenic PCR may be heavily biased toward certain types of base substitutions, which may create desired variants at a significantly lowered frequency.

Even if variants are initially created randomly with equal frequency, clonal cells containing different variant reporters may replicate at different rates, resulting in under-representation of certain sequences. This effect is particularly notable with techniques which involve solid-phase growth as colonies, where colony size heterogeneity can have huge effects on the proportion of each sequence. Molecular cloning is commonly performed by transformation of microbial cells with plasmid DNA followed by selection and growth on agar plates. Variant libraries may be constructed by following this traditional protocol followed by pooling of colonies by scraping and resuspension in liquid media. Protocols may be modified to perform selection only in liquid culture, which may not enrich as strongly for transformed cells if non-lethal (bacteriostatic) selection is used. The potential for presence of cells lacking the a variant reporter should be weighed against the benefit of more even variant representation. If all-batch methods are used, it is important to determine the efficiency of each step by estimating the total yield, perhaps using colony-forming unit assays on small aliquots.

In this chapter, we will describe an approach to quantify the effect of the initial nucleotide sequence, beginning at the +1 position, of a reporter transcript on its activity. We will use a constitutive promoter expressing a *yfp* gene found on a shuttle-vector plasmid which integrates as a single copy into the *Bacillus subtilis* genome (230). We generate a small library containing all 5-mer (4^5 = 1024) variants from +1 to +5 beginning at the first base transcribed by the normal promoter sequence using Q5 Site-Directed Mutagenesis which involves circular ligation of a PCR product of the full plasmid using a primer containing five completely random basepairs in the correct position and transformation into *E. coli*. Finally, the resultant plasmid library is extracted and use to transform *B. subtilis* cells, integrating into the genome, and pooled (Figure A1-1). We additionally transform this library into two derivative strains, one containing a marker-replacement of RppH (BKE30630), and one containing a deletion of RppH along with a xylose-inducible RppH complementation cassette (JG260).

Site-directed mutagenesis of reporter plasmid

Materials

- >1 ng/µL pJG019 YFP reporter plasmid (GenBank: KX499653.1)
- PCR Primers to amplify the entire reporter plasmid with degenerate nucleotide region
- NEB Q5 Site-Directed Mutagenesis kit
- Nuclease-free Water
- Chemical- or electro-competent *E. coli*
- 14 mL Snap-cap culture tubes
- SOC Broth Media
- LB-Agar Selection Plates (100 µg/mL Ampicillin or 5 µg/mL Chloramphenicol)
- 50% Glycerol
- Plasmid Midi-prep kit (ZymoPURE II Plasmid Midiprep Kit)
- *B. subtilis* transformation medium (25 g/L K2HPO4·3H20, 6 g/L KH2PO4, 1 g/L trisodium citrate, 0.2 g/L MgSO4·7H20, 2 g/L Na2SO4 (pH 7.0), 50 M FeCl3, 2 M MnSO4, 0.4% glucose, 0.2% glutamate)
- *B. subtilis* 168 and derivative strains BKE30630, JG260

Generation of randomized reporter library

Test of reporter plasmid mutagenesis efficiency

To begin, we will use a molecular cloning approach to generate randomized sequence reporter constructs (Figure A1-2). As the efficiency of these steps can vary depending on protocol and materials, we will initially perform a small test run to quantify the number of independent randomized sequences generated.

1. Assemble a PCR reaction to generate the linear reporter plasmid construct with randomized base pairs:
    - 12.5 µL 2x Q5 Hot-Start Master Mix
    - 1.25 µL Primer 1
    - 1.25 µL Primer 2
    - 1 µL Plasmid (1-20 ng/µL)
    - 9 µL Nuclease-free Water
2. Mix reagents and perform PCR in a thermocycler with the following program:
    - Initial Denaturation: 98 °C for 30 seconds
    - PCR Cycle:
        - 98 °C for 10 seconds
        - 50-72 °C for 15 seconds
        - 72 °C for 30 seconds/kilobase
    - Final Extension: 72 °C for 2 minutes
3. Hold at 4 °C or on ice until next step.
4. Assemble the following reaction to prepare linearized plasmid ends for ligation (kinase), ligate, and degrade (DpnI) input plasmid template. Additionally, prepare a negative control reaction that does not include the 10x KLD Enzyme Mix.
    - 1 µL PCR Product
    - 5 µL 2x KLD Reaction Buffer
    - 1 µL 10x KLD Enzyme Mix
    - 3 µL Nuclease-free water
5. Mix reaction and incubate at room temperature for 5 minutes.
6. Transform both the KLD reaction and negative control into competent *E. coli* with appropriate protocol. Example for standard chemically-competent *E. coli* cloning cells:

a. Thaw 100 µL aliquot of chemically-competent cells on ice

b. Add 5 µL of KLD reaction to the cells and mix by gently flicking tube

c. Incubate for 5-30 minutes on ice

d. Heat shock cells for exactly 30 second at 42 °C

e. Place back on ice for 5 minutes

f. Add 900 µL of room temperature SOC media to tube

g. Incubate, shaking at 250 rpm, at 37 °C for 45-60 minutes

h. Spread 100 µL onto LB-agar selection plate and incubate overnight at 37 °C

7. The next day, count the number of colonies present on the selection plates and determine how much mutagenesis is necessary to obtain sufficient sequence diversity.

Depending on the efficiency of the KLD ligation reaction and the quality of the competent *E. coli* cells a range of *E. coli* transformants may result. Approximately 10-fold more distinct reporter sequences will be generated if the entire transformation outgrowth is used. Very few colonies should be present on the negative control plate. Ideally, a few individual colonies should be re-streaked and grown for plasmid miniprep, and the region targeted for mutagenesis sequenced by Sanger sequencing to verify that the library is being formed properly.

Generation of randomized reporter plasmid

At this point, two options are available, either the outgrowth can be continued in selective media overnight, or the entire volumes may be plated on large selective plates and the colonies resuspended. If plating and resuspension is not performed, you may skip to the next section. Depending on the requirements of the experiment, tens or hundreds of thousands of distinct sequences may be targeted. For this example experiment, we targeted 25,000 mutants, requiring four volumes of KLD reaction. To generate the plasmids, the above protocol was repeated exactly until the plating of the outgrown, transformed, *E. coli* cells.

1. Plate 500 µL of outgrown cells on lightly-dried 150 mm LB-agar selection plates
2. Incubate overnight at 37 °C.
3. Pipette 1 mL of LB broth onto each plate and gently resuspend the colonies in broth using a serological pipette, glass spreader device, or other smooth instrument.
4. Pipette up the resuspended cells, and pool the liquid together.
5. Mix well, and place 600 µL in 2 mL snap-cap tubes or cryovials.
6. Add 900 µL 50% glycerol to each tube and gently mix by inversion.
7. Store aliquots at -80 °C for further use.
8. The randomized plasmid pool (remainder of the resuspended colonies) will be midi-prepped.

Purification of randomized reporter plasmid

As this experiment is being done in a different bacteria, the randomized sequence plasmid library needs to be purified for subsequent transformation of *B. subtilis*. Any plasmid midi-prep kit may be used. For this example experiment, we used the Zymo ZymoPURE II Plasmid Midiprep Kit. As an alternative to larger-scale purification kits, a modification of standard plasmid minipreps can be used to obtain similar DNA yields (231).

Preparation of final reporter library

Much as in the initial preparation of the randomized plasmid pool, the efficiency of the transformation of bacteria by the prepared plasmid library can vary. We perform an initial test transformation to determine how much DNA must be transformed.

1. Inoculate a 5 mL culture each of *B. subtilis* transformation medium with *B. subtilis* 168, BKE30630, and ACS006.
2. Incubate, standing, at 37 °C overnight or, alternatively, shaking at room temperature overnight.
3. Incubate culture at 37 °C shaking at 250 rpm until the culture reaches a density of OD at 600nm of 0.4-0.8 (1.5-3 hours).
4. Transfer 1 mL of each culture to a new culture tube, and add 4 µg of midi-prepped plasmid DNA.
5. Incubate cultures at 37 °C shaking at 250 rpm for 40 minutes.
6. Add 1 mL of SOC Broth with 0.1 µg/mL chloramphenicol to each culture and incubate shaking for an additional 45 minutes.
7. Spread 100uL of each culture onto an LB-chloramphenicol selection plate and incubate overnight at 37 °C.
8. The next day, count the number of colonies present on the selection plates.

With *B. subtilis* 168, we expect approximately 500 transformants per plate under these conditions, for about 10,000 total transformants per transformation. For this experiment, we target 100,000 total transformants, larger than the initial number of *E. coli* transformants to help retain a relatively even proportion of the total poor for each sequence in the library.

9. Inoculate a 5 mL culture of *B. subtilis* transformation medium with each strain.

10. Incubate, standing, at 37 °C overnight or, alternatively, shaking at room temperature overnight.
11. Incubate cultures at 37 °C shaking at 250 rpm until the culture reaches a density of OD at 600nm of 0.4-0.8 (1.5-3 hours).
12. Split the cultures by pipetting 1 mL each into five new culture tubes, and add 4 µg of midi-prepped plasmid DNA to each.
13. Incubate cultures at 37 °C shaking at 250 rpm for 40 minutes.
14. Add 1 mL of SOC Broth with 0.1 µg/mL chloramphenicol to the each culture and incubate shaking for an additional 45 minutes.
15. Centrifuge the cultures at 5000 x G in a centrifuge at room temperature for 5 minutes.
16. Decant the supernatant and resuspend the cell pellet in the remaining media (~150-200 mL)
17. Spread each tube of resuspended cells onto a 150 mm LB-chloramphenicol selection plates and incubate overnight at 37 °C.
18. Pipette 1 mL of LB broth onto each plate and gently resuspend the colonies in broth using a serological pipette, glass spreader device, or other smooth instrument.
19. Pipette up the resuspended cells, and pool the liquid from each strain together.
20. Mix well, and aliquot into 600 µL aliquots in 2 mL snap-cap tubes or cryovials.
21. Add 900 µL of 50% glycerol to each tube and store at -80 °C for future use.

Flow Cytometry

The core of the Sort-Seq method involves flow cytometry and fluorescence-activated cell sorting (FACS) to analyze and split the input reporter library into subpopulations, each generated from cells falling within a specific range of fluorescence intensity values (gates). The output subpopulations are subjected to high-throughput sequencing to determine the frequency of each variant in each subpopulation. The eventual goal is to use the distribution of variants among subpopulations to infer the average fluorescence intensity of cells containing each variant, and perhaps the cell-to-cell variability for each isogenic variant.

As a result of the large number of factors which affect the abundance of a fluorescent reporter molecule, fluorescence can vary across an extremely large range. For this reason, subpopulations are usually selected such that the gates are evenly spaced on a log scale. Any number of gates can be used as long as at least two subpopulations are generated. Additional gates provide greater accuracy at increased cost and time expenditure. Generally, each variant is treated as having log-normal distributed fluorescence (cite). The accuracy of estimations of mean intensity is greater for more gates, but is also dependent on the variability for each sequence variant. Greater cell-to-cell variability increases the likelihood that a single sequence will be spread across multiple subpopulations The accuracy of Sort-Seq for individual variants with different number of gates has previously been analyzed, and in general, mean estimates are accurate for variants where the standard deviation of that variant is similar to or smaller than the gate width (229). In some cases, acceptable results can be obtained with only a single gate (229).

The desired number of gates can be determined from an estimate of the cell-to-cell variability for the reporter system, but the width of gates must be determined from the range of reporter intensities present in the variant library. Depending on the experimental setup, this may be determined either empirically by flow cytometry of the library itself, to determine the minimum and maximum fluorescence present. Additional controls may be desired, especially if it is necessary to validate whether the reporter library construction was successful. Negative controls may include cells lacking the reporter construct, or lacking a particular effector. Positive controls likely depend on experimental setup, but should consist of cells exhibiting fluorescence values representative of bright variants.

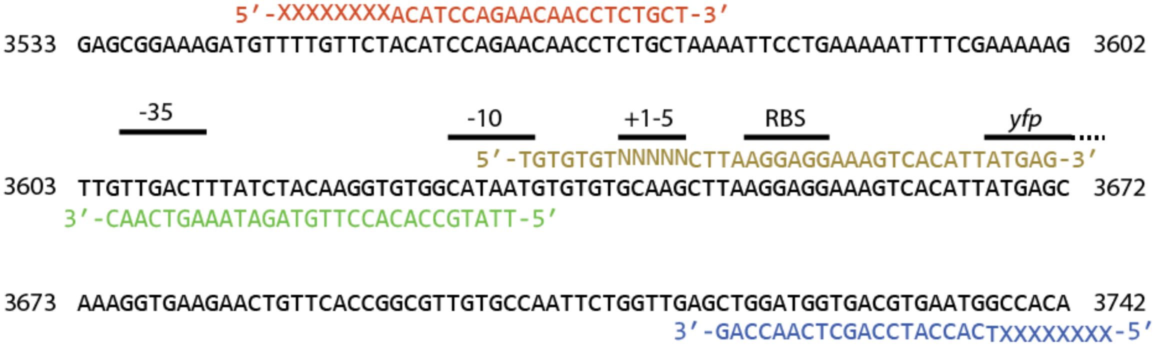

5′-XXXXXXXXACATCCAGAACAACCTCTGCT-3′
3533 GAGCGGAAAGATGTTTTGTTCTACATCCAGAACAACCTCTGCTAAAATTCCTGAAAAATTTTCGAAAAAG 3602
-35
-10
+1-5
RBS
yfp
5′-TGTGTGTNNNNNCTTAAGGAGGAAAGTCACATTATGAG-3′
3603 TTGTTGACTTTATCTACAAGGTGTGGCATAATGTGTGTGCAAGCTTAAGGAGGAAAGTCACATTATGAGC 3672
3′-CAACTGAAATAGATGTTCCACACCGTATT-5′
3673 AAAGGTGAAGAACTGTTCACCGGCGTTGTGCCAATTCTGGTTGAGCTGGATGGTGACGTGAATGGCCACA 3742
3′-GACCAACTCGACCTACCACTXXXXXXXX-5′

Figure A1-2: Sequence details of the randomized transcript start sequence.
The sequence of the constitutive promoter sequence used in the Sort-Seq experiment. The green and gold primers are used in site-directed mutagenesis to create randomized sequences at the +1-5 positions. Later, the red and blue primers are used to amplify this region for high-throughput sequencing.

To continue our use of Sort-Seq to determine the effect of the initial transcribed sequence on fluorescent reporter, we will grow our variant library, prepare the cells for flow cytometry, choose our subpopulation gates, and finally sort a sufficient number of cells for later analysis (Figure A1-3A). In this experiment, we will grow each strain library to mid-exponential phase as three replicates in rich media. Fifteen minutes before collection, one set of complementation strain cultures will be induced with xylose to express RppH. We will then wash the cells in buffer to remove rich media which can interfere with quantification of our yellow fluorescent protein. We choose to sort our cells into four subpopulations, partly because the BD FACSAria we use is capable of simultaneous sorting of four subpopulations, allowing the sorting to take place in a single run (Figure A1-3B). We chose to keep our cell suspension on ice while sorting in order to limit the amount of protein production and turnover which could affect reporter levels in cells sorted towards the end of the run. We finally return the cells to rich media for an outgrowth to generate a larger number of cells for extraction of genomic DNA for later high-throughput sequencing (Figure A1-3C).

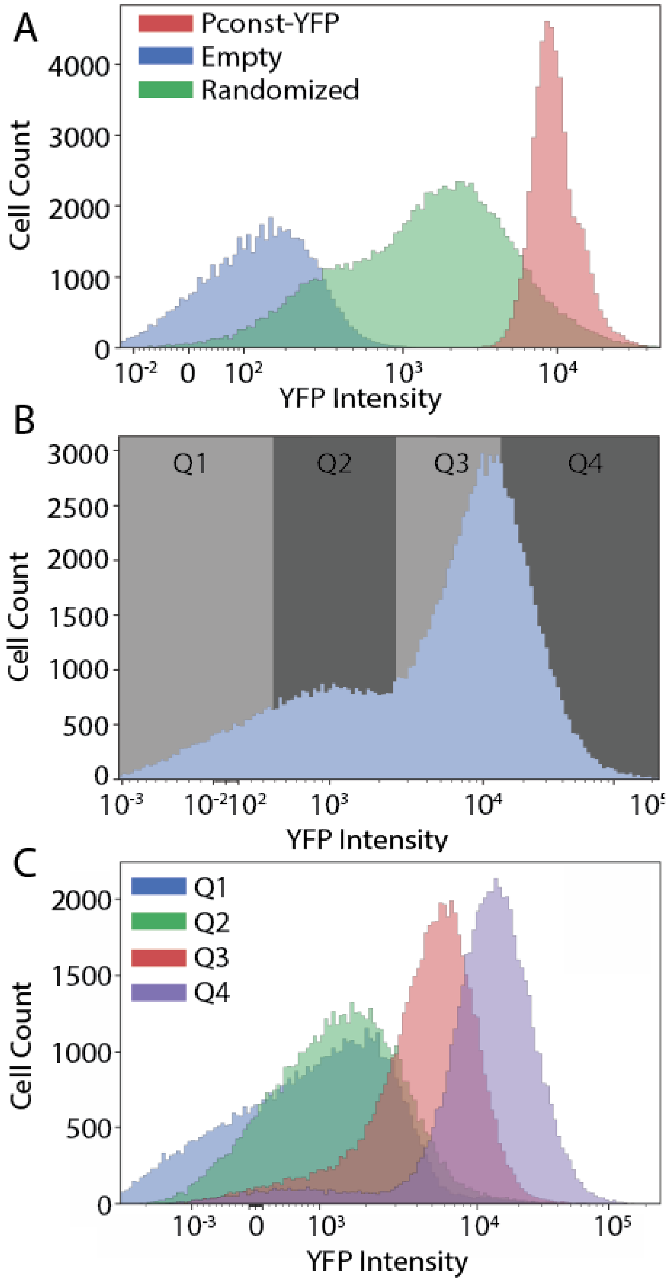


Figure A1-3: Progress of fluorescence distributions along the Sort-Seq process.
(A) Histograms showing the distribution of YFP fluorescence for the starting constitutive promoter-YFP containing strain (Pconst-YFP), a strain lacking this reporter (Empty) and the mutagenized reporter library (Randomized). (B) Histogram of the flow cytometry sorting data illustrating the gate locations used to separate the randomized mutant library into four subpopulations. (C) Histograms of each of the four subpopulations after re-growth and flow cytometry of the output subpopulations illustrating the spread of the fluorescence distribution.

FACS Separation of Reporter Library

Materials

- LB Broth
- 14 mL Snap-cap culture tubes
- 5 mL culture tubes
- Phosphate-buffered saline (NaCl 8 g/L, KCl 0.2 g/L, Na_2HPO_4 1.44 g/L, KH_2PO_4 0.24 g/L)

Preparation of cultures for flow cytometry

To start the first step of the Sort-Seq assay itself, we will prepare a cell culture grown in the appropriate conditions to be sorted into equally log-spaced gates by flow cytometry. To gather the information needed to set up this experiment, we will perform initial flow cytometric readings of some control reporters as well as the input reporter library to determine what settings are required.

1. The day before flow cytometry is to be performed, inoculate 5 mL cultures in LB selective broth of a positive control (*B. subtilis* 168 with integrated starting reporter plasmid pJG019, JG040) and a negative control lacking any YFP reporter construct and incubate, shaking, overnight at 37 °C and 250 rpm.
2. The next morning, re-inoculate the two control strains into new 5 mL cultures of LB broth with 50 µL of overnight culture.
3. Thaw one glycerol stock aliquot of each prepared reporter library and mix entire aliquot with 15 mL of LB selective broth and split between three culture tubes. Prepare a second set of three cultures with the inducible complementation strain.

4. Incubate cultures, shaking, at 37 °C and 250 rpm until they reach mid-exponential phase, OD at 600nm around 0.5-0.6.
5. Add 50 µL of 50% xylose solution to one set of three inducible complementation strain cultures and incubate an additional 15 minutes.
6. Aliquot 1 mL of the reporter libraries into a centrifuge tube and centrifuge at 5000 x G and freeze pellet at -80 °C for later genomic DNA extraction.
7. Centrifuge the cultures at 5000 x G in a centrifuge at room temperature for 5 minutes.
8. Decant the supernatant and resuspend the cell pellet in 5 mL phosphate-buffered saline (PBS).
9. Repeat the centrifugation step and resuspend the pellet again.
10. Store the cell pellets in the culture tube on ice until flow cytometry.

Sort-seq FACS of reporter library

Now that we have prepared cell cultures in PBS we are ready to sort our reporter library into subpopulations in the first half of the Sort-Seq assay. This requires a cell sorter with the appropriate laser and filters for the reporters used in the library as well as the appropriate equipment and methods for the cell types used. In our experiment we use a BD FACSAria II cell sorter with the smallest nozzle and the 488 nm laser with the 530/30 bandpass filter appropriate for YFP-containing bacterial cells. The exact flow cytometry approach may vary depending on the instrument and characteristics of the reporter library used. Individual steps will depend on the instrument used, so here we provide an overview of the process. We begin by analyzing our controls and the reporter library to decide on both scattering filter gates to reduce noise as well as the fluorescence gates used for the Sort-seq approach itself.

1. Begin by running blank PBS and the negative fluorescence control sample through the cell sorter and adjust the instrument settings to obtain an appropriate threshold rate and side- and front-scattering gates to discriminate true cells from noise present in the PBS blank.
2. Run the positive control reporter through the cell sorter and adjust the laser and detector settings to ensure optimal detection of the fluorescence signal.

   The cell-to-cell variability for the positive and negative control samples should be representative of cell-to-cell variability for these culture and sorter conditions for reporter genotypes not affected by additional sources of noise, and provide a lower estimate for expected variability.
3. Run a reporter library through the cell sorter to generate the observed distribution of reporter intensity.

   At this point, utilize the information generated to decide on explicit values for the sorting gates as discussed above and in (229). These steps can also be done on previous days if more convenient.
4. Prepare the sorting equipment with four empty 5 mL culture tubes.
5. Sort each reporter library into four output subpopulations with appropriate gates. We chose to sort until each subpopulation reached 1 million sort events, approximately 15 minutes at 20,000 total threshold events per second.
6. Repeat steps 4 and 5 for each of the strains and replicate cultures.
7. Mix the output subpopulations with an equal volume of LB broth and culture overnight at 37 °C and 250 rpm.
8. Prepare a glycerol stock of each population by mixing 600 µL of culture with 900 µL of 50% glycerol and store at -80 °C.
9. Aliquot 1 mL of each subpopulation culture into a centrifuge tube and centrifuge at 5000 x G and freeze pellet at -80 °C for later genomic DNA extraction.

High-throughput sequencing library preparation

The final essential step in any Sort-Seq approach is the high-throughput sequencing of the output variants in each subpopulation. This sequencing can utilize any variety of high-throughput sequencing technology. Perhaps most commonly, Illumina short-read sequencing is used to cheaply obtain high numbers of accurate reads. In order to generate sequencing-ready libraries, the variant sequences from the reporters must be amplified and appropriate barcodes and adapter sequences added to enable high-throughput sequencing.

Although DNA can be extracted directly from the flow cytometry-sorted cells and subjected to PCR, extraction of small amounts of DNA can be inefficient, and any loss of material can have a large impact on the number of variants recovered. Generally, sorted subpopulations are re-grown to larger quantity, and DNA extracted by normal methods for that organism. Traditional PCR is sufficient to amplify the region targeted for mutation, although care must be taken to include sufficient input DNA so that enough copies of the reporter are present to properly represent the sorted populations. Experimenters should target the amplified region carefully to ensure that the nucleotides of interest will be appropriately placed so they are present in sequence generated by the sequencing primers, preferably in both reads if paired-read sequencing is used. Finally, as this process involves significant amounts of PCR amplification, potentially of small amounts in input sequence, care must be taken to avoid contamination by similar sequences which may be present in laboratories in large amounts if similar reporters are commonly used.

To prepare sequencing-ready DNA libraries, multiple methods are available. Adapters may be added directly with PCR, either in a single step during the initial amplification of the reporter sequence, or in a secondary PCR using the initial amplified target as the template. Alternatively, after initial amplification, adapters and barcodes may be added using ligation approaches using standard adapters used for other library preparation methods. Each of these methods has advantages and disadvantages. Single-step PCR with primers carrying sequencing adapters is fast and minimizes experimental work, but requires many long primers, each with a unique barcode and specific to that particular reporter. PCR with these extremely large primers can be unreliable, and may require troubleshooting to ensure proper amplification. Performing a primary PCR, perhaps with primers adding a short, general, adapter sequence followed by a secondary PCR adding different barcode sequences along with sequencing adapters, can simplify primer design and enable primer re-use for multiple amplicons. Performing two PCRs, however, can increase the risk of sequence bias resulting in uneven amplification of different variant sequences, biasing results. Finally, PCR-amplified DNA may be directly ligated to barcoded double-stranded DNA adapters. This adds an additional enzymatic step, but can utilize standard sequencing adapters, either commercially available or home-made, and can minimize additional sequencing bias and difficulty in performing PCR with long 5′-extension containing primers.

Standard considerations for amplicon sequencing on the chosen sequencing platforms needs to be addressed. For instance, with Illumina sequencing, registration of sequence clusters on the flow cell requires considerable sequence variation in the first few rounds of sequencing. If all sequences have the same sequence immediately downstream of the sequencing primer, sequencing may fail without modifications to the protocol. This may be addressed either in design of the sequencing library, for instance by including an internal barcode sequence on the initial PCR primer which will be immediately downstream of the sequencing primer. A similar effect can also be accomplished during sequencing by including a large proportion of a more standard sequencing library containing random starting sequences, for instance the readily-available PhiX Illumina sequencing library available to sequencing facilities.

To continue our analysis of initial transcribed sequences in *B. subtilis* transcripts, we extract genomic DNA from our sorted and re-grown subpopulations using phenol-chloroform extraction and ethanol precipitation. We then perform an initial PCR using primers which add carefully-chosen barcodes to the 5′ end of each strand using a high-fidelity polymerase (Figure A1-2). At this point each library contains distinct barcodes sequences and we mix these to create an equimolar pool to simplify further steps. We then A-tail and phosphorylate the ends of our PCR products and ligate these to barcoded, forked, Illumina sequencing adapters (Kapa Biosystems). We finally sequence these on an Illumina NextSeq 500 machine with paired-end sequencing, ensuring reads of both internal barcodes as well as coverage of the five nucleotide randomized sequence by both reads to increase sequence accuracy.

Amplicon Sequencing of Reporter Library Subpopulations

Materials

- Phenol, buffer-saturated pH 8
- Chloroform
- 1:1 Phenol:Chloroform
- Saline-EDTA (0.15 M NaCl, 0.01 M EDTA pH 8.0)
- Lysozyme (8 mg/mL in saline-EDTA)
- 95-100% Ethanol

- Nuclease-free water
- PCR polymerase system
- Barcoded primers to target the randomized reporter region (10 µM)
  - Forward primers: JRG629/JRG631-637
  - Reverse primers: JRG630/JRG638-644
- SPRI Purification Beads (Beckman Coulter AMPure XP)
- Magnetic stand
- UV spectrophotometer
- Klenow Fragment (3′ -> 5′ exo-) (NEB)
- dATP (10 mM)
- NEBNext Blunt/TA Ligation Mix
- KAPA Single-Indexed Adapter Kit (30 µM)

Extraction of genomic DNA

In the previous section we froze cell pellets containing overnight cultured cells from the initial reporter library as well as from each of the sorted subpopulations. Initially, we will extract and purify genomic DNA from these cells to use as a template for PCR to amplify the particular reporter region that we randomized for our reporter library. We perform the following steps in parallel for each subpopulation as well as the input populations. Care should be taken to prevent cross-contamination at all steps.

1. Resuspend the frozen cell pellets in 360 µL of saline-EDTA.
2. Add 40 µL of 8 mg/mL lysozyme in saline-EDTA and mix.
3. Incubate samples for 35 minutes at 37 °C.
4. Add 400 µL phenol and invert by hand for two minutes.
5. Centrifuge for 15 minutes at >12,000 x G.
6. Transfer the supernatant to a fresh centrifuge tube, avoiding the interphase.

7. Add 400 µL phenol:chloroform mixture and invert by hand for two minutes.
8. Centrifuge for 5 minutes at >12,000 x G.
9. Transfer the supernatant to a fresh centrifuge tube, avoiding the interphase.
10. Add 400 µL chloroform and invert by hand for two minutes.
11. Centrifuge for 5 minutes at >12,000 x G.
12. Transfer the supernatant to a fresh centrifuge tube, avoiding the interphase.
13. Add 800 µL of 95-100% ethanol and invert repeatedly.
14. Centrifuge for 15 minutes and wash DNA pellet with 400 µL 70% ethanol.
15. Remove ethanol and resuspend pellet in 200 µL nuclease-free water.

Amplification of target reporter region

Next, we will target out particular region of interest by amplification with PCR. We utilize standard PCR with high-fidelity polymerase (NEB Q5 Polymerase) to generate our amplicons. For our experiment, we introduce 8-nucleotide barcode sequences at either end of the PCR to allow direct demultiplexing and to add sequence complexity in the first rounds of sequencing. This addition of barcodes is optional, and could easily be replaced by using distinct barcoded adapters for each sample in the adapter ligation step. We finally purify our PCR product using SPRI beads to prepare the DNA for A-tailing and ligation.

1. Assemble a PCR reaction to generate the randomized reporter amplicon for each sample. Each reaction should have a different pair of primers to generate distinct barcode pairs.
    - 25 µL 2x Q5 Hot-Start Master Mix
    - 2.5 µL Primer 1 (10 µM)
    - 2.5 µL Primer 2 (10 µM)
    - 1 µL Genomic DNA

- 19 µL Nuclease-free Water

2. Mix reagents and perform PCR in a thermocycler with the following program:
    - Initial Denaturation: 98 °C for 30 seconds
    - PCR Cycle (25 cycles):
        - 98 °C for 10 seconds
        - 50-72 °C for 15 seconds
        - 72 °C for 30 seconds
    - Final Extension: 72 °C for 2 minutes
3. Hold at 4 °C or on ice until next step.
4. Add 1 volume (50 µL) of SPRI bead mix to the PCR reaction and pipette 90% of the total volume up and down 20 times.
5. Incubate mixture at room temperature for 10 minutes.
6. Place the tubes on the magnetic stand for 5 minutes or until clear.
7. Remove the supernatant and add 200 µL of 80% ethanol.
8. Let stand on magnets for 30 seconds.
9. Remove the supernatant and add 200 µL of 80% ethanol.
10. Let stand on magnets for 30 seconds.
11. Remove the supernatant and let stand for 30 seconds. Use a fine pipette tip to remove all residual ethanol.
12. Let pellets air-dry for 5 minutes.
13. Remove tubes from magnetic stand and resuspend magnetic beads in 27.5 µL nuclease-free water.
14. Incubate at room temperature for 2 minutes.
15. Place tubes back on magnetic stand for 5 minutes or until clear.
16. Remove 25 µL supernatant containing purified PCR product to fresh tubes.
17. Store on ice until next steps or at -20 °C for longer storage.

Sample pooling

After the PCR step, each amplicon is separately tagged with a unique pair of 8-nucleotide sequences on each end which will enable us to demultiplex each sample while only using a single barcoded Illumina adapter. This both simplifies the remainder of library preparation as well as limiting use of expensive commercial adapters and enabling further multiplexing of the Sort-Seq pool with additional sequencing experiments to better utilize the extremely large number of reads available with Illumina technology. We combine these samples at an equal molar ratio to ensure similar read depth for each sample.

At this step, if internal barcodes were not added during target amplification, samples should not be pooled until after barcoded adapter ligation in the next step.

1. Quantify each PCR product by UV absorption.
2. Run all samples on a 2% TAE agarose gel. Stain and image with fluorescent intercalating dye (ethidium bromide, SYBR Green, etc.).
3. Estimate intensity of each band by densitometry and check for correspondence with quantities determined by UV absorbance.
4. (Optional: Quantify individual libraries by qPCR)
5. Normalize PCR products to a concentration of 10 ng/µL in 20 µL nuclease-free water.
6. Mix equal volumes of each normalized PCR product to total approximately 1000 ng.

   Depending on the total number of samples, different amounts of each should be added. For instance, if 48 total samples at 10 ng/µL are pooled, 2 µL of each should be pooled.
7. Add nuclease-free water to 50 µL if the pool volume is below 50 µL.

8. Add 1 volume of SPRI bead mix to the pooled samples and pipette 90% of the total volume up and down 20 times.
9. Incubate mixture at room temperature for 10 minutes.
10. Place the tubes on the magnetic stand for 5 minutes or until clear.
11. Remove the supernatant and add 200 µL of 80% ethanol.
12. Let stand on magnets for 30 seconds.
13. Remove the supernatant and add 200 µL of 80% ethanol.
14. Let stand on magnets for 30 seconds.
15. Remove the supernatant and let stand for 30 seconds. Use a fine pipette tip to remove all residual ethanol.
16. Let pellets air-dry for 5 minutes.
17. Remove tubes from magnetic stand and resuspend magnetic beads in 22.5 µL nuclease-free water.
18. Incubate at room temperature for 2 minutes.
19. Place tubes back on magnetic stand for 5 minutes or until clear.
20. Remove 20 µL supernatant containing purified PCR product to fresh tubes.
21. Store pooled samples on ice until next steps or at -20 °C for longer storage.

Preparation of sequencing-ready DNA libraries

In order to sequence our amplicons with Illumina sequencing, we must add adapters containing sequencing primer binding sites as well as flow-cell adapter sequence. This can be done either with additional PCR rounds or by directly ligating prepared dsDNA adapters. We will ligate commercial barcoded dT-tailed adapters from Kapa Biosystems. In the previous step we amplified our reporter using a high-fidelity PCR enzyme which results in primarily blunt-ended DNA products. In order to generate the final Illumina-ready libraries, we will first A-tail our PCR products using Klenow fragment, and finally directly ligate these A-tailed products to the barcoded adapters.

A-tailing of PCR products

To dA-tail the PCR products, we utilize the property of purified Klenow fragment of DNA polymerase of non templated addition of single nucleotides to the 3′ end of dsDNA.

1. Prepare the following enzymatic reaction on ice:
    a. 20 µL Pooled PCR Product (~25 ng/µL)
    b. 5 µL NEBuffer 2 (10X)
    c. 0.5 µL dATP (10 mM)
    d. 2 µL Klenow Fragment (3´→ 5´ exo–)
    e. 22.5 µL Nuclease-free water
2. Incubate in a thermocycler for 30 minutes at 37 °C.
3. Add 1 volume (50 µL) of SPRI bead mix to the PCR reaction and pipette 90% of the total volume up and down 20 times.
4. Incubate mixture at room temperature for 10 minutes.
5. Place the tubes on the magnetic stand for 5 minutes or until clear.
6. Remove the supernatant and add 200 µL of 80% ethanol.
7. Let stand on magnets for 30 seconds.
8. Remove the supernatant and add 200 µL of 80% ethanol.
9. Let stand on magnets for 30 seconds.
10. Remove the supernatant and let stand for 30 seconds. Use a fine pipette tip to remove all residual ethanol.
11. Let pellets air-dry for 5 minutes.

12. Remove tubes from magnetic stand and resuspend magnetic beads in 17.5 µL nuclease-free water.
13. Incubate at room temperature for 2 minutes.
14. Place tubes back on magnetic stand for 5 minutes or until clear.
15. Remove 15 µL supernatant containing purified PCR product to fresh tubes.
16. Store on ice until next steps or at -20 °C for longer storage.

Ligation of Illumina adapters to generate sequencing-ready libraries

Preparation of the sequencing libraries is simplified by direct ligation of dT-tailed commercial adapters using an enzymatic mix optimized for dA/dT-tailed ligations. After this ligation with complete adapters and purification, the pooled library is ready for sequencing, and should be at a concentration of ≥25 nM in 17.5 µL.

1. Freshly prepare 15 µM barcoded adapters in Adapter Dilution Buffer.
   Adapters should be prepared fresh in Adapter Dilution Buffer or 10 mM Tris-HCl, pH 8, 10 mM NaCl, 1 mM EDTA. Do not prepare more than required.
2. Prepare the following enzymatic reaction for each PCR amplicon:
   a. 15 µL A-Tailed PCR Product (~500 ng total)
   b. 5 µL 15 µM Adapter DNA
   c. 20 µL 2x Blunt/TA Ligation Mix
3. Incubate reactions for 15 minutes at room temperature.
4. Place reactions on ice until next step.
5. Add 10 µL nuclease-free water to each reaction.
6. Add 0.9 volumes (45 µL) of SPRI bead mix to the PCR reaction and pipette 90% of the total volume up and down 20 times.

7. Incubate mixture at room temperature for 10 minutes.
8. Place the tubes on the magnetic stand for 5 minutes or until clear.
9. Remove the supernatant and add 200 µL of 80% ethanol.
10. Let stand on magnets for 30 seconds.
11. Remove the supernatant and add 200 µL of 80% ethanol.
12. Let stand on magnets for 30 seconds.
13. Remove the supernatant and let stand for 30 seconds. Use a fine pipette tip to remove all residual ethanol.
14. Let pellets air-dry for 5 minutes.
15. Remove tubes from magnetic stand and resuspend magnetic beads in 17.5 µL nuclease-free water.
16. Incubate at room temperature for 2 minutes.
17. Place tubes back on magnetic stand for 5 minutes or until clear.
18. Remove 15 µL supernatant containing purified PCR product to fresh tubes.
19. Store on ice at -20 °C.
20. Submit for standard Illumina library quality control and sequencing.

    Depending on the efficacy of the SPRI bead purification at size-selection as well as the amount of adapter-dimer formed during ligation, a second round of bead cleanup/size-selection may need to be performed to minimize the amount of adapter-dimer in the final library.

Data Processing and Analysis

The final step in a Sort-Seq experiment is the processing and analysis of the high-throughput sequencing data. ultimately, we want to use estimates of the activity of all of the different reporter variants to infer biologically meaningful conclusions. Before that point, we must process the raw sequencing reads to determine which reads came from which sample and subpopulation, identify accurate reads coming from functional reporter variants, identify the reporter variant in each read, and tabulate the frequency of each reporter in each subpopulation. These steps are highly dependent on the particular library construction and sequencing methods used, but are all standard bioinformatic methods for which a variety of software tools at different levels of complexity or specificity are available (232).

Once the data has been processed and the proportion of reads in each subpopulation are determined, Sort-Seq seeks to estimate the mean value and perhaps variability for each reporter variant, simultaneously quantifying thousands of reporters. This is accomplished by using the number of reads in each subpopulation as well as the fluorescence characteristics of that subpopulation. Many methods have been used for this estimation, and details of these methods are beyond the scope of this chapter, but several are detailed in (229). Perhaps the simplest method involves simply assigning each read the average fluorescence intensity value for its subpopulation, and averaging these values for every read of a particular reporter variant. This approach often works well, although methods incorporating more complicated statistics can be used to increase accuracy, especially in experiments with fewer gates. Finally, the estimates of individual reporter variant activity may be used to fit quantitative models or otherwise draw conclusions about the biological system interrogated by the experiment.

In our experimental case, we initially perform some filtering on our read quality scores, followed by identification of reads containing proper promoter sequence to eliminate variants containing unintended insertions or deletions, as well as reads derived from sources other than our reporter. We then identify the five base pair region of interest, and tabulate those numbers across all subpopulations. Finally, we include an example exploratory analysis showing the effect of variants at different transcript positions on reporter activity.

Read Processing and Data Extraction

1. Filter read pairs to eliminate low-quality sequences using a tool such as Trimmomatic, Prinseq, or AfterQC.
2. Align the read library to a reference amplicon sequence using a short-read aligner such as Bowtie2 or BWA and discard reads without any match to the amplicon sequence.
3. Next, filter the individual read pairs determine which arose from valid reporter constructs containing the proper reporter sequence with no insertions, deletions, or substitutions beyond the randomized portion.

   In this case, this filtering was done using a custom Python script utilizing the HTSeq library to read the Bowtie2 BAM alignment output.

   a. Check each read for the expected promoter sequence at the beginning of the read to determine which strand the read arose from.
   b. Check both the sequence between the barcodes and before and after the randomized region against the expected sequence and discard reads with improper lengths or mismatches before or after the randomized sequence and discard improper reads.
4. Extract the barcode sequences from each read and compare these to the barcodes used for the sequence libraries to assign each read to a particular sample.

   If carefully chosen barcodes are used, techniques can be used to correct single-base read errors

5. Extract the 5-mer randomized sequence for each read and generate a count table of reads containing each possible sequence in each sample.
6. Optional: Adjust the read counts to account for overrepresenation of rare sequences due to mutations in very abundant sequences using empirical error frequencies.

Count Data Analysis and Fluorescence Intensity Estimation

1. To simplify downstream analysis, eliminate sequences which are not found present in all of the pre-sorting sequencing samples.
2. At this point, a variety of quality control visualizations can be made to validate the Sort-Seq process and begin to compare difference conditions.
    a. Visualizing distributions of sequence counts in each sample using histograms or kernel density estimate plots allow for analysis of the number of sequences present in sufficient quantity to estimate fluorescence intensities.
    b. Running principal component analysis (PCA) on either raw or normalized count data for each sample to visualize between-sample, between-replicate, and between-gate similarity can help verify that samples that should behave similarly or differently actually do and that the gated subpopulations are distinct.
3. With the count data for each library for subpopulations of each sample condition, estimate the fluorescence intensity of each reporter sequence. This can be done in multiple ways.
    a. A simple method consists of a weighted mean calculated by multiplying the number of cells containing each sequence (proportion of reads multiplied by cells sorted into that subpopulation) in each subpopulation

by the fluorescence intensity of that quartile (for instance the median of the output subpopulation intensities or the midpoint of the gate on a log-scale) and dividing by the total number of cells.

b. A more complicated method involves maximum-likelihood estimation of the mean and variance of each reporter sequence fluorescence using minimization algorithms as described in (223, 229).

With the estimates of reporter fluorescence intensity for each sequence, the Sort-Seq assay is complete and the data is ready for downstream experiment-specific analysis.

Results

Sequencing of the SortSeq libraries yielded 69,693,688 total read pairs across 54 total samples. Of these reads, 97.24% properly aligned to the reference sequence. Of these 67,770,670 read pairs, approximately 50% or 34,173,909 contained sequences that passed all filters and were counted for downstream analysis. For this analysis, data was not compensated for read count error induced by high abundance sequences. The number of reads counted for each sample ranged from a minimum of 254,370 to a maximum of 2,262,065, although the majority ranged from 400,000 to 900,000.

For our mutant library consisting of 5 randomized positions, there are a total of 1024 possible sequences available. In the pre-FACS samples, a total of 718 sequences were present with at least 10 counts in all three starting strains. The read counts for each sequence in the starting pools varied across three orders of magnitude, with the 5%, 50% and 95% quantiles being approximately 20, 160, and 2450 respectively. This is perhaps fewer sequences and more variability in sequence abundance than would be expected by a random selection process, indicating that there were significant bottlenecks or bias during the library preparation steps. Libraries prepared from pre-sort cultures either after pooling or after overnight growth before flow cytometry show very similar proportions of each sequence, with correlation coefficients for sequence counts above 0.997 for each strain.

Principal component analysis (PCA) on the proportion of counts for each sequence shows a very strong relationship between the sorting selection for different strains (Figure A1-4A). Each sorted subpopulation clusters together, with the most variable being the two darkest subpopulations. This indicates that much of the variability between conditions is due to general, between-sequence, variability of fluorescence intensity, and also that selection for different intensities was likely reproducible between samples.

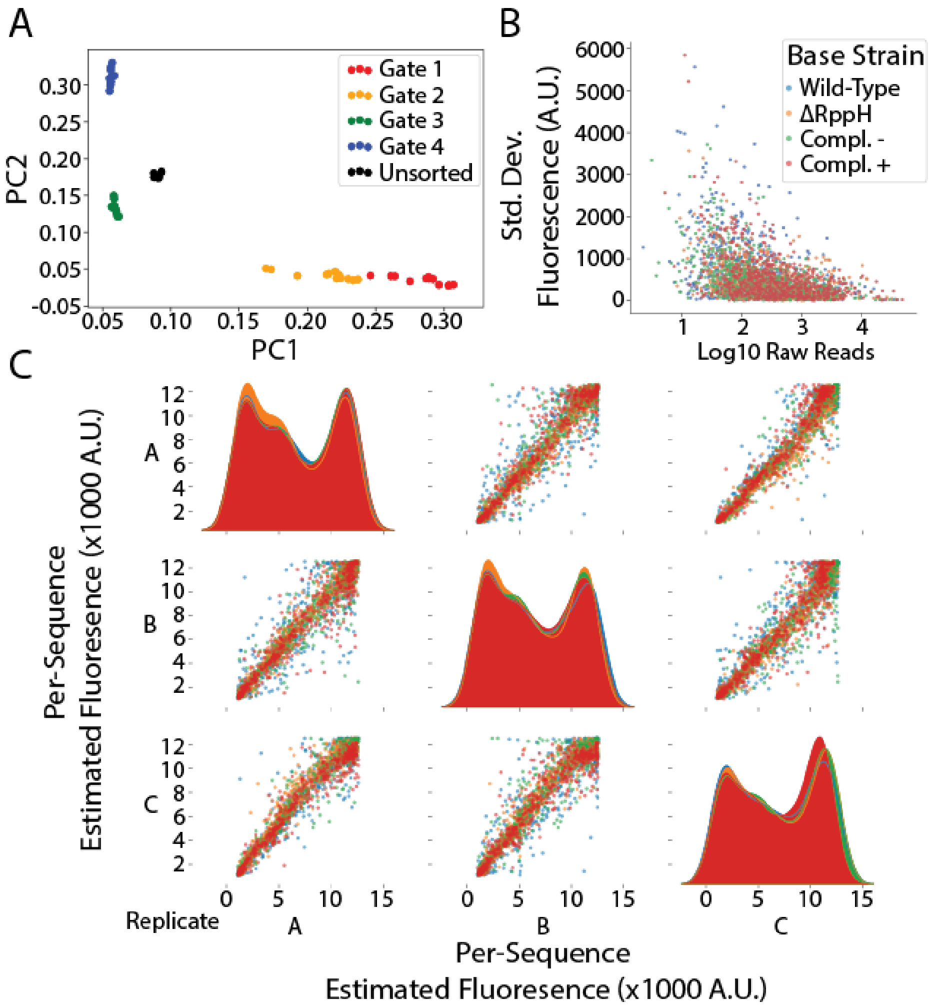


Figure A1-4: Sort-Seq performance analysis.

(A) Principal component analysis (PCA) plot showing the relationship between the individual sequence counts for the four output subpopulations and the input randomized library. (B) Scatter plot showing the relationship between the variability in estimated fluorescence and the starting abundance of each individual mutant sequence. (C) Individual scatter plots showing the reproducibility in estimated fluorescence intensity for each mutant sequence across three biological replicates: A, B, and C.

Fluorescence intensity for each sequence in each sample was estimated by the weighted average-of-averages method as described above. We used the average subpopulation intensity for each of the four subpopulations as measured by flow cytometry of the output populations after overnight growth (Figure A1-3C). We tested estimation of fluorescence intensity using read counts either from individual replicates or by using mean read count across our three replicates. Comparison of estimated intensities from different replicates reveals a strong correlation between replicates A, B, and C, albeit with significant noise (Figure A1-4B). This noise is dramatically more severe for sequences with low read count (Figure A1-4C). As the process of estimating the fluorescence intensity further distorts and biases the experimental noise, we used the mean read count to estimate fluorescence intensities. ultimately, a more complex statistical model could be used to properly estimate the variance of the intensity estimates.

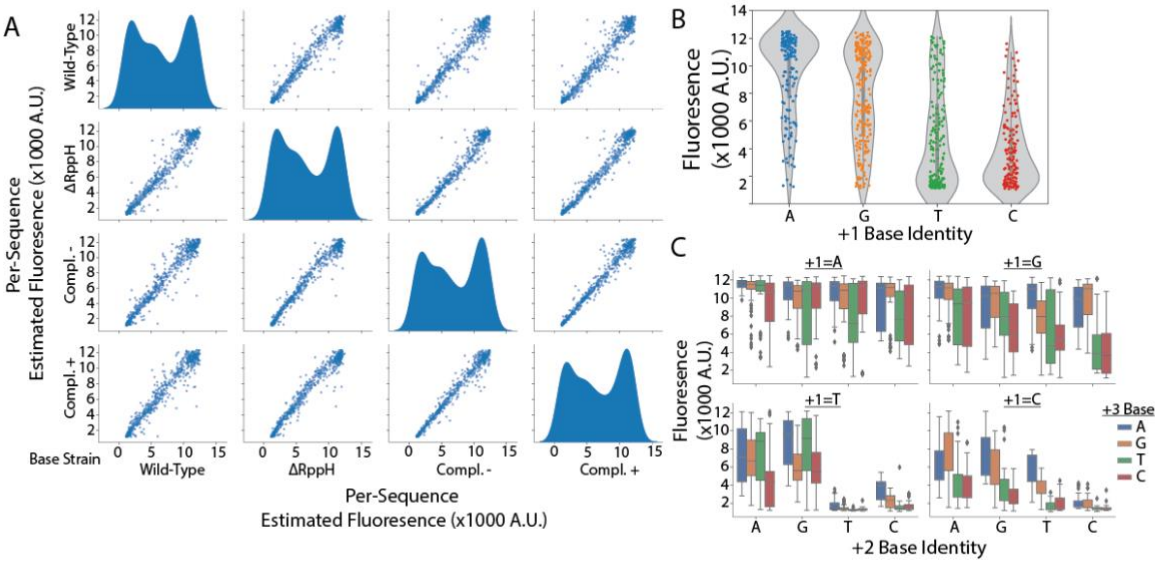


Figure A1-5: Sort-Seq sequence analysis.
(A) Individual scatter plots showing the reproducibility in estimated fluorescence intensity for each mutant sequence across the four biological conditions in the study. Conditions include wild-type *B. subtilis*, a RppH-knockout strain, and an inducible RppH complementation strain with and without inducer. (B) Violin plot showing the distribution of predicted fluorescence intensity for mutant sequence with different nucleotides in the +1 position. (C) Box plots showing the values of predicted fluorescence intensity for mutant sequence with different nucleotides in the +1-3 positions.

The fluorescence intensity estimated for each sequence remained remarkably consistent across all conditions (Figure A1-5A). Pairwise correlation coefficients for fluorescence intensity estimates between all four conditions were above 0.97. This correspondence is increased if data is limited to sequences which average at least 100 reads in the pre-sort samples. The Sort-Seq estimates of fluorescence intensity are accurate, with a close relationship between the estimated intensity of individual mutants as measured by flow cytometry and as estimated by the Sort-Seq method (Figure A1-6).

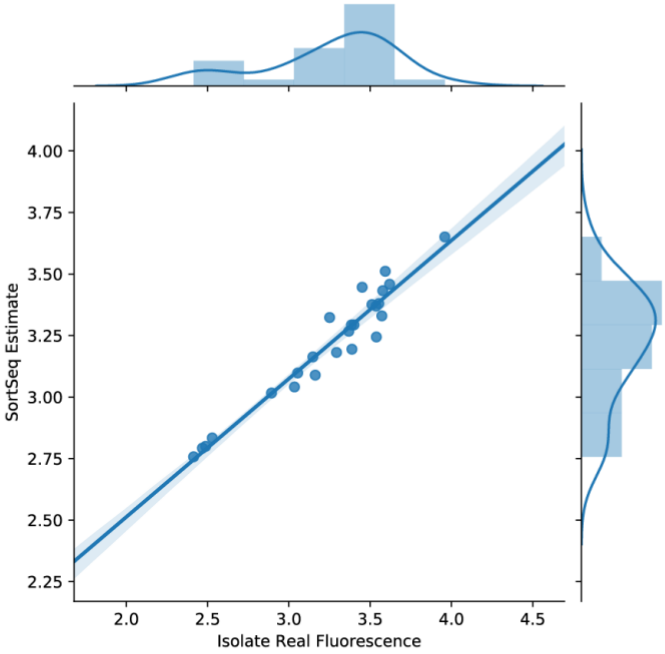


Figure A1-6: Accuracy of Sort-Seq fluorescence estimates compared to isolate fluorescence.
Scatter plot illustrating the relationship between the actual median log-intensity of 25 randomly selected isolate sequences as measured by flow cytometry and the Sort-Seq estimated log-fluorescence.

Although very few differences were observed between conditions, the total dataset represents a quantification of the expression from 718 distinct promoters containing different sequences at the transcription start site. The five-nucleotide region randomized for this experiment begins at the +1 position of the transcript for the native Pconst sequence. In *E. coli*, transcription from the sigma-70 promoter is very dependent on the presence of a purine in the optimal position for transcription initiation(233). The optimal transcription start site is the 11th base pair from the -10 position of the promoter, the position of the initial guanine residue in the Pconst promoter. In the absence of a purine in the optimal position, RNA polymerase may either initiate transcription from a pyrimidine with low efficiency, or from a downstream purine, shifting the transcription start site and lowering efficiency. Additionally, in *B. subtilis* several elements are known to degrade RNA in processes involving the 5′ terminus, and differential efficiencies for these processes may also affect RNA half-life and, correspondingly, translation. This SortSeq approach does not enable determination of the true transcription start site, but does reflect the aggregate sequence effects on transcription efficiency and RNA stability.

As the weighted-mean approach to estimation of intensity clips the minimum and maximum values for estimated intensity to the average intensity of the highest and lowest subpopulations (a limitation which does not apply to the maximum-likelihood approach), we can observe intensities ranging from 1000 to 12500 artificial units (A.U.). Due to the known preference of bacterial RNA polymerase for initiation at purine residues, we expect to see a bias in estimated intensity for those sequences containing a purine residue at the +1 position. This effect is present in the data (Figure A1-5B). Although there are many high-expression +1 purifine sequences and low-expression +1 pyrimidine sequences, there are still many sequences which do not follow this pattern. RNA polymerase is capable of shifting the transcription start site to a downstream purine, so further stratification of the data to examine the effect of downstream nucleotides. Indeed, among sequences containing +1 pyrimidines, +2 purines enable high expression, while +2 pyrimidine sequences are limited to expression levels below 4000 A.U. (Figure A1-5C).

Although disappointing from a perspective hoping for condition-dependent differences in fluorescence intensity for certain sequences, this dataset reflects the reproducibility of the fluorescence estimates generated from this SortSeq experiment and illustrates the effectiveness for observing sequence-specific effects on genetic processes.

Bibliography

1. Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356.

2. Walsh C. 2003. Regulation of Antibiotic Biosynthesis in Producer Organisms, p. 159–174. *In* Antibiotics: Actions, Origins, Resistance1st Edition. ASM Press.

3. Waters LS, Storz G. 2009. Regulatory RNAs in bacteria. Cell 136:615–628.

4. Ray-Soni A, Bellecourt MJ, Landick R. 2016. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. Annu Rev Biochem 85:319–347.

5. Merino E, Yanofsky C. 2005. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21:260–264.

6. Gusarov I, Nudler E. 1999. The mechanism of intrinsic transcription termination. Mol Cell 3:495–504.

7. Greenblatt J, McLimont M, Hanly S. 1981. Termination of transcription by nusA gene protein of Escherichia coli. Nature 292:215–220.

8. Mondal S, Yakhnin AV, Sebastian A, Albert I, Babitzke P. 2016. NusA-dependent transcription termination prevents misregulation of global gene expression. Nature Microbiology 1:15007.

9. Epshtein V, Cardinale CJ, Ruckenstein AE, Borukhov S, Nudler E. 2007. An allosteric path to transcription termination. Mol Cell 28:991–1001.

10. Holmes WM, Platt T, Rosenberg M. 1983. Termination of transcription in E. coli. Cell 32:1029–1032.

11. Epshtein V, Dutta D, Wade J, Nudler E. 2010. An allosteric mechanism of Rho-dependent transcription termination. Nature 463:245–249.

12. Breaker RR. 2011. Prospects for riboswitch discovery and analysis. Mol Cell 43:867–879.

13. Proshkin S, Mironov A, Nudler E. 2014. Riboswitches in regulation of Rho-dependent transcription termination. Biochim Biophys Acta 1839:974–977.

14. DebRoy S, Gebbie M, Ramesh A, Goodson JR, Cruz MR, van Hoof A, Winkler WC, Garsin D a. 2014. A riboswitch-containing sRNA controls gene expression by sequestration of a response regulator. Science 345:937–940.

15. Sedlyarova N, Shamovsky I, Bharati BK, Epshtein V, Chen J, Gottesman S, Schroeder R, Nudler E. 2016. sRNA-Mediated Control of Transcription Termination in E. coli. Cell 167:111–121.e13.

16. Montange RK, Batey RT. 2008. Riboswitches: emerging themes in RNA structure and function. Annu Rev Biophys 37:117–33.

17. Ceres P, Garst AD, Marcano-Velázquez JG, Batey RT. 2013. Modularity of select riboswitch expression platforms enables facile engineering of novel genetic regulatory devices. ACS Synth Biol 2:463–472.

18. Weisberg RA, Gottesman ME. 1999-1. Processive Antitermination. J Bacteriol 181:359–367.

19. Santangelo TJ, Artsimovitch I. 2011-5. Termination and antitermination: RNA polymerase runs a stop sign. Nat Rev Microbiol 9:319–329.

20. Artsimovitch I, Landick R. 2002. The transcriptional regulator RfaH stimulates RNA

chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. Cell 109:193–203.

21. Said N, Krupp F, Anedchenko E, Santos KF, Dybkov O, Huang Y-H, Lee C-T, Loll B, Behrmann E, Bürger J, Mielke T, Loerke J, Urlaub H, Spahn CMT, Weber G, Wahl MC. 2017. Structural basis for λN-dependent processive transcription antitermination. Nat Microbiol 2:17062.

22. Gusarov I, Nudler E. 2001. Control of intrinsic transcription termination by N and NusA: the basic mechanisms. Cell 107:437–449.

23. Cardinale CJ, Washburn RS, Tadigotla VR, Brown LM, Gottesman ME, Nudler E. 2008. Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. Science 320:935–938.

24. Burmann BM, Knauer SH, Sevostyanova A, Schweimer K, Mooney R a., Landick R, Artsimovitch I, Rösch P. 2012. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. Cell 150:291–303.

25. Goodson JR, Winkler WC. 2018. Processive Antitermination. Microbiol Spectr 6.

26. Friedman DI. 1988. Regulation of phage gene expression by termination and antitermination of transcription. The bacteriophages 2:263–319.

27. Patterson TA, Zhang Z, Baker T, Johnson LL, Friedman DI, Court DL. 1994. Bacteriophage lambda N-dependent transcription antitermination. Competition for an RNA site may regulate antitermination. J Mol Biol 236:217–228.

28. Mogridge J, Legault P, Li J, Van Oene MD, Kay LE, Greenblatt J. 1998. Independent ligand-induced folding of the RNA-binding domain and two functionally

distinct antitermination regions in the phage lambda N protein. Mol Cell 1:265–275.

29. Legault P, Li J, Mogridge J, Kay LE, Greenblatt J. 1998. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. Cell 93:289–299.

30. Thapar R, Denmon AP, Nikonowicz EP. 2014. Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins. Wiley Interdiscip Rev RNA 5:49–67.

31. Mogridge J, Mah T-F, Greenblatt J. 1998. Involvement of boxA Nucleotides in the Formation of a Stable Ribonucleoprotein Complex Containing the Bacteriophage λ N Protein. J Biol Chem 273:4143–4148.

32. Nodwell JR, Greenblatt J. 1993. Recognition of boxA antiterminator RNA by the E. coli antitermination factors NusB and ribosomal protein S10. Cell 72:261–268.

33. Mason SW, Li J, Greenblatt J. 1992. Host factor requirements for processive antitermination of transcription and suppression of pausing by the N protein of bacteriophage lambda. J Biol Chem 267:19418–19426.

34. Rees WA, Weitzel SE, Das A, von Hippel PH. 1997. Regulation of the elongation-termination decision at intrinsic terminators by antitermination protein N of phage lambda. J Mol Biol 273:797–813.

35. Nudler E, Gottesman ME. 2002. Transcription termination and anti-termination in E. coli. Genes Cells 7:755–768.

36. Liu K, Zhang Y, Severinov K, Das A, Hanna MM. 1996. Role of Escherichia coli RNA polymerase alpha subunit in modulation of pausing, termination and anti-

termination by the transcription elongation factor NusA. EMBO J 15:150–161.

37. Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R. 2012. Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes Dev 26:2621–2633.

38. Valabhoju V, Agrawal S, Sen R. 2016. Molecular Basis of NusG-mediated Regulation of Rho-dependent Transcription Termination in Bacteria. J Biol Chem 291:22386–22403.

39. Herskowitz I, Signer ER. 1970. A site essential for expression of all late genes in bacteriophage lambda. J Mol Biol 47:545–556.

40. Yarnell WS, Roberts JW. 1992. The phage lambda gene Q transcription antiterminator binds DNA in the late gene promoter as it modifies RNA polymerase. Cell 69:1181–1189.

41. Lowery C, Richardson JP. 1977. Characterization of the nucleoside triphosphate phosphohydrolase (ATPase) activity of RNA synthesis termination factor p. II. Influence of synthetic RNA …. J Biol Chem.

42. Guérin M, Robichon N, Geiselmann J, Rahmouni AR. 1998. A simple polypyrimidine repeat acts as an artificial Rho-dependent terminator in vivo and in vitro. Nucleic Acids Res 26:4895–4900.

43. Li SC, Squires CL, Squires C. 1984. Antitermination of E. coli rRNA transcription is caused by a control region segment containing lambda nut-like sequences. Cell 38:851–860.

44. Squires CL, Greenblatt J, Li J, Condon C, Squires CL. 1993. Ribosomal RNA

antitermination in vitro: requirement for Nus factors and one or more unidentified cellular components. Proc Natl Acad Sci U S A 90:970–974.

45. Arnvig KB, Zeng S, Quan S, Papageorge A, Zhang N, Villapakkam AC, Squires CL. 2008. Evolutionary comparison of ribosomal operon antitermination function. J Bacteriol 190:7251–7257.

46. Berg KL, Squires C, Squires CL. 1989. Ribosomal RNA operon anti-termination. Function of leader and spacer region box B-box A sequences and their conservation in diverse micro-organisms. J Mol Biol 209:345–358.

47. Singh N, Bubunenko M, Smith C, Abbott DM, Stringer AM, Shi R, Court DL, Wade JT. 2016. SuhB Associates with Nus Factors To Facilitate 30S Ribosome Biogenesis in Escherichia coli. MBio 7:e00114.

48. Baniulyte G, Singh N, Benoit C, Johnson R, Ferguson R, Paramo M, Stringer AM, Scott A, Lapierre P, Wade JT. 2017. Identification of regulatory targets for the bacterial Nus factor complex. Nat Commun 8:2027.

49. Clerget M, Jin DJ, Weisberg RA. 1995. A zinc-binding region in the beta’ subunit of RNA polymerase is involved in antitermination of early transcription of phage HK022. J Mol Biol 248:768–780.

50. King RA, Banik-Maiti S, Jin DJ, Weisberg RA. 1996. Transcripts that increase the processivity and elongation rate of RNA polymerase. Cell 87:893–903.

51. Banik-Maiti S, King RA, Weisberg RA. 1997. The antiterminator RNA of phage HK022. J Mol Biol 272:677–687.

52. Robert J, Sloan SB, Weisberg RA, Gottesman ME, Robledo R, Harbrecht D. 1987.

The remarkable specificity of a new transcription termination factor suggests that the mechanisms of termination and antitermination are similar. Cell 51:483–492.

53. Hung SC, Gottesman ME. 1997. The Nun protein of bacteriophage HK022 inhibits translocation of Escherichia coli RNA polymerase without abolishing its catalytic activities. Genes Dev 11:2670–2678.

54. King RA, Weisberg RA. 2003. Suppression of factor-dependent transcription termination by antiterminator RNA. J Bacteriol 185:7085–7091.

55. Vitiello CL, Kireeva ML, Lubkowska L, Kashlev M, Gottesman M. 2014. Coliphage HK022 Nun protein inhibits RNA polymerase translocation. Proc Natl Acad Sci U S A 111:E2368–75.

56. Oberto J, Clerget M, Ditto M, Cam K, Weisberg RA. 1993. Antitermination of early transcription in phage HK022. Absence of a phage-encoded antitermination factor. J Mol Biol 229:368–381.

57. Irnov I, Winkler WC. 2010. A regulatory RNA required for antitermination of biofilm and capsular polysaccharide operons in Bacillales. Mol Microbiol 76:559–575.

58. Bailey MJ, Hughes C, Koronakis V. 1996. Increased distal gene transcription by the elongation factor RfaH, a specialized homologue of NusG. Mol Microbiol 22:729–737.

59. Bailey MJ, Koronakis V, Schmoll T, Hughes C. 1992. Escherichia coli HlyT protein, a transcriptional activator of haemolysin synthesis and secretion, is encoded by the rfaH (sfrB) locus required for expression of sex factor and lipopolysaccharide genes. Mol Microbiol 6:1003–1012.

60. Bailey MJ, Hughes C, Koronakis V. 1997. RfaH and the ops element, components of a novel system controlling bacterial transcription elongation. Mol Microbiol 26:845–851.

61. Reay P, Yamasaki K, Terada T, Kuramitsu S, Shirouzu M, Yokoyama S. 2004. Structural and sequence comparisons arising from the solution structure of the transcription elongation factor NusG from Thermus thermophilus. Proteins 56:40–51.

62. Martinez-Rucobo FW, Sainsbury S, Cheung ACM, Cramer P. 2011. Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. EMBO J 30:1302–1310.

63. Liu B, Steitz TA. 2017. Structural insights into NusG regulating transcription elongation. Nucleic Acids Res 45:968–974.

64. Kang JY, Mooney RA, Nedialkov Y, Saba J, Mishanina TV, Artsimovitch I, Landick R, Darst SA. 2018. Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. Cell 173:1650–1662.e14.

65. Ponting CP. 2002. Novel domains and orthologues of eukaryotic transcription elongation factors. Nucleic Acids Res 30:3643–3652.

66. Kyrpides NC, Woese CR, Ouzounis CA. 1996. KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. Trends Biochem Sci 21:425–426.

67. Mooney RA, Schweimer K, Roesch P, Gottesman M, Landick R. 2009. Two structurally independent domains of E. coli NusG create regulatory plasticity via distinct interactions with RNA polymerase and regulators. J Mol Biol 391:341–358.

68. Burova E, Hung SC, Sagitov V, Stitt BL, Gottesman ME. 1995. Escherichia coli NusG protein stimulates transcription elongation rates in vivo and in vitro. J Bacteriol 177:1388–1392.

69. Yakhnin AV, Murakami KS, Babitzke P. 2016. NusG is a Sequence-specific RNA Polymerase Pause Factor that Binds to the Non-template DNA Within the Paused Transcription Bubble. J Biol Chem.

70. Weixlbaumer A, Leon K, Landick R, Darst SA. 2013. Structural basis of transcriptional pausing in bacteria. Cell 152:431–441.

71. Kang JY, Mishanina TV, Bellecourt MJ, Mooney RA, Darst SA, Landick R. 2018. RNA Polymerase Accommodates a Pause RNA Hairpin by Global Conformational Rearrangements that Prolong Pausing. Mol Cell 69:802–815.e1.

72. Crickard JB, Fu J, Reese JC. 2016. Biochemical Analysis of Yeast Suppressor of Ty 4/5 (Spt4/5) Reveals the Importance of Nucleic Acid Interactions in the Prevention of RNA Polymerase II Arrest. J Biol Chem 291:9853–9870.

73. Nedialkov Y, Svetlov D, Belogurov GA, Artsimovitch I. 2018. Locking the non-template DNA to control transcription. Mol Microbiol.

74. Guo G, Gao Y, Zhu Z, Zhao D, Liu Z, Zhou H, Niu L, Teng M. 2015. Structural and biochemical insights into the DNA-binding mode of MjSpt4p:Spt5 complex at the exit tunnel of RNAPII. J Struct Biol 192:418–425.

75. Ehara H, Yokoyama T, Shigematsu H, Yokoyama S, Shirouzu M, Sekine S-I. 2017. Structure of the complete elongation complex of RNA polymerase II with basal factors. Science 357:921–924.

76. Turtola M, Belogurov GA. 2016. NusG inhibits RNA polymerase backtracking by stabilizing the minimal transcription bubble. Elife 5.

77. Beutin L, Manning PA, Achtman M, Willetts N. 1981. sfrA and sfrB products of Escherichia coli K-12 are transcriptional control factors. J Bacteriol 145:840–844.

78. Zuber PK, Artsimovitch I, NandyMazumdar M, Liu Z, Nedialkov Y, Schweimer K, Rösch P, Knauer SH. 2018. The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. Elife 7.

79. Belogurov GA, Vassylyeva MN, Svetlov V, Klyuyev S, Grishin NV, Vassylyev DG, Artsimovitch I. 2007. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. Mol Cell 26:117–129.

80. Yakhnin AV, Yakhnin H, Babitzke P. 2008. Function of the Bacillus subtilis transcription elongation factor NusG in hairpin-dependent RNA polymerase pausing in the trp leader. Proc Natl Acad Sci U S A 105:16131–16136.

81. Belogurov G a., Mooney R a., Svetlov V, Landick R, Artsimovitch I. 2009. Functional specialization of transcription elongation factors. EMBO J 28:112–22.

82. Yakhnin AV, Babitzke P. 2014. NusG/Spt5: are there common functions of this ubiquitous transcription elongation factor? Curr Opin Microbiol 18:68–71.

83. Burmann BM, Schweimer K, Luo X, Wahl MC, Stitt BL, Gottesman ME, Rösch P. 2010. A NusE:NusG complex links transcription and translation. Science 328:501–504.

84. Strauß M, Vitiello C, Schweimer K, Gottesman M, Rösch P, Knauer SH. 2016. Transcription is regulated by NusA:NusG interaction. Nucleic Acids Res 44:5971–

5982.

85. Tomar SK, Artsimovitch I. 2013. NusG-Spt5 proteins - universal tools for transcription modification and communication. Chem Rev 113:8604–8619.

86. Saxena S, Myka KK, Washburn R, Costantino N, Court DL, Gottesman ME. 2018. Escherichia coli transcription factor NusG binds to 70S ribosomes. Mol Microbiol 108:495–504.

87. Kohler R, Mooney RA, Mills DJ, Landick R, Cramer P. 2017. Architecture of a transcribing-translating expressome. Science 356:194–197.

88. Demo G, Rasouly A, Vasilyev N, Svetlov V, Loveland AB, Diaz-Avalos R, Grigorieff N, Nudler E, Korostelev AA. 2017. Structure of RNA polymerase bound to ribosomal 30S subunit. Elife 6.

89. Proshkin S, Rahmouni AR, Mironov A, Nudler E. 2010. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328:504–508.

90. Banerjee S, Chalissery J, Bandey I, Sen R. 2006. Rho-dependent transcription termination: more questions than answers. J Microbiol 44:11–22.

91. Núñez B, Avila P, de la Cruz F. 1997. Genes involved in conjugative DNA processing of plasmid R6K. Mol Microbiol 24:1157–1168.

92. Paitan Y, Orr E, Ron EZ, Rosenberg E. 1999. A NusG-like transcription anti-terminator is involved in the biosynthesis of the polyketide antibiotic TA of Myxococcus xanthus. FEMS Microbiol Lett 170:221–7.

93. Chatzidaki-Livanis M, Weinacht KG, Comstock LE. 2010. Trans locus inhibitors limit

concomitant polysaccharide synthesis in the human gut symbiont Bacteroides fragilis. Proc Natl Acad Sci U S A 107:11976–11980.

94. Goodson JR, Klupt S, Zhang C, Straight P, Winkler WC. 2017. LoaP is a broadly conserved antiterminator protein that regulates antibiotic gene clusters in *Bacillus amyloliquefaciens*. Nat Microbiol 2:17003.

95. Arutyunov D, Arenson B, Manchak J, Frost LS. 2010. F plasmid TraF and TraH are components of an outer membrane complex involved in conjugation. J Bacteriol 192:1730–1734.

96. Jones CS, Osborne DJ, Stanley J. 1993. Molecular comparison of the IncX plasmids allows division into IncX1 and IncX2 subgroups. J Gen Microbiol 139:735–741.

97. NandyMazumdar M, Artsimovitch I. 2015. Ubiquitous transcription factors display structural plasticity and diverse functions: NusG proteins - Shifting shapes and paradigms. Bioessays 37:324–334.

98. Varon M, Fuchs N, Monosov M, Tolchinsky S, Rosenberg E. 1992. Mutation and mapping of genes involved in production of the antibiotic TA in Myxococcus xanthus. Antimicrob Agents Chemother 36:2316–21.

99. Simunovic V, Zapp J, Rachid S, Krug D, Meiser P, Müller R. 2006. Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methylglutaryl-CoA synthases, and trans-acting acyltransferases. Chembiochem 7:1206–20.

100. Chatzidaki-Livanis M, Coyne MJ, Comstock LE. 2009. A family of transcriptional antitermination factors necessary for synthesis of the capsular polysaccharides of

Bacteroides fragilis. J Bacteriol 191:7288–7295.

101. Newman DJ, Cragg GM. 2016. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod 79:629–661.

102. Sidebottom AM, Carlson EE. 2015. A reinvigorated era of bacterial secondary metabolite discovery. Curr Opin Chem Biol 24:104–111.

103. Roberts JW, Shankar S, Filter JJ. 2008. RNA polymerase elongation factors. Annu Rev Microbiol 62:211–233.

104. Behnken S, Lincke T, Kloss F, Ishida K, Hertweck C. 2012. Antiterminator-Mediated Unveiling of Cryptic Polythioamides in an Anaerobic Bacterium. Angew Chem Int Ed 51:2425–2428.

105. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.

106. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285.

107. Walsh C. 2003. Antibiotics: Actions, Origins, Resistance1 edition. ASM Press.

108. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH. 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43:W237–W243.

109. Koumoutsi A, Chen X-H, Henne A, Liesegang H, Hitzeroth G, Franke P, Vater J,

Borriss R. 2004. Structural and functional characterization of gene clusters directing nonribosomal synthesis of bioactive cyclic lipopeptides in Bacillus amyloliquefaciens strain FZB42. J Bacteriol 186:1084–1096.

110. Herzner AM, Dischinger J, Szekat C, Josten M, Schmitz S, Yakéléba A, Reinartz R, Jansen A, Sahl H-G, Piel J, Bierbaum G. 2011. Expression of the lantibiotic mersacidin in Bacillus amyloliquefaciens FZB42. PLoS One 6:e22389.

111. Wu L, Wu H, Chen L, Lin L, Borriss R, Gao X. 2014. Bacilysin overproduction in Bacillus amyloliquefaciens FZB42 markerless derivative strains FZBREP and FZBSPA enhances antibacterial activity. Appl Microbiol Biotechnol.

112. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6:343–345.

113. Jarmer H, Berka R, Knudsen S, Saxild HH. 2002. Transcriptome analysis documents induced competence of Bacillus subtilis during nitrogen limiting conditions. FEMS Microbiol Lett 206:197–200.

114. Chowdhury SP, Hartmann A, Gao X, Borriss R. 2015. Biocontrol mechanism by root-associated Bacillus amyloliquefaciens FZB42 – a review. Front Microbiol 6.

115. Chen X-H, Vater J, Piel J, Franke P, Scholz R, Schneider K, Koumoutsi A, Hitzeroth G, Grammel N, Strittmatter AW, Gottschalk G, Süssmuth RD, Borriss R. 2006. Structural and functional characterization of three polyketide synthase gene clusters in Bacillus amyloliquefaciens FZB 42. J Bacteriol 188:4024–36.

116. Fan B, Li L, Chao Y, Förstner K, Vogel J, Borriss R, Wu X-Q. 2015. dRNA-Seq Reveals Genomewide TSSs and Noncoding RNAs of Plant Beneficial

Rhizobacterium Bacillus amyloliquefaciens FZB42. PLoS One 10.

117. Irnov I, Sharma CM, Vogel J, Winkler WC. 2010. Identification of regulatory RNAs in Bacillus subtilis. Nucleic Acids Res 38:6637–51.

118. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. 2011-7. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39:e90.

119. Chen XH, Koumoutsi A, Scholz R, Schneider K, Vater J, Süssmuth R, Piel J, Borriss R. 2009. Genome analysis of Bacillus amyloliquefaciens FZB42 reveals its potential for biocontrol of plant pathogens. J Biotechnol 140:27–37.

120. Drögemüller J, Schneider C, Schweimer K, Strauß M, Wöhrl BM, Rösch P, Knauer SH. 2017. Thermotoga maritima NusG: domain interaction mediates autoinhibition and thermostability. Nucleic Acids Res 45:446–460.

121. Frey S, Görlich D. 2014. A new set of highly efficient, tag-cleaving proteases for purifying recombinant proteins. J Chromatogr A 1337:95–105.

122. Coulon A, Ferguson ML, de Turris V, Palangat M, Chow CC, Larson DR. 2014. Kinetic competition during the transcription cycle results in stochastic RNA processing. Elife 3:1–22.

123. Vvedenskaya IO, Vahedian-Movahed H, Bird JG, Knoblauch JG, Goldman SR, Zhang Y, Ebright RH, Nickels BE. 2014. Interactions between RNA polymerase and the “core recognition element” counteract pausing. Science 344:1285–1289.

124. Fiore JL, Nesbitt DJ. 2013. An RNA folding motif: GNRA tetraloop-receptor

interactions. Q Rev Biophys 46:223–264.

125. Cilley CD, Williamson JR. 2003. Structural mimicry in the phage phi21 N peptide-boxB RNA complex. RNA 9:663–676.

126. Mitra P, Ghosh G, Hafeezunnisa M, Sen R. 2017. Rho Protein: Roles and Mechanisms. Annu Rev Microbiol 71:687–709.

127. Bhavsar AP, Zhao X, Brown ED. 2001-1. Development and Characterization of a Xylose-Dependent System for Expression of Cloned Genes in Bacillus subtilis: Conditional Complementation of a Teichoic Acid Mutant. Appl Environ Microbiol 67:403–410.

128. Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM. 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339:62–66.

129. Matz MV, Wright RM, Scott JG. 2013. No control genes required: Bayesian analysis of qRT-PCR data. PLoS One 8:e71448.

130. Aronesty E. 2013. Comparison of Sequencing Utility Programs. Open Bioinforma J 7:1–8.

131. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

132. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.

133. Landy M, Warren GH. 1948. Bacillomycin; an antibiotic from Bacillus subtilis active against pathogenic fungi. Proc Soc Exp Biol Med 67:539–541.

134. Chen XH, Scholz R, Borriss M, Junge H, Mögel G, Kunz S, Borriss R. 2009. Difficidin and bacilysin produced by plant-associated Bacillus amyloliquefaciens are efficient in controlling fire blight disease. J Biotechnol 140:38–44.

135. Barrick JE, Breaker RR. 2007. The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol 8:R239.

136. Stülke J. 2002. Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures. Arch Microbiol 177:433–440.

137. van Tilbeurgh H, Declerck N. 2001. Structural insights into the regulation of bacterial signalling proteins containing PRDs. Curr Opin Struct Biol 11:685–693.

138. Babitzke P, Gollnick P. 2001. Posttranscription Initiation Control of Tryptophan Metabolism in Bacillus subtilis by the trp RNA-Binding Attenuation Protein (TRAP), anti-TRAP, and RNA Structure. J Bacteriol 183:5795–5802.

139. Shu CJ, Zhulin IB. 2002. ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins. Trends Biochem Sci 27:3–5.

140. O'Hara BP, Norman RA, Wan PT, Roe SM, Barrett TE, Drew RE, Pearl LH. 1999. Crystal structure and induction mechanism of AmiC-AmiR: a ligand-regulated transcription antitermination complex. EMBO J 18:5175–5186.

141. Del Papa MF, Perego M. 2008. Ethanolamine Activates a Sensor Histidine Kinase Regulating Its Utilization in Enterococcus faecalis. J Bacteriol 190:7147–7156.

142. Chai W, Stewart V. 1998. NasR, a novel RNA-binding protein, mediates nitrate-responsive transcription antitermination of the Klebsiella oxytoca m5al nasF operon

leader in Vitro1. J Mol Biol 283:339–351.

143. Wilson SA, Wachira SJ, Norman RA, Pearl LH, Drew RE. 1996. Transcription antitermination regulation of the Pseudomonas aeruginosa amidase operon. EMBO J 15:5907–5916.

144. Norman RA, Poh CL, Pearl LH, O'Hara BP, Drew RE. 2000. Steric Hindrance Regulation of the Pseudomonas aeruginosa Amidase Operon. J Biol Chem 275:30660–30667.

145. Boudes M, Lazar N, Graille M, Durand D, Gaidenko TA, Stewart V, van Tilbeurgh H. 2012. The structure of the NasR transcription antiterminator reveals a one-component system with a NIT nitrate receptor coupled to an ANTAR RNA-binding effector. Mol Microbiol 85:431–444.

146. Chai W, Stewart V. 1999. RNA sequence requirements for NasR-mediated, nitrate-responsive transcription antitermination of the Klebsiella oxytoca M5al nasF operon leader1. J Mol Biol 292:203–216.

147. Ann M. Stock, Victoria L. Robinson, Goudreau APN. 2000. Two-Component Signal Transduction. Annu Rev Biochem 69:183–215.

148. West AH, Stock AM. 2001. Histidine kinases and response regulator proteins in two-component signaling systems. Trends Biochem Sci 26:369–376.

149. Galperin MY. 2006-6. Structural Classification of Bacterial Response Regulators: Diversity of Output Domains and Domain Combinations. J Bacteriol 188:4169–4182.

150. Garsin DA. 2010-4. Ethanolamine Utilization in Bacterial Pathogens: Roles and Regulation. Nat Rev Microbiol 8:290–295.

151. Fox KA, Ramesh A, Stearns JE, Bourgogne A, Reyes-Jara A, Winkler WC, Garsin DA. 2009. Multiple posttranscriptional regulatory mechanisms partner to control ethanolamine utilization in Enterococcus faecalis. Proc Natl Acad Sci U S A 106:4435–4440.

152. Ramesh A, DebRoy S, Goodson JR, Fox KA, Faz H, Garsin DA, Winkler WC. 2012. The mechanism for RNA recognition by ANTAR regulators of gene expression. PLoS Genet 8:e1002666.

153. Hall B, Arshad S, Seo K, Bowman C, Corley M, Jhaveri SD, Ellington AD. 2009. In vitro selection of RNA aptamers to a protein target by filter immobilization. Curr Protoc Mol Biol Chapter 24:Unit 24.3.

154. Hoinka J, Berezhnoy A, Sauna ZE, Gilboa E, Przytycka TM. 2014. AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. Res Comput Mol Biol 8394:115–128.

155. Hoinka J, Berezhnoy A, Dao P, Sauna ZE, Gilboa E, Przytycka TM. 2015. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. Nucleic Acids Res 43:5699–5707.

156. Kupakuwana GV, Crill JE 2nd, McPike MP, Borer PN. 2011. Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. PLoS One 6:e19395.

157. Soldevilla MM, Hervas S, Villanueva H, Lozano T, Rabal O, Oyarzabal J, Lasarte JJ, Bendandi M, Inoges S, López-Díaz de Cerio A, Pastor F. 2017. Identification of LAG3 high affinity aptamers by HT-SELEX and Conserved Motif Accumulation (CMA). PLoS One 12:e0185169.

158. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpää MJ, Bonke M, Palin K, Talukder S, Hughes TR, Luscombe NM, Ukkonen E, Taipale J. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res 20:861–873.

159. Ogawa N, Biggin MD. 2012. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. Methods Mol Biol 786:51–63.

160. Pei S, Slinger BL, Meyer MM. 2017. Recognizing RNA structural motifs in HT-SELEX data for ribosomal protein S15. BMC Bioinformatics 18:298.

161. Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. PLoS Comput Biol 5:e1000590.

162. Ruan S, Swamidass SJ, Stormo GD. 2017. BEESEM: estimation of binding energy models using HT-SELEX data. Bioinformatics 33:2288–2295.

163. Morth JP, Feng V, Perry LJ, Svergun DI, Tucker PA. 2004. The Crystal and Solution Structure of a Putative Transcriptional Antiterminator from Mycobacterium tuberculosis. Structure 12:1595–1605.

164. Barrick JE. 2009. Predicting Riboswitch Regulation on a Genomic Scale, p. 1–13. *In* Serganov, A (ed.), Riboswitches: Methods and Protocols. Humana Press, Totowa, NJ.

165. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. Nucleic Acids Res 31:439–441.

166. Barquist L, Burge SW, Gardner PP. 2016. Studying RNA Homology and

Conservation with Infernal: From Single Sequences to RNA Families. Curr Protoc Bioinformatics 54:12.13.1–12.13.25.

167. Tsoy O, Ravcheev D, Mushegian A. 2009. Comparative Genomics of Ethanolamine Utilization. J Bacteriol 191:7157–7164.

168. Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–1337.

169. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

170. Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.

171. Yao Z, Weinberg Z, Ruzzo WL. 2006. CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 22:445–452.

172. Ramesh A, DebRoy S, Goodson JR, Fox K a., Faz H, Garsin D a., Winkler WC. 2012. The mechanism for RNA recognition by ANTAR regulators of gene expression. PLoS Genet 8:e1002666.

173. Drew R, Lowe N. 1989. Positive control of Pseudomonas aeruginosa amidase synthesis is mediated by a transcription anti-termination mechanism. J Gen Microbiol 135:817–823.

174. Zhao Y, Li L, Zheng G, Jiang W, Deng Z, Wang Z, Lu Y. 2018. CRISPR/dCas9-Mediated Multiplex Gene Repression in Streptomyces. Biotechnol J 13:e1800121.

175. Gebbie MP. 2017. Analysis of Genetic Regulatory Mechanisms that Control

Ethanolamine Utilization in Enterococcus faecalis. Digital Repository at the University of Maryland.

176. Murphy MB, Fuller ST, Richardson PM, Doyle SA. 2003. An improved method for the in vitro evolution of aptamers and applications in protein detection and purification. Nucleic Acids Res 31:e110.

177. Peng Q, Vijaya Satya R, Lewis M, Randad P, Wang Y. 2015. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC Genomics 16:589.

178. Illumina. 2014. Low-Diversity Sequencing on the Illumina HiSeq Platform. 770-2014-035.

179. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.

180. Phuong D, Hoinka J, Wang Y, Takahashi M, Zhou J, Costa F, Rossi J, Burnett J, Backofen R, Przytycka TM. 2016. AptaTRACE: Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments.

181. Release Notes - MEME Suite.

182. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J. 2013. DNA-binding specificities of human transcription factors. Cell 152:327–339.

183. Orenstein Y, Shamir R. 2015. HTS-IBIS: fast and accurate inference of binding site motifs from HT-SELEX data. bioRxiv.

184. Hoinka J, Backofen R, Przytycka TM. 2018. AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. Mol Ther Nucleic Acids 11:515–517.

185. Hoinka J, Przytycka T. 2016. AptaPLEX - A dedicated, multithreaded demultiplexer for HT-SELEX data. Methods 106:82–85.

186. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. 2010. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput Biol 6:e1000832.

187. Houman F, Diaz-Torres MR, Wright A. 1990. Transcriptional antitermination in the bgl operon of E. coli is modulated by a specific RNA binding protein. Cell 62:1153–1163.

188. Aymerich S, Steinmetz M. 1992. Specificity determinants and structural features in the RNA target of the bacterial antiterminator proteins of the BglG/SacY family. Proc Natl Acad Sci U S A 89:10410–10414.

189. Smith C, Heyne S, Richter AS, Will S, Backofen R. 2010. Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA. Nucleic Acids Res 38:W373–W377.

190. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474.

191. Bouvet P. 2009. Identification of nucleic acid high-affinity binding sequences of proteins by SELEX. Methods Mol Biol 543:139–150.

192. Andrews S, Others. 2010. FastQC: a quality control tool for high throughput sequence data.

193. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. 2017. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18:80.

194. Ginolhac A, Jarrin C, Robe P, Perrière G, Vogel TM, Simonet P, Nalin R. 2005. Type I polyketide synthases may have evolved through horizontal gene transfer. J Mol Evol 60:716–725.

195. Ridley CP, Lee HY, Khosla C. 2008. Evolution of polyketide synthases in bacteria. Proc Natl Acad Sci U S A 105:4595–4600.

196. Behnken S, Hertweck C. 2012. Cryptic Polyketide Synthase Genes in Non-Pathogenic Clostridium SPP. PLoS One 7:e29609.

197. Czyz A, Mooney RA, Iaconi A, Landick R. 2014. Mycobacterial RNA Polymerase Requires a U-Tract at Intrinsic Terminators and Is Aided by NusG at Suboptimal Terminators. MBio 5:e00931–14.

198. Shi D, Svetlov D, Abagyan R, Artsimovitch I. 2017. Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. Nucleic Acids Res 45:8835–8843.

199. Lawson MR, Ma W, Bellecourt MJ, Artsimovitch I, Martin A, Landick R, Schulten K, Berger JM. 2018. Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein. Mol Cell 71:911–922.e4.

200. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence

similarity searching. Nucleic Acids Res 39:W29–W37.

201. Steiner T, Kaiser JT, Marinkoviç S, Huber R, Wahl MC. 2002. Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities. EMBO J 21:4641–4653.

202. Drögemüller J, Stegmann CM, Mandal A, Steiner T, Burmann BM, Gottesman ME, Wöhrl BM, Rösch P, Wahl MC, Schweimer K. 2013. An autoinhibited state in the structure of Thermotoga maritima NusG. Structure 21:365–375.

203. Chalissery J, Muteeb G, Kalarickal NC, Mohan S, Jisha V, Sen R. 2011. Interaction surface of the transcription terminator Rho required to form a complex with the C-terminal domain of the antiterminator NusG. J Mol Biol 405:49–64.

204. Mayerle M, Woodson SA. 2013. Specific contacts between protein S4 and ribosomal RNA are required at multiple stages of ribosome assembly. RNA 19:574–585.

205. Meyer PA, Li S, Zhang M, Yamada K, Takagi Y, Hartzog GA, Fu J. 2015. Structures and Functions of the Multiple KOW Domains of Transcription Elongation Factor Spt5. Mol Cell Biol 35:3354–3369.

206. Hauser M, Steinegger M, Söding J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 32:1323–1330.

207. Schultz JE, Bruder S, Schultz A, Martinez SE, Zheng N, Beavo JA. 2005. Bacterial GAF domains. BMC Pharmacol 5:S17.

208. Taylor BL, Zhulin IB. 1999-6. PAS Domains: Internal Sensors of Oxygen, Redox

Potential, and Light. Microbiol Mol Biol Rev 63:479–506.

209. Henry JT, Crosson S. 2011. Ligand-binding PAS domains in a genomic, cellular, and structural context. Annu Rev Microbiol 65:261–286.

210. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34:W604–608.

211. Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195.

212. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–80.

213. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One 5:e9490.

214. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

215. Dunbar KL, Büttner H, Molloy EM, Dell M, Kumpfmüller J, Hertweck C. 2018. Genome Editing Reveals Novel Thiotemplated Assembly of Polythioamide Antibiotics in Anaerobic Bacteria. Angew Chem Int Ed Engl 57:14080–14084.

216. Dotu I, Adamson SI, Coleman B, Fournier C, Ricart-Altimiras E, Eyras E, Chuang JH. 2018. SARNAclust: Semi-automatic detection of RNA protein binding motifs from immunoprecipitation data. PLoS Comput Biol 14:e1006078.

217. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp A-C, Munschauer M, Ulrich A, Wardle GS,

Dewell S, Zavolan M, Tuschl T. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141:129–141.

218. Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, Sorek R. 2016. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science 352:aad9822.

219. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854–858.

220. Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow C. 2009. Quantitative phenotyping via deep barcode sequencing. Genome Res 19:1836–1842.

221. Kinney JB, Murugan A, Callan CG Jr, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A 107:9158–9163.

222. Herzenberg LA, Sweet RG, Herzenberg LA. 1976. Fluorescence-activated cell sorting. Sci Am 234:108–117.

223. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol 30:521–530.

224. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM. 2013. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc Natl Acad Sci U S A 110:14024–14029.

225. Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, Wang CL. 2014. Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol Syst Biol 10:748.

226. Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res 24:1698–1706.

227. Peterman N, Lavi-Itzkovitz A, Levine E. 2014. Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. Nucleic Acids Res 42:12177–12188.

228. Rohlhill J, Sandoval NR, Papoutsakis ET. 2017. Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated Escherichia coli Growth on Methanol. ACS Synth Biol 6:1584–1595.

229. Peterman N, Levine E. 2016. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. BMC Genomics 17:206.

230. Guérout-Fleury AM, Frandsen N, Stragier P. 1996. Plasmids for ectopic integration in Bacillus subtilis. Gene 180:57–61.

231. Pronobis MI, Deuitch N, Peifer M. 2016. The Miraprep: A Protocol that Uses a Miniprep Kit and Provides Maxiprep Yields. PLoS One 11:e0160509.

232. Normand R, Yanai I. 2013. An Introduction to High-Throughput Sequencing Experiments: Design and Bioinformatics Analysis, p. 1–26. *In* Shomron, N (ed.), Deep Sequencing Data Analysis. Humana Press, Totowa, NJ.

233. Lewis DEA, Adhya S. 2004. Axiom of determining transcription start points by RNA polymerase in Escherichia coli. Mol Microbiol 54:692–701.