Title
Designing and Engineering Complex Behavior in Living Machines

Author
Temme, Karsten Louis

Publication Date
2011

Peer reviewed|Thesis/dissertation

Designing and Engineering Complex Behavior in Living Machines

By

Karsten Louis Temme

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Joint Doctor of Philosophy
with the University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Associate Professor, Christopher A. Voigt, Chair
Professor, Krishna K. Niyogi
Assistant Professor, J. Christopher Anderson

Fall 2011

Abstract

Designing and Engineering Complex Behavior in Living Machines

by

Karsten Louis Temme

Doctor of Philosophy in Bioengineering

University of California, Berkeley

Associate Professor Christopher A. Voigt, Chair

Living organisms exhibit many fascinating behaviors that profoundly impact the world surrounding us. Some of these functions have been harnessed within biotechnology and redirected towards solving problems of industrial relevance, *e.g.* the microbial synthesis of pharmaceuticals and biofuels. However, these behaviors tend to be encoded by relatively simple genetics and constrained to a handful of lab-friendly organisms. To gain access to diverse, more complex behaviors, I have developed a number of methods for understanding and reprogramming sophisticated genetic networks inside cells. First, I reverse engineered and predictively modified the control of the Type III secretion system in *Salmonella* using mathematical modeling and high-throughput gene expression measurements in single cells. Next, a "refactoring" methodology was developed to entirely reconstruct and specify the genetics of a complex behavior using synthetic parts. I applied this strategy to reprogram and modularize the agriculturally relevant behavior of nitrogen fixation from *Klebsiella*. Then, classical engineering methods for nonlinear systems optimization were adapted to guide the selection of parts and optimize performance in well-specified genetic systems. Finally, orthogonal genetic wires were developed to engineer multiple behaviors in a single cell. I utilized these wires to construct orthogonal networks for sensing environmental conditions, performing computation, and synthesizing red and green pigments in *E. coli*. Together, these techniques and engineering principles represent a major step towards the design and synthesis of entire genomes based solely on genetic information found in sequence databases.

Table of Contents

List of Figures

List of Tables

Acknowledgements

To my advisor, Christopher Voigt, who has inspired me to think of biology as an engineering platform. Thank you for teaching me to pursue, refine, and communicate great ideas, while consistently encouraging me to think bigger.

To my dissertation advisors, Krishna Niyogi, and J. Christopher Anderson, for their advice and support, thoughtful questions and comments.

To my friends in science, who make my time in the lab more than just work. From tinkering with DNA to tasting wine, you have made every day more rewarding. Science is a bumpy road, and it has been a pleasure to weather the journey with you.

To my friends outside of academia, who have shared the joys and heartaches of this journey. You have been a blessing in my life, even when I have been too wrapped up in my work to be there for you.

To my family, for loving me unconditionally and encouraging me beyond imagination. I would not be here without you. Your support is a debt I can never fully repay.

To my wife, Lauren, who has been a constant source of joy and love. Thank you for the big and the small, especially the thankless ways you keep me going. You are a continual inspiration for me.

To my parents, who have always encouraged my curiosity. You have consistently sacrificed to give me opportunities to excel and succeed in life. Your love has been unwavering. Thank you.

Chapter 1 – Synthetic Biology as a New Engineering Discipline

In the past decade, synthetic biology has emerged as a new field at the confluence of biology and engineering, academia and industry. It represents a transition within genetic engineering, from an age where technology limited the pace of scientific inquiry, to a future focused on discovering biological design principles amidst an exploding wealth of data about the living world.

Genetic engineering grew from technological advances, namely the discovery of restriction enzymes and the invention of PCR. These tools ushered in an era in which the DNA of organisms can be edited and copied. However, the scope of genetic engineering has been limited by the availability of naturally occurring sequences and a need to have physical access to source DNA.

Synthetic biology is also being enabled primarily by two technologies. Sequencing has generated extensive databases of information about diverse organisms, while synthesis now enables the creation of any arbitrary DNA sequence. Together, these provide the tools for reprogramming cells with novel behaviors. Databases can be scoured for parts that give rise to desirable functions. Using these parts, genetic programs can be designed on a computer, with the ability to rewrite underlying DNA in any way specified. Programs can then be printed via DNA synthesis, and installed into the organism of our choosing.

Synthetic biologists have made significant advances in reprogramming organisms. Genetic circuits have been constructed that mimic their electronic counterparts producing cells that behave as oscillators,[1] logic gates,[2] counters,[3] toggle switches,[4] and cameras.[5] Microbes have been transformed into chemical factories for the manufacture of drugs,[6] fuels,[7] and nutritional supplements.[8] The future of healthcare may involve tumor-killing bacteria[9] and viruses that destroy biofilms.[10] Despite these advances, the scale and complexity of genetic programs remains consistently simple.[11]

How can we repackage naturally-occurring, complex behaviors and successfully integrate them into sophisticated genetic programs? This dissertation will discuss efforts to apply engineering principles to gain access to complex and unique biological behaviors in order to repurpose them for the betterment of mankind.

Transforming behaviors into reusable parts for synthetic biology begins by understanding how they are wired into the cells where they are found. The following chapter describes characterization of the dynamics of regulatory interactions that lead *Salmonella* to build and operate a nanostructure similar to a hypodermic needle, the type III secretion system. It is with this structure that *Salmonella* secrete proteins that allow them to invade our bodies. In this work, I gained an appreciation for the complexity and elegance of regulation that emerges through evolution. The genetic program was manipulated to change the operation of the needle in predictable ways. However, studying the genetic control of the needle remained a complicated process, fraught with knowledge that we might not fully understand the extent of involved regulation.

What if we could rebuild a complex behavior from scratch, armed with only digital information and a toolbox of synthetic parts? In Chapter 3, a "refactoring" methodology is developed that enables rewriting the DNA encoding clusters of genes to be controlled by new genetic programs. With this technology, we refactor the nitrogen

fixation gene cluster, a grouping of 20 genes responsible for the assembly of the most sophisticated metal-based enzyme in biology and the chemical conversion of atmospheric nitrogen to ammonia - chemistry critical to all life on earth. Via a principled approach similar to software development, electronic data are transformed into a working biological machine. These efforts produced the largest rationally designed and engineered genetic program to date. All of the cornerstones of synthetic biology, including computational design methods, part building and standardization, and DNA synthesis were required for successful completion of this project.

Despite years of work to refactor the nitrogen fixation gene cluster, the activity of the refactored system is less than the naturally occurring system. While the synthetic system carries the benefit of being fully defined and free of unknown regulation, it also begs the question of how to improve upon our initial design. The refactored gene cluster provides a clean platform to study questions about gene cluster design principles and optimization of complex networks. We explore these in Chapter 4, where a method for combinatorially constructing DNA programs is developed. Further, a classical method for optimizing performance of non-linear systems is applied to a subset of the nitrogen fixation pathway. Together, these efforts form the foundation for developing a future platform for refactoring and optimizing complex behaviors.

Finally, Chapter 5 introduces tools and design principles that could facilitate engineering genetic systems. Orthogonal genetic wires are developed from the RNA polymerase from the T7 phage. These enable rationally specified gene expression that does not interact with host transcription processes. The library of polymerases, promoters and terminators serves as a reusable set of parts that can function in many organisms to implement genetic programs. Further, a centralized "controller" is developed to link environmental sensors and computational genetic circuits to the execution of refactored functions. Together, the orthogonal wires and controller are used to engineer production of red (lycopene) and green (deoxychromoviridans) pigments in a single cell under response to varying environmental cues.

The advent of genetic engineering led to products created by microbes and plants that touch our lives daily, including pharmaceuticals such as insulin and human growth hormone, crops that are resistant to pests, and petroleum alternatives like ethanol. A new class of biotechnology companies uses synthetic biology to produce improved biofuels, industrial chemicals, and materials like spider silk.

Aided by new tools and techniques like those described in this dissertation, synthetic biologists can expand the scope of their efforts, engineering organisms to perform radical behaviors. In the coming years, we may be witness to plants that sense and remediate pollution or crops that do not require fertilizer. We will soon embark on a path where future living machines are released from the lab and operate outside bioreactors. Future innovation promises the routine design and manufacture of living machines capable of fascinating behaviors.

Chapter 2 - Reverse Engineering Complex Biological Behavior

Portions of this chapter are reproduced from:
Temme, K. et al. (2008). Induction and Relaxation Dynamics of the Regulatory Network Controlling the Type III Secretion System Encoded within Salmonella Pathogenicity Island 1, J. Mol. Biol. 377, 47-61.

Abstract

Bacterial pathogenesis requires the precise spatial and temporal control of gene expression, the dynamics of which are controlled by regulatory networks. A network encoded within Salmonella Pathogenicity Island (SPI) 1 controls the expression of a type III protein secretion system involved in the invasion of host cells. The dynamics of this network are measured in single cells using promoter-green fluorescent protein (GFP) reporters and flow cytometry. During induction, there is a temporal order of gene expression, with transcriptional inputs turning on first, followed by structural and effector genes. The promoters show varying stochastic properties, where graded inputs are converted into all-or-none and hybrid responses. The relaxation dynamics are measured by shifting cells from inducing to noninducing conditions and by measuring fluorescence decay. The *gfp* expressed from promoters controlling the transcriptional inputs (*hilC* and *hilD*) and structural genes (*prgH*) decay exponentially, with a characteristic time of 50–55 min. In contrast, the *gfp* expressed from a promoter controlling the expression of effectors (*sicA*) persists for 110±9 min. This promoter is controlled by a genetic circuit, formed by a transcription factor (InvF), a chaperone (SicA), and a secreted protein (SipC), that regulates effector expression in response to the secretion capacity of the cell. A mathematical model of this circuit demonstrates that the delay is due to a split positive feedback loop. This model is tested in a *ΔsicA* knockout strain, where *sicA* is complemented with and without the feedback loop. The delay is eliminated when the feedback loop is deleted. Furthermore, a robustness analysis of the model predicts that the delay time can be tuned by changing the affinity of SicA:InvF multimers for an operator in the *sicA* promoter. This prediction is used to construct a targeted library, which contains mutants with both longer and shorter delays. This combination of theory and experiments provides a platform for predicting how genetic perturbations lead to changes in the global dynamics of a regulatory network.

1 Introduction

Type III secretion is a virulence mechanism shared by many gram-negative pathogens [12,13]. It consists of a needle-like structure that crosses both the inner and the outer bacterial membranes and extends 50 nm from the cell surface [14]. It functions to translocate effector proteins from the bacterial cytoplasm to the host cell, where they hijack signaling processes and reorganize actin to facilitate endocytosis [15]. The type III secretion system (T3SS) is a complex molecular machine that requires more than 40 structural, effector, and chaperone genes, most of which are encoded together on a

genomic island [16]. An internal regulatory network controls when and where each of the genes is expressed [17].

Salmonella enterica contains two T3SS that facilitate the different stages of an infection. The needle encoded in SPI-1 can be induced in culture and is required for bacteria to invade mammalian cells [18,19]. Once in the intracellular environment, SPI-1 is strongly repressed and SPI-2 is activated, encoding a different type III apparatus and new effectors [20]. Additional effectors exist at other genomic locations and in the *spv* operon on a virulence plasmid [21,22].

The network is organized such that environmental and cell state regulators converge on three transcription factors (HilC, HilD, and HilA) whose regulation is highly interconnected (Figure 2-1) [23-27]. These transcription factors form a commitment circuit and differentially control the transcription of SPI-1 operons. HilA is required for the transcription of the *prg* operon, whose genes form the inner membrane ring and needle shaft [28-30]. All three *hil* transcription factors induce the long *invFGEABCspaMNOPQRS sicAsipBCDAiacPsicPsptP* operon, which encodes the outer membrane ring and functional components, chaperones, translocators, and effectors [28,30,31]. Two upstream promoters, which are activated by HilC/HilD and HilA, respectively, control the *inv* operon [21,31,32]. In addition, there is an internal promoter upstream of the *sicA* gene, whose activity is controlled by InvF transcription factor and SicA chaperone [33].

The SPI-1 network contains several regulatory motifs whose dynamics have been previously characterized [34]. The network is centered on a four-tiered regulatory cascade [17]. Transcriptional cascades have been shown to alter the timing and stochastic properties of gene expression [35-37]. A coherent feedforward motif occurs when an input activates an intermediate regulator and together they turn on a downstream process (e.g., HilD activates HilC and together they activate HilA) [38]. These motifs can filter short pulses of input and can delay the activation and relaxation of the circuit [39,40]. The SPI-1 network contains several such interlocked feedforward motifs [41].

Several promoters also have interesting architectures that can convert the signaling properties of a pathway. The *hilA*, *invF*, and *prgH* promoters contain multiple cooperative binding sites [29]. In addition, the activator-binding sites of the *hilA* promoter can also be bound by repressor in a mutually exclusive manner [27,42]. These properties have been shown previously to result in all-or-none activation, where individual cells in a population either fully express a gene or do not express it at all [43-45]. The *inv* operon is also controlled by multiple promoters, which can act to integrate the properties of different input signals [31,32].

Positive feedback loops can amplify expression and introduce irreversibility into a switch [46-48]. The commitment circuit formed by HilC/HilD/HilA contains several positive feedback loops [26,49]. Another positive feedback loop is present in a SPI-1 circuit that upregulates the expression of effector proteins in response to the capacity of the cell to export proteins (Figure 2-1) [33,50,51]. At the core of the circuit is the InvF transcription factor, which is only active when bound to the SicA chaperone. Prior to the completion of functional needles, SipC binds to and sequesters SicA [52]. Once the needle is functional, SipC is exported and SicA is free to bind InvF. The InvF:SicA complex then activates the *sicA* promoter, thereby upregulating the expression of additional chaperones and effectors, including some that are encoded outside SPI-1 [21,51]. This circuit forms a "split"

positive feedback loop, where SicA is internal to the loop but InvF is encoded outside the loop.

The cascade, feedforward/feedback motifs, and promoter architectures collectively determine the dynamics of the regulatory pathway. To characterize these dynamics, we construct plasmid-based reporters with transcriptional fusions between SPI-1 promoters and green fluorescent protein. Cells are shifted from noninducing to inducing conditions, and activation dynamics are measured in single cells using flow cytometry. These experiments reveal a temporal ordering in the activation of each promoter and a conversion between graded and all-or-none responses. After induction, the cells are shifted into noninducing conditions, and the decay of fluorescence is measured. There is a pronounced delay in the decay of GFP expressed from the promoter controlling effector expression (*sicA*). In contrast, the GFP from the other SPI-1 promoters decays rapidly and exponentially. Genetic mutants and a mathematical model are used to demonstrate that the delay is due to the split positive feedback loop formed by InvF and SicA. The model is used to predict how the delay can be varied by changing a biochemical parameter. This prediction is confirmed by applying random mutagenesis to an operator and by measuring the change in relaxation delay. Together, the experiments and theory provide a quantitative framework for studying the dynamics of this virulence pathway.

2 Results

2.1 Activation dynamics of the SPI-1 regulatory network

A series of reporter plasmids consisting of transcriptional fusions between four SPI-1 promoters and GFP is constructed. The *hilC*, *hilD*, *prgH*, and *sicA* promoters were chosen to be representative of transcriptional inputs, T3SS structural genes, and effectors, respectively. These plasmids are used in parallel growth experiments[53] in *Salmonella typhimurium* SL1344 to measure the network dynamics in response to shifts in environmental conditions.

The SPI-1 genes are induced when Salmonella enters stationary phase in Luria–Bertani (LB) broth [54,55]. This induction is dependent on the concentration of NaCl. SPI-1 is not induced in 0.085 M NaCl (L-broth) and is maximally induced in 0.3 M NaCl (inducing medium). To induce the SPI-1 genes, Salmonella cultures are grown overnight in L-broth and then diluted into the inducing medium. As the culture grows, aliquots are taken every half hour, and expression is stopped by adding kanamycin and by putting the cells on ice. The samples are then analyzed for fluorescence using flow cytometry. Single-cell experiments are critical when only a subset of the population is active, as has been observed previously for SPI-1 genes [56,57].

The average gated fluorescence is shown in Figure 2-2 as a function of cell density (OD600). As the cell density increases, the *hilD* and then the *hilC* promoters turn on first. Next, the *prgH* promoter, which controls the expression of T3SS structural components, is activated. This delay is consistent with the transcriptional cascade and feedforward loops, which can act to implement an input threshold that must be reached before downstream promoters are activated [34]. Finally, the *sicA* promoter, which controls the expression of chaperones, effectors, and translocators, is induced. A similar

transcriptional ordering has been observed during the assembly of the flagella basal body in *Escherichia coli* [53] and *Caulobacter crescentus* [58].

Different promoters in the SPI-1 network also generate different stochastic effects (Figure 2-3). Both transcriptional inputs (HilC and HilD) follow a graded induction, where the entire population uniformly turns on expression under inducing conditions. In contrast, the *prgH* promoter, which controls structural genes, is activated in an all-or-none manner, where individuals in the population either fully express this operon or do not express it at all [59]. The fraction of the population expressing *prg* genes increases as the population is induced. The architectures of the *hilA* and *prgH* promoters contain motifs that have been previously shown to be able to convert a graded input into an all-or-none output [27,43,44]. The *sicA* promoter, which controls the expression of effectors, has both graded and all-or-none properties. Only a fraction of the population turns on, but that population increases in fluorescence by 200-fold as cells are induced.

2.2 Relaxation dynamics of the SPI-1 regulatory network

The regulatory motifs present in the SPI-1 network could also impact the inactivation of the network when the input signal is removed [34]. In particular, the presence of positive feedback loops and feedforward motifs can impact relaxation time [39,46]. Thus, experiments were performed to determine how the network relaxes when cells are shifted into an environment where SPI-1 is repressed. After SPI-1 is fully induced, an aliquot of bacteria is washed and rediluted into noninducing L-broth, and time points are taken every half hour. As before, the fluorescence in individual cells is analyzed by flow cytometry.

After dilution, the fluorescent reporters expressed from the *hilC*, *hilD*, and *prgH* promoters decay rapidly, whereas expression from the *sicA* promoter persists (Figure 2-3). The GFP expressed from the *hilC* and *hilD* promoters decays with a half-life of 55 ± 6 and 50 ± 14 min, respectively (Figure 2-4). The GFP expressed from the *prgH* promoter, which controls the expression of structural genes, also degrades with a half-life of 50 ± 11 min. Because a stable variant of GFP is used, the primary contributor to fluorescence decay is dilution due to cell division. Indeed, for all three promoters, the decay fits a simple exponential model, which appears as a linear fit on a semilog plot.

The *sicA* promoter exhibits approximately a 1 hr delay, with a half-life of 110±9 min (Figure 2-4). In contrast to the upstream promoters, the decay does not fit an exponential model, implying that it follows nonlinear kinetics, as would be expected from a positive feedback loop. It is noteworthy that the maximum fluorescence from the *hilC* and *sicA* promoters are nearly identical prior to dilution, thus eliminating any potential nonlinear effects due to fluorescence as the source of the delay.

2.3 Reconstitution of *sicA* with and without positive feedback

There are two regulatory mechanisms that together lead to persistence in effector expression. First, the *sicA* promoter is activated by SicA:InvF multimers [50]. This forms a split positive feedback loop because InvF is encoded upstream of the *sicA* promoter and outside the loop (Figs. 2-1 and 2-6). Positive feedback loops can lead to a delay in the deactivation of a circuit. A second mechanism is continued leaky secretion due to a delay in the loss of functional T3SS structures from the cells. Leaky secretion continues to

allow free SicA to be available for binding InvF. It is expected that both leaky secretion and a feedback loop are required for the effector expression to persist.

To assess the role of the feedback loop in the delay, a *ΔsicA* knockout strain, which eliminates the feedback loop, was obtained [50]. In this knockout strain, expression from the *sicA* promoter is completely abolished (Figure 2-5b, PBAD30). The feedback loop can be reconstituted by introducing the *sicA* promoter and the *sicA* gene on a plasmid (Figure 2-5a, P30.psicA.sicA). This plasmid is based on the p15a origin of replication and is cotransformed with the ColE1-based *sicA* reporter plasmid. The reconstituted feedback loop recovers wild-type expression from the *sicA* promoter (Figure 2-5b). The delay that occurs when the network relaxes is slightly longer (165±22 min) than the wild-type feedback loop, which may be the result of reconstituting the loop on a multicopy (p15a origin) plasmid. However, the maximum fluorescence observed from the *sicA* reporter does not increase significantly.

A second plasmid was designed to complement the expression of *sicA*, but to eliminate the positive feedback loop. To complement *sicA*, a plasmid that contains the *sicA* gene under the control of an arabinose-inducible promoter was constructed (Figure 2-5a, PBAD30.sicA). Here, *sicA* is no longer under the control of the *sicA* promoter; thus, the feedback loop has been broken but the network can still be activated by adding arabinose to the media. The network is induced by growing cells in the presence of arabinose, and relaxation is measured by shifting cells into media lacking arabinose. An advantage of using the arabinose-inducible system is that this strain retains the arabinose utilization operon. Thus, the addition of arabinose produces a pulse of expression that reaches a peak 60 min after induction and is completely off 140 min after induction. This eliminates any potential positive feedback effects that could occur in the inducible system itself. The pulse of activity is uniform throughout the cell population.

Constructing the arabinose-inducible plasmid required variation of the ribosome-binding site (rbs) controlling *sicA* expression. Initially, a strong rbs was used. This construct produced a constitutively active variant where, in the absence of arabinose, sufficient SicA was expressed through leaky transcription to turn on the *sicA* promoter on the reporter plasmid. A weak rbs was then tested (BBa_B0033), producing the opposite problem where, even in the presence of arabinose, too little SicA was expressed and the network did not turn on. Finally, an intermediate rbs (BBa_B0032), which was just right, was used. In the absence of arabinose, there is only a slight shift in the fluorescence recorded from the psicA_gfp reporter (Figure 2-5b). When 1.3 mM arabinose is added to the inducing media, the expression from the *sicA* promoter reporter reaches the same fluorescence as wild type.

Growth induction and dilution experiments were performed to assay the impact of the positive feedback loop on the relaxation of expression from the *sicA* promoter. The arabinose inducer is added to the media after 150 min. This roughly mimics an alternative regulatory structure where *sicA* is regulated by a cascade with InvF as an upstream regulator, as opposed to one containing a positive feedback loop (Figure 2-6, Topology 2). After the network is fully induced, the cells are diluted into L-broth lacking arabinose, and the relaxation dynamics are measured. While the inducible construct achieves the same level of activation during induction (Figure 2-5b), the fluorescence rapidly decays after cells are shifted into media lacking the inducer. The delay fits an exponential model with a half-life of 51±5 min, which is equivalent to what is observed for the *hilC*, *hilD*,

and *prgH* promoters (Figure 2-7). The delay returns when the full feedback loop is reconstituted on the plasmid. Together, these experiments demonstrate that the split positive feedback loop is required for the observed relaxation delay.

2.4 Mathematical model of the SicA:InvF genetic circuit

A simple mathematical model of the SicA:InvF positive feedback loop is constructed to quantify the observed relaxation dynamics. A schematic of the circuit and biochemical parameters is shown in Figure 2-6. The model is abstracted to describe a genetic circuit containing three proteins. The transcription factor X (InvF) is activated when bound by chaperone Y (SicA). The chaperone also binds to the effector protein to be secreted, Z (SipC). The formation of protein complexes is assumed to be fast with respect to transcription and translation. This leads to a set of three ordinary differential equations (ODEs) that capture the change in the total number of each protein (X_T, Y_T, and Z_T) in the cell:

$$\frac{dX_T}{dt} = \alpha_x k_0 - \gamma_x (X_T - C_{XY}) - \gamma_{xy} C_{XY} \quad (1)$$

$$\frac{dY_T}{dt} = a_y \left(\eta k_0 + k_1 \frac{C_{XY}}{K + C_{XY}}\right) - \gamma_y (Y_T - C_{XY} - C_{YZ}) - \gamma_{xy} C_{XY} - \gamma_{yz} C_{YZ} \quad (2)$$

$$\frac{dZ_T}{dt} = a_z \left(\eta k_0 + k_1 \frac{C_{XY}}{K + C_{XY}}\right) - \gamma_z (Z_T - C_{YZ}) - \gamma_{yz} C_{YZ} - k_z C_{YZ} \quad (3)$$

where the parameters are defined in Table 2-1 and shown in Figure 2-6. The number of X:Y and Y:Z complexes is given by the relationships $C_{XY} = K_{XY} X_0 Y_0$ and $C_{YZ} = K_{YZ} Y_0 Z_0$, where K_{XY} and K_{YZ} are equilibrium constants, and X_0, Y_0, and Z_0 are the numbers of unbound proteins. In addition, there are three conservation relationships for the total number of each protein: $X_T = X_0(K_{XY} Y_0 + 1)$, $Y_T = Y_0 (K_{XY} X_0 + K_{YZ} Z_0 + 1)$, and $Z_T = Z_0 (K_{YZ} Y_0 + 1)$. Here, we assume that X (InvF) and Y (SicA) form a 1:1 complex. It has been proposed that SicA could interact with InvF as a dimer [51]. If this turns out to be the case, then bistability could be introduced into the model, although even then the bistability occurs in a very narrow slice of parameter space due to the split nature of the feedback loop.

To solve the equations, the network is initially induced by setting $k_0 = 1.0$ transcripts/min. Once the network reaches steady state, the shift to noninducing conditions is simulated by stopping transcription from the upstream promoter ($k_0 = 0$). The relaxation of the network is then recorded. The presence of the positive feedback loop yields the observed delay (Figure 2-7). Importantly, the activity of the feedback loop requires continued secretion from the T3SS ($k_z > 0$). When there is no secretion, the delay is eliminated (as the secretion rate constant $k_z \rightarrow 0$ in Figure 1-8).

When *sicA* is knocked out of the genome and reconstituted on the P_{BAD} inducible plasmid, this produces a circuit architecture resembling a cascade (Figure 2-6, Topology 2). The feedback loop encoded in the *sicA* promoter is not present, and it is predicted that this would reduce the observed delay. To simulate P_{BAD} data, the first term of Eq. (2) is changed to $\alpha_y k_0$; otherwise, the equations remain the same. This has the effect of decreasing the delay, which is observed in the experiments (Figure 2-7).

2.5 Robustness and evolvability of the SicA:InvF genetic circuit

The mathematical model is used to perform a robustness analysis for the circuit [60]. Starting from the set of nominal parameter values (Table 2-1), each parameter is independently varied through a physiologically relevant range. For each parameter set, Eqs. (1)–(3) are solved, and secretion-dependent induction and timescale for the delay in deactivation are measured (Figure 2-8).

Some combinations of parameters result in a circuit that can no longer be induced when secretion becomes active. To keep track of this, the expression from the *sicA* promoter is recorded for cells that are secreting (upper green lines) and compared to the expression when cells are not secreting (lower red lines). When both of these curves are high or low, the circuit is either always on or always off, independent of secretion. The difference between these curves indicates the degree to which the circuit can be induced by secretion. Some of the parameters, such as the degradation rate of free SicA (γ_Y), are robust (Figure 2-8). In contrast, other parameters only have narrow regions that result in functional circuits, including the degradation rate of SicA:SipC complexes (γ_{YZ}), which needs to be slow.

The delay in the deactivation dynamics of each parameter set is also recorded. The timescale is defined as the time required for half of the maxi- mum fluorescence to degrade ($\tau_{1/2}$). Some of the parameters that are relatively robust, such as the expression rate of SipC (a_Z), have almost no affect on the delay (Figure 2-8). The only parameter that has a significant effect on the delay is the binding constant (K) for InvF:SicA binding to the internal *sicA* promoter. By varying this parameter over a physiologically relevant range, the model predicts that the delay can be varied from 1 hr (where the delay to feedback is eliminated) to 4 hr and still produce a secretion-inducible circuit. In essence, this parameter could act as an evolutionary "tuning knob" where mutations could control how long protein expression persists after the network is deactivated.

To test this prediction of the model, the binding constant K was varied experimentally by mutating the known SicA:InvF operator in the *sicA* promoter [51]. Targeted random mutagenesis was used to vary the sequence of the operator on the plasmid containing the reconstituted feedback loop (Figure 2-5a, P30.psicA.sicA) in the *ΔsicA* strain. The relaxation dynamics of seven sequenced mutants were measured (Figure 2-9a). A remarkable diversity of relaxation times was obtained from this small library. The fastest mutant has $\tau_{1/2} = 78 \pm 17$ min, which is only slightly slower than for the *hilC*, *hilD*, *prgH*, and PBAD.sicA decay times (Figure 2-9b). Interestingly, mutants with slower decay times were also obtained, with the slowest being $\tau_{1/2} = 210 \pm 33$ min. Each relaxation trajectory was fitted with the model to obtain the predicted binding constant K (Figure 2-9a). It is intriguing that such a large variety of delay times could be obtained from such a small mutant library.

3 Discussion

Pathogens interact with their hosts through the timed release of multiple protein and chemical effectors. This timing has to be carefully controlled to reflect the specific requirements of particular microenvironments encountered during an infection. Complex molecular machines, such as the T3SS, also require regulation to coordinate their self-

assembly and function. Here, we have experimentally characterized the induction and relaxation dynamics of the regulatory network controlling the *S. typhimurium* T3SS encoded within SPI-1. This network contains a series of integrated genetic circuits that control the temporal and stochastic properties of gene expression when the secretion system is induced. In addition, a split positive feedback loop is shown to affect the relaxation dynamics, where effector expression persists after cells are shifted into noninducing conditions.

Upon induction, the SPI-1 promoters are activated in a manner that is consistent with the self-assembly of the needle. The overall dynamics are very similar to that observed for flagella [53], despite significant divergence with the SPI-1 operon structure and regulatory interactions. In addition, when induced, only a fraction of bacteria turn on the SPI-1 genes [56,57,61]. Single-cell experiments are critical for characterizing the dynamics of networks that are only active in a subset of the population [62].

Other *Salmonella* genes have been shown to be induced within a subpopulation. For example, intracellular bacteria differ in their induction of SPI-2 and other virulence genes [57,63,64]. Stochastic gene expression may lead to different physiological roles; for example, individuals within a single host cell have been observed to bifurcate into a static or a rapid growth state [65]. Population heterogeneity has been observed for other intracellular pathogens, such as the presence of both static and motile *Mycobacterium* phenotypes within a single host cell [66]. The stochastic control of virulence genes enables individuals in a population to probabilistically vary their physiological behavior at each stage of an infection [67].

Within the SPI-1 regulatory network, there is a genetic circuit that links the expression of effectors with the secretion state of the cell [51]. A unique aspect of this circuit is that the positive feedback loop is split (Figure 2-6, Topology 1). One can imagine two alternate network topologies. Each topology can be secretion-inducible, but they differ in the way the network relaxes. First, both the chaperone and the transcription factor could be encoded upstream of the internal promoter (Figure 2-6, Topology 2). This would produce a cascade, as is often found in regulatory networks [38]. This topology produces a significantly shorter relaxation delay, and it is difficult to increase the delay by changing the model parameters. This topology is roughly simulated when the *sicA* gene is placed under inducible control, in which case the network is similarly induced, but the delay is almost eliminated (Figure 2-7).

The SicA:InvF circuit, including the split feedback loop, is conserved in *Yersinia enterocolitica* (YsaE, SycB, and YspC are homologous to InvF, SicA, and SipC) [68]. Interestingly, *Shigella flexneri* also has homologues to these proteins (MxiE, IpgC, and IpaC); however, the genomic organization and the resulting genetic circuit differ considerably [69]. In this circuit, the IpgC:MxiE complexes activate outside promoters controlling effector expression, but they do not upregulate their own production [70]. Therefore, there is no positive feedback loop, and the circuit resembles a simple cascade (Topology 2). It will be interesting to compare the induction and relaxation dynamics of the *Salmonella* circuit with those encoded in *Yersinia* and *Shigella*.

A third topology, where both the transcription factor and the chaperone are downstream of the internal promoter, is also possible (Figure 2-6, Topology 3). This forms a complete positive feedback loop, where all of the required genes are internal to the loop. If there is higher-order cooperativity in the nonlinear term, this topology can

produce a bistable switch, which is difficult to deactivate once it is turned on [46-48]. A similar topology is present in the regulatory network controlling the expression of flagella [71]. Within the network, a genetic circuit controls the expression of flagellin, which forms the filament. After the basal body is constructed, the sigma factor FliA (analogous to *X*) activates its own expression and an anti-sigma factor FlgM (analogous to *Y*), which is secreted. After the filament is constructed, FlgM accumulates in the cytoplasm and deactivates the positive feedback loop by binding to FliA. The irreversibility intrinsic to this topology with respect to the input signal enables the circuit to continue the formation of filament until the flagellum is fully constructed once the pathway is initiated. This irreversibility would be disadvantageous in the case of SPI-1 because the effectors are temporally induced to promote invasion, after which their expression needs to be downregulated.

The SicA:InvF split feedback loop and the resulting relaxation delay may have a physiological role in controlling a physiological bifurcation that occurs immediately post-invasion. After invading a macrophage, Salmonella either persists and grows in the intracellular environment or rapidly induces apoptosis within an hour [55,72]. Rapid apoptosis is dependent on the SipB effector (transcribed from the *sicA* promoter) and caspase-1 [72,73]. Transcription from the *sicA* promoter has been shown to persist during the first hours after invasion [61]. The inability to turn off SipB expression post-invasion (via a *Δlon* mutant) has been shown to increase the probability of apoptosis [74,75]. The length of effector expression post-invasion could alter the ratio of persistent cells to apoptotic cells, and, ultimately, the speed of pathogenesis [76]. The SicA:InvF operator in the *sicA* promoter could act as an evolutionary tuning knob by which the probability of inducing apoptosis could be controlled. This hypothesis could be tested by perturbing the strength of the feedback loop and by measuring the fraction of infected macrophages that enter an apoptotic state.

The combination of theory with experiments provides a quantitative platform for understanding the dynamics of regulatory networks. These tools enable more refined methods to interrogate networks by perturbing their underlying biochemical mechanisms. Here, this approach is applied to the regulatory network controlling SPI-1 in two ways. First, we were able to rewire the network to convert a split positive feedback loop into a cascade. Second, mutations were made to a transcription-binding site. In both cases, the model was used both to predict the effect of making these genetic changes and then to quantitatively frame the resulting change in dynamics. This represents a powerful approach to perturbing and understanding how sets of biochemical interactions lead to the observed complex global dynamics.

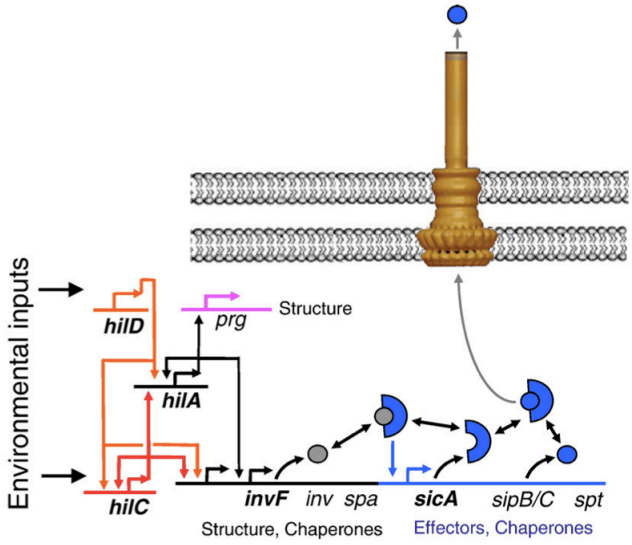

Figure 2-1 The regulatory network encoded within SPI-1 is shown. The SPI-1 T3SS is induced by a number of environmental and cell state signals, such as osmolarity, Mg2+, and extracellular phosphate, which differentially affect the regulators *hilD*, *hilC*, and *hilA* [77]. These transcription factors coregulate each other and control different operons internal to SPI-1. Only HilA regulates the expression of genes forming the inner membrane ring and shaft (*prg*), but all three converge at two promoters controlling a long operon that begins with the *invF* gene. There is a promoter internal to this operon that is activated by InvF when bound to the SicA chaperone [33,50,51]. The chaperone is sequestered by the SipC protein until it is exported, after which SicA can bind to InvF, leading to the up-regulation of additional effector proteins. The colors correspond to the temporal data shown in Figure 1-2. The cryo-electron microscopy structure of the needle complex is shown [14].

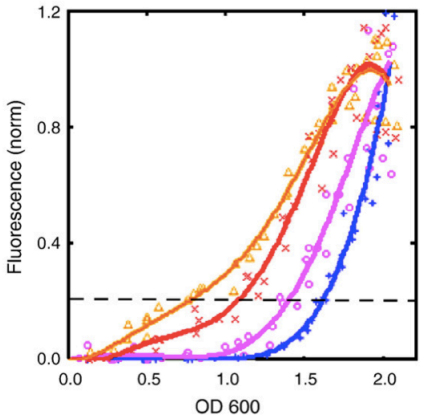

Figure 2-2 The SPI-1 promoters are induced in a temporal order. The *hilD* (orange), *hilC* (red), *prgH* (purple), and *sicA* (blue) promoters are transcriptionally fused to *gfp* and introduced into cells on a plasmid. Cells are grown in noninducing media (L-broth) and diluted into inducing media, and fluorescence is measured in single cells as a function of density (OD600). The average gated fluorescence of the population is calculated for the fraction of the population that turns on at the end of the induction experiment. The average is then normalized by the maximum [53]. The data represent four replicate experiments performed on different days for each promoter. Each data point is an aliquot from these 16 growth experiments. The lines are drawn for each promoter using a polynomial averaging algorithm. The dashed line is drawn as a reference for the point when the promoter reaches 20% of its maximum.

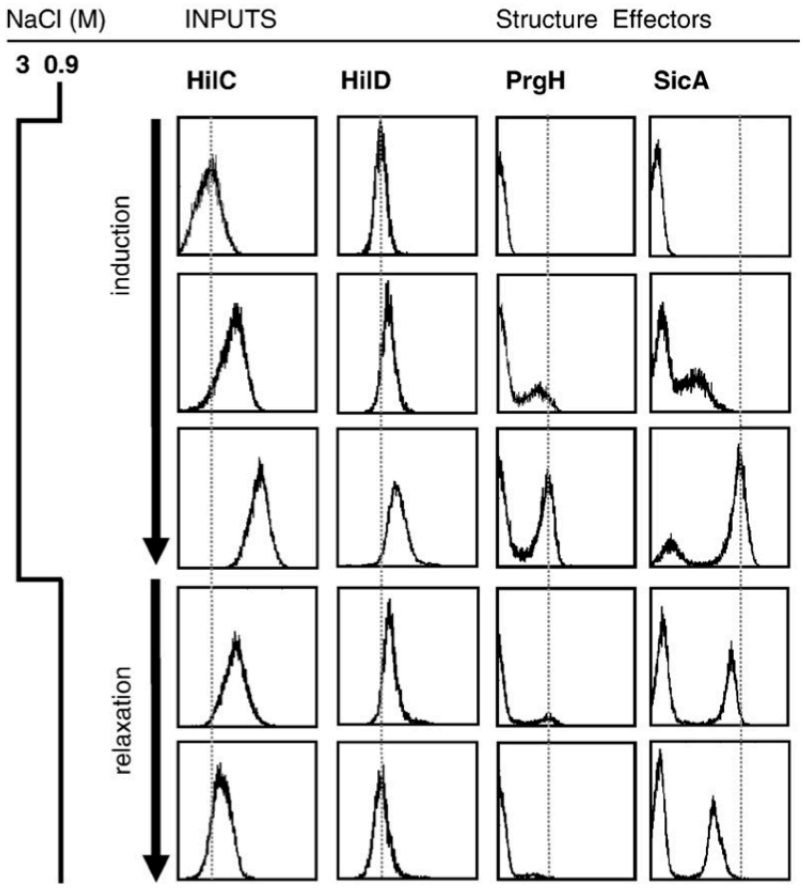

Figure 2-3 SPI-1 promoters follow graded, all-or-none, or hybrid induction dynamics. Gated cytometry data are shown for the *hilC*, *hilD*, *prgH*, and *sicA* promoters transcriptionally fused to *gfp*. Cells are grown in noninducing media (L-broth with 0.085 M NaCl) and then shifted into inducing media (LB broth with 0.3 M NaCl). Data are shown for each promoter preinduction (first row), intermediate induction (second row), and full induction (third row). The OD600 values of these samples are: *hilC* (0.13, 0.80, and 1.9), *hilD* (0.11, 0.75, and 1.8), *prgH* (0.12, 1.5, and 2.06), and *sicA* (0.12, 1.4, and 2.07). Both the *hilC* and the *hilD* promoters have a graded response to the shift in condition, where the entire population is induced uniformly. In contrast, the *prgH* promoter has an all-or-none response, and the *sicA* promoter has a hybrid response with features of both all-or-none and graded dynamics. After the cells are fully induced, they are shifted back into noninducing conditions. Fluorescence from the *hilD*, *hilC*, and *prgH* promoters decays rapidly to the preinduced level. However, expression from the *sicA* promoter persists strongly. After dilution in L-broth, time points are shown after 50 and 140 min. The dashed lines are drawn as a guide to the eye.

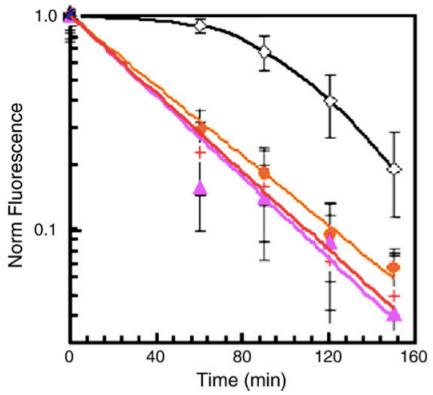

Figure 2-4 Expression from the *sicA* promoter persists after cells are shifted into noninducing conditions. Cells are grown in inducing media (0.3 M NaCl LB broth) until the network is fully activated after 400 min. Cells are then shifted into noninducing media (0.085 M NaCl LB broth) and grown for 150 min. The decay of fluorescence is shown after the cells are shifted into noninducing conditions (t=0 min). The fluorescence values are normalized by the maximum fluorescence prior to dilution. On a semilog plot, exponential decay appears as a straight line, as is the case for the *hilC* (orange circles), *hilD* (red +), and *prgH* (pink triangles) promoters. There is a significant delay in the decay of the *sicA* promoter (white diamonds). The straight lines are the best fitted to an exponential equation, as determined using linear regression algorithm. It is noteworthy that the *sicA* and *hilC* promoters have almost equal fluorescence prior to the shift to noninducing conditions. Each data point represents the average of four experiments performed on different days, and the standard deviation of these experiments is shown as error bars.

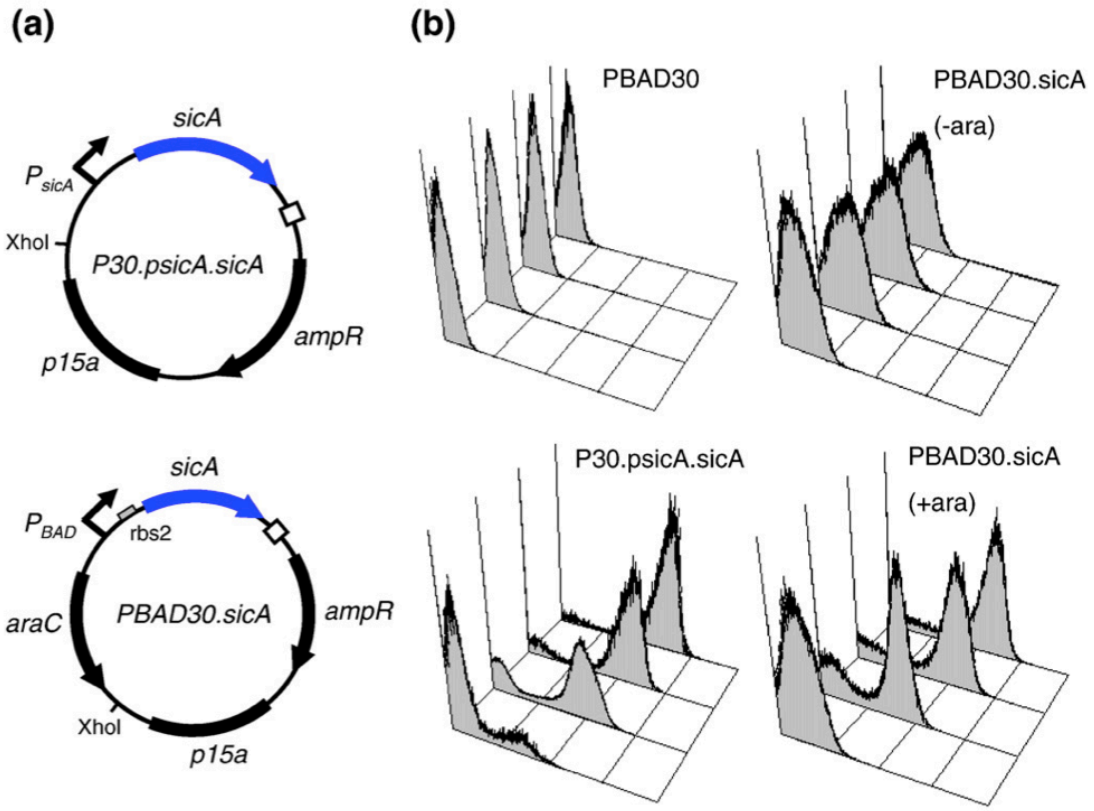

Figure 2-5 An inducible *sicA* plasmid was constructed and tested for complementation in a *ΔsicA* knockout strain. (a) To reconstitute the feedback loop, the *sicA* promoter and gene were cloned together from the Salmonella genome and ligated into PBAD30 at an introduced XhoI site that eliminates the *araC* gene and the PBAD promoter (top). To create an inducible system without the feedback loop, the *sicA* gene was cloned into the PBAD30 plasmid (bottom). Different rbs were tested to obtain an inducible construct. (b) Cytometry data are shown for the *ΔsicA* knockout containing the psicA_gfp reporter on a ColE1 plasmid. Data are shown (bottom to top) 2, 4, 6, and 8 hr after the shifting of the cells into inducing conditions. No activity from the *sicA* promoter was observed when *ΔsicA* cells carried the empty PBAD30 plasmid as a control (top left). When the *sicA* gene is driven by the wild-type *sicA* promoter (P30.psicA.sicA), induction is recovered (bottom left). When the *sicA* gene is fused to an arabinose-inducible promoter (PBAD30.sicA), the *sicA* promoter is induced in the presence of 1.33 mM arabinose (bottom right), but not in the absence of arabinose (top right).

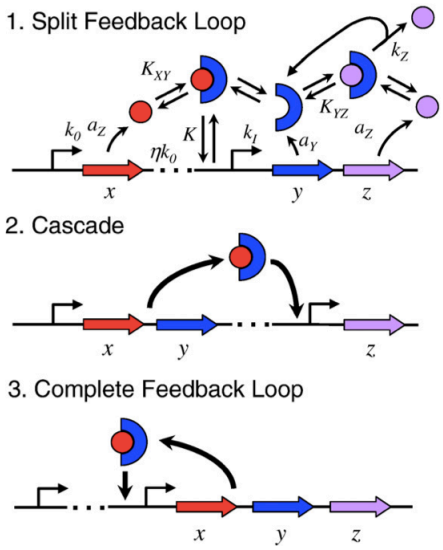

Figure 2-6 A generalized schematic of the secretion-control circuit is shown for the topology that appears in *Salmonella* and for two alternative topologies (Topologies 2 and 3). Topology 1 shows the split positive feedback loop motif, where the transcription factor *X* (InvF) appears external to the feedback loop, but the activating chaperone *y* (SicA) is internal to the loop. The parameters used in the model are shown (Eqs. (1)–(3) and Table 1-1). Topology 2 is a simple cascade where both the transcription factor and the chaperone are outside the internal promoter, thus eliminating any feedback. This topology is roughly recreated when *sicA* is placed under inducible control and inducer is added at the time when *invF* is expressed (Figs. 1-5a and 1-7). Topology 3 is a complete feedback loop, where all of the necessary components are internal to the loop.

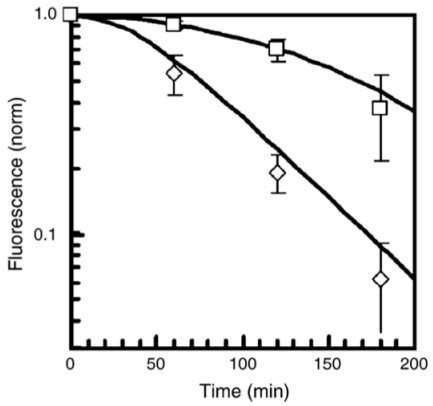

Figure 2-7 The persistence of expression from the *sicA* promoter is eliminated by disruption of the positive feedback loop. Data are shown for the *ΔsicA* deletion mutant, where the *sicA* gene is complemented either by an inducible construct (Figure 1-5a, PBAD30.sicA) or under the control of the *sicA* promoter (P30.psicA.sicA). Cells are grown in inducing conditions and then shifted into noninducing conditions (t = 0 min), and the decay in fluorescence is measured. When the *sicA* promoter is under the control of an inducible promoter, the total fluorescence from the reporter reaches approximately the same maximum (Figure 1-5b), but the delay is eliminated (diamonds). When the feedback loop is reconstituted on the plasmid, the delay is recovered (squares). The lines represent the fit to the model (Eqs. (1)–(3) and Table 1-1). Each data point is the average of six experiments performed on different days, and error bars represent the standard deviation. The data are normalized by the maximum fluorescence under inducing conditions.

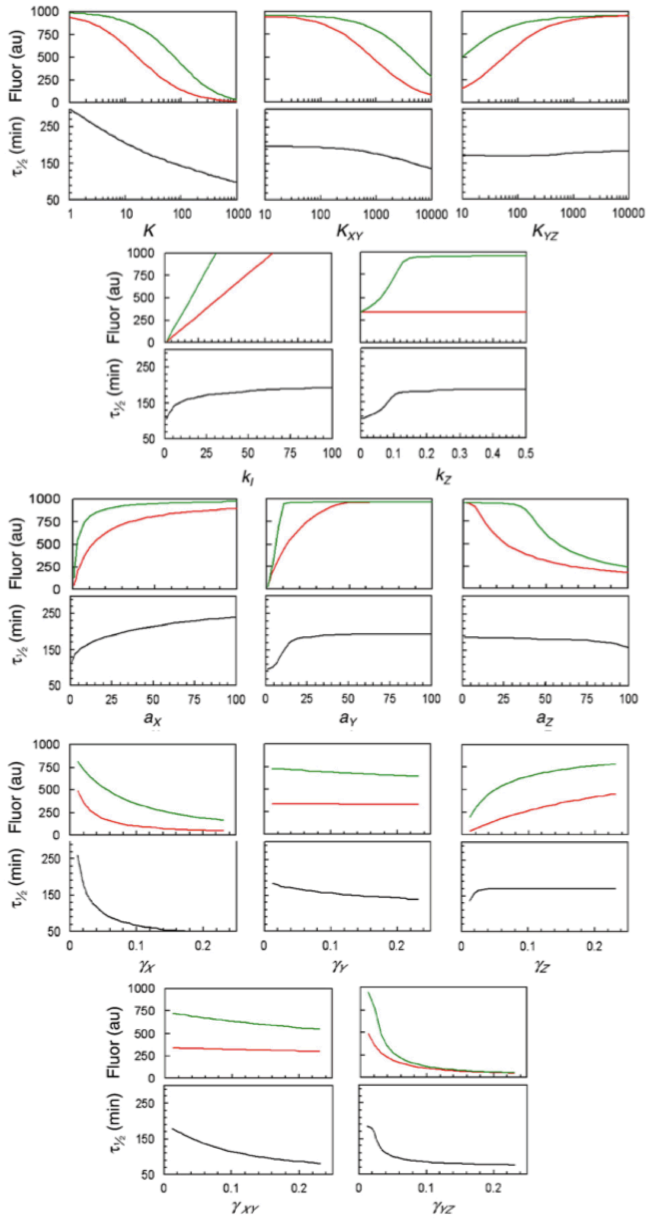

Figure 2-8 A robustness analysis of the model is performed. Each parameter is independently varied from the nominal set (Table 1-1). For each set of parameters, Eqs. (1)–(3) are solved, and three values are recorded from the simulation: the maximum fluorescence without secretion (top graph, red line), the maximum fluorescence with secretion (top graph, green line), and the half-life for the decay of fluorescence ($\tau_{1/2}$; bottom graph). Very few of the parameters have the capacity to affect the decay half-life without resulting in a circuit that cannot be induced by secretion. Note that when there is no secretion ($k_z = 0$), the positive feedback loop is broken, and the delay is eliminated. The units for the parameters are provided in Table 1-1.

(a)

	SicA:InvF	-35	$\tau_{1/2}$ (min)	*K* (nM)
wt sicA	ATTGGGCGTTTTTTGAATGTTC		162 ± 22	32 ± 14
wt sopE	A**AA**GG**A**C**T**TTTTTTGAA**A**GTTC		165 ± 8	35 ± 10
library	A**NN**GG**N**C**T**TTTTTTGAA**N**GTTC			
A.1	A**AA**GG**A**C**T**TTTTTTGAATGTTC		210 ± 33	2 ± 8
A.2	ATTGG**A**C**T**TTTTTTGAATGTTC		190 ± 20	18 ± 8
A.3	A**A**TGGGC**T**TTTTTTGAATGTTC		173 ± 26	32 ± 18
A.4	A**GG**GG**A**C**T**TTTTTTGAA**A**GTTC		143 ± 20	48 ± 9
A.5	A**GA**GG**T**C**T**TTTTTTGAA**C**GTTC		137 ± 13	56 ± 10
A.6	A**GA**GG**A**C**T**TTTTTTGAA**C**GTTC		132 ± 8	58 ± 6
A.7	A**CG**GGGC**T**TTTTTTGAA**C**GTTC		78 ± 17	146 ± 81

(b)

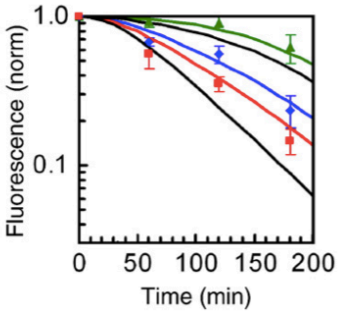

Figure 2-9 Mutations to the SicA:InvF-binding site in the internal promoter vary the strength of the positive feedback loop, resulting in different relaxation times. (a) The -50 to -30 region of the *sicA* promoter, which includes the previously identified SicA:InvF-binding site, is shown [51]. For each sequence, the resulting empirical half-life ($\tau_{1/2}$) is shown in addition to the binding constant (*K*) derived from fitting the mutant data to the model. Error bars are reported as the standard deviation of six experiments performed on different days. When the binding site from the *sicA* promoter is mutated to the *sopE* site, there is no change in the observed half-life. A small library, where the differences between the *sicA* and *sopE* operators were randomly mutated (library), was then created. The resulting mutants have a significant diversity of decay times (A.1–A.7). (b) The decay data are shown for three representative mutants. The wild-type sicA promoter (top black line) and the PBAD30.sicA (lower black line) data are shown as reference (**Figure 2-**). The mutant with the longest decay time (A.1, red triangles) and two mutants with the fastest decay times (A.6, blue diamonds; A.7, red squares) are shown. Error bars represent the standard deviation of six experiments performed on different days.

Table 2-1 SPI-1 Model Parameters

Name	Description	Value	Units
k_0	Transcription rate of external promoter	1	transcripts min^{-1}
k_1	Transcription rate of internal promoter[a]	22	transcripts min^{-1}
η	Efficiency of transcriptional read-through[b]	0.01	
α_x	X proteins produced per transcript[a]	12	proteins $transcript^{-1}$
α_y	Y proteins produced per transcript[a]	14	proteins $transcript^{-1}$
α_z	Z proteins produced per transcript[c]	90	proteins $transcript^{-1}$
γ_x, γ_{xy}	Degradation rate of X and XY complexes[d]	0.022	min^{-1}
γ_Y	Degradation rate of Y[a,e]	0.046	min^{-1}
γ_z	Degradation rate of Z[f]	0.139	min^{-1}
γ_{zy}	Degradation rate of ZY complexes[f,g]	0.022	min^{-1}
K_{xy}	Equilibrium dissociation constant for XY dimers[a,h]	1700	nM
K_{yz}	Equilibrium dissociation constant for YZ dimers[h,i]	28	nM
k_z	Secretion rate of Z[j]	0.1	min^{-1}
K	Equilibrium dissociation constant of XY to the internal promoter[h]	33	nM

[a]These data were fitted to the experimental data shown in Figure 2-7. The remainder was fixed to values estimated from the literature.
[b]This parameter captures the probability that an mRNA transcribed from the upstream promoter also encodes the *sicA* gene. The two genes are encoded 25 kb apart.
[c]The production rate is set to reach a steady state of 6,000–10,000 SipC proteins/cell. The intracellular concentration of SipA has been measured to be 6,000±3,000 molecules/cell [78].
[d]Corresponds to a ~32-min in vivo half-life. SicA does not affect the stability of InvF in vivo [51].
[e]The degradation rate of SicA is set to a half-life of 15 min.
[f]In the absence of chaperone, the SipC protein degrades rapidly in vivo. This rate has been shown to range between 2 and 10 min in vivo in *Salmonella*[52] and *Shigella* [79]. Here, the rate is chosen to yield a half-life of 5 min.
[g]The SipC protein is stabilized when bound to SicA [52]. In the presence of SicA, the in vivo half-life of SipC has been measured to be ~30 min. The monomeric form of SipC interacts with SipB, greatly increasing its degradation rate. Dilution is the primary mode of degradation, so this half-life reflects the cell division time. For the analogous system in *Shigella*, IpaC degrades in the cytoplasm with a half-life of ~5 min, but is stabilized when bound to IpgC (>20 min) [79,80].
[h]Assumes an intracellular volume of 10^{-15} L. The units are converted from concentration to molecules for Eqs. (1)–(3).
[i]The SipC:SicA binding affinity has been measured [81].
[j]The rate of IpaC (homologous to SipC) secretion has directly been measured to be 0.17 min^{-1} in Yersinia [82]. The secretion rate of SipA was similarly measured and found to be 0.06 min^{-1} [78].

Chapter 3 - Refactoring Biological Behavior via Synthetic Biology

Abstract

Bacterial genes associated with a single trait are often grouped in a contiguous unit of the genome, known as a gene cluster. It is difficult to genetically manipulate many gene clusters due to complex, integrated host regulation. We developed a systematic approach to completely remove native regulation and specify the genetics of a gene cluster by rebuilding it from the bottom-up using only synthetic, well-characterized parts. We applied this approach to an agriculturally relevant gene cluster from *Klebsiella oxytoca* encoding the nitrogen fixation pathway for converting atmospheric N_2 to ammonia. The native gene cluster consists of 20 genes in 7 operons and encoded in 23.5kb of DNA. We constructed a "refactored" gene cluster sharing little DNA sequence identity with wild-type and for which the function of every component genetic part is defined. This work demonstrates the potential for synthetic biology tools to rewrite the genetics encoding biological functions to facilitate its access, engineering, and transferability.

1 Introduction

Many functions of interest for biotechnology are encoded in gene clusters, including metabolic pathways, nanomachines, nutrient scavenging mechanisms, and energy generators [83]. Gene clusters can be very large, encoding dozens of genes and spanning more than 100kb of DNA. Often, the cluster contains internal regulation that is highly embedded in the global regulatory network of the organism. Promoters and 5'-UTRs are complex and integrate many regulatory inputs [84,85]. Regulation is highly redundant; for example, the existence of multiple embedded feedforward and feedback loops [86]. Regulation can also be internal to genes, including promoters, pause sites, small RNAs, and sensitivity to supercoiling [87,88]. Further, genes often physically overlap and regions of DNA can have multiple functions [89]. Together, the redundancy and extent of this regulation can make it difficult to manipulate a gene cluster to break its control by native environmental stimuli, optimize its function, or transfer it between organisms. As an example, many clusters that have been identified due to genomic efforts are cryptic, meaning that laboratory conditions cannot be identified in which they are active [90].

Gene clusters have been controlled from the top-down by manipulating the native regulation or adding synthetic regulation. For example, either knocking out a repressor or overexpressing an activator has turned on clusters encoding biosynthetic pathways [91-95]. When the cluster is a single operon, it has been shown that a promoter can be inserted upstream to induce expression [96]. Similar approaches have been used to produce pharmaceutical compounds in a heterologous host using DNA obtained from environmental, metagenomic, microbiome, and otherwise unculturable organisms [97]. The entire echinomycin biosynthetic pathway was transferred from *S. lasaliensis* into *E. coli* by placing each native gene under the control of a synthetic promoter [98]. Synthetic DNA has been used to rebuild complete gene clusters and genomes [99-101], and this enables the

production of physical DNA based on information from sequence databases. However, the genetics of clusters encoding valuable functions are often poorly characterized, making accessing the function difficult.

We have developed a bottom-up process to systematically remove the regulation of a gene cluster and replace it with synthetic genetic parts (Figure 3-1). This is referred to as "refactoring," a term borrowed from software development where code is rewritten to achieve some goal (e.g., stability), without affecting the function of the program [102]. Our process involves the removal of all native regulation, including that which is unknown or uncharacterized. It begins by removing all non-coding DNA, regulatory proteins, and non-essential genes. The codons of essential genes are changed to create a DNA sequence to be as different as possible from the wild-type gene. The recoded genes are computationally scanned to eliminate internal regulation. They are organized into operons and placed under the control of synthetic parts (promoters, ribosome binding sites, and terminators) that are functionally separated by insulator parts. Finally, a controller consisting of genetic sensors and circuits regulates the conditions and dynamics of gene expression. As a demonstration, we have applied this process to refactor the gene cluster encoding nitrogen fixation in *Klebsiella oxytoca* strain M5al (also historically known as *Klebsiella pneumoniae* strain M5al).

Nitrogen fixation is the conversion of atmospheric N_2 to ammonia (NH_3), such that it can enter metabolism [103]. Industrial nitrogen fixation through the Haber-Bosch process is used to produce fertilizer and is a major use of energy in agriculture. Many microorganisms are able to fix nitrogen using the nitrogenase enzyme. Typically, the genes necessary for nitrogen fixation occur together in a gene cluster, including the nitrogenase subunits, the biosynthesis of metalloclusters (FeMo-co, P-cluster [8Fe-7S] and [4Fe-4S]), e- transport, and regulator proteins (Figure 3-2A) [104,105]. FeMo-co is one of the most complex metalloclusters known in biology [106-108]. The gene cluster from *K. oxytoca* has been a model system for studying nitrogen fixation and consists of 20 genes encoded in 23.5kb of DNA (Figure 3-1, top) [109]. The complete cluster has been transferred to *E. coli*, thus demonstrating that it has all of the genes necessary for nitrogen fixation [110]. The encoding of this function is complex, many of the genes overlap, and operons are oriented in opposite directions.

2 Results

2.1 Characterization of the Native Gene Cluster

Prior to refactoring a pathway, it is useful to perform a robustness analysis to determine the tolerances of a gene or set of genes to changes in expression level (Figure 3-2B). This informs the grouping of genes into operons, the identification of non-essential genes to be excluded from the final cluster, and the selection of synthetic parts to obtain desired expression levels. The robustness analysis is performed by knocking out individual genes or sets of genes and complementing them on a plasmid under inducible control. The tolerance is obtained by measuring nitrogenase activity as a function of the activity of the inducible promoter.

It is noteworthy that this pathway is generally sensitive to small changes in expression and has a clear optimum (Figure 3-2B). The genes that form nitrogenase

(*nifHDK*) are sufficient to recover activity in the absence of *nifY* [111] or *nifT* [112]. The chaperone NifY is required to achieve full activity and broadens the tolerance to changes in expression level. NifT did not have an effect on activity, as observed previously [112]. The genes controlling electron transport (*nifJ* and *nifF*) need to be expressed at low levels, and activity falls rapidly as expression increases. The optima for genes participating in the metal cluster biosynthetic pathways vary. The *nifUSVWZM* operon, which encodes proteins for early Fe-S cluster formation (NifUS) and proteins for component maturation (NifVWZ for Component I and NifM for Component II), needs to be expressed at low levels, whereas *nifBQ*, encoding proteins for FeMo-co core synthesis (NifB) and molybdenum integration (NifQ), need to be expressed 10-fold higher. NifEN is tolerant to varied expression levels. However, activity is quickly lost with the inclusion of *nifX* [113]. The native cluster also includes the regulatory proteins NifL and NifA, which integrate environmental signals and ensure that nitrogen fixation only occurs at low oxygen and in the absence of fixed nitrogen [114]. The genes *nifT*, *nifX,* and *nifLA* are not included in the refactored cluster.

2.2 Design of Synthetic Genes

The DNA sequence of each gene to be included in the refactored cluster is recoded to remove any internal regulation, including that which may be undiscovered. Gene sequences were chosen by selecting codons that satisfy constraints for expression in *Klebsiella* as well as produce a DNA sequence that is as divergent as possible from the wild-type gene. Computational methods were used to scan gene sequences for RBSs, σ54 and σ70 promoters, and terminators [115,116]. If identified, different codons were chosen in these regions. The outcomes of this process are genes that share only an average of 48% codon identity with their wild-type counterparts. This introduces point mutations throughout the gene, which reduces the likelihood that internal regulation is preserved. The objective of this process is to generate genes that are inert aside from encoding the wild-type amino acid sequence.

2.3 Design of Synthetic Regulation

The expression level of each recoded gene is controlled by a synthetic RBS, synthetic promoter, and its order within an operon. In the wild-type cluster, genes often overlap, RBSs appear within the coding region of the upstream gene, and operons appear in different orientations (Figure 3-1). In the refactored cluster, each genetic function is encoded within a distinct "part," insulators are included to increase the separation between these functions. Synthetic RBSs were chosen to match the strength of the wild-type sequence. To characterize wild-type RBS strength, the region surrounding the start codon of each gene (-60 bp to +90 bp) was transcriptionally fused to red fluorescent protein (mRFP), and fluorescence was measured from a plasmid (low-copy pSC101* origin, Figure 3-2C). Synthetic RBSs of sufficient length to capture the full ribosome footprint (~35bp) were generated with the RBS Calculator [116], and the strength of each was experimentally measured using the same method. If the synthetic and wild-type RBSs differed by more than 3-fold in expression, new RBS sequences were generated

and screened. Insulator parts consisting of ~50bp of random DNA precede each synthetic RBS. One benefit of this organization is that it facilitates physical DNA manipulation and assembly, particularly when changing gene order or swapping gene or part variants.

2.4 Construction and Debugging of Synthetic Operons

The recoded genes and their synthetic RBSs are assembled into refactored operons under control of synthetic promoters and terminators (Figure 3-3A). The operons were designed *in silico* and computationally scanned to remove unintended regulatory sequences that might be created by part junctions. The DNA encoding each operon was chemically synthesized, and nitrogenase activity was measured by complementing the corresponding knockout strains from a plasmid (low copy pSC101* origin). Some of the initial designs for refactored operons showed little or no activity. In these situations, it is challenging to identify the problem because so many changes have been made simultaneously. To rapidly identify the problem, a debugging method was developed that can be generalized when refactoring different functions (Figure 3-3A). Chimeric operons were created by replacing a synthetic region of DNA (*e.g.*, a gene) with its wild-type counterpart. The function of each chimera in this library was assessed to identify which region of synthetic DNA caused a loss of activity. New chimeras would then be constructed with increasingly fine-resolution changes between synthetic and wild-type DNA. This approach "zooms in" on the problematic region of DNA, which can then be fixed. The most common problem is due to errors in the reference DNA sequence (Genbank, X13303.1) [109]. Refactored genes were designed using only the amino acid sequence information from the database; thus, they were sensitive to sequencing errors leading to missense mutations that reduced or eliminated activity. Indeed, 18 such mutations were identified and confirmed by carefully resequencing the wild-type cluster (Appendix B, Table B1). Fifteen of the 18 mutations occurred in refactored operons that required debugging and were corrected (Figure 3-3B). This demonstrates the challenge of reconstituting biological functions using only database information and DNA synthesis [117].

Modifying synthetic RBS strength was also important to debugging. The function of the synthetic *nifUSVWZM* operon was significantly improved by changing RBSs to match a 1:1 ratio of NifU:NifS . The initial selection of RBSs led to an observed 10:1 ratio in their respective RBS strengths. After debugging, *nifU* and *nifS* RBS strength was better balanced (1.25:1), and this improved activity. For one RBS, the measurement method proved to be inaccurate. We found the measured strength of the wild-type *nifQ* RBS was extremely low (Figure 3-2C), and the synthetic *nifBQ* operon showed low activity when the synthetic *nifQ* RBS was matched to the measured strength. In contrast, the robustness analysis showed a requirement for high expression level of the *nifBQ* operon (Figure 3-2B). Thus, a strong synthetic RBS near the strength of the *nifB* RBS was used and significantly improved *nifBQ* operon activity. In one case, our initial recoded *nifH* gene did not express well using either wild-type or synthetic regulation (Appendix B, Figure B3). We designed a new synthetic gene, requiring that it diverge from both the native and first synthetic DNA sequences and found that the new synthetic gene both expressed well and recovered activity.

2.5 Assembling and Controlling the Synthetic Gene Cluster

As with the robustness analysis, the nitrogenase activities of the refactored operons were measured as a function of the IPTG-inducible P_{tac} promoter (Figure 3-3C). Each operon has a different optimum. To combine the operons, the P_{tac} promoters were replaced with T7 promoters that have a strength close to the measured optimum (Figure 3-3D and Appendix A). The nitrogenase genes (*nifHDK*) are highly expressed in *Klebsiella* under fixing conditions (up to 10% of cell protein) [118], so the strongest promoter was used to control this operon (121 SPU). A long operon was built to include the *nifEN* and *nifJ* genes, where the lower expression required for *nifJ* was achieved through transcriptional attenuation. The *nifF* gene was encoded separately under the control of a medium strength promoter (17 SPU). Finally, the *nifUSVWZM* and *nifBQ* operons were controlled by weak promoters (7 SPU). Each of the individual refactored operons under the control of a T7 promoter was able to recover the activity observed from the P_{tac} promoter and corresponding optimal IPTG concentration (Figure 3-3E).

Transitioning the control to T7 promoters and moving the inducible expression of an engineered T7 RNA polymerase (T7*, Appendix A) to a separate "controller" plasmid facilitates the assembly of the complete cluster from refactored operons. We first assembled half-clusters using Gibson Assembly [119] and verified their function in strains with the corresponding genes deleted. The first half-cluster consisted of the *nifHDKYENJ* operon. The second half-cluster was assembled from the *nifBQ*, *nifF*, and *nifUSVWZM* operons. The half-clusters were able to recover 18% ± 0.7% and 26% ± 8.4% of wild-type activity, respectively. The full synthetic cluster was assembled from both half-clusters (Figure 3-4), and its activity measured in a strain where the full cluster is deleted. The synthetic gene cluster recovers nitrogenase activity at 7.4% ± 2.4% of the wild-type (Figure 3-5A). Strains carrying the synthetic gene cluster grew by nitrogen fixation, albeit 3.5-fold slower than the wild-type strain (Appendix B, Figure B2), and incorporated ^{15}N-labelled nitrogen into 24% ± 1.4% of their cellular nitrogen content, as measured by isotope ratio mass spectronomy (IRMS) (Figure 3-5B). The complete refactored cluster consists of 94 genetic parts, including a controller (Figure 3-4), and the function of each part is defined and characterized. Therefore, the genetics of the refactored system are complete. While it retains significant function, the process of simplification reduces activity as expected [102].

2.6 Changing Controllers to Execute Novel Regulatory Programs

The modular organization of the controller and refactored cluster simplifies changing the regulation of the system. This can be achieved by simply transforming a different controller; so long as the dynamic range of the T7* RNAP expression is the same, the activity of the system will be preserved. To demonstrate this, we constructed two additional controllers (Figure 3-5A). Controller #2 changes the chemical that induces the system by placing the expression of T7* RNAP under the control of the aTc-inducible P_{tet} promoter. When induced, Controller #2 produces nitrogenase activity identical to Controller #1 (7.2% ± 1.7%). The controller can also serve as a platform to encode genetic circuits to control regulatory dynamics or to integrate multiple sensors. To this end, Controller #3 contains two inducible systems (IPTG and aTc) and an inverter (cI)

[120], the output of which is integrated by a transcriptional NOR gate [2]. Similar logic gates have appeared in numerous synthetic genetic programs, including those involved in pattern formation [121,122]. By introducing operator sites OR1 and OR2 into the P_{tac} promoter, we created a chimeric promoter that could be repressed by both cI and LacI; thus producing the ANDN function. In the presence of IPTG and the absence of aTc, nitrogen fixation is 6.6% ± 1.7% of wild-type activity. These controllers represent the simplicity by which the regulation of the refactored cluster can be changed.

In addition to making it possible to add new regulation, the process of refactoring eliminates the native regulation of the cluster. This is demonstrated through the decoupling of nitrogenase activity from the environmental signals that normally regulate its activity. For example, ammonia is a negative regulator that limits overproduction of fixed nitrogen [114]. In the presence of 17.5 mM ammonia, no nitrogenase activity is observed for the wild-type cluster (Figure 3-5C). In contrast, the refactored gene cluster maintains activity in the presence of ammonia (1.1% ± 0.5%). Interestingly, this 7-fold reduction of activity is not due to residual regulation present in the system. Rather, it occurs because the addition of ammonia to the media reduces the output of the controller by 4.5-fold (Figure 3-5C). In theory, this could be fixed by increasing the expression level of T7* RNAP, but it speaks to the need to create genetic circuits that are robust to environmental context (one goal the recently established CAGEN competition aims to achieve [123]).

3. Discussion

In this study, we developed a methodology for recapitulating biological behaviors using only well-characterized, synthetic parts. Our approach is predicated on characterizing parts in isolation and using them to construct genetic circuits with predictable performance. Indeed, we observed consistent and reliable function by most parts, particularly when we utilized them in conjunction with insulator parts. The unexpected behavior of a few parts is most likely due to contextual anomalies or measurement error. Moreover, part composability requires that parts be orthogonal and their combination not produce off-target effects. We sought to maximize composability through our part design and selection: transcriptional regulation using phage components, computationally designed RBS sequences, and codon randomization for gene sequences. Future efforts may expand on our use of orthogonal transcription (T7 RNAP) to also utilize orthogonal translation.

When intended performance of an engineered system is not observed, identification and correction of faulty components is critical. Indeed, debugging strategies are integrated into the development process for software and engineered systems. Here, we have employed a methodology for dissecting a poorly performing genetic network in order to locate the causative part(s). We did not attempt to alter the behavior of the part. Instead, we simply replaced it with an alternative component. Similar strategies have been successfully adopted in the engineering and maintenance of electrical circuits, computer networks, and distribution systems. In exceedingly large networks, principal component analysis has proven valuable in debugging. Genetic debugging may require comparable strategies as designed genetic programs grow in scale.

A hallmark of engineered systems is the presence of modularity between components. In designing modularity into the refactored gene cluster, we considered future genetic architectures as well as applications of the host. We sought to construct a self-contained genetic program that could readily be integrated into a larger genetic program or newly engineered host. Additionally, we envisioned the need to change the broader regulatory context under which nitrogen fixation operates. The adoption of a controller uniquely afforded us the ability to satisfy both these goals. Focusing on the controller as a design element provides a common point for turning on refactored gene clusters and enables the decoupled development of regulatory programs. As such, the controller may serve as a unifying element in the future assembly of complex genetic circuits.

Genetic systems that emerge from evolutionary processes exhibit high redundancy, efficiency of information coding, and layers of regulation that rely on different biochemical mechanisms [34,124,125]. This impedes the study and harnessing of biological functions by tweaking their natural genetics. A research area of synthetic biology has been to build synthetic gene circuits to study regulatory dynamics that are ideally in isolation of host regulation [4,126]. In essence, we have applied this philosophy to rebuild a complex biosynthetic pathway. As with many synthetic circuits, the first design has reduced activity and is sensitive to changes in environmental conditions. However, it represents a platform to cleanly study those regulatory interactions that are required to recover these attributes. Further, from an engineering perspective, simplification enables model-guided manipulation of the pathway, and decoupling from natural regulation may ease combining biological functions as part of genome design.

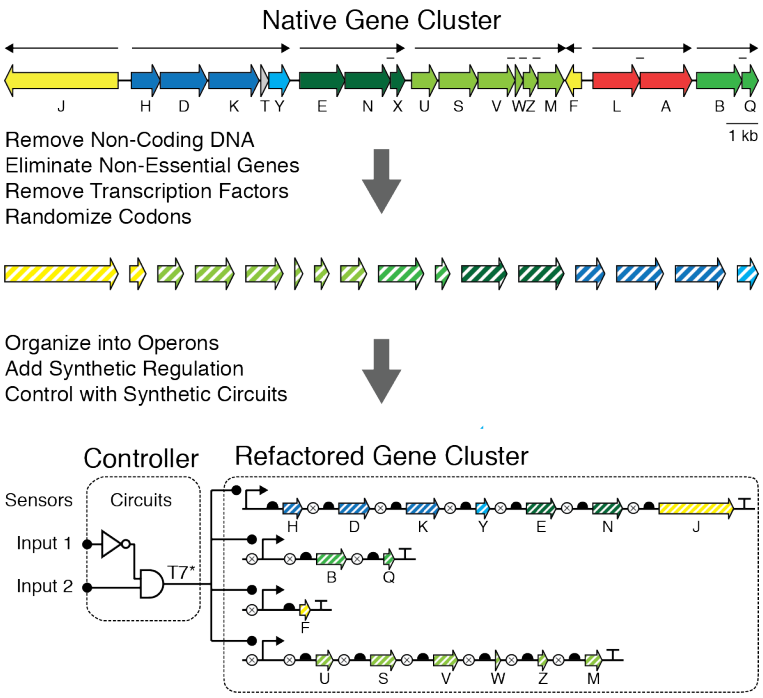

Figure 3-1 The process of refactoring a gene cluster. The wild-type *K. oxytoca* nitrogen fixation gene cluster is shown at top. The genes are coloured by function: blue (nitrogenase), green (co-factor biosynthesis, shading corresponds to operons), yellow (e- transport), and grey (unknown). The thin arrows show the length and orientation of the seven operons and a horizontal bar indicates overlapping genes. The recoded genes are shown as dashed lines. The symbols used to define the refactored cluster and controller are defined in Figs. 3-4 and 3-5, respectively.

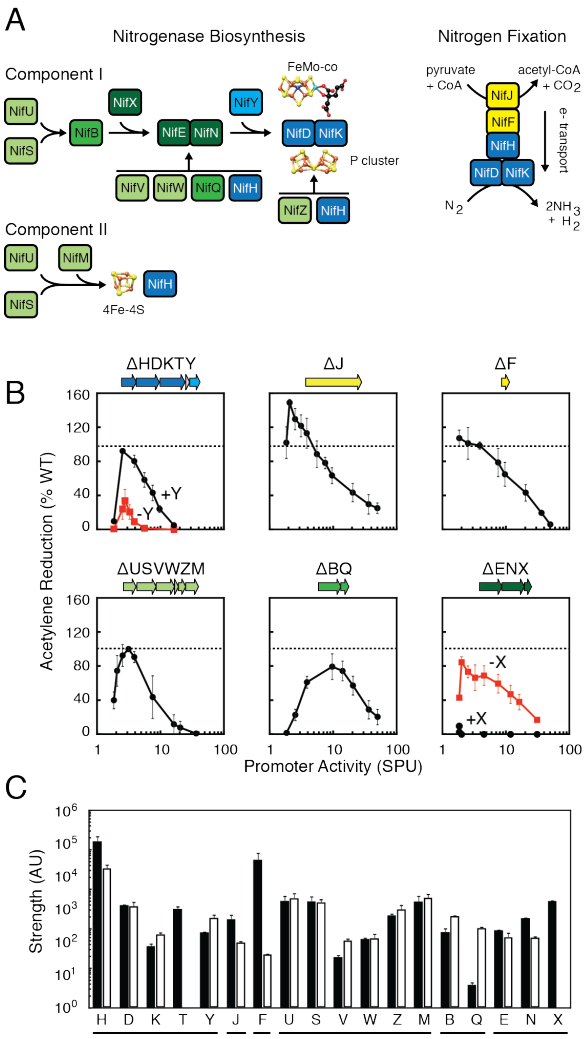

Figure 3-2 The robustness of the nitrogen fixation pathway to changes in the expression of component proteins. (A) The pathway for nitrogenase maturation is shown and proteins are coloured by function (Figure 2-1). The metal clusters are synthesized by the biosynthetic pathway: the P-cluster and FeMo-co in component I (NifDK, tetramer) and the 4Fe-4S cluster in component II (NifH, dimer) [104,105]. Nitrogen fixation catalyzed by the matured nitrogenase is shown with its *in vivo* electron transport chain. (B) The tolerance of nitrogenase activity to changes in the expression of component proteins are shown. Activity is measured via an acetylene reduction assay and the % compared to wild-type *K. oxytoca* is presented. Wild-type operons are expressed from a P_{tac} promoter on a low copy plasmid. The promoter activity is calculated as the output of the P_{tac} promoter at a given concentration of IPTG and compared to a constitutive promoter. The effect of not including NifY (-Y) and NifX (-X) are shown in red. (C) The comparison of the strength of wild-type (black) and synthetic (white) ribosome binding sites (RBSs) is shown. The RBSs were measured through an in-frame transcriptional fusion (-60 to +90) with mRFP. The strength is measured as the geometric average from a distribution of cells measured by flow cytometry. The synthetic RBSs of *nifF* and *nifQ* are not intended to match the wild-type measurement. No synthetic RBSs were generated for *nifT* and *nifX*. Error bars represent the standard deviation of at least three experiments performed on different days.

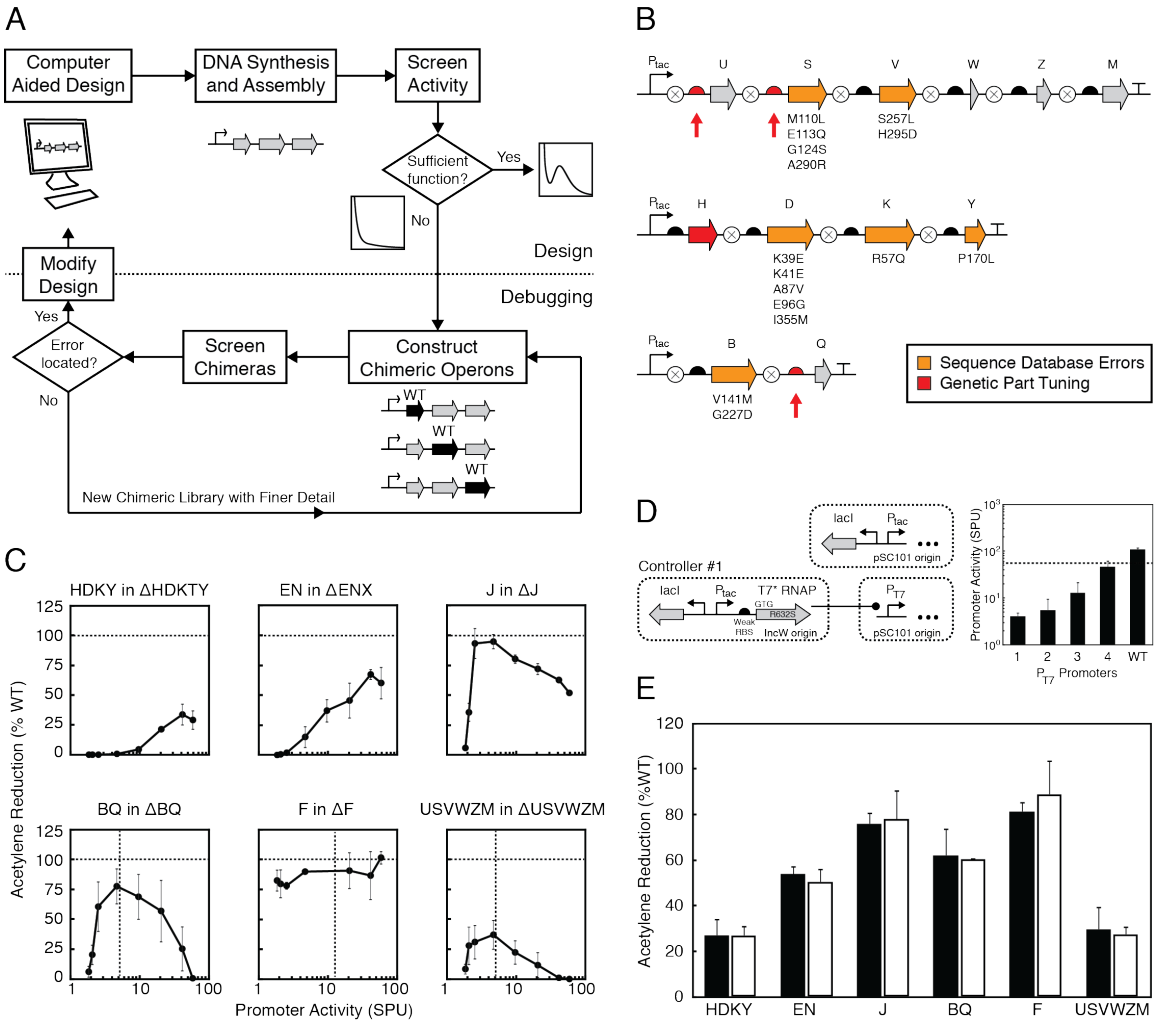

Figure 3-3 Debugging the refactored operons. (A) The process is shown for the identification of problem sequences within a refactored operon. After design and synthesis, the problematic DNA is crossed with wild-type to create a chimeric library, which is screened. This is done iteratively to reduce the size of the problematic region until the specific errors are identified. (B) The debugging process led to the correction of RBS strengths (red arrows), the recoded sequence of *nifH*, and numerous nucleotide errors found in the sequenced cluster in the database (Appendix B, Table B1). (C) Nitrogenase activity is shown as a function of promoter strength for each refactored operon in respective *K. oxytoca* knockout strains (Δ*nifHDKTY*, Δ*nifENX*, Δ*nifJ*, Δ*nifBQ*, Δ*nifF*, and Δ*nifUSVWZM*). Vertical dashed lines indicate strength of the mutant T7 promoter that controls each operon in the complete refactored gene cluster (for HDKY, EN, and J, the promoter is equivalent to 121 SPU). (D) A controller plasmid decouples operon expression from the inducible promoter. A T7 RNAP variant (T7*) was designed to reduce toxicity. A set of 4 mutated T7 promoters were used to control the expression of each operon. P_{tac} activity under 1mM IPTG induction is indicated by a dashed horizontal line. (E) Nitrogenase activity is compared for each refactored operon under the control of the P_{tac} promoter at the optimal IPTG concentration (black) and the controller (part D) with 1mM IPTG and expression controlled by different T7 promoters (white). The T7 promoters used are WT for operons HDKY, EN and J; promoter 2 for operons BQ and USVWZM; and promoter 3 for F. Error bars represent the standard deviation of at least three experiments performed on different days.

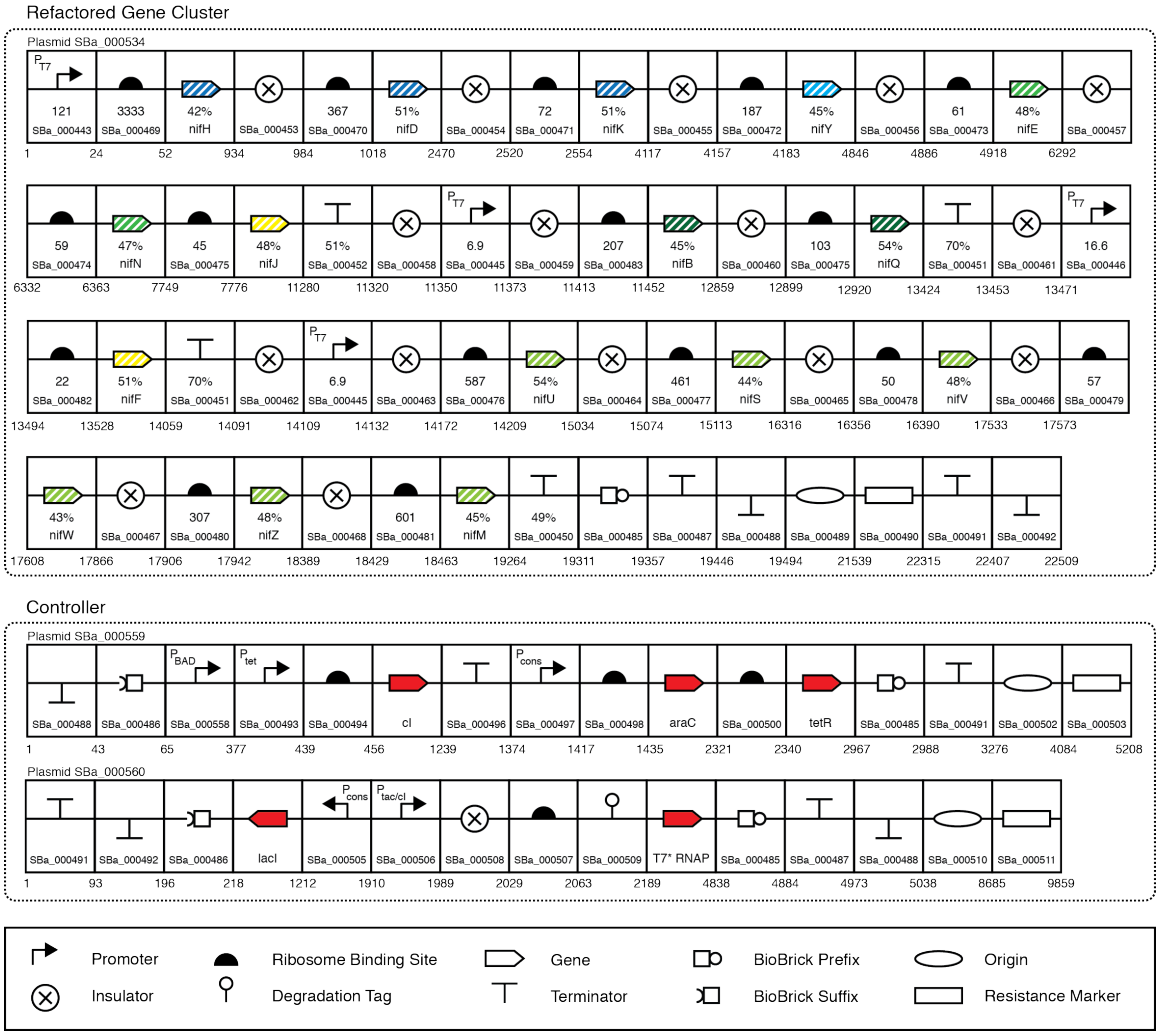

Figure 3-4 Comprehensive schematic of the refactored gene cluster and controller. Each of the 94 parts is represented according to the SBOL visual standard (www.sbolstandard.org), and the SynBERC Registry part number (registry.synberc.org) and part activity are shown. T7 promoter strength is reported as the fluorescence of a RFP reporter and presented as SPU (Appendix A). Terminator strengths are measured in a reporter plasmid and reported as the percentage reduction in RFP expression when compared to a reporter without a terminator. The RBS strength is reported in as arbitrary units of expression from the induced P_{tac} promoter (1mM IPTG) and a fusion gene between the first 90 nucleotides of the gene and red fluorescent protein. The nucleotide numbers for the plasmids containing the refactored cluster and controller are shown. The codon identity of each recoded gene as compared to wild-type is shown as a percent.

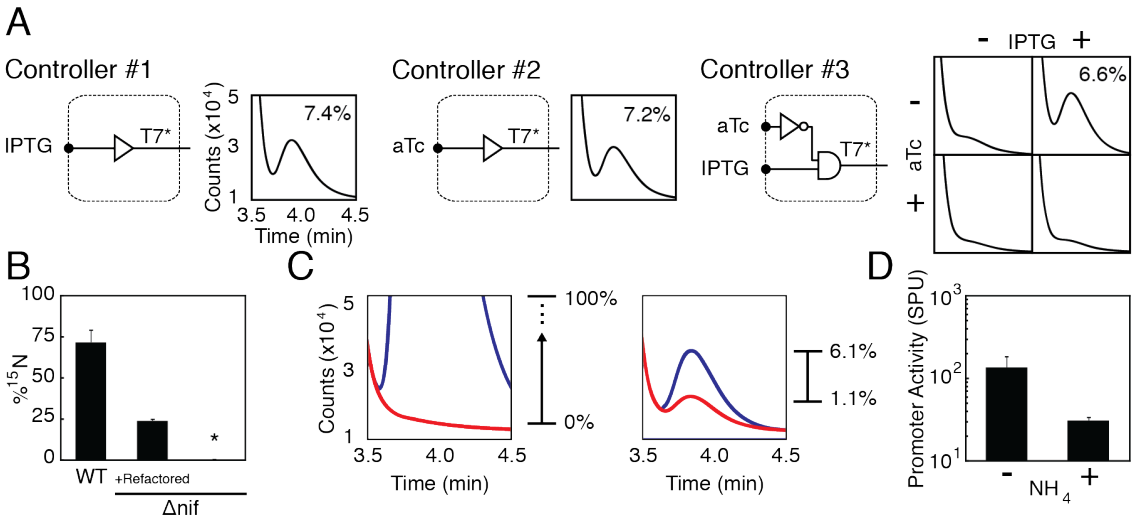

Figure 3-5 Regulation of the complete refactored gene cluster. (A) Nitrogenase activity for the three controllers are shown: IPTG-inducible, aTc-inducible, and IPTG ANDN aTc logic. The gas chromatography trace is shown for each as well as the calculated percent of wild-type activity, (7.4%±2.4%, 7.2%±1.7% and 6.6%±1.7% respectively). Standard deviation is calculated using data from at least two experiments performed on different days. (B) ^{15}N incorporation into cell biomass is shown. Nitrogen fixation from N_2 gas by the refactored gene cluster was traced using $^{15}N_2$ and measured using isotope ratio mass spectronomy (IRMS). Data are represented as the fraction of cellular nitrogen that is ^{15}N. The standard deviation represents two experiments performed on different days. (C) The effect of ammonia on regulation of nitrogenase expression is shown. Acetylene reduction traces shown with (red) and without (blue) addition of 17.5 mM ammonium acetate for wild type cells (left) and cells bearing synthetic *nif* system (right). The synthetic system was induced by Controller #1 using 1mM IPTG and exhibited nitrogenase activity of 1.1%±0.5% and 6.1%±0.4% with and without ammonium acetate respectively. (D) T7* RNAP expression of Controller #1 corresponding to Part C is shown. Strains carrying Controller #1 and a RFP reporter plasmid were characterized under 1mM IPTG induction with or without addition of ammonium acetate.

Chapter 4 - Optimizing the Performance of Biological Behaviors

1 Introduction

Refactoring complex gene clusters and engineering metabolic pathways requires numerous iterations between design, construction and evaluation in order to improve a desired system property, *e.g.* higher product titers, lower toxicity, or improved nitrogen fixation. One common way to affect these properties is to modify gene expression levels within the system, even if the direct relationship between gene expression and the system property is unknown. Making quantitative changes to gene expression can be achieved through the use of regulatory elements, *e.g.* promoters or ribosome binding sites, that exhibit rationally predictable behavior. Further, recent technological developments surrounding DNA modification provide a platform for automating the assembly of many genetic parts. Here, our efforts to develop techniques for rapidly editing and optimizing genetic programs are described.

2 Automating Edits of DNA Programs

Synthetic biologists often use DNA as a medium for programming new functionality into biological systems. From the earliest use of restriction enzymes for genetic engineering, to the modern use of DNA synthesis to produce genes from digital information, the ability to manipulate DNA sequence has influenced the rate of industrial and scientific innovation [11]. Currently, a major bottleneck in synthetic biology is the physical revision or significant alteration of DNA sequences in large, complex genetic programs. It is necessary to develop an automated pipeline to enable arbitrary modifications to large, complex genetic programs in a high-throughput and reliable manner. A number of state-of-the-art DNA manipulation methodologies exist, and we evaluate the suitability of these for automation. Further, a technology platform capable of performing multiple, simultaneous DNA sequence modifications is demonstrated. Future automation of this platform will require improvements in assembly efficiency.

3 Existing DNA Assembly Technologies

Several DNA manipulation methodologies have recently been described (see associated Table 4-1):

1. Biobrick: In this form of restriction enzyme-based cloning, each DNA part is flanked by a standard set of restriction sites. When parts are ligated together, the standard restriction sites are erased in the junction between parts and regenerated outside the parts. The advantage of this methodology is that new parts can readily be added to the assembly as long as each part contains the standard restriction sites [127]. There are two major drawbacks of this method. First, only two DNA parts can be combined in each step. Successive steps of digestion and ligation have to be performed to introduce additional parts. Second, the method places constraints on the physical location for part addition. If the order of parts needs to be changed, an entire reconstruction of the plasmid may be required.

2. Type IIs restriction enzymes: Type IIs restriction enzymes are a unique class of restriction enzyme whereby the cut site resides outside of the recognition site of the enzyme. As a result, the overhang created from this digestion can be defined arbitrarily. By designing DNA parts with appropriate overhang sequences, multiple parts can be ligated simultaneously in a defined order. Furthermore, the single-stranded overhang sequences can be designed to be non-palindromic, thus removing unwanted by-products, *e.g.* self-ligation and self-concatemerization, typically present in traditional restriction enzyme cloning. For this reason, Type IIs restriction enzymes exhibit increased efficiency compared to traditional restriction enzyme-based cloning.
3. Homology-based assembly: In this method, multiple parts can be assembled together in a single reaction. To guide assembly, each DNA part contains 20-30 bp homology to neighboring parts. Parts are generated via PCR or restriction enzyme digest. In a subsequent enzymatic reaction, the ends of the double-stranded DNA are digested to expose single-stranded DNA containing homology to other parts. The single-stranded sequences from multiple parts can anneal to each other and then be ligated together to create a single contiguous DNA molecule. Due to the long homology region, this method is highly suitable for assembly using large DNA parts. Recent embodiments of the method include Gibson assembly [119], GeneArt Seamless Cloning, InFusion, or SLIC [128]. These mainly differ in the enzymatic reaction used to create the exposed homology regions.
4. DNA synthesis: Large DNA molecules of specified sequence can be synthesized by annealing chemically synthesized oligonucleotides. This method offers the greatest flexibility in terms of information content, as every single base pair in the DNA molecule can be arbitrarily defined. However, this method is highly costly and requires the longest turnaround time, making it unsuitable for many applications. The method is mainly appropriate in situations whereby there is no template available to use as a starting point for assembly, as in the case of codon-optimization for gene sequences [129].

4 A Cloning Method to Facilitate Automated Optimization

The construction of large genetic programs (>10kbp in size) in the Voigt lab has been facilitated primarily by homology-based assembly and Type IIs enzymes. In particular, the methods developed by Gibson *et. al.*[119] played a foundational role throughout the projects described in this dissertation. To aid the design and assembly of new DNA molecules, a web-based CAD tool enabling interactive design and an associated calculator for oligos were developed (Figure 4-1). This software has been readily shared throughout the synthetic biology community and has been used to design and construct over 1,000 DNA assemblies to date. Recently, others have developed software offering similar functionality and design capabilities [130-132]. Together, software that aids both the design and automation of synthetic biology is critical to future scaling and commercialization efforts.

To enable optimization of genetic programs at the DNA level via part selection, we extended the Gibson method to perform one-pot assembly of combinatorial libraries. This technique was utilized in Chapter 5 to generate libraries of lycopene biosynthetic

pathways. It enables dynamic part selection by combining regulatory elements encoded in oligos and genes as static double stranded DNA parts (Figure 4-2). Recently, a group from JCVI described a similar method for assembling combinatorial libraries [133].

5 The Nelder-Mead Algorithm

It is possible to utilize numerical optimization methods to guide selection of regulatory elements in order to alter gene expression and to improve desired system properties. One relevant algorithm is the Nelder-Mead method, a nonlinear optimization algorithm that minimizes an objective function in multidimensional space [134]. We use the Nelder-Mead method to optimize a system property where each dimension in algorithmic space corresponds to expression of a gene in the engineered system. Points in this space represent a particular combination of expression levels for the genes in the system. As a result, each point may be considered a uniquely engineered strain. The algorithm is used to suggest new coordinates in space that improve the system property. New strains can be engineered by modifying regulatory elements to attain the suggested levels of gene expression. After evaluating the performance of the new strains, the algorithm can be used to predict subsequent modifications. This process iterates until the system property has been improved a desired amount.

The Nelder-Mead method relies on the concept of a simplex, which is an object in N dimensional space having N+1 vertices. The objective function is evaluated at each vertex of the simplex, and the algorithm uses this information to propose a sequence of new coordinates for evaluation. New coordinates are proposed according to the following process:

1. Order the simplex vertices: $f(x_1) \le f(x_2) \le \ldots \le f(x_{n+1})$
2. Calculate x_0, the center of gravity of all points except x_{n+1}.
3. Calculate a Reflection coordinate: $x_r = x_o + \alpha(x_o - x_{n+1})$
4. Calculate an Expansion coordinate: $x_e = x_o + r(x_o - x_{n+1})$
5. Calculate a Contraction coordinate: $x_c = x_{n+1} + \rho(x_o - x_{n+1})$
6. Calculate Reduction coordinates: $x_i = x_1 + \sigma(x_i - x_1)$ for all $i \in \{2,\ldots,n+1\}$

The objective function is evaluated at these points and used to determine a new simplex according to the following criteria:

1. If the Reflection, Expansion or Contraction coordinates are better than the worst simplex point, x_{n+1}, define a new simplex by replacing the worst simplex point with the best of the three (Reflection, Expansion or Contraction).
2. Otherwise, define a new simplex by combining the best simplex point with the Reduction coordinates.

These steps constitute an iteration of the algorithm. The newly defined simplex becomes the seed for generating new coordinates during the next iteration of the algorithm. Iterations typically continue until one of the coordinates in the simplex crosses a desired threshold for objective function evaluation.

6 Optimization of a Nitrogen Fixation Operon

We have optimized the performance of a nitrogen fixation operon by varying the selection of promoters that control expression of individual genes. We initially refactored the *nifEN* operon so that each gene was expressed under the control of a unique T7 promoter (Figure 4-3). To assess the impact of refactoring the *nifEN* operon, we quantitatively measured the capacity of the synthetic operon to complement a *nifEN* knockout strain and recover the ability to fix nitrogen. Our refactored system showed limited ability to fix nitrogen (20% of wild-type activity).

We subsequently applied the Nelder-Mead method to optimize *nifE* and *nifN* gene expression with the goal of improving nitrogen fixation rates. Our algorithmic space consisted of two dimensions, *nifE* and *nifN* expression. Our coordinate system was scaled to the strength of the promoters controlling these genes. To enable varied levels of gene expression, we generated and characterized a library of mutant T7 promoters (Table 4-2). Our library covers three order of magnitude of gene expression We then randomly selected mutants from the library of T7 promoters to generate two additional strains with rationally altered levels of *nifE* and *nifN* expression. The strength of T7 promoters used in these three strains defined our initial simplex. We evaluated nitrogen fixation for each strain in the simplex (strain 1: 20%, strain 2: 9%, strain 3: 12%) and used the algorithm to calculate Reflection coordinates (Figure 4-4). To construct the strain that matched the Reflection coordinates, we chose promoters from our library nearest to the coordinates in strength. We evaluated nitrogen fixation in this Reflection strain and found that it significantly outperformed (52%) our initial strains (Table 4-3).

We stopped iteration of the algorithm at this point for two reasons. Firstly, our improved strain surpassed expectations and performed sufficiently for downstream applications. Secondly, constructing the next predicted strain requires promoters stronger than exist in our library. To reach these higher levels of gene expression, we could attempt to engineer stronger promoters. Alternatively, we could target complimentary changes to multiple regulatory elements, *e.g.* the promoter and ribosome binding site for a given gene, to achieve desired expression levels. This would require describing the strengths of each type of element in common units of expression.

Table 4-1 Comparison of Current DNA Modification Technologies

Method	Cost	Strengths	Weaknesses	IP Status
Biobrick	$	Easy Protocol Robust Reaction	Rigid Architecture Limited to 2 Parts	Open
Type IIs	$	Excellent Efficiency for Small Parts Arbitrary Junction Between Parts	Poor Efficiency for Large Parts Requires Design Step	Open
Homology	$$	Excellent Efficiency for Large Parts Arbitrary Junction Between Parts	Poor Efficiency for Small Parts Requires Design Step	Available as a Kit
DNA Synthesis	$$$	Any Sequence can be Synthesized	High Cost Slow Turnaround Time	Open

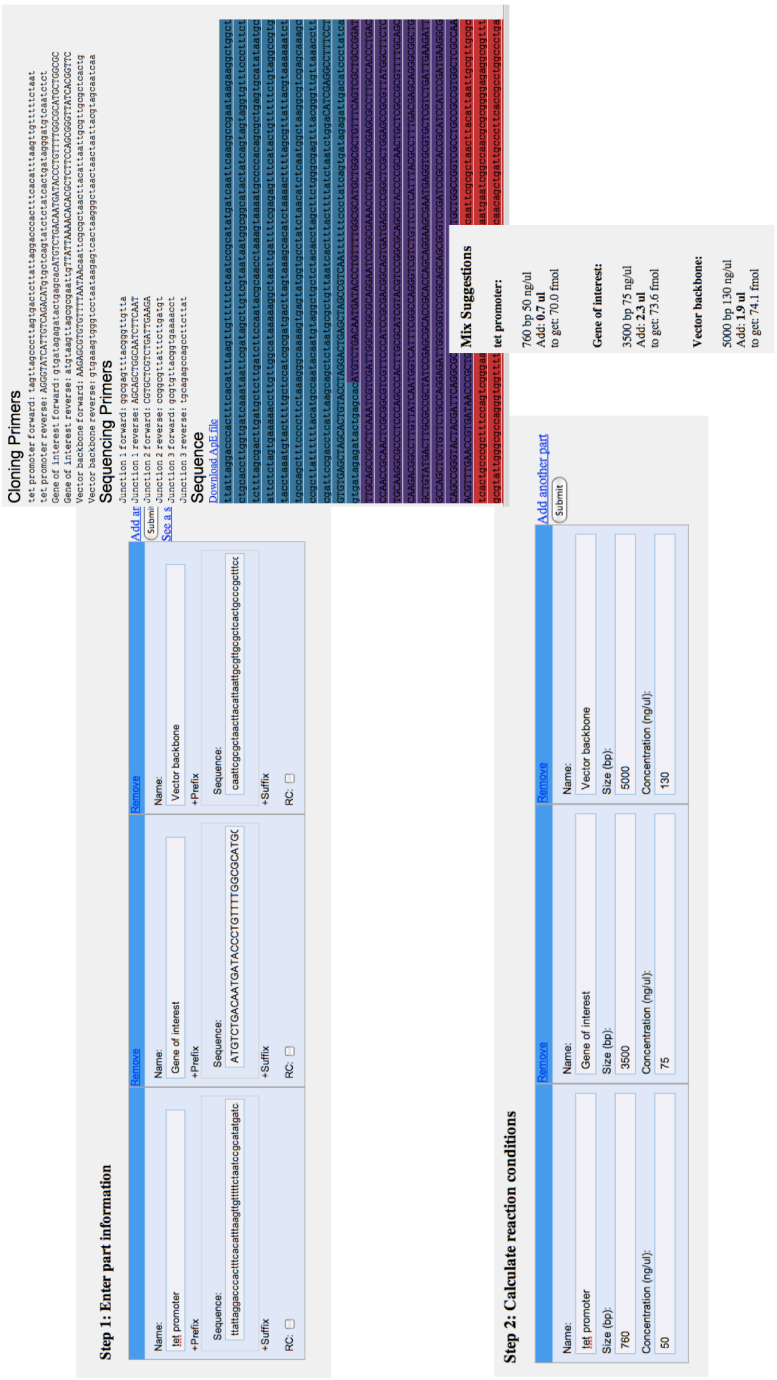

Figure 4-1 Screenshots of a CAD for Gibson Assembly. An assembly is designed by specifying the number of parts and relevant details like sequence and junction information. A drag-and-drop interface enables rearrangement of the assembly composition. The application designs cloning primers as well as sequencing primers for assembly verification. Finally, a calculator is presented that aids in the preparation of the assembly reaction.

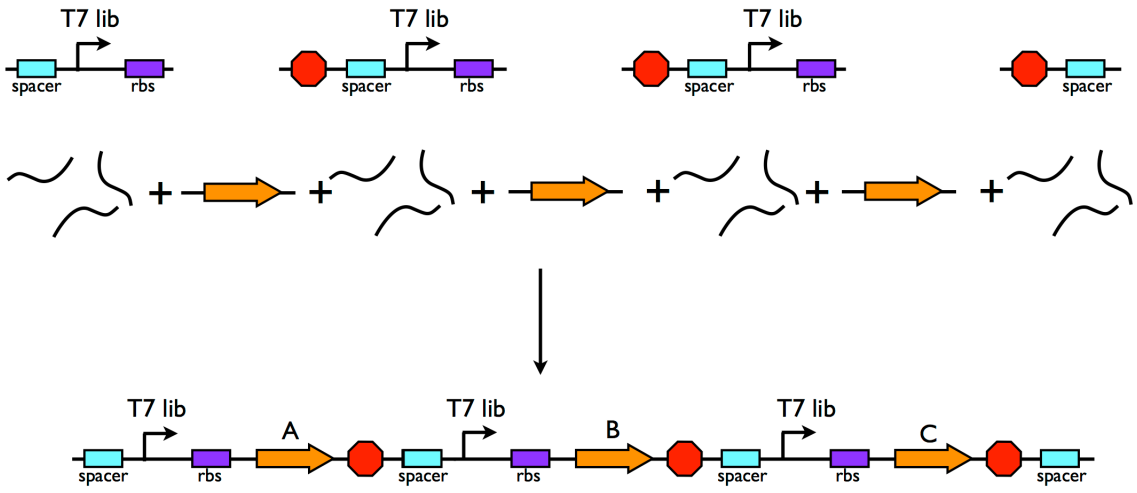

Figure 4-2 Combinatorial assembly of genetic programs. Dynamically selected parts are encoded by oligo libraries, enabling rapid modification of elements used in a given assembly. Oligos are pooled together with double stranded static elements in a design and assembled in a one-pot, modified Gibson reaction.

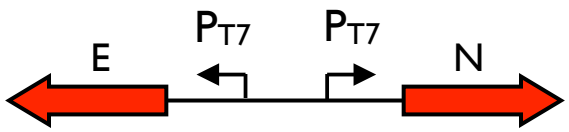

Figure 4-3 Synthetic *nifEN* operon under control of T7 promoters.

Table 4-2 T7 Promoter Library

Promoter Name	**Sequence** TAATACGACTCACTANNNNNAGA	**Strength**
WT	TAATACGACTCACTATAGGGAGA	13780
Mut1	TAATACGACTCACTACAGGCAGA	147
Mut2	TAATACGACTCACTAGAGAGAGA	752
Mut3	TAATACGACTCACTAATGGGAGA	1346
Mut4	TAATACGACTCACTATAGGTAGA	2127
Mut5	TAATACGACTCACTAAAGGGAGA	3738
Mut6	TAATACGACTCACTATTGGGAGA	6710

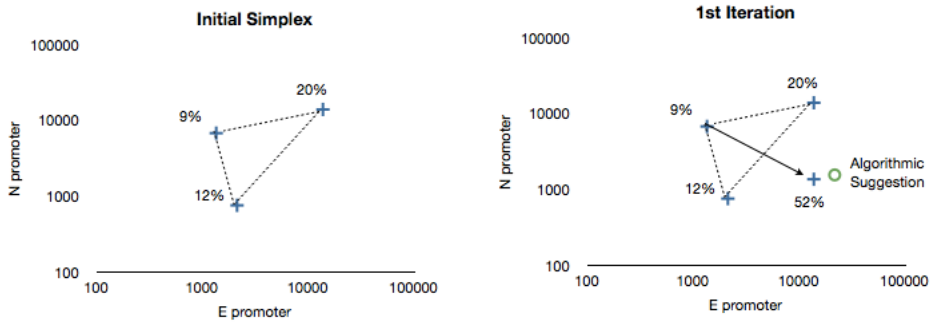

Figure 4-4 Use of the Nelder-Mead algorithm to optimize the *nifEN* operon. Three initial operons were constructed using different T7 promoters. After measuring their activity, the algorithm suggested a new construct (green circle). Promoters were selected that most closely matched the suggested point. The new strain exhibited significantly improved activity.

Table 4-3 Results of Nelder Mead Optimization

Strain Name	E Promoter	N Promoter	Fixation
#1 (N259)	WT	WT	20%
#2 (N438)	Mut 4	Mut 5	9%
#3 (N307)	Mut 4	Mut 2	12%
Reflection (N319)	WT	Mut 3	52%

Chapter 5 - Orthogonal Control of Multiple Behaviors

Abstract

T7 RNA polymerase (RNAP) has been used extensively in biotechnology for overexpressing genes. However, T7 RNAP exhibits significant host toxicity, and it is difficult to tune output expression levels. Moreover, the growing complexity of genetic programs requires increasing numbers of diverse parts. To address these drawbacks, we have engineered a toolbox of non-toxic, orthogonal RNAPs, cognate promoters and transcriptional terminators. Further, we demonstrate the utility of these tools through the inducible, tuneable, and simultaneous transcriptional control of two metabolic pathways.

1 Introduction

Rational and precise control of gene expression is critical to biotechnology and the emerging field of synthetic biology [135,136]. Towards this goal, T7 RNAP and its short, specific 23bp promoter have been valuable tools for gene expression *in vitro* and in a variety of hosts, including gram-negative and gram-positive bacteria, plant chloroplasts, and mammalian cells [137-140]. Recently, T7 RNAP has been used as a standardized genetic part in synthetic biology applications, such as genetic networks that count and orthogonal transcription-translation systems [3,141]. The wild-type T7 RNAP and consensus T7 promoter are frequently used to direct gene expression in an all-or-none manner. However, previous work has demonstrated that promoter mutations yield different strengths of expression [142,143]. Likewise, T7 RNAP can be mutated so that it recognizes novel promoters [144,145]. In fact, a single mutation (N748D) confers on T7 RNAP the ability to recognize the T3 phage promoter [146]. Finally, while T7 RNAP transcribes DNA nearly eightfold faster than native *E. coli* polymerases [147], a property that may lead to transcript instability, active site mutations can reduce its transcription elongation rate [148]. Building on these efforts, we developed a library of non-toxic RNAPs, promoters and terminators that enable orthogonal and tuneable control of gene expression.

2 Results

As a first step in engineering more useful polymerases, we controlled T7 RNAP concentration and processivity through a combination of mechanisms to reduce its toxicity (Figure 5-1A). The wild-type T7 RNAP was first placed under control of an IPTG inducible promoter in a low copy plasmid, and co-transformed with a second plasmid containing a T7 promoter and fluorescent reporter gene. Host toxicity was assessed by plating strains on inducing media and counting colonies that formed after 24 hours of growth. Colonies were smaller and fewer in number under toxic conditions, due to increased transcription by T7 RNAP and subsequent translation of the transcripts. In the initial construct, we observed significant toxicity at high levels of T7 RNAP expression, even in the absence of T7 promoters (Figure 5-1B, RNAP v1). A Lon-mediated, N-terminal degradation tag from the *umuD* gene in *E. coli* [149] was added to limit polymerase concentration, but little reduction in toxicity was observed (Figure 5-

1B, RNAP v2). Next, polymerase processivity was reduced with a novel, serendipitous active site mutation (R632S), resulting in constructs with minimal host toxicity (Figure 5-1B, RNAP v3). Finally, translation of the RNAP was controlled using a weak ribosome binding site and GTG start codon (Figure 5-1B, RNAP v4). The fully modified T7 RNAP backbone (T7* RNAP) exhibited no host toxicity under all conditions tested.

To generate orthogonal RNAP-promoter interactions, we implemented a mutation strategy inspired by the diversity of sequenced phage genomes; similar to comparative genomics approaches that have been used to generate orthogonal transcription factor-DNA interactions [150]. In the T7 RNAP, promoter binding is determined by a β-hairpin, known as the specificity loop (Figure 5-1C) [151]. The specificity loop contacts the binding domain of the promoter (-12bp and -7bp). Changes to the specificity loop sequence confer the ability to recognize different promoter sequences [144-146]. However, random point mutations can disrupt the β-hairpin structure and eliminate RNAP activity. We hypothesized that the β-hairpin might be a modular unit that could be swapped between functional RNAP variants while retaining activity. To test this, 43 T7 RNAP homologs were identified from NCBI (protein BLAST, E-value < 1^{-100}) and those viruses with fully sequenced genomes were selected for further analysis. Alignments revealed 13 RNAP subfamilies based on similarity of specificity loop sequence (Appendix C, Figure C1). We then identified promoters in each phage genome and constructed consensus promoter sequences for each phage. Consensus promoters clustered into subfamilies with perfect correlation to specificity loop subfamilies (Appendix C, Figure C1). Novel RNAPs were generated by grafting the subfamily sequences in place of the wild-type specificity loop in our non-toxic T7* RNAP. Similarly, we replaced the binding domain of the T7 promoter with subfamily consensus sequences. Promoter recognition by mutant RNAPs was assessed by co-transforming the plasmid libraries and measuring activity (Figure 5-1D). This strategy yielded a set of 4 highly orthogonal RNAP-promoter pairs, named after a representative phage in each subfamily (T7*-P_{T7}, T3*-P_{T3}, K1F*-P_{K1F}, N4*-P_{N4}). An asterisk is used to denote that each RNAP is non-toxic, although the toxicity controls vary in each plasmid (Appendix C, Figure C2). It is interesting to note that the mutations made to generate these RNAPs and promoters have not been previously described in literature. In one case, the N4* RNAP contains 10 altered residues within the specificity loop. However, we found that this RNAP needs to be expressed more highly than the other RNAPs to achieve comparable levels of gene expression.

We extended the orthogonal RNAP-promoter expression system to facilitate tuneable control of gene expression through the creation of promoter libraries of varying strength. The T7 promoter is modular, consisting of an RNAP binding domain and a strength-determining domain [151-153]. Mutations to the strength-determining domain should alter promoter activity without affecting RNAP specificity. We tested this modularity by creating a library of different strength promoters by mutating the wild-type T7 promoter from -2bp to +3bp and screening for activity, similar to recent work by Ellington and co-workers [143]. Promoters with a range of strengths were selected and sequenced to generate a diverse P_{T7} library. We mutated the strength-determining domain of P_{T3} to match the P_{T7} library. When orthogonal RNAP (T7* or T3*) are co-transformed with the promoter libraries (P_{T7} or P_{T3}), orthogonality is retained while the general trend for strength persists (Figure 5-1E). These results support the premise of two modular domains within the T7

promoter and suggest a general strategy for varying strength in any orthogonal RNAP-promoter system.

Finally, we attempted to maximize the sequence diversity of transcriptional terminators available in the orthogonal RNAP system. The reuse of repetitive DNA sequences in constructing genetic networks can be very challenging. Sequence diversity can aid physical DNA manipulation techniques while reducing the occurrence of homologous recombination. However, T7 RNAP is commonly used with a single unique terminator, leading to significant repetition when utilizing multiple polymerases and transcriptional targets. To facilitate the creation of multi-operon networks with the T7 expression system, we generated a library of transcriptional terminators based on the wild-type T7 terminator sequence (Appendix A, Figure A4). We created a degenerate terminator library to mimic the stem-loop structure present in the wild-type terminator sequence. Functional terminators were obtained by co-transforming the library with T7* RNAP and screening for a reduction in gene expression. Using this approach, we obtained 8 terminators that exhibited termination efficiencies better than or equal to the wild-type T7 counterpart.

Our RNAP toolbox enables the simultaneous control of different biological functions within a cell. To demonstrate this, we constructed a genetic program for the inducible production of two small-molecule pigments, lycopene (red) and deoxychromoviridans (green). First, we engineered a plasmid containing two RNAP (T7* and K1F*) under control of multiple inducing signals and a logic circuit (Figure 5-2A). The T7* RNAP is induced by IPTG, while the K1F* RNAP is controlled by an *AND* gate that is active only in the presence of both IPTG and aTc. We co-transformed this plasmid with P_{T7} and P_{K1F}* reporter plasmids and measured their activity to verify proper operation of the genetic circuit (Figure 5-2D). The T7* RNAP exhibits 7.8 fold induction, while the K1F* RNAP is induced 9.2 fold (Figure 5-2D, where fold induction is reported as the minimum expression in an "ON" state divided by the maximum expression in an "OFF" state). Next, we utilized our promoter and terminator libraries to optimize expression of the two biosynthetic pathways for pigment production. We sought to enable inducible expression of each pigment and to obtain strong contrast between induction states. It was necessary to improve the yield of one pigment, lycopene, while eliminating leaky expression of the other, deoxychromoviridans.

E. coli produces small amounts of lycopene through the 1-deoxy-D-xylulose-5-phosphate (DXP) pathway following the introduction of the carotenoid genes, *crtE*, *crtB*, and *crtI* (Figure 5-2B) [154]. Previous studies found lycopene production improved following overexpression of two genes, *dxs* and *idi*, which function at critical metabolism branch points [8,135]. We designed a synthetic network consisting of *dxs*, *idi*, *crtE*, *crtB*, and *crtI* genes under control of T7* RNAP. Each gene was inserted into a unique synthetic operon consisting of an insulator, the T7 promoter library, synthetic ribosome binding sites, the gene, and a T7 terminator mutant. The synthetic network was assembled combinatorially from the five synthetic operons [133], co-transformed with T7* RNAP, and plated on inducing media. Colonies varied in colour, and five of the darkest colonies were selected to confirm absence of expression without inducer. One strain exhibited excellent pigment production and contrast.

Violacein, a purple pigment, is synthesized from L-tryptophan by the genes *vioABCDE* and serves as the basis for our deoxychromoviridans engineering [155,156]. We

initially cloned *vioABCDE* into our T7* RNAP expression system and found leaky expression yielded visible pigment, even in the absence of a T7* RNAP. To reduce leaky expression, we transitioned to the K1F* RNAP, the weakest of the orthogonal library. Next, we split the genes into two operons, separating *vioA* from the remainder of the genes with an intervening terminator. Finally, elimination of *vioC* and *vioD* yields the green pigment, deoxychromoviridans (Figure 5-2B). We found that leaky expression was eliminated and observed production of dark green pigment when induced.

When the two biosynthetic pathways are co-transformed with our multiple-RNAP plasmid, we observe minimal pigment production in the absence of inducers and visibly significant accumulation when properly induced (Figure 5-2C). Lycopene is synthesized in the presence of IPTG, regardless of the presence or absence of aTc. Deoxychromoviridans, in contrast, is only synthesized when both inducing molecules are present. Further quantification was performed by extracting pigments and measuring relative absorbance under the varying inducer conditions (Figure 5-2E). Together, this system demonstrates the utility of T7-derived parts for tuning gene expression and the orthogonal control of complex pathways.

Naturally occurring genetic networks exhibit high degrees of modularity in part due to orthogonal subcomponents [157,158]. Further, the study of biological design principles can be advanced by the isolation of cellular processes [159,160]. T7 RNAP provides an excellent platform for these efforts because it is orthogonal to host transcription machinery. We have expanded the tools for engineering orthogonal genetic systems by generating a toolbox of non-toxic, orthogonal RNAPs as well as promoters and terminators for tuneable gene expression. Combined with synthetic metagenomics approaches, our strategy may be useful in generating large libraries of orthogonal parts [117]. Finally, our T7 toolbox may prove valuable in efforts to refactor biological functions decoupled from host regulation or in bottom-up, modular genome design [161].

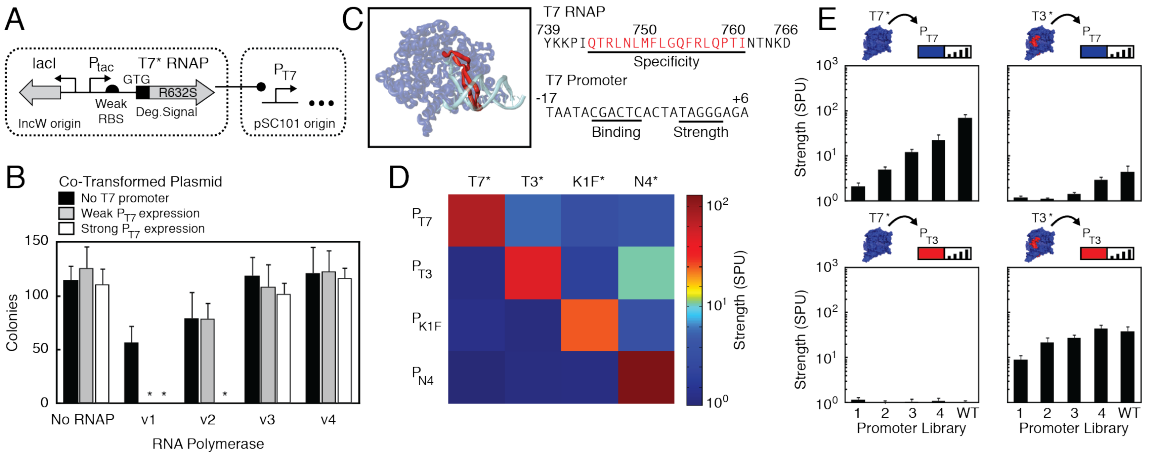

Figure 5-1 Generation of orthogonal T7 RNA polymerases and promoters. (A) The non-toxic T7 expression system is shown under control of IPTG inducible regulation (P_{tac} and *lacI*). T7 RNAP and T7 promoters reside on low copy plasmids (IncW and pSC101 origins respectively). Expression of T7 RNAP is minimized through use of a weak RBS, N-terminal degradation tag, and GTG start codon. Polymerase processivity is reduced by an active site mutation (R632S). (B) Toxicity of T7 RNAP variants is shown. Each variant was co-transformed with plasmids bearing T7 promoters, and cells were grown on plates containing inducer to assess toxicity. RNAP variants contain different cumulative modifications (v1 = WT T7 RNAP in low-copy plasmid, v2 = degradation tag, v3 = active site mutation, v4 = weak RBS and GTG start codon). The three promoter plasmids produce different levels of T7-mediated expression (black = no T7 promoter, gray = weak P_{T7} expression, white = strong P_{T7} expression). Following, T7* is used to denote the non-toxic RNAP (v4). (C) Interaction of the T7 RNAP and T7 promoter is shown. A β-hairpin in T7 RNAP, known as the "specificity" loop, is coloured red. The specificity loop interacts with the 5' end of the promoter that is highlighted pink. The 3' region in the promoter determines transcription initiation rate. Wild-type T7 sequences for the specificity loop and promoter are shown. (D) Activity of orthogonal RNAPs and promoters is shown. Orthogonal RNAPs were generated by grafting the specificity loops from different phage subfamilies into the T7* RNAP. These RNAPs are named after a representative phage in the subfamily and are denoted with an asterisk. Promoters were similarly constructed by mutating the binding region in the wild-type T7 promoter to match the consensus promoter sequence of each phage subfamily. RNAPs were co-transformed with promoter mutants controlling a fluorescent reporter protein, and strength of expression under 1mM IPTG induction is shown as the geometric mean of a distribution of cells measured by flow cytometry. (E) Modularity of the specificity and strength regions within the T7 promoter is demonstrated. A promoter library was generated by mutating the strength-determining region of the wild-type promoter (top left). Mutants are numbered 1 through 4. RNAP-promoter specificity was altered by changing the binding region in each member of the library (bottom right). When the library was co-transformed with non-cognate RNAP, minimal activity was observed (top right and bottom left). Data are reported in standard promoter units (SPU, Appendix A). Error bars represent the standard deviation of three experiments from different days.

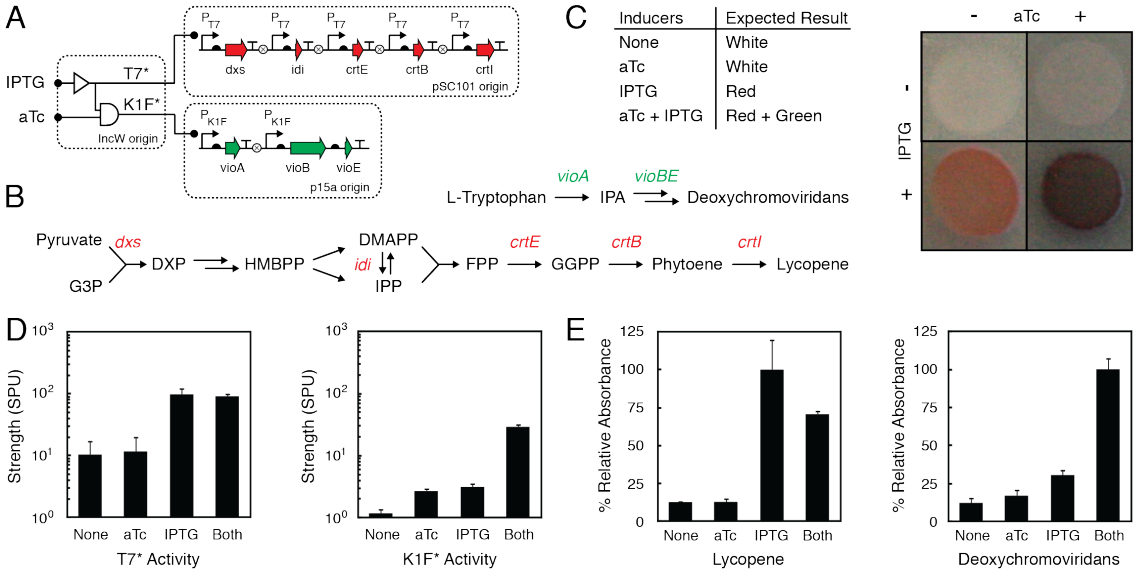

Figure 5-2 Integration of orthogonal RNAPs for the control of multiple biosynthetic pathways. (A) A genetic program relying on two orthogonal RNAPs to coordinate inducible production of green and red pigments is shown. The lycopene biosynthesis pathway is controlled by T7* RNAP in response to induction by IPTG. K1F* RNAP controls expression of the green pigment, deoxychromoviridans, and is expressed only in the presence of both IPTG and aTc. Operons are drawn using SBOL symbols: RBSs are solid half-circles, insulators are open circles marked with an *X*, and terminators are in the shape of a *T*. (B) Biosynthetic pathways for lycopene and deoxychromoviridans are shown. Intermediate products are shown in black, and genes expressed by the engineered system are shown in italics. (C) Expression of pigments under different inducing conditions is shown. *E. coli* strain MG1655 was co-transformed with plasmids encoding the genetic program described in part A. Cells were grown overnight in non-inducing conditions, diluted to an OD_{600} of 1.0, and spotted onto plates containing inducers. In the presence of IPTG and aTc, both pigments are synthesized. The combination of green and red pigments appears brownish-purple. Photos of representative colonies are shown. (D) Characterization of the dual RNAP system. T7* and K1F* activity in the integrated circuit was measured using P_{T7} and P_{K1F} fluorescent protein reporters. Data are reported in standard promoter units (SPU, Appendix A). (E) Quantitative assessment of pigment production. Cells were grown for 24 hours in LB media containing the indicated inducers. Lycopene and deoxychromoviridans were extracted, and absorbance measured (lycopene, 580nm, deoxychromoviridans, 650nm). Data are reported as the fraction of maximum absorbance observed. Error bars represent standard deviation of at three experiments from different days.

References

1 Stricker, J. *et al.* A fast, robust and tunable synthetic gene oscillator. *Nature* **456**, 516-519, (2008).

2 Tamsir, A., Tabor, J. J. & Voigt, C. A. Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'. *Nature* **469**, 212-215, (2011).

3 Friedland, A. E. *et al.* Synthetic gene networks that count. *Science* **324**, 1199-1202, (2009).

4 Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339-342, (2000).

5 Levskaya, A. *et al.* Synthetic biology: engineering Escherichia coli to see light. *Nature* **438**, 441-442, (2005).

6 Ro, D. K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940-943, (2006).

7 Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463**, 559-562, (2010).

8 Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898, (2009).

9 Anderson, J. C., Clarke, E. J., Arkin, A. P. & Voigt, C. A. Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol* **355**, 619-627, (2006).

10 Lu, T. K. & Collins, J. J. Dispersing biofilms with engineered enzymatic bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11197-11202, (2007).

11 Purnick, P. E. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol* **10**, 410-422, (2009).

12 Cornelis, G. R. The type III secretion injectisome. *Nature reviews. Microbiology* **4**, 811-825, (2006).

13 Galan, J. E. & Collmer, A. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284**, 1322-1328, (1999).

14 Marlovits, T. C. *et al.* Structural insights into the assembly of the type III secretion needle complex. *Science* **306**, 1040-1042, (2004).

15 Galan, J. E. & Zhou, D. Striking a balance: modulation of the actin cytoskeleton by Salmonella. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 8754-8761, (2000).

16 Groisman, E. A. & Ochman, H. Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87**, 791-794, (1996).

17 Lucas, R. L. & Lee, C. A. Unravelling the mysteries of virulence gene regulation in Salmonella typhimurium. *Mol Microbiol* **36**, 1024-1033, (2000).

18 Clark, M. A., Jepson, M. A., Simmons, N. L. & Hirst, B. H. Preferential interaction of Salmonella typhimurium with mouse Peyer's patch M cells. *Res Microbiol* **145**, 543-552, (1994).

19 Jones, B. D., Ghori, N. & Falkow, S. Salmonella typhimurium initiates murine infection by penetrating and destroying the specialized epithelial M cells of the Peyer's patches. *J Exp Med* **180**, 15-23, (1994).

20 Eriksson, S., Lucchini, S., Thompson, A., Rhen, M. & Hinton, J. C. Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica. *Mol Microbiol* **47**, 103-118, (2003).
21 Eichelberg, K. & Galan, J. E. Differential regulation of Salmonella typhimurium type III secreted proteins by pathogenicity island 1 (SPI-1)-encoded transcriptional activators InvF and hilA. *Infect Immun* **67**, 4099-4105, (1999).
22 Matsui, H. *et al.* Virulence plasmid-borne spvB and spvC genes can replace the 90-kilobase plasmid in conferring virulence to Salmonella enterica serovar Typhimurium in subcutaneously inoculated mice. *Journal of bacteriology* **183**, 4652-4658, (2001).
23 Bajaj, V., Lucas, R. L., Hwang, C. & Lee, C. A. Co-ordinate regulation of Salmonella typhimurium invasion genes by environmental and regulatory factors is mediated by control of hilA expression. *Mol Microbiol* **22**, 703-714, (1996).
24 Boddicker, J. D., Knosp, B. M. & Jones, B. D. Transcription of the Salmonella invasion gene activator, hilA, requires HilD activation in the absence of negative regulators. *Journal of bacteriology* **185**, 525-533, (2003).
25 Lucas, R. L. & Lee, C. A. Roles of hilC and hilD in regulation of hilA expression in Salmonella enterica serovar Typhimurium. *Journal of bacteriology* **183**, 2733-2745, (2001).
26 Olekhnovich, I. N. & Kadner, R. J. DNA-binding activities of the HilC and HilD virulence regulatory proteins of Salmonella enterica serovar Typhimurium. *Journal of bacteriology* **184**, 4148-4160, (2002).
27 Schechter, L. M. & Lee, C. A. AraC/XylS family members, HilC and HilD, directly bind and derepress the Salmonella typhimurium hilA promoter. *Mol Microbiol* **40**, 1289-1299, (2001).
28 Kubori, T. *et al.* Supramolecular structure of the Salmonella typhimurium type III protein secretion system. *Science* **280**, 602-605, (1998).
29 Lostroh, C. P. & Lee, C. A. The HilA box and sequences outside it determine the magnitude of HilA-dependent activation of P(prgH) from Salmonella pathogenicity island 1. *Journal of bacteriology* **183**, 4876-4885, (2001).
30 Sukhan, A., Kubori, T., Wilson, J. & Galan, J. E. Genetic analysis of assembly of the Salmonella enterica serovar Typhimurium type III secretion-associated needle complex. *Journal of bacteriology* **183**, 1159-1167, (2001).
31 Lostroh, C. P., Bajaj, V. & Lee, C. A. The cis requirements for transcriptional activation by HilA, a virulence determinant encoded on SPI-1. *Mol Microbiol* **37**, 300-315, (2000).
32 Akbar, S., Schechter, L. M., Lostroh, C. P. & Lee, C. A. AraC/XylS family members, HilD and HilC, directly activate virulence gene expression independently of HilA in Salmonella typhimurium. *Mol Microbiol* **47**, 715-728, (2003).
33 Darwin, K. H. & Miller, V. L. InvF is required for expression of genes encoding proteins secreted by the SPI1 type III secretion apparatus in Salmonella typhimurium. *Journal of bacteriology* **181**, 4949-4954, (1999).
34 Alon, U. *An introduction to systems biology : design principles of biological circuits*. (Chapman & Hall/CRC, 2007).

35 Hooshangi, S., Thiberge, S. & Weiss, R. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3581-3586, (2005).
36 Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965-1969, (2005).
37 Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962-1965, (2005).
38 Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-68, (2002).
39 Kalir, S., Mangan, S. & Alon, U. A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli. *Mol Syst Biol* **1**, 2005 0006, (2005).
40 Mangan, S., Zaslaver, A. & Alon, U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J Mol Biol* **334**, 197-204, (2003).
41 Ellermeier, C. D., Ellermeier, J. R. & Slauch, J. M. HilD, HilC and RtsA constitute a feed forward loop that controls expression of the SPI1 type three secretion system regulator hilA in Salmonella enterica serovar Typhimurium. *Mol Microbiol* **57**, 691-705, (2005).
42 Schechter, L. M., Jain, S., Akbar, S. & Lee, C. A. The small nucleoid-binding proteins H-NS, HU, and Fis affect hilA expression in Salmonella enterica serovar Typhimurium. *Infect Immun* **71**, 5432-5435, (2003).
43 Biggar, S. R. & Crabtree, G. R. Cell signaling can direct either binary or graded transcriptional responses. *Embo J* **20**, 3167-3176, (2001).
44 Rossi, F. M., Kringstein, A. M., Spicher, A., Guicherit, O. M. & Blau, H. M. Transcriptional control: rheostat converted to on/off switch. *Mol Cell* **6**, 723-728, (2000).
45 Spudich, J. L. & Koshland, D. E., Jr. Non-genetic individuality: chance in the single cell. *Nature* **262**, 467-471, (1976).
46 Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *Embo J* **20**, 2528-2535, (2001).
47 Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & Van Oudenaarden, A. Multistability in the lactose utilization network of Escherichia coli. *Nature* **427**, 737-740, (2004).
48 Xiong, W. & Ferrell, J. E., Jr. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* **426**, 460-465, (2003).
49 De Keersmaecker, S. C. *et al.* Microarray analysis and motif detection reveal new targets of the Salmonella enterica serovar Typhimurium HilA regulatory protein, including hilA itself. *Journal of bacteriology* **187**, 4381-4391, (2005).
50 Darwin, K. H. & Miller, V. L. The putative invasion protein chaperone SicA acts together with InvF to activate the expression of Salmonella typhimurium virulence genes. *Mol Microbiol* **35**, 949-960, (2000).

51 Darwin, K. H. & Miller, V. L. Type III secretion chaperone-dependent regulation: activation of virulence genes by SicA and InvF in Salmonella typhimurium. *Embo J* **20**, 1850-1862, (2001).
52 Tucker, S. C. & Galan, J. E. Complex function for SicA, a Salmonella enterica serovar typhimurium type III secretion-associated chaperone. *Journal of bacteriology* **182**, 2262-2268, (2000).
53 Kalir, S. *et al.* Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* **292**, 2080-2083, (2001).
54 Lee, C. A. & Falkow, S. The ability of Salmonella to enter mammalian cells is affected by bacterial growth state. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 4304-4308, (1990).
55 Lundberg, U., Vinatzer, U., Berdnik, D., von Gabain, A. & Baccarini, M. Growth phase-regulated induction of Salmonella-induced macrophage apoptosis correlates with transient expression of SPI-1 genes. *Journal of bacteriology* **181**, 3433-3437, (1999).
56 Bumann, D. Examination of Salmonella gene expression in an infected mammalian host using the green fluorescent protein and two-colour flow cytometry. *Mol Microbiol* **43**, 1269-1283, (2002).
57 Hautefort, I., Proenca, M. J. & Hinton, J. C. Single-copy green fluorescent protein gene fusions allow accurate measurement of Salmonella gene expression in vitro and during infection of mammalian cells. *Appl Environ Microbiol* **69**, 7480-7491, (2003).
58 Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M. & Shapiro, L. Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **290**, 2144-2148, (2000).
59 Novick, A. & Weiner, M. Enzyme Induction as an All-or-None Phenomenon. *Proceedings of the National Academy of Sciences of the United States of America* **43**, 553-566, (1957).
60 Voigt, C. A., Wolf, D. M. & Arkin, A. P. The Bacillus subtilis sin operon: an evolvable network motif. *Genetics* **169**, 1187-1202, (2005).
61 Boddicker, J. D. & Jones, B. D. Lon protease activity causes down-regulation of Salmonella pathogenicity island 1 invasion gene expression after infection of epithelial cells. *Infect Immun* **72**, 2002-2013, (2004).
62 Suel, G. M., Garcia-Ojalvo, J., Liberman, L. M. & Elowitz, M. B. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* **440**, 545-550, (2006).
63 Detweiler, C. S., Monack, D. M., Brodsky, I. E., Mathew, H. & Falkow, S. virK, somA and rcsC are important for systemic Salmonella enterica serovar Typhimurium infection and cationic peptide resistance. *Mol Microbiol* **48**, 385-400, (2003).
64 Drecktrah, D., Knodler, L. A., Ireland, R. & Steele-Mortimer, O. The mechanism of Salmonella entry determines the vacuolar environment and intracellular gene expression. *Traffic* **7**, 39-51, (2006).
65 Abshire, K. Z. & Neidhardt, F. C. Growth rate paradox of Salmonella typhimurium within host macrophages. *Journal of bacteriology* **175**, 3744-3748, (1993).

66 Stamm, L. M. *et al.* Mycobacterium marinum escapes from phagosomes and is propelled by actin-based motility. *J Exp Med* **198**, 1361-1368, (2003).
67 Wolf, D. M. & Arkin, A. P. Motifs, modules and games in bacteria. *Curr Opin Microbiol* **6**, 125-134, (2003).
68 Walker, K. A. & Miller, V. L. Regulation of the Ysa type III secretion system of Yersinia enterocolitica by YsaE/SycB and YsrS/YsrR. *Journal of bacteriology* **186**, 4056-4066, (2004).
69 Mavris, M. *et al.* Regulation of transcription by the activity of the Shigella flexneri type III secretion apparatus. *Mol Microbiol* **43**, 1543-1553, (2002).
70 Le Gall, T. *et al.* Analysis of virulence plasmid gene expression defines three classes of effectors in the type III secretion system of Shigella flexneri. *Microbiology* **151**, 951-962, (2005).
71 Chilcott, G. S. & Hughes, K. T. Coupling of flagellar gene expression to flagellar assembly in Salmonella enterica serovar typhimurium and Escherichia coli. *Microbiol Mol Biol Rev* **64**, 694-708, (2000).
72 Monack, D. M., Detweiler, C. S. & Falkow, S. Salmonella pathogenicity island 2-dependent macrophage death is mediated in part by the host cysteine protease caspase-1. *Cell Microbiol* **3**, 825-837, (2001).
73 Hersh, D. *et al.* The Salmonella invasin SipB induces macrophage apoptosis by binding to caspase-1. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2396-2401, (1999).
74 Takaya, A., Kubota, Y., Isogai, E. & Yamamoto, T. Degradation of the HilC and HilD regulator proteins by ATP-dependent Lon protease leads to downregulation of Salmonella pathogenicity island 1 gene expression. *Mol Microbiol* **55**, 839-852, (2005).
75 Takaya, A. *et al.* Derepression of Salmonella pathogenicity island 1 genes within macrophages leads to rapid apoptosis via caspase-1- and caspase-3-dependent pathways. *Cell Microbiol* **7**, 79-90, (2005).
76 Guiney, D. G. The role of host cell death in Salmonella infections. *Curr Top Microbiol Immunol* **289**, 131-150, (2005).
77 Ellermeier, J. R. & Slauch, J. M. Adaptation to the host environment: regulation of the SPI1 type III secretion system in Salmonella enterica serovar Typhimurium. *Curr Opin Microbiol* **10**, 24-29, (2007).
78 Schlumberger, M. C. *et al.* Real-time imaging of type III secretion: Salmonella SipA injection into host cells. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12548-12553, (2005).
79 Menard, R., Sansonetti, P., Parsot, C. & Vasselon, T. Extracellular association and cytoplasmic partitioning of the IpaB and IpaC invasins of S. flexneri. *Cell* **79**, 515-525, (1994).
80 Birket, S. E. *et al.* Preparation and characterization of translocator/chaperone complexes and their component proteins from Shigella flexneri. *Biochemistry* **46**, 8128-8137, (2007).
81 Osiecki, J. C. *et al.* IpaC from Shigella and SipC from Salmonella possess similar biochemical properties but are functionally distinct. *Mol Microbiol* **42**, 469-481, (2001).

82 Enninga, J., Mounier, J., Sansonetti, P. & Tran Van Nhieu, G. Secretion of type III effectors into host cells in real time. *Nature methods* **2**, 959-965, (2005).
83 Fischbach, M., Voigt, C.A. Prokaryotic gene clusters: A rich toolbox for synthetic biology. *Biotechnol. J.* **5**, 1277-1296, (2010).
84 Ishihama, A. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol Rev* **34**, 628-645, (2010).
85 Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* **5**, 451-463, (2004).
86 Temme, K. *et al.* Induction and relaxation dynamics of the regulatory network controlling the type III secretion system encoded within Salmonella pathogenicity island 1. *J Mol Biol* **377**, 47-61, (2008).
87 Georg, J. & Hess, W. R. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev* **75**, 286-300, (2011).
88 Guell, M., Yus, E., Lluch-Senar, M. & Serrano, L. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nature reviews. Microbiology* **9**, 658-669, (2011).
89 Johnson, Z. I. & Chisholm, S. W. Properties of overlapping genes are conserved across microbial genomes. *Genome Res* **14**, 2268-2272, (2004).
90 Zazopoulos, E. *et al.* A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nature biotechnology* **21**, 187-190, (2003).
91 Gottelt, M., Koi, S., Gomez-Escribano, J.P., Bibb, M., Takano, E. Deletion of a regulatory gene within the cpk gene cluster reveals novel antibacterial activity in Steptomyces coelicolor A3(2). *Microbiology* **156**, 2343-2353, (2010).
92 Lombo, F., Brana, A.F., Mendez, C., Salas, J.A. The mithramycin gene cluster of Steptomyces argillaceus contains a positive regulatory gene and two repeated DNA sequences that are located at both ends of the cluster. *J. Bacteriol.* **181**, 642-647, (1999).
93 Medema, M. H., Bretiling, R., Takano, E. Synthetic biology in Steptomyces bacteria. *Methods in Enzymology* **497**, 485-502, (2011).
94 Pickens, L. B., Tang, Y., Chooi, Y-H. Metabolic engineering for the production of natural products. *Annu. Rev. Chem. Biomol. Eng.* **2**, 1-26, (2011).
95 Smanski, M. J., Peterson, R.M., Rajski, S.R., Shen, B. . Engineered Streptomyces platensis strains that overproduce antibiotics platensimycin and platencin. *Antimicrob. Agents Chemother.* **53**, 1299-12304, (2009).
96 Biggins, J. B., Liu, X., Feng, Z., Brady, S.F. Metabolites from the induced expression of crypic single operons found in the genome of Burkolderia pseudomallei. *JACS* **133**, 1638-1641, (2011).
97 Banik, J. J., Brady, S.F. Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr. Opin. Microb.* **13**, 603-609, (2010).
98 Watanabe, K., Hotta, K., Praseuth, A.P., Koketsu, K., Migita, A., Boddy, C.N., Wang, C.C.C., Oguri, H., Oikawa, H. Total biosynthesis of antitumor nonribosomal peptides in Escherichia coli. *Nature Chemical Biology*, 1-6, (2006).
99 Dymond, J. S. *et al.* Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477**, 471-476, (2011).
100 Isaacs, F. J. *et al.* Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* **333**, 348-353, (2011).

101 Kodumal, S. J., Patel, K.G., Reid, R., Menzella, H.G., Welch, M., Santi, D.V. Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl. Acad. Sci. USA* **101**, 15573-15578, (2004).
102 Chan, L. Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol Syst Biol* **1**, 2005 0018, (2005).
103 Burris, R. H. Nitrogenases. *J Biol Chem* **266**, 9339-9342, (1991).
104 Hu, Y., Fay, A. W., Lee, C. C., Yoshizawa, J. & Ribbe, M. W. Assembly of nitrogenase MoFe protein. *Biochemistry* **47**, 3973-3981, (2008).
105 Rubio, L. M. & Ludden, P. W. Maturation of nitrogenase: a biochemical puzzle. *J Bacteriol* **187**, 405-414, (2005).
106 Einsle, O. *et al.* Nitrogenase MoFe-protein at 1.16 A resolution: a central ligand in the FeMo-cofactor. *Science* **297**, 1696-1700, (2002).
107 Hoover, T. R. *et al.* Identification of the V factor needed for synthesis of the iron-molybdenum cofactor of nitrogenase as homocitrate. *Nature* **329**, 855-857, (1987).
108 Shah, V. K. & Brill, W. J. Isolation of an iron-molybdenum cofactor from nitrogenase. *Proc Natl Acad Sci U S A* **74**, 3249-3253, (1977).
109 Arnold, W., Rump, A., Klipp, W., Priefer, U. B. & Puhler, A. Nucleotide sequence of a 24,206-base-pair DNA fragment carrying the entire nitrogen fixation gene cluster of Klebsiella pneumoniae. *J Mol Biol* **203**, 715-738, (1988).
110 Dixon, R. A. & Postgate, J. R. Genetic transfer of nitrogen fixation from Klebsiella pneumoniae to Escherichia coli. *Nature* **237**, 102-103, (1972).
111 Homer, M. J., Paustian, T. D., Shah, V. K. & Roberts, G. P. The nifY product of Klebsiella pneumoniae is associated with apodinitrogenase and dissociates upon activation with the iron-molybdenum cofactor. *J Bacteriol* **175**, 4907-4910, (1993).
112 Simon, H. M., Homer, M. J. & Roberts, G. P. Perturbation of nifT expression in Klebsiella pneumoniae has limited effect on nitrogen fixation. *J Bacteriol* **178**, 2975-2977, (1996).
113 Gosink, M. M., Franklin, N. M. & Roberts, G. P. The product of the Klebsiella pneumoniae nifX gene is a negative regulator of the nitrogen fixation (nif) regulon. *J Bacteriol* **172**, 1441-1447, (1990).
114 Dixon, R. & Kahn, D. Genetic regulation of biological nitrogen fixation. *Nat Rev Microbiol* **2**, 621-631, (2004).
115 Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* **8**, R22, (2007).
116 Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27**, 946-950, (2009).
117 Bayer, T. S. *et al.* Synthesis of Methyl Halides from Biomass Using Engineered Microbes. *J Am Chem Soc* **131**, 6508-6515, (2009).
118 Orme-Johnson, W. H. Molecular basis of biological nitrogen fixation. *Annu Rev Biophys Biophys Chem* **14**, 419-459, (1985).

119 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345, (2009).
120 Yokobayashi, Y., Weiss, R. & Arnold, F. H. Directed evolution of a genetic circuit. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 16587-16591, (2002).
121 Basu, S., Gerchman, Y., Collins, C. H., Arnold, F. H. & Weiss, R. A synthetic multicellular system for programmed pattern formation. *Nature* **434**, 1130-1134, (2005).
122 Tabor, J. J. *et al.* A synthetic genetic edge detection program. *Cell* **137**, 1272-1281, (2009).
123 Murray, R. *CAGEN: Critical Assessment of Genetically Engineered Networks - OpenWetWare*, <http://openwetware.org/wiki/CAGEN> (2011).
124 Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-1664, (2002).
125 Palsson, B. *Systems biology : properties of reconstructed networks*. (Cambridge University Press, 2006).
126 Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335-338, (2000).
127 Shetty, R. P., Endy, D. & Knight, T. F., Jr. Engineering BioBrick vectors from BioBrick parts. *J Biol Eng* **2**, 5, (2008).
128 Li, M. Z. & Elledge, S. J. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nature methods* **4**, 251-256, (2007).
129 Welch, M., Villalobos, A., Gustafsson, C. & Minshull, J. Designing genes for successful protein expression. *Methods in Enzymology* **498**, 43-66, (2011).
130 Densmore, D. *et al.* Algorithms for automated DNA assembly. *Nucleic Acids Res* **38**, 2607-2616, (2010).
131 Leguia, M., Brophy, J., Densmore, D. & Anderson, J. C. Automated assembly of standard biological parts. *Methods in Enzymology* **498**, 363-397, (2011).
132 Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J. & Govindarajan, S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **7**, 285, (2006).
133 Ramon, A. & Smith, H. O. Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. *Biotechnol Lett* **33**, 549-555, (2011).
134 Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *The Computer Journal* **7**, 308-313, (1965).
135 Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12678-12683, (2005).
136 Ellis, T., Wang, X. & Collins, J. J. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature biotechnology* **27**, 465-471, (2009).
137 Studier, F. W. & Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* **189**, 113-130, (1986).
138 McBride, K. E., Schaaf, D. J., Daley, M. & Stalker, D. M. Controlled expression of plastid transgenes in plants based on a nuclear DNA-encoded and plastid-

targeted T7 RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 7301-7305, (1994).

139 Wyatt, L. S., Moss, B. & Rozenblatt, S. Replication-deficient vaccinia virus encoding bacteriophage T7 RNA polymerase for transient gene expression in mammalian cells. *Virology* **210**, 202-205, (1995).

140 Conrad, B., Savchenko, R. S., Breves, R. & Hofemeister, J. A T7 promoter-specific, inducible protein expression system for Bacillus subtilis. *Mol Gen Genet* **250**, 230-236, (1996).

141 An, W. & Chin, J. W. Synthesis of orthogonal transcription-translation networks. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8477-8482, (2009).

142 Ikeda, R. A., Warshamana, G. S. & Chang, L. L. In vivo and in vitro activities of point mutants of the bacteriophage T7 RNA polymerase promoter. *Biochemistry* **31**, 9073-9080, (1992).

143 Davidson, E. A., T, V. A. N. B., Levy, M. & Ellington, A. D. Emulsion based selection of t7 promoters of varying activity. *Pac Symp Biocomput*, 433-443, (2010).

144 Chelliserrykattil, J., Cai, G. & Ellington, A. D. A combined in vitro/in vivo selection for polymerases with novel promoter specificities. *BMC Biotechnol* **1**, 13, (2001).

145 Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499-503, (2011).

146 Raskin, C. A., Diaz, G. A. & McAllister, W. T. T7 RNA polymerase mutants with altered promoter specificities. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 3147-3151, (1993).

147 Iost, I., Guillerez, J. & Dreyfus, M. Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo. *Journal of bacteriology* **174**, 619-622, (1992).

148 Bonner, G., Lafer, E. M. & Sousa, R. Characterization of a set of T7 RNA polymerase active site mutants. *The Journal of biological chemistry* **269**, 25120-25128, (1994).

149 Gonzalez, M., Frank, E. G., Levine, A. S. & Woodgate, R. Lon-mediated proteolysis of the Escherichia coli UmuD mutagenesis protein: in vitro degradation and identification of residues required for proteolysis. *Genes Dev* **12**, 3889-3899, (1998).

150 Desai, T. A., Rodionov, D. A., Gelfand, M. S., Alm, E. J. & Rao, C. V. Engineering transcription factors with novel DNA-binding specificity using comparative genomics. *Nucleic Acids Res* **37**, 2493-2503, (2009).

151 Cheetham, G. M., Jeruzalmi, D. & Steitz, T. A. Structural basis for initiation of transcription from an RNA polymerase-promoter complex. *Nature* **399**, 80-83, (1999).

152 Rong, M., He, B., McAllister, W. T. & Durbin, R. K. Promoter specificity determinants of T7 RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 515-519, (1998).

153 Bandwar, R. P., Jia, Y., Stano, N. M. & Patel, S. S. Kinetic and thermodynamic basis of promoter strength: multiple steps of transcription initiation by T7 RNA

polymerase are modulated by the promoter sequence. *Biochemistry* **41**, 3586-3595, (2002).

154 Cunningham, F. X., Jr., Sun, Z., Chamovitz, D., Hirschberg, J. & Gantt, E. Molecular structure and enzymatic function of lycopene cyclase from the cyanobacterium Synechococcus sp strain PCC7942. *Plant Cell* **6**, 1107-1121, (1994).

155 Balibar, C. J. & Walsh, C. T. In vitro biosynthesis of violacein from L-tryptophan by the enzymes VioA-E from Chromobacterium violaceum. *Biochemistry* **45**, 15444-15457, (2006).

156 Sanchez, C., Brana, A. F., Mendez, C. & Salas, J. A. Reevaluation of the violacein biosynthetic pathway and its relationship to indolocarbazole biosynthesis. *Chembiochem* **7**, 1231-1240, (2006).

157 Ramos, J. L. *et al.* The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* **69**, 326-356, (2005).

158 Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52, (1999).

159 Filipovska, A. & Rackham, O. Building a Parallel Metabolism within the Cell. *ACS Chem Biol* **3**, 51-63, (2008).

160 Rackham, O. & Chin, J. W. Cellular logic with orthogonal ribosomes. *J Am Chem Soc* **127**, 17584-17585, (2005).

161 Fischbach, M. & Voigt, C. A. Prokaryotic gene clusters: a rich toolbox for synthetic biology. *Biotechnol J* **5**, 1277-1296, (2010).

162 Kelly, J. R. *et al.* Measuring the activity of BioBrick promoters using an in vivo reference standard. *Journal of biological engineering* **3**, 4, (2009).

163 Philippe, N., Alcaraz, J. P., Coursange, E., Geiselmann, J. & Schneider, D. Improvement of pCVD442, a suicide plasmid for gene allele exchange in bacteria. *Plasmid* **51**, 246-255, (2004).

164 Jacob, G. S., Schaefer, J., Garbow, J. R. & Stejskal, E. O. Solid-state NMR studies of Klebsiella pneumoniae grown under nitrogen-fixing conditions. *The Journal of biological chemistry* **262**, 254-259, (1987).

165 Mason, C. A. & Hamer, G. Cryptic Growth in Klebsiella-Pneumoniae. *Appl Microbiol Biot* **25**, 577-584, (1987).

166 Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* **38**, W695-699, (2010).

167 Lavigne, R., Sun, W. D. & Volckaert, G. PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. *Bioinformatics* **20**, 629-635, (2004).

168 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190, (2004).

Appendix A - Characterizing and Measuring Part Function

1. Promoter and Terminator Characterization

Standard Promoter Unit

Endy and co-workers proposed that an *in vivo* reference standard could reduce variation in the measurement of promoter activity [162]. We define one standard promoter unit (SPU) as the RFP expression from plasmid N110 (SBa_000564) when measured in the standard characterization assay. We calculated SPU as follows:

$$Activity\ of\ Promoter\ \phi\ (SPU) = \frac{\phi\ Activity\ (Arbitrary\ Fluorescence\ Units)}{N110\ Activity\ (Arbitrary\ Fluorescence\ Units)}$$

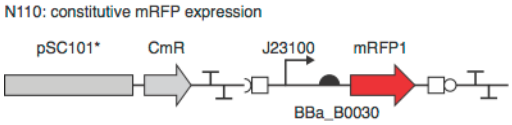

Figure A1 - Standard Promoter Unit Vector. Our reference standard is based on the Kelly standard [162]. However, we utilized a different constitutive promoter (J23100 vs J23101), RBS (BBa_0030 vs BBa_B0032), fluorescent reporter (mRFP1 vs GFP), and plasmid backbone (pSC101 and CmR vs p15a and KmR).

Promoter Characterization and T7 Promoter library

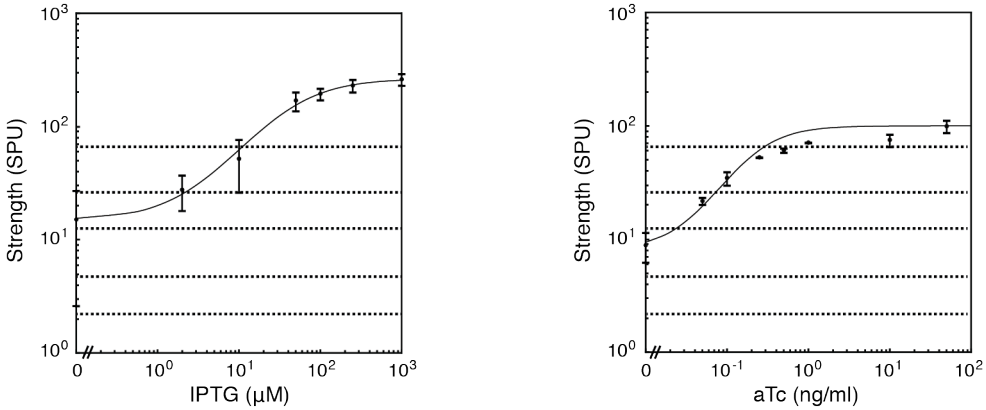

Figure A2 - Promoter characterization for P_{tac} promoter (left) and P_{tet} promoter (right). The promoter strengths of P_{tac} promoter and P_{tet} promoter were measured under varied concentrations of inducers (IPTG or aTc). The strengths of T7 promoters (T7 wild-type and mutants, Figure 1E) are shown as horizontal dotted lines. The strengths of T7 promoter mutants were characterized under 1mM IPTG induction from T7* RNAP (plasmid N249). The measured fluorescence (AU) for each mutant was converted to SPU.

Termination Efficiency Calculation

We created plasmid N292 (shown below) to test termination efficiency. It contains GFP and RFP, separated by the wild-type T7 terminator. We eliminated the terminator to produce control plasmid N287 to serve as a reference for our efficiency calculation. Mutant terminators are cloned in place of the wild-type terminator from N292. We calculate termination efficiency as follows:

$$Termination\ Efficiency\ (\%) = 1 - \frac{Terminator\ Activity\ (Arbitrary\ Fluorescence\ Units)}{N287\ Activity\ (Arbitrary\ Fluorescence\ Units)}$$

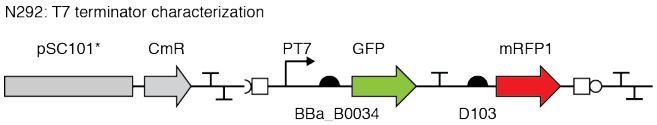

Figure A3 - Standard Terminator Characterization Vector.

Terminator Library and Characterization

Terminator plasmids were co-transformed with plasmid N249 and characterized under 1mM IPTG induction of T7* RNAP. RFP expression was measured for each terminator, and data are reported as the percentage reduction in measured fluorescence when compared to a derivative of N292 carrying no terminator.

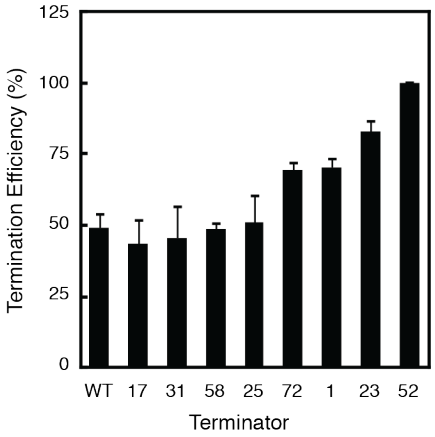

Figure A4 - Terminator characterization. A terminator library was cloned into plasmid N292 and co-transformed with the T7* RNAP (N249). Cells were induced by 1mM IPTG, and expression of a fluorescent protein downstream each terminator was measured by flow cytometry. Termination efficiency was calculated by comparing fluorescence to a plasmid that does not carry a terminator. Data reported are the results of three experiments performed on multiple days.

Appendix B - Nitrogen Fixation Gene Cluster Characterization and Engineering

1. Native Gene Cluster Re-Sequencing

The *nif* gene cluster in *K. oxytoca* Ma5L was re-sequenced from PCR fragments. The re-sequenced DNA sequence was compared to the reference sequence from Genbank, X13303.1 [109]. Sequence differences are listed in the below table. The nucleotide locations are numbered relative to X13303.1.

Table B1. DNA sequence errors in *nif* cluster sequence X13303.1

Location	X13303.1	This Study	Impact
5187	a	g	NifD K39E
5193	a	g	NifD K41E
5332	ct	tc	NifD A87V
5359	a	g	NifD E96G
6137	c	g	NifD I355M
6140	a	c	silent
6749	ga	ag	NifK R57Q
7371	t	c	silent
7371	t	c	silent
8168	cgg	ggc	*nifT* 5' UTR
8919	c	t	NifY P170L
9341	-	insert t	*nifE* 5' UTR
9498	cg	gc	NifE A35G
9697	ac	ct	silent
9891	c	t	NifE P166L
12541	ca	ac	NifX Q131T
12770	t	c	*nifU* 5' UTR
13646	gc	cg	*nifS* 5' UTR
13999	a	c	NifS M110L
14008	g	c	NifS E113Q
14037	a	g	silent
14041	g	a	NifS G124S
14539	gc	cg	NifS A290R
15657	tc	ct	NifV S257L
15770	gc	cg	NifV H295D
16739	g	a	silent
16814	c	t	silent
22050	g	a	NifB V141M
22309	g	a	NifB G227D

2. *K. oxytoca* Knockout Strains

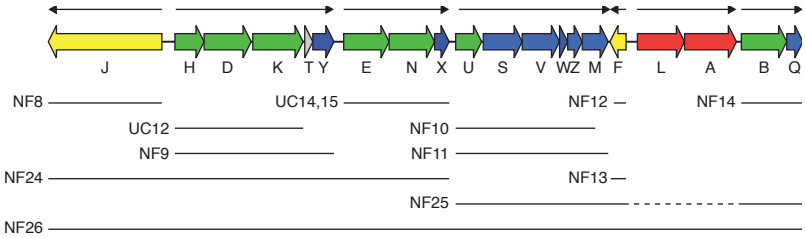

Figure B1 - The *nif* operon deletions are shown. The solid lines show the region of deleted *nif* operons. The dashed line in NF25 shows the retained *nifLA* operon.

Table B2. Construction and verification of all *K.oxytoca nif* gene deletion mutants

Name	Genotype	Note	Parental Strain	Allelic Exchange Plasmid[4]	DNA sequence of the deletion region[5]
NF8	Δ*nifJ, Km*	operon deletion	M5aL	pNif16	gcaggagaactaaaggcccg**GCATGC** (Km cassette)**GCATGC**aagcgcccatggccccggca
UC12	Δ*nifHDK*	in-frame deletion	M5aL	pRHB296	cataaacaggcacggctggt**GGTACC**ccatcaggtgccccgcgtca
NF9	Δ*nifHDKTY*	operon deletion	M5aL	pNif9	atggccccggcaggcgcaat**GCATGC**gacgctcttccccacgttac
UC14	Δ*nifENX*	operon deletion	M5aL	pRHB294	gacgctcttccccacgttac**GCATGC**gatccggacccgcgccgcta
UC15	Δ*nifENX*	operon deletion	M5aL	pRHB295	ctcttccccacgttacgctc**GCATGC**gatccggacccgcgccgcta
NF10	Δ*nifUSVWZM*	operon deletion[1]	M5aL	pNif10	cggacccgcgccgctagccg**GCATGC**atctttggcagcagagccag
NF11	Δ*nifUSVWZM*	operon deletion	M5aL	pNif11	cggacccgcgccgctagccg**GCATGC**agcctcggcggctacccgtt
NF12	Δ*nifF*	operon deletion[2]	M5aL	pNif12	actggttatcctgatccagc**GCATGC**ctttgcactaccgcggccca
NF13	Δ*nifF*	operon deletion	M5aL	pNif13	ccgttaacgcctacagcacg**GCATGC**ctttgcactaccgcggccca
NF14	Δ*nifBQ*	operon deletion	M5aL	pNif14	attcagggacgcgggttgcc------atgtgattatgcgacgtctt
NF24	Δ*nifJ,HDKTY, ENX*	operon deletion[3]	M5aL	pNif32	gtcaaacaccagaatattga**GAGCTC**tgtcgtttctgtgacaaagc
NF25	Δ*nifUSVWZM,F,BQ*	operon deletion	NF14	pNif51	cggacccgcgccgctagccg**GCATGC**ctttgcactaccgcggccca
NF26	Δ*nifJ,HDKTY,ENX, USVWZM,F,LA,BQ*	whole *nif* deletion[3]	NF24	pNif54	gtcaaacaccagaatattga**CTCGAG**atgtgattatgcgacgtctt

1) 83bp of *nifM* 3' retained
2) 82bp of *nifF* 3' retained
3) 500bp of *nifJ* 3' retained
4) The backbone for all allelic exchange plasmids is pDS132[163].
5) The bold region shows the restriction site sequence that replaces the deletion region. The 20 nucleotides flanking each site of the deletion region are shown.

3. Growth by Nitrogen Fixation

Cells capable of nitrogen fixation should exhibit measurable growth on media that lacks nitrogen by utilizing atmospheric N_2 as a source of nitrogen. Conversely, cells incapable of nitrogen fixation should not grow on nitrogen-free media.

In parallel to the $^{15}N_2$ incorporation assay, we monitored strain growth under nitrogen-limited media conditions and 100% $^{15}N_2$. Cells were grown on derepression media as used in the Nitrogenase Activity Assay. Derepression media is not strictly nitrogen-free, containing 1.43 mM serine in order to promote ribosomal RNA production and hasten nitrogenase biosynthesis [164].

Strains containing Controller #1 and the refactored gene cluster grew nearly 30% as much as wild-type strains. In contrast, minimal growth was observed in *Δnif* strains, consistent with the limited nitrogen available from serine and cell lysis products [165].

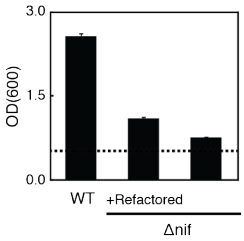

Figure B2 - Cell growth supported by nitrogen fixation. The dotted line indicates initial seeding density of OD600 0.5. Wild-type *Klebsiella* grew to an OD600 2.57 ± 0.07 after 36 hours of incubation in depression conditions (Methods, N_2-dependent Growth Assay). Eliminating the full *nif* cluster severely inhibited cell growth (*Δnif*, OD600 0.76 ± 0.02). Complementing the knockout strain with the refactored cluster and Controller #1 under 1mM IPTG induction yielded growth of OD600 1.10 ± 0.03.

4. Western Blot Assay for Synthetic *nifH* Expression

The first synthetic *nifHDK* did not exhibit nitrogenase activity under induction ranging from 0 to 1mM IPTG, and the *nifH* gene (synthetic $nifH_{v1}$) was identified as a problematic part using the debugging protocol shown in Figure 3-3B. However there was no mutation found. Western blots were further used to confirm problematic synthetic *nifH* expression.

A western blot for NifH protein in Figure B3 (left) showed that wild type *nifH* expressed well with either synthetic *nifD* or *nifK* (construct N10, N12, N14), whereas synthetic $nifH_{v1}$ was not expressed regardless of the context of *nifDK* (construct N1 and N19). A second synthetic *nifH* (synthetic $nifH_{v2}$) was used to replace synthetic $nifH_{v1}$. The western blot in Figure B3 (right) showed the synthetic $nifH_{v2}$ (construct N38) expressed well.

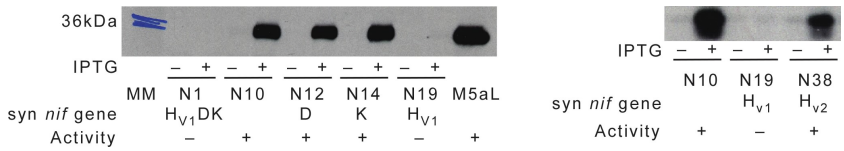

Figure B3 - Expression of synthetic *nifH* variants. Western blot assay to detect the expression of synthetic $nifH_{v1}$ (left) and synthetic synthetic $nifH_{v2}$ (right). All constructs bore P_{tac}-*nifHDK* with the synthetic gene indicated. Cultures were induced with 50μM IPTG.

Appendix C - Generation of Orthogonal Parts

1. Identification of RNAP and Promoter Subfamilies

Inspired by comparative genomics, we utilized a bioinformatics approach to guide our generation of orthogonal RNAP:promoter interactions. Initially, we searched for homologs of the T7 RNAP via a BLAST serach of the NCBI database. We identified 43 proteins with an E-value lower than 1^{-100}. From this set, we identified 28 phage RNAP with fully sequenced genomes and selected these for further analysis. We aligned the RNAPs using ClustalW (default parameters, except gap opening penalty of 100) to identify the specificity loop (T7 G732-P780) in each protein (Figure C1) [166]. We find that the 28 phage RNAP cluster into 13 subfamilies based on their specificity loop sequences.

We then identified promoters in each phage genome, using available annotations or the PHIRE software package [167]. WebLogo was used to generate sequence logos for each phage [168], and the regions from -12bp to -7bp in each logo were aligned. We find that the promoter sequence logos cluster into 11 subfamilies. Each family directly correlates with one RNAP subfamily, as shown in Figure C1.

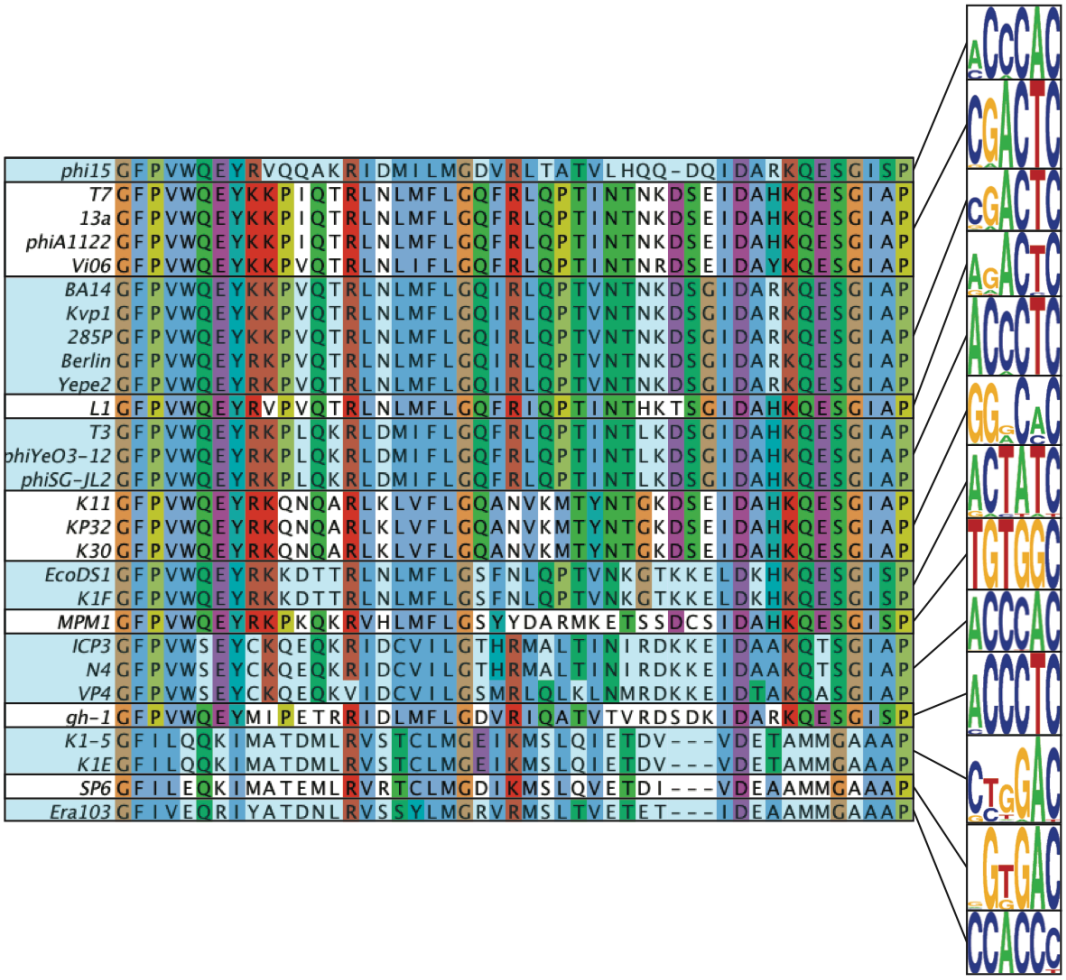

Figure C1 - Identification of RNAP subfamilies based on specificity loop sequence alignment. Homologs to T7 RNAP were identified from NCBI, and their specificity loops were aligned. The RNAPs cluster into 13 subfamilies. Promoters were identified in each phage genome and used to construct consensus sequence logos for each phage. Consensus sequences were aligned and clustered into 11 subfamilies. RNAP subfamilies perfectly correlate with consensus promoter subfamilies, meaning each member RNAP recognizes the same consensus promoter sequence.

2. Generation of Orthogonal RNAP:Promoter Pairs

Based on specificity loop subfamilies and consensus promoter subfamilies, we mutated the T7* RNAP and wild-type T7 promoter. In all cases, we replaced the 6bp between -12 and -7 in the wild-type T7 promoter. For RNAP mutants, we made 2 or 3 derivatives for each subfamily in which the size of the specificity loop varied. The derivative RNAPs exhibiting activity were mutated in the intervening region between residues 744 and 761 (Figure C2). In all cases, we found the activity of derivative RNAP to be quite weak in the N249 backbone. When we made the same mutations to earlier versions of the non-toxic RNAP (Figure 5-1), we found much higher derivative RNAP activity.

	739 750 760 766	ΔT7*	Promoter	Backbone
T7*	YKKPIQTRLNLMFLGQFRLQPTINTNKD	-	CGACTC	v4
T3*	YKKPIQKRLDMIFLGQFRLQPTINTNKD	4	ACCCTC	v2
K1F*	YKKPIQTRLNLMFLGSFNLQPTVNTNKD	3	ACTATC	v2
N4*	YKKPIQTRIDCVILGTHRMALTINTNKD	10	ACCCAC	v1

Figure C2 - Specificity loop and promoter specificity region sequences. Orthogonal RNAP were generated by grafting novel specificity loops into the T7* RNAP between residues 744 and 761. Orthogonal promoters were generated by replacing in the binding domain in the wild-type T7 promoter (between -12bp to -7bp) with the consensus sequence from a given phage subfamily in a fluorescent reporter plasmid (N155). RNAP and promoters were co-transformed, grown in 1mM IPTG and their activity measured by flow cytometry. To improve activity of orthogonal RNAP, mutant specificity loops conferring activity were transferred to earlier versions of the non-toxic T7* RNAP system. The optimal RNAP version for each orthogonal RNAP is shown.