Title: Computational design to efficiently search, map, and optimize multi-protein genetic systems in gram-negative and gram-positive bacteria

Authors: Iman Farasat[1], Manish Kushwaha[2], Jason Collens[2], Michael Easterbrook[1], Matthew Guido[1], Howard M. Salis†[1,2]

[1]Department of Chemical Engineering and [2]Department of Biological Engineering, Pennsylvania State University, University Park, PA 16802
†Corresponding author: salis@psu.edu

Abstract

Engineering multi-protein genetic systems to maximize their performance remains a combinatorial challenge, particularly when measurement throughput is limited. We have developed a computational design and modeling approach to build predictive models and identify optimal expression levels, while circumventing combinatorial explosion. Maximally informative genetic system variants are first designed by the RBS Library Calculator, an algorithm that optimizes the smallest ribosome binding site library to efficiently search the expression space across a >10,000-fold range with tailored search resolutions, sequence constraints, and well-predicted translation rates. We validated the algorithm's predictions using a 644 sequence data-set, within single and multi-protein genetic systems, modifying plasmids and genomes, and in *Escherichia coli* and *Bacillus subtilis*. We then combined the search algorithm with kinetic modeling to map the mechanistic relationship between sequence, expression, and overall activity for a 3-enzyme biosynthesis pathway, requiring only 73 measurements to forward design highly productive pathway variants. The combination of sequence design and systems modeling accelerates the optimization of many-protein systems, and allow previous measurements to quantitatively inform future designs.

Main Text

Engineering metabolic pathways and genetic circuits requires the systematic tuning of protein expression levels to identify a genetic system variant that delivers a desired behavior (Ajikumar et al, 2010; Du et al, 2012; Makino et al, 2011; Moon et al, 2012; Paddon et al, 2013; Quan et al, 2011; Santos et al, 2012; Tseng & Prather, 2012; Xu et al, 2013; Yim et al, 2011; Zelcbuch et al, 2013; Zhang et al, 2012; Zhao et al, 2013). Understanding the relationship between a genetic system's expression levels and its phenotype enables rational optimization of its behavior, though this relationship is difficult to determine particularly when many proteins are interacting together. The optimal expression levels to maximize a targeted behavior will also

vary according to the proteins' activities and the system's biomolecular interactions; efficient approaches are needed to optimize diverse genetic systems.

Expression optimization of genetic systems has been performed by introducing a library of genetic parts with different DNA sequences to vary protein expression, followed by characterization of the genetic system's behavior to identify the variants that functioned best (Du et al, 2012; Lee et al, 2013; Pfleger et al, 2006; Sandoval et al, 2012; Santos et al, 2012; Torella et al, 2013; Wang et al, 2009; Xu et al, 2013; Zelcbuch et al, 2013). Recent advances have dramatically improved our ability to assemble multiple DNA fragments, or introduce targeted DNA mutations, into several locations within plasmids and genomes, enabling combinatorial expression optimization of larger genetic systems (Cong et al, 2013; Maresca et al, 2013; Urnov et al, 2010; Wang et al, 2009). In particular, engineered ribosome binding sites (RBSs) are commonly used to control a mRNA's translation rate, and its corresponding protein expression level, due to their non-repetitive sequences, their proximity to the protein's coding sequence, their control over individual proteins in bacterial operons, and the option of combining them with existing promoters to dynamically regulate expression (Bonnet et al, 2012; Lou et al, 2010; Moon et al, 2012; Mutalik et al, 2013; Salis, 2011; Wang et al, 2009; Zelcbuch et al, 2013).

These advances in the synthesis, assembly, and mutagenesis of large genetic systems have enabled the targeting of a much larger set of proteins, and access to a wider range of engineered behaviors (Yadav et al, 2012). However, several limitations constrain our ability to apply combinatorial expression optimization to large genetic systems, understand the relationship between expression and behavior, and find variants with the best possible behavior. The construction of genetic system libraries is limited by the yield and breadth of the DNA modification technique as well as the maximum number of modified plasmids or genomes that can be maintained inside the population of host organisms. The number of characterized genetic system variants is also limited by the throughput of the assay used to quantify its performance. Finally, as the number of proteins targeted for optimization increases, the number of combinations of different expression levels for each protein will rapidly grow, diminishing the fraction of combinations that can be constructed and characterized using the same approach. As a result of this under-sampling, especially when using random mutagenesis of genetic parts, the chance of finding the best possible genetic system variant with near-optimal expression levels greatly decreases as the size of the expression space grows (**Figure 1A**). In particular, the combinatorial expansion of the expression level space can not be addressed by developing improved methods for assembling DNA libraries or increasing assay throughput.

Here, we present a computational design approach to overcome the limitations of combinatorial expression optimization, identify expression-activity relationships, and find the

optimal expression levels in a multi-protein genetic system. First, we formulate the mini-max expression optimization problem whose solution is the smallest genetic part library that maximally searches a genetic system's expression level space. To solve this optimization problem for bacterial genetic systems, we developed an automated algorithm, called the RBS Library Calculator, to design the smallest synthetic RBS library that uniformly increases a protein's expression level across a selected translation rate range on a >100,000-fold proportional scale. This algorithm combines a predictive biophysical model of bacterial translation with a genetic optimization algorithm. Through iterations of *in silico* mutation, recombination, prediction, and selection, synthetic RBS library sequences using the 16-letter degenerate alphabet are designed to maximize the search coverage of a selected translation rate space, while minimizing the number of RBS variants in the library (**Figure 1B**). The algorithm has several modes: *Search* to cover the widest possible expression space; *Genome Editing* to constrain expression optimization towards using the fewest, consecutive genome mutations; and *Zoom* to target a narrow expression range towards the optimal levels. These modes enable one to control the search space, search resolution, and sequence design constraints to be configured according to the number of proteins targeted for combinatorial expression optimization, the DNA mutagenesis technique, and the assay's throughput.

Notably, the RBS Library Calculator uses a biophysical model to design an unlimited number of synthetic RBS libraries, on-demand, with well-predicted translation initiation rates. Biophysical models can account for the several mechanisms controlling translation rate (Espah Borujeni et al; Gingold & Pilpel, 2011; Na et al, 2010; Salis, 2011; Salis et al, 2009; Zouridis & Hatzimanikatis, 2007). In particular, our biophysical model calculates the ribosome's binding free energy to mRNAs, which is responsible for controlling its translation initiation rate. The free energy model is determined by 16S rRNA hybridization, canonical and non-canonical Shine-Dalgarno sequences, differences in start codons and spacer regions, the unfolding of mRNA structures, upstream standby site accessibility, and the presence of long-range RNA interactions. Our model can also account for the differences in ribosomes between bacterial species, enabling the prediction of translation initiation rates in valuable, but less well-studied, organisms (Jaschke et al, 2011; Medina et al, 2011; Ravasi et al, 2012). Here, we show that these predictions are accurate in both *Escherichia coli* and *Bacillus subtilis* as model organisms for gram-negative and gram-positive bacteria. We also discuss the differences between using a biophysical model to design genetic parts in contrast to utilizing a toolbox of previously characterized genetic parts.

Second, to overcome the limits of combinatorial expansion, we use system-level mechanistic modeling to construct and validate a quantitative relationship between a genetic system's RBS sequences, translation rates, and phenotype, which is then used to predict the expression levels and RBS sequences that maximize the genetic system's performance. We demonstrate this

approach on a 3-enzyme carotenoid biosynthesis pathway, showing that characterization of only 73 pathway variants was necessary to develop a predictive kinetic model. Using this quantitative relationship, we then forward design pathway variants with optimal enzyme expression levels, achieving high carotenoid productivities of up to 441 µg/gDCW/hr.

Results

Solving the protein expression mini-max optimization problem

To validate the RBS Library Calculator's *Search* mode, three optimized RBS libraries were designed using high, medium, or low search resolutions with 36, 16, or 8-variants per library, respectively, to control reporter protein expression on a multi-copy plasmid in *E. coli* DH10B (**Table I**). Degenerate RBS sequences primarily utilized 2-nucleotide degeneracies (S, K, R, B, and M) with only one instance of a 3-nucleotide degeneracy (M). None contained a 4-nucleotide degeneracy (N). *Search* mode inserted degenerate nucleotides 5 to 19 nucleotides upstream of the start codon to modulate both the 16S rRNA binding affinity and the unfolding energetics of inhibitory mRNA structures. We quantified the optimized RBS libraries' search ranges, coverages, and translation rate predictions by measuring reporter protein expression levels from individual RBS variants within each library. Fluorescence measurements were taken during 24-hour cultures maintained in the early exponential growth phase by serial dilutions. All DNA sequences, translation rate predictions, and fluorescence measurements are provided in the **Supplementary Table 1**.

Fluorescence measurements show that the optimized RBS libraries searched the 1-dimensional (1 protein) expression level spaces with high coverages, high dynamic ranges, and accurate translation rate predictions. The 36-variant RBS library systematically increased *mRFP1* expression from low to high levels with a 49,000-fold dynamic range and 94% search coverage (**Figure 1C**), while the 16-variant RBS library uniformly increased *sfGFP* expression across a 169,000-fold scale, from below the detection limit to nearly the maximum allowable cytometer signal, with only a small coverage gap at 100 au (79% search coverage) (**Figure 1D**). The lowest resolution RBS library contained only 8 variants, but uniformly increased *sfGFP* expression between the selected translation rate range, yielding protein expression levels from 63 to 49,000 au (299-fold dynamic range) with a high 99% search coverage (**Figure 1E**).

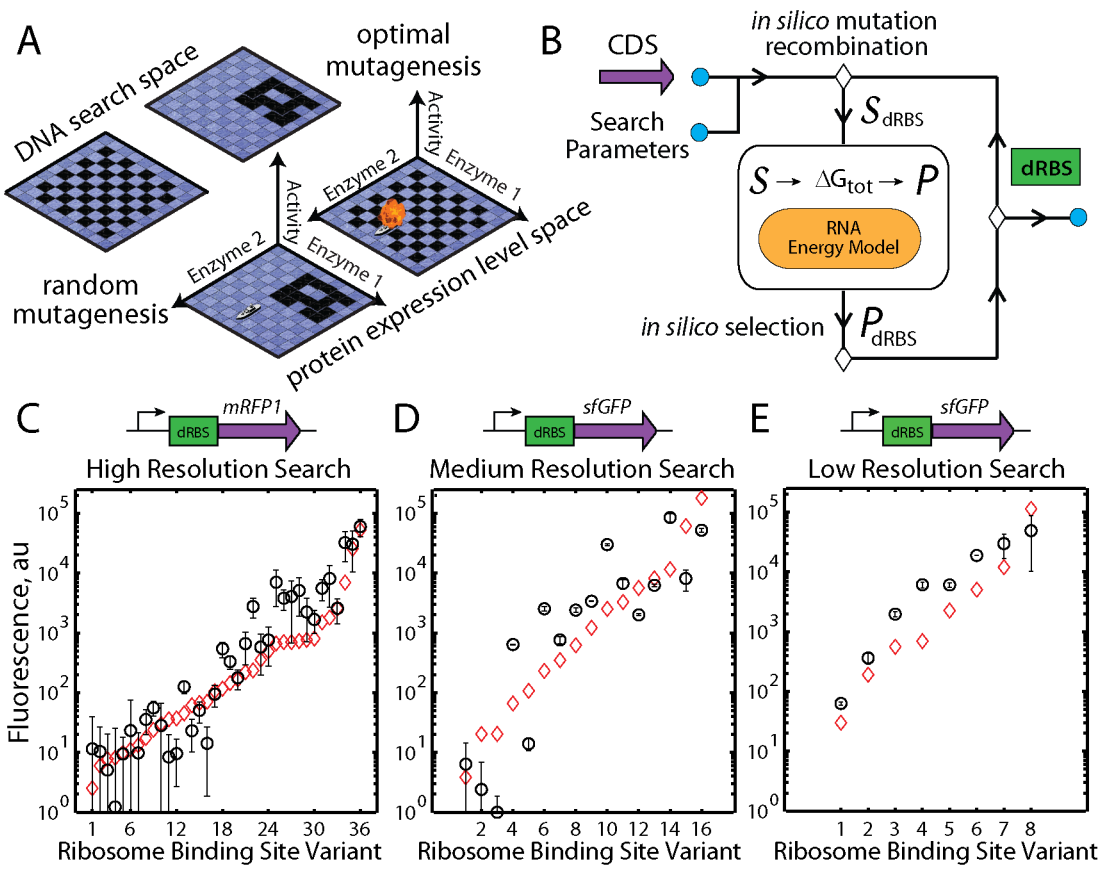

Figure 1: Validation of the RBS Library Calculator's *Search* mode in *E. coli*. (A) A conceptual illustration that random DNA mutagenesis produces clustering in expression level space, due to limitations in library capacity and characterization throughput. Black squares represent covered points in sequence or expression space. Optimal mutagenesis – the winning strategy to the game of Battleship – creates fewer sequence variants that maximally cover the expression level space. (B) The algorithm combines a biophysical model of translation with a genetic algorithm to identify the smallest degenerate RBS sequences with maximal search coverage. (C, D, E) Optimized RBS libraries search a 1-dimensional expression level space with 94%, 79%, 99% search coverages at high, medium, and low search resolutions, resp. Translation initiation rate predictions (red diamonds) are compared to measurements (Pearson R^2 is 0.88, 0.79, and 0.89, respectively). Data averages and standard deviations from 6 measurements.

Table 1: Characteristics of optimized and random RBS libraries.

Degenerate RBS Sequence	Protein	Min TIR (model scale)	Max TIR (model scale)	Res	# Seq
***Search* Mode**					
AACGACGTCGACGATCACAACTTSAKGDBGTATTC	**mRFP1**	**2.5**	**53000**	**0.2**	**36**
ACTGATCTAGGGAAAGCATTASGSASGTCRAAAGA	**sfGFP**	**3.8**	**180000**	**0.3**	**16**
CGTAAAGTTAAACCGMGCGAAATTAGKASGTATTA	**sfGFP**	**30**	**113000**	**0.35**	**8**
AACGAAGACMATGATCACAACTTAAKGASGTATTC	**CFP**	**8.0**	**46000**	**0.35**	**8**
AACGCCGTCGACGKTCACAACTTCAGGASGTMTTC	**mRFP1**	**70**	**35000**	**0.35**	**8**

ACTAGGTTTATACCACAAAACAAGKGGKWTAAAAA	GFPmut3b	10	46000	0.35	8
CCAATATACCAAATAAAGAGTYGMGGMSGTCAAGG	CrtE	68	72000	0.3	16
AACGTACACACACAATTATACKAAGSRGRTCCGAA	CrtB	3.3	20000	0.3	16
TAAACCCAACAAATTAGACTATAAKKAGKYTAATA	CrtI	97	203000	0.3	16
***Zoom* Mode**					
AATTCGATTTTTAGGAACAGTTAAGGRGGHTAATA	CrtE	32000	305000	0.35	6
AGAGTACAATAGAMATYAAAATMAGGAGGTCAACA	CrtB	1800	233000	0.35	8
TTAGATTTTAAATAACAATACTMAKGAGGTSCAAC	CrtI	26000	1347000	0.35	8
***Genome Editing* Mode**					
TGTGAGCGGATAACAATTTTWAVGASGAAACAGCT	LacZ	20	55000	0.3	12
GAGACGCAAATAWGGMGKTWCTCGAATTCGAATTC	mRFP1	2.3	21000	0.25	16
ATACMTAACACAAAGWGGAGGYAGAATTCGAATTC	mRFP1	2500	96000	0.2	8
Random RBS Libraries					
AACGAAGACAATGATCACAACTTANNNNNNTATTC	CFP	0.4	34000	--	4096
TCACAACTTAACGCCGTCGACGGCNNNNNNTATAT	mRFP1	6.7	148000	--	4096
ACTAGGTTTATACCACAAAACAANNNNNNTAACAA	GFPmut3b	5.1	58000	--	4096

The biophysical model of bacterial translation accurately predicted the translation initiation rates from the 60 RBS variants with an average error $\Delta\Delta G_{total}$ of 1.74 kcal/mol, which is equivalent to predicting the measured translation initiation rate to within 2.2-fold. The biophysical model's predictions were particularly accurate for the high and low resolution libraries (average $\Delta\Delta G_{total}$ = 1.05 kcal/mol, R^2 = 0.88; and average $\Delta\Delta G_{total}$ = 0.46 kcal/mol, R^2 = 0.89, respectively) in contrast to the medium resolution library that contains several outliers at low expression levels (average $\Delta\Delta G_{total}$ = 3.71 kcal/mol, R^2 = 0.72).

Several types of user-selected sequence constraints can be incorporated into the solution. In the simplest example, RBS libraries may be designed to include restriction enzyme recognition sites, homologous overlap sequences, or other desired nucleotide sequences at specified locations that enable their facile molecular cloning into a genetic system. Next, we introduce a set of RBS sequence constraints that optimizes directed genome mutagenesis to efficiently control and search chromosomally-encoded protein expression levels.

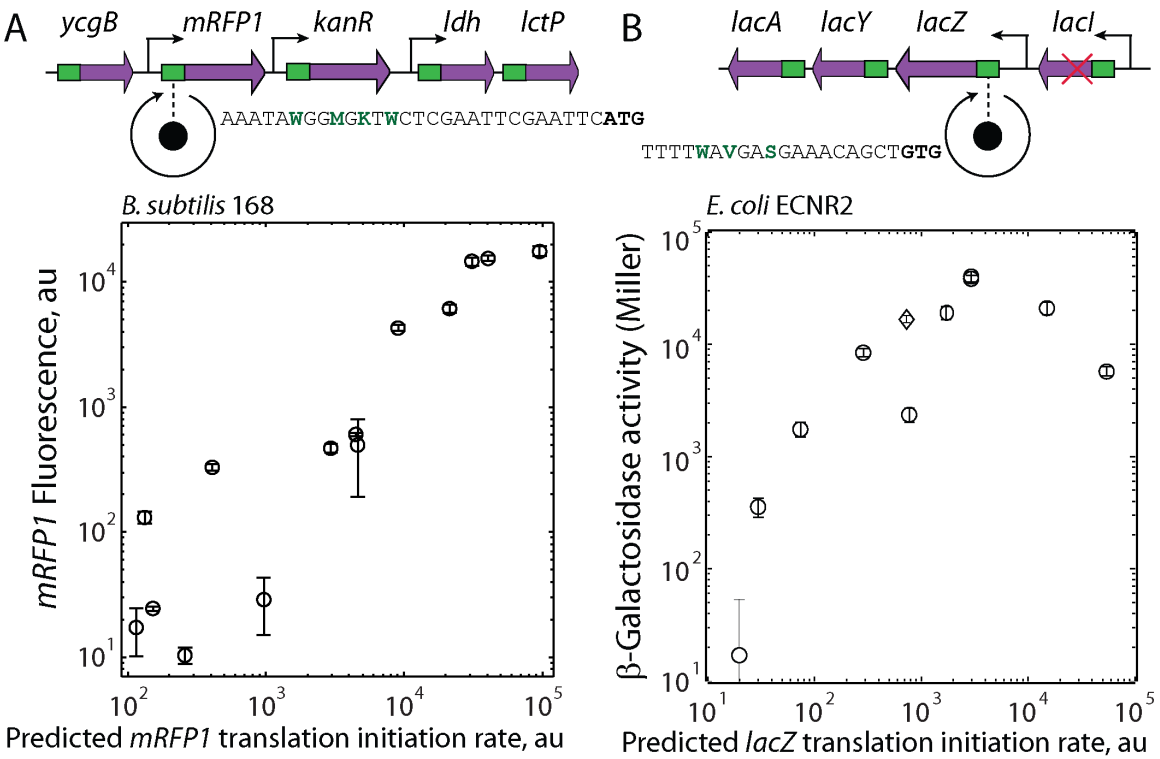

Figure 2: Validation of the RBS Library Calculator's *Genome Editing* mode in *E. coli* and *B. subtilis*. (A) Two RBS libraries were optimized to control the expression of a genomic single-copy of *mRFP1*, incorporated into the *amyE* locus of *B. subtilis*. Fluorescence measurements from 14 clones were compared to their predicted translation initiation rates (Pearson R^2 is 0.81). The expression space was searched with 76% coverage. Data averages and standard deviations from 3 measurements. (B) A 12-variant RBS library was optimized to control genomic *lacZ* expression. Predicted translation initiation rates are compared to measured *lacZ* activities (circles), including the wild-type (diamond), showing a linear relationship below the activity plateau (Pearson R^2 is 0.93). The expression space was searched with 84% coverage. Data averages and standard deviations from 4 measurements.

Automated search for optimal protein expression levels encoded in gram-positive and gram-negative bacterial genomes

Genome engineering techniques enable the targeted mutagenesis of genomic DNA, either by employing oligo-mediated allelic recombination, homologous recombination, or site-directed non-homologous end joining (Cho et al, 2013; Cong et al, 2013; Esvelt & Wang, 2013; Mali et al, 2013; Sharan et al, 2009; Urnov et al, 2010; Wang et al, 2009). Targeted mutations can modulate genomic protein expression levels, though the number and breadth of nucleotide variants that can be inserted in a single pass is limited by several factors, including the requirements for homology and specificity, the number and location of the targeted loci, the efficiency of DNA repair mechanisms, and the activity of helper recombinases or endonucleases. To search chromosomally-encoded protein expression levels, we introduced a *Genome Editing* version of the algorithm that identifies the minimal number of genomic RBS

mutations that uniformly increases a protein's expression level across a wide range. Optimization is initialized using the wild-type genomic RBS and protein coding sequences, and the solution is directly used to perform genome mutagenesis. The effectiveness of this approach was evaluated in both gram-negative and gram-positive bacteria to illustrate how the biophysical model's ability to predict translation rates in diverse organisms can be used for finding optimal expression levels.

First, we employed homologous recombination to introduce an optimized library of heterologous cassettes into the *Bacillus subtilis* 168 genome, using *Genome Editing* mode to optimize two RBS libraries that control expression of the reporter *mRFP1* with translation initiation rates from 100 to 96000 au on the model's proportional scale (**Table I**). Translation rate predictions use ACCUCCUUU as the 3' end of the *B. subtilis* 16S rRNA. Fluorescence measurements of 14 single clones from the libraries show that single-copy *mRFP1* expression varied from 10 and 17600 au with a search coverage of 76%, well-predicted translation initiation rates that were proportional to the measured expression levels (R^2 = 0.81), and with a low error in the calculated ribosomal interactions (average $\Delta\Delta G_{total}$ = 1.77 kcal/mol) (**Figure 2A**). This example demonstrates that the physical interactions that control translation initiation in gram-negative bacteria are sufficiently similar in gram-positive bacteria to provide the ability to accurately predict translation rates and efficiently find optimal expression levels.

Second, we employed MAGE mutagenesis on the *E. coli* MG1655-derived *EcNR2* genome (Wang et al, 2009), targeting its *lacI-lacZYA* locus and controlling *lacZ* protein expression levels (**Figure 2B**). We first conducted three rounds of MAGE mutagenesis to introduce an in-frame stop codon into the *lacI* repressor coding sequence (**Supplementary Tables 2 and 3**). Using the algorithm's *Genome Editing* mode, we then designed a 12-variant degenerate oligonucleotide with 7 consecutive mutated positions to target the *lacZ* RBS sequence and uniformly increase its translation initiation rate from 10 to 100,000 au (**Table I**). We conducted twenty rounds of MAGE mutagenesis to introduce the 12 sets of RBS mutations into the genome, and selected 16 colonies for sequencing of the *lacZ* genomic region. 10 of these colonies harbored genomes with unique mutated RBS sequences controlling *lacZ* translation.

lacZ activities from the derivative EcNR2 genomes were individually measured using Miller assays after long-time cultures maintained in the early exponential growth phase (**Figure 2C**). The measured *lacZ* expression levels varied across a 2400-fold range, searched the expression space with 84% coverage, and were well predicted by the biophysical model's predicted translation initiation rates up to 3000 au on the model's proportional scale (R^2 = 0.93). Though, interestingly, increasing the *lacZ* translation initiation rate beyond 3000 au, which is 4-fold over its wild-type rate, did not further increase *lacZ* activity. Specific growth rates were recorded and did not change as translation initiation rates were increased (**Supplementary Table 1**). The

plateau in protein expression suggests that there is a critical point where translation initiation may no longer be the rate-limiting step in protein expression, which we call the maximum translation rate capacity. Measuring the maximum translation rate capacity of a coding sequence becomes essential when proportional control over protein expression is desired, particularly at high levels. Next, we use the RBS Library Calculator to measure the maximum translation rate capacity of a codon-optimized protein coding sequence.

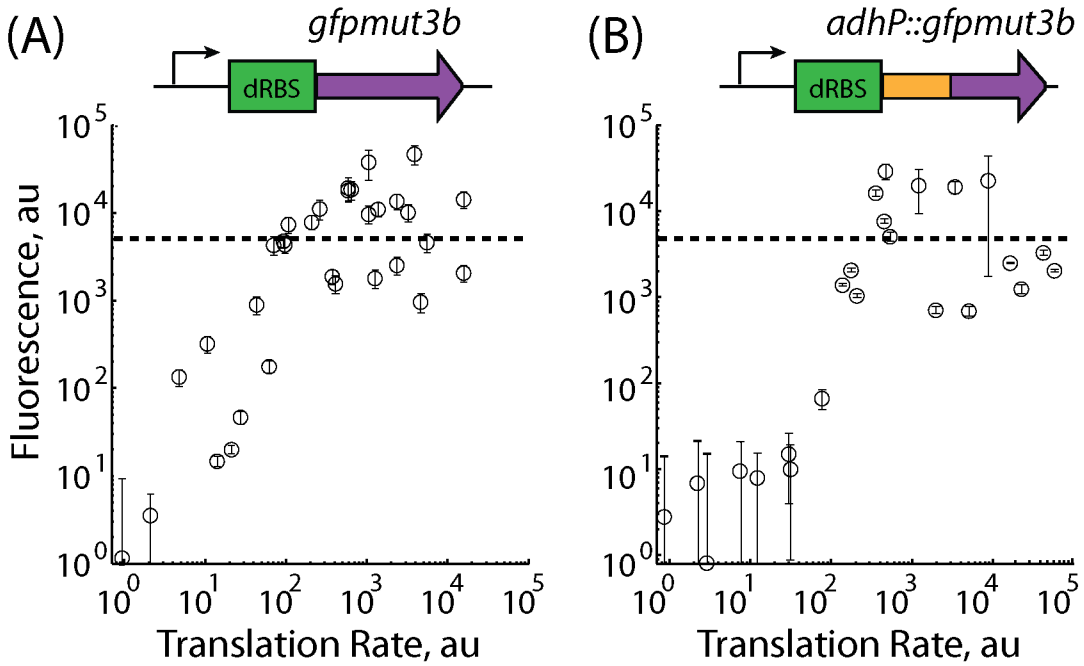

Figure 3: Measuring the translation rate capacity of a codon-optimized gene. The translation initiation rates of (A) *gfpmut3b* and (B) *adhp::gfpmut3* fusion proteins were uniformly increased across a 10,000-fold scale using an optimized RBS library to identify the critical point where translation initiation is no longer the rate-limiting step in protein expression and folding. The expression of both proteins reached the same plateau (dashed line) at similar translation initiation rates. Data averages and standard deviations from 3 measurements.

Determining the maximum translation rate capacity of codon-optimized genes

The translation rate of a protein coding sequence is determined by both its translation initiation and elongation rates; whichever step is slowest becomes the rate-limiting step for the overall translation process. Systematic increases in a gene's translation initiation rate allow one to identify the critical point when translation elongation becomes the rate-limiting step, labeled the maximum translation rate capacity. While codon optimization is a commonly used approach to improving a gene's translation elongation rate, the maximum translation rate capacity of codon-optimized genes has never been quantified.

We employed the RBS Library Calculator to determine the maximum initiation-limited translation rate capacity of a codon-optimized protein *gfpmut3b* in rich media conditions. A 32-variant RBS library was optimized with *Search* mode to vary translation initiation rates between 1 and 100,000 au. Fluorescence measurements of 12 RBS variants with translation initiation rates from 1.2 to 644 au proportionally controlled protein expression levels according to the biophysical model's predictions (R^2 = 0.93) (**Figure 3A**). However, with additional increases in translation initiation beyond this critical point, fluorescent protein expression levels reached a sporadic plateau. We then investigated whether the observed plateau was intrinsic to the *gfpmut3b* coding sequence affecting translation elongation and protein folding, or to the regulatory sequences controlling translation initiation. We created an N-terminal fusion between *adhP* and *gfpmut3b*, and designed a new 24-variant optimized RBS library to control the translation initiation rate of the *adhP*::*gfpmut3b* gene. Fluorescence measurements of individual RBS variants revealed a plateau with the same average fluorescence (**Figure 3B**) at a similar predicted translation initiation rate of 400 au on the model's proportional scale. Therefore, the plateau in protein expression is intrinsic to the *gfpmut3b* protein coding sequence, which occurs at a critical translation initiation rate. This critical point is an accurate quantitative metric of the gene's maximum initiation-limited translation rate capacity for a selected growth condition.

The best approach to codon optimization has remained unclear, particularly as many mechanisms for altering translation elongation rates have been demonstrated (Li et al, 2012; Menzella, 2011; Plotkin & Kudla, 2010). The application of the RBS Library Calculator to measure a gene's maximum translation rate capacity differentiates between proposed codon optimization approaches while verifying that a gene has truly been codon-optimized. Importantly, for valuable multi-protein systems where maximum overexpression is desired, independently verifying each gene's translation rate capacity with optimized RBS libraries, using either gene fusions or employing translational coupling (Mendez-Perez et al, 2012), is both time-efficient and prudent.

Efficient search in multi-dimensional expression level spaces

Most complex genetic systems express multiple proteins that work together to carry out their function. Optimization of multi-protein systems is particularly difficult, as it requires searching a larger combinatorial protein expression level space. We next evaluated *Search* mode's ability to explore a 3-dimensional expression space by constructing a bacterial operon encoding *cfp*, *mRFP1*, and *gfpmut3b* reporter proteins, and introducing either optimized or randomly mutagenized RBS libraries to control their translation initiation rates. The optimized RBS libraries were designed using a low search resolution. The resulting 8-variant RBS libraries contained 2-nucleotide degeneracies at distributed positions from 4 to 26 nucleotides upstream

of the start codon, including positions far from the Shine-Dalgarno sequence, that varied predicted translation rates across a 5000-fold range (**Supplementary Table 4**). To create the randomly mutagenized RBS libraries, we introduced 4-nucleotide degeneracies in a six nucleotide region within the Shine-Dalgarno sequence, a commonly used approach to vary translation rates, to create a 4096-variant RBS library with widely different translation rates (**Supplementary Table 5**). In both cases, 3-part combinatorial assembly of DNA fragments was employed to construct a library of bacterial operons (Gibson et al, 2009), generating 512 operon variants when using optimized RBS libraries, and up to 68.7 billion operon variants when using random RBS libraries. The extent of DNA library assembly is limited, however, and only a sub-sample of the randomized bacterial operon variants will ever be constructed or selected for characterization.

We compared search coverages when using either optimized or randomly mutagenized RBS libraries in the 3-protein bacterial operon. For each case, 500 strains with operon variants were randomly selected, individually cultured, and their CFP, mRFP1, and GFPmut3b fluorescences were quantified by color-corrected flow cytometry. The optimized RBS libraries searched the 3-dimensional protein expression level space across a 20,000-fold range with a 42% search coverage (**Supplementary Figure 1**). In contrast, the randomly mutagenized RBS library only partly covered the expression level space, showing a high degree of clustering that is responsible for decreasing its search coverage to 14% (**Supplementary Figure 2**), which agrees with computationally predicted search coverage of 14.7% using Monte Carlo sampling (**Supplementary Methods**). Similar search coverages of 16.7% and 19.1% are computationally predicted for a 4096-variant library NNNGGANNN (Mutalik et al, 2013) and a 23328-variant library DDRRRRRDDDD (Wang et al, 2009), respectively. A minority of randomly generated operon variants expressed higher or lower levels that would be necessary for many applications. Using the algorithm's *Search* mode, higher-dimensional expression spaces may be efficiently sampled with high coverages at targeted resolutions (**Supplementary Figure 3**).

As even more proteins are targeted for optimization, it becomes increasingly difficult to use a single combinatorial library to search for optimal expression levels with a sufficiently high resolution and coverage to extract useful knowledge. Instead, the RBS Library Calculator can be used in an iterative fashion to narrow down the optimal protein expression levels within a large multi-dimensional space. First, optimized RBS libraries are designed in *Search* mode with a low search resolution to cover the largest translation rate space, and combinatorially cloned to create a library of genetic systems. After sequencing and measuring the performance of a small number of genetic variants, their RBS variants are fed to the biophysical model of translation, which predicts the translation rates of each protein coding sequence. Second, these performance measurements and translation rates are combined with system-level modeling to predict the optimal translation rates that will maximize the genetic system's performance.

Finally, the RBS Library Calculator in *Zoom* mode is used to optimize an RBS library to target these optimal translation regions, creating improved genetic systems with higher performances. Importantly, the use of system-level modeling to predict the relationship between translation rate and performance allows one to carry out global optimization of the genetic system, identifying the best possible variant while avoiding local maxima. In the next section, we demonstrate this strategy by carrying out global optimization of a 3-enzyme biosynthesis pathway.

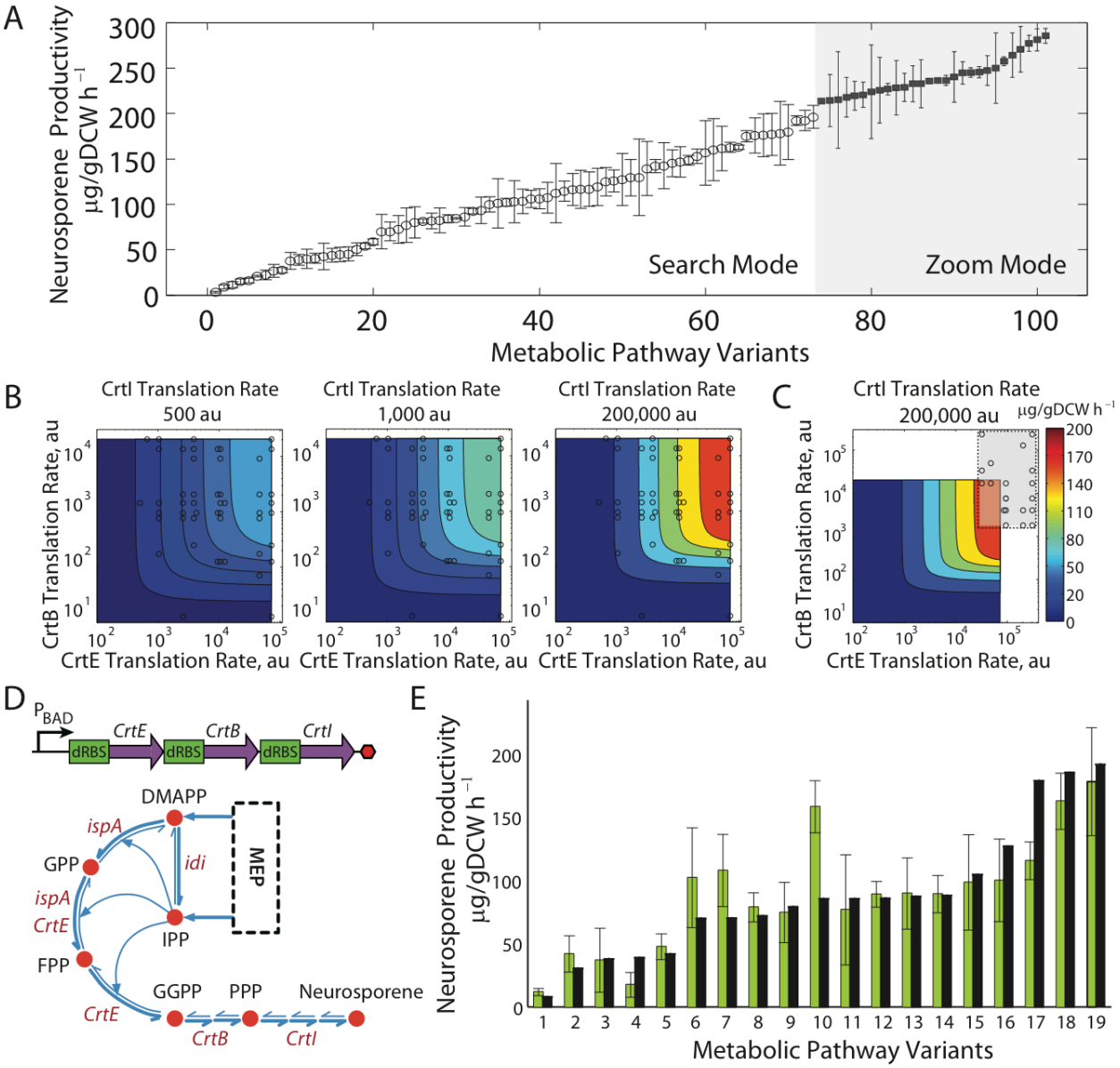

Figure 4: Efficient optimization of multi-enzyme pathways. (A) Characterization of two libraries of neurosporene biosynthesis pathway variants, using optimized RBS libraries designed by *Search* mode (left) or *Zoom* mode (right). Averages and standard deviations from at least 3 measurements of neurosporene productivities. (B) Measurement data and translation rate predictions (circles) from *Search* mode are used to parameterize a kinetic model of the

pathway's reaction rates, showing the relationship between *crtEBI* translation rates and neurosporene productivity. (C) According to model results, a translation rate region (gray box) is targeted in *Zoom* mode. Translation rate predictions from selected pathway variants are shown (circles). (D) A schematic of the bacterial operon encoding CrtEBI, and their corresponding reactions and metabolites. Cofactors are not shown. (E) Characterization of 19 additional CrtEBI pathway variants, comparing the kinetic model's predicted neurosporene productivities (black bars) and measurements (green bars). Data averages and standard deviations from 2 measurements.

Pathway Mapping and Optimization Using Kinetic Modeling

We next applied the RBS Library Calculator to carry out mapping and optimization of a multi-enzyme pathway, while minimizing the number of constructed pathway variants and measurements. Our approach combines three steps: using the algorithm's *Search* mode to determine the relationship between RBS sequence, translation rate, and pathway activity; applying modeling to predict the optimal translation rate regions with maximal pathway activity; and designing improved pathway variants using the algorithm's *Zoom* mode for sweeping a narrow targeted region of translation space that exhibits high activity.

Search mode was employed to vary the translation rates of a 3-enzyme carotenoid biosynthesis pathway from *R. sphaeroides*. Three 16-variant optimized RBS libraries were designed to vary *crtE*, *crtB*, and *crtI* from 445 to 72000 au, 3 to 20000 au, and 97 to 203000 au, respectively (**Supplementary Table 6**). 3-part combinatorial DNA assembly onto a ColE1 vector resulted in up to 4096 clonal pathway variants, transcribed by the arabinose-induced P_{BAD} promoter. 73 clones containing unique pathway variants were randomly selected, sequenced, transformed into *E. coli* MG1655-derived EcHW2f strain (**Supplementary Table 3**), and cultured for a 7 hour post-induction period. Their neurosporene contents were determined by hot acetone extraction and spectrophotometry. Within a single library, the pathways' neurosporene productivities were uniformly varied between 3.3 to 196 µg/gDCW/hr (**Figure 4A** and **Supplementary Table 7**). Using optimized RBS libraries yielded a large continuum of pathway activities with the smallest number of measurements.

Biophysical model predictions from sequenced RBSs indicate that the translation rates broadly explored the selected 3-dimensional space (**Figure 4B**), which eliminates redundant measurements and thus maximizes the measurements' information content. As *crtEBI* translation rates were increased, pathway productivities did not reach a plateau, suggesting that translation initiation remains the rate-limiting step. The biophysical model identifies where each pathway variant exists within the translation rate space, which can be combined with systems-level modeling to understand the relationship between enzyme expression and metabolic flux, and to direct the design of pathway variants with improved productivities.

To demonstrate this approach, we developed a kinetic model of the pathway's reaction network, listing 24 reversible elementary reactions that describe the enzymatic conversion of isoprenoid precursors (DMAPP and IPP) to neurosporene, including enzymes' binding to substrates, and the release of products (**Figure 4D** and **Supplementary Figure 4**). Differential equations describing the metabolic dynamics have 48 unknown kinetic parameters. Mole balances on each enzyme and flux constraints reduced the equations to having 33 unknown parameters (**Supplementary Methods**). We use an ensemble modeling approach (Contador et al, 2009; Tran et al, 2008) that combines model reduction and dimensional analysis to compare the pathway variants' calculated fluxes to a reference pathway, and to relate the kinetic model's predicted neurosporene production fluxes to measured neurosporene productivities. Changing a pathway variant's translation rates proportionally control the kinetic model's total enzyme concentrations, which alters the predicted neurosporene productivity.

We then employed model identification to determine the kinetic model parameters that reproduced the measured neurosporene productivities for the 72 non-reference pathway variants, across ten independent and randomly initialized optimization runs (**Supplementary Table 8**). The resulting kinetic model relates *crtEBI* translation rates to neurosporene productivities across a 10,000-fold, 3-dimensional translation rate space (**Figure 4B**). To test the kinetic model's predictive ability, we characterized 19 additional pathway variants, used the biophysical model to predicted their *crtEBI* translation rates from sequenced RBSs, and applied the kinetic model to calculate each variant's productivity (**Figure 4E**). The kinetic model correctly determined how changing the enzymes' translation rates controlled the pathway's productivity (24% error across a 100-fold productivity range) (**Supplementary Table 9**). Overall, kinetic model predictions were more accurate at higher *crtEBI* translation rates (**Supplementary Figure 5**). In general, a high *crtE* translation rate was necessary for high biosynthesis rates, while low *crtB* and high *crtI* translation rates were sufficient to balance the pathway.

To target *crtEBI* translation rates towards higher pathway productivities, we employed the algorithm in *Zoom* mode to design low resolution, 8-variant RBS libraries with translation rate ranges predicted by the kinetic model; 32000 to 305000 au for *crtE*, 1800 to 232000 au for *crtB*, and 26000 to 1347000 au for *crtI* (**Figure 4C** and **Supplementary Table 10**). After combinatorial DNA assembly, 28 clones containing unique pathway variants were randomly selected, sequenced, and cultured for a 7 hour post-induction period. The resulting neurosporene productivities improved up to 286 µg/gDCW/hr (**Figure 4A**) (**Supplementary Table 11**). The best pathway variant's neurosporene productivity was further increased to 441 µg/gDCW/hr when the strain was grown in improved media and aeration conditions (Alper et al, 2006) (**Supplementary Figure 6**). Alternative, non-mechanistic models may also be employed to guide system optimization; for example, using computational geometry to identify optimal expression levels resulted in a 15% predicted productivity error (**Supplementary Information**).

Discussion

A key challenge to successfully engineering cellular organisms has been the combinatorial vastness of their genetic instruction space, and the complex relationship between organism genotype and phenotype. Some DNA mutations have no effect, while others dramatically alter an organism's behavior. The ability to identify the best DNA sequence for a desired phenotype is complicated by the sheer number of design choices. The number of possible sequences in 150 base pairs of DNA, sufficient to encode at most ten promoters or ribosome binding sites, is more than the number of atoms in the Universe. While advances in DNA synthesis, library assembly, and high-throughput mutagenesis have accelerated the engineering design-build-test cycle for genetic systems, understanding the relationship between genotype and phenotype is essential to engineer large genetic systems where it is simply not feasible to exhaustively construct, characterize, or screen for a desired behavior.

To engineer genetic systems, a common synthetic biology approach has been to utilize a toolbox of previously characterized genetic parts to control protein expression (Babiskin & Smolke, 2011; Blount et al, 2012; Mutalik et al, 2013). Significant characterization effort is needed to ensure that these parts are modular, orthogonal, and can vary the expression of many individual proteins uniformly across a >100,000-fold range. Several non-repetitive genetic parts are also needed to express multiple proteins at similar levels as repetitive sequences are known to induce homologous recombination, particularly when the genetic system places a significant burden on the host's growth rate (Lovett, 2004; Sleight et al, 2010). The toolbox itself is a static list of DNA sequences and measured functions, and can not incorporate additional design criteria or sequence constraints *ex post facto* without additional characterization to ensure similar function. This limitation inhibits the use of new DNA assembly and genome modification techniques that rely on required sequence lengths, properties, or motifs (Gibson et al, 2009; Jiang et al, 2013; Wang et al, 2009; Wang et al, 2012). These measured functions also depend on the host organism, and significant re-characterization is needed to engineer genetic systems in alternative hosts.

In contrast, our computational approach designs an unlimited number of non-repetitive ribosome binding site library sequences to uniformly control translation initiation across the physiologically possible range, while incorporating flexible sequence lengths and constraints. RBS library sequences are optimized using experimentally validated biophysical rules that account for the differences in ribosomes across bacteria, and can predict translation initiation rates in both gram-negative and gram-positive bacteria. Potentially confounding interactions that affect protein expression are minimized by eliminating long single-stranded RNA regions or long RNA duplexes that may reduce mRNA stability, by ensuring that translation elongation is not rate-limiting, and by ensuring that mRNAs are always translated to protect them from

RNAse activity. By incorporating these design rules into the engineering of bacterial operons, one can achieve proportional control of protein expression by manipulating only RBS sequences. As additional biophysical rules continue to be developed (Espah Borujeni et al), they are incorporated into the forward design process, and can improve the accuracy of predictions on previously designed sequences. Thus, computational design can evolve concomitantly with our understanding of gene expression and the development of new DNA assembly, genome mutagenesis, and genome synthesis techniques to accelerate the engineering of large genetic systems.

Regardless of the source of the genetic parts, our ability to engineer many-protein genetic systems is limited by the expansion of the combinatorial expression space, particularly when proteins work synergistically to control system behavior. System-level models can both explain and inform the relationship between a genetic system's expression levels and its overall function (Westerhoff & Palsson, 2004), but identifying and validating the model's parameter values has often required extensive characterization of the system using several methods (Prill et al, 2010). Here, we created a predictive kinetic model of a biosynthesis pathway using only 73 measurements of the pathway's final product to determine the unknown kinetic parameters, and a further 19 measurements of new pathway variants to validate their significance. A key outcome of this work is the use of computational design to maximize the information content of each measurement, while minimizing the presence of confounding variables. We introduced a small number of genetic perturbations into a rationally designed genetic system that maximally affected protein expression, eliminating redundancy and context effects. We also apply model reduction to eliminate non-independent model parameters and time-scale dependence. As a result, the number of unknown model parameters grows linearly with the genetic system's complexity; modeling a pathway with an additional enzyme-catalyzed reaction will add 3 independent parameters, requiring 6 maximally informative measurements to identify them. This approach is broadly applicable to diverse types of models.

An analysis of the *crtEBI* pathway's kinetic model explains why metabolic optimization efforts have been generally laborious. First, each enzyme has the potential to be a rate-limiting step in the pathway. Distributed control over the pathway's flux requires that all enzyme expression levels must be tuned achieve high productivities. In particular, large changes in enzyme expression levels are needed to exert control; small changes in enzyme levels are buffered by compensating changes in metabolite concentrations (Fendt et al, 2010). This principle illustrates the need for genetic parts that maximally change protein expression levels across a wide range.

Second, though pathway optimization efforts aim to achieve optimal expression levels, the definition of optimality has remained elusive. Metabolic pathways are balanced when their

intermediate metabolites do not accumulate to toxic levels; however, the overall net flux through a balanced pathway can vary dramatically, and highly productive pathways are desired. We use the kinetic model to demonstrate a new definition of optimality that incorporates the use of flux control coefficients (FCCs) from the field of Metabolic Control Analysis (Fell, 1992; Kholodenko & Westerhoff, 1993).

Shown in **Supplementary Figure 8**, the *crtEBI* pathway's FCCs succinctly quantify how differential changes in enzyme expression control the pathway's overall productivity *P*, according to the partial derivatives $\frac{\partial \log P}{\partial \log[crtE]}$, $\frac{\partial \log P}{\partial \log[crtB]}$, and $\frac{\partial \log P}{\partial \log[crtI]}$, which are evaluated across the 3-dimensional expression space. High FCCs indicate where increasing an enzyme's expression will increase pathway's productivity, while low FCCs show regions where increasing expression does not lead to a significant improvement in productivity. Negative FCCs show regions where excess enzyme expression causes growth toxicity, due to overconsumption of the IPP and DMAPP precursors or saturation of the host's protein synthesis capacity. Using FCCs, we can determine when a pathway is considered balanced or optimally balanced. A pathway is balanced when differential increases in enzyme expression all have the same effect on pathway productivity, which occurs when the enzymes' FCCs are equal. A balanced pathway may have a low pathway productivity if the FCCs are all equally high; increases in all the pathway's enzymes will increase the pathway's productivity. In contrast, an optimally balanced pathway will have nearly zero FCCs; increasing the enzymes' expression levels has a minimal impact on pathway productivity. According to the summation rule for FCCs, if control over a pathway's productivity is reduced at one step, it is correspondingly increased at another. An optimally balanced pathway has shifted control of its flux over to the upstream metabolic module controlling precursor biosynthesis. These criteria form the foundation for designing optimally balanced metabolic modules that synergistically work together to maximize product biosynthesis rates.

Altogether, the ability to rationally search expression spaces, parameterize models to predict phenotype from genotype, and target optimal expression levels enables the efficient mapping and optimization of a multi-protein genetic system towards quantitatively defined optimally balanced criteria. The solution to the expression mini-max optimization problem, and the modeling of genetic systems, provides a quantitative design and diagnostic framework to identify the most important DNA mutations that will best improve a system's performance. Our proposed approach will dramatically accelerate the rational engineering of genetic systems by reducing characterization efforts and generating an understandable relationship between sequence, expression, and system performance.

A software implementation of the RBS Library Calculator is available at http://www.salis.psu.edu/software, online since 2011. As of October 2013, 350 unaffiliated researchers have designed 3015 optimized RBS libraries for diverse biotechnology applications.

Materials and Methods

Strains and Plasmid Construction

All strains and plasmids are listed in **Supplementary Table 3**.

To construct plasmid-based RBS libraries in *Escherichia coli* strain DH10B, CDS sequences (mRFP1 or sfGFP) were PCR amplified from pFTV1 or pFTV2 using mixed primers that encode optimized degenerate RBSs. The gel-purified PCR product was joined with digested, gel-purified vector backbone using a 2-part chew-back anneal-repair (CBAR) reaction (Gibson et al, 2009) to create the pIF1, pIF2, and pIF3 expression plasmids. Plasmids were transformed into *E. coli* DH10B, selected on chloramphenicol, and verified by sequencing. Expression plasmids contain a ColE1 origin of replication, a chloramphenicol resistance marker, the J23100 sigma^{70} constitutive promoter, the optimized degenerate ribosome binding site, and the selected reporter gene.

To construct genomic RBS libraries in *Bacillus subtilis* strain 168, a *Bacillus* integration vector pDG1661 was modified by replacing the spoVG-*lacZ* region with an mRFP expression cassette, containing the pVeg constitutive promoter from Bacillus, an RBS sequence flanked by BamHI and EcoRI restriction sites, the *mRFP1* coding sequence, and a T1 terminator. A mixture of annealed oligonucleotides containing optimized RBS libraries were inserted between the BamHI and EcoRI sites by ligation. The integration vector was integrated into the *amyE* genomic locus of *Bacillus subtilis 168* using the standard protocol and selected on 5 µg/mg chloramphenicol. Integration was verified by sequencing PCR amplicons containing the RBS and *mRFP1* expression cassette.

To construct genomic RBS libraries in *Escherichia coli* EcNR2 (Wang, 2009), 90mer oligonucleotides were designed to have minimal secondary structure at their 5' and 3' ends, and were synthesized with 5’ phosphorothioate modifications and 2' fluro-uracil to improve their allelic replacement efficiencies (Integrated DNA Technologies, Coralville, Iowa). Their concentrations were adjusted to 1 uM in water. The EcNR2 strain was incubated overnight in LB broth with antibiotic (50 µg/ml Ampicillin or chloramphenicol) at 30 ˚C and with 200 RPM orbital shaking. The culture was then diluted to early exponential growth phase (OD_{600}=0.01) in 5 ml SOC, reaching mid-exponential growth phase within 2 to 3 hours. When reaching an OD_{600} of 0.5 to 0.7, the culture was warmed to 42 ˚C for 20 minutes and then placed on ice. 1 mL culture was centrifuged for 30 seconds at >10,000 g and the supernatant was discarded. The

cell pellet was washed twice with chilled water, dissolved in the oligo aqueous solution, and electroporated using an Eppendorf electroporator (model 2510) at 1800 V. The culture was recovered by incubation in pre-warmed SOC at 37 ˚C until reaching an OD_{600} of 0.5 to 0.7. The culture was then used for an additional cycle of mutagenesis, plated on LB agar to obtain isogenic clones, or pelleted to make glycerol stocks. Mutagenesis was verified by sequencing PCR amplicons of the *lacZ* locus.

To combinatorially assemble 3-reporter operons in *Escherichia coli* strain DH10B, PCR amplicons containing Cerulean, mRFP1, and GFPmut3b/vector backbone were amplified from pFTV3 using mixed primers containing optimized degenerate RBS sequences and 40 bp overlap regions. The PCR products were Dpn1 digested, gel purified, and joined together into the pFTV vector using a 3-part CBAR assembly reaction (Gibson et al, 2009), using the existing J23100 constitutive promoter. The library of plasmids was transformed into *E. coli* DH10B and selected on LB plates with 50 µg/ml chloramphenicol. To combinatorially assemble *crtEBI* operons driven by a P_{BAD} promoter, the crtE coding sequence was first sub-cloned into a FTV3-derived vector that replaced the constitutive J23100 promoter with an *araC*-P_{BAD} cassette, followed by PCR amplification of crtE, crtB, and crtI/vector using mixed primers containing optimized degenerate RBS sequences and 40 bp overlap regions. PCR products were joined together using a 3-part CBAR assembly reaction to create a library of plasmids, which was transformed into *E. coli* DH10B, selected on LB plates with 50 µg/ml chloramphenicol. Isolated pathway variants were verified by sequencing. *crtEBI* coding sequences originated from *Rhodobacter sphaeroides* 2.4.1 and were codon-optimized and synthesized by DNA 2.0 (Menlo Park, CA).

Growth and Measurements

To record fluorescence measurements from RBS variants controlling reporter expression, transformed strains and a wild-type DH10B strain were individually incubated overnight at 37 ˚C, 200 RPM in a 96 deep well plate containing 750 µL LB broth and 50 µg/ml chloramphenicol, or 50 µg/ml streptomycin for the DH10B strain. 5 µl of the overnight culture was diluted into 195 µL M9 minimal media supplemented with 0.4 g/L glucose, 50 mg/L leucine, and 10 µg/ml antibiotic in a 96-well micro-titer plate. The plate was incubated in a M1000 spectrophotometer (TECAN) at 37 ˚C until its OD_{600} reached 0.20. Samples were extracted, followed by a 1:20 serial dilution of the culture into a second 96-well micro-titer plate containing fresh M9 minimal media. A third plate was inoculated and cultured in the same way to maintain cultures in the early exponential phase of growth for 24 hours. The fluorescence distribution of 100,000 cells from culture samples was recorded by a LSR-II Fortessa flow cytometer (BD biosciences). Protein fluorescences were determined by taking fluorescence distributions' averages and subtracting DH10B's average auto-fluorescence.

To record fluorescence measurements from 3-reporter operon libraries, 500 colonies were randomly selected and grown individually using LB Miller media with 50 µg/ml chloramphenicol, for 16 hrs at 37 ˚C with 200 RPM orbital shaking, inside a 96 deep-well plate. Cultures were then diluted 1:20 into fresh supplemented LB Miller media within a 96-well micro-titer plate, incubated at at 37 ˚C in a M1000 spectrophotometer (TECAN) until the maximum OD_{600} reached 0.20. The blue, red, and green fluorescence distributions of samples were recorded using flow cytometry, applying a previously calibrated color correction to remove cross-fluorescence. The average blue, red, and green fluorescence is determined by subtracting average DH10B autofluorescence.

To record *lacZ* activities using Miller assays, *E. coli* EcNR2 genome variants containing *lacI* knockouts and *lacZ* RBS mutations were grown overnight at 30 ˚C with 250 RPM orbital shaking in a 96 deep-well plate containing LB Miller and 50 ug/ml chloramphenicol. Cultures were then diluted into fresh supplemented LB Miller media and cultured at 30˚C to an OD_{600} of 0.20. 20 µL of cultures were diluted into 80 µL permeabilization solution and incubated at 30˚C for 30 minutes. 25 µL samples were then transferred into a new microplate to perform Miller assays. 150 µL of ONPG solution was added and absorbances at 420, 550 were recorded by the M1000 for a three hour period. Using this data, Miller units were calculated by finding the average value of (OD_{420} - 1.75 OD_{550}) / OD_{600} during the times when the product synthesis rate was constant.

To measure neurosporene productivities, pathway variants were incubated for 16hrs at 30˚C, 250 RPM orbital shaking in 5 ml culture tubes, then washed with PBS, dissolved in fresh LB miller (50 µg/ml chloramphenicol, and 10mM arabinose), and grown for another 7 hours. Cells were centrifuged (Allegra X15R at 4750 RPM) for 5 minutes, washed with 1 ml ddH2O, and dissolved in 1 ml acetone. The samples were incubated at 55 ˚C for 20 minutes with intermittent vortexing, centrifuged for 5 minutes, and the supernatants transferred to fresh tubes. Absorbance was measured at 470 nm using NanoDrop 2000c spectrophotometer and converted to µg Neurosporene (x 3.43 µg/nm absorbance). The remaining pellet was heated at 60 ˚C for 48 hrs to determine dry cell weight. Neurosporene content was calculated by normalizing Neurosporene production by dry cell weight. Neurosporene productivity was determined by dividing by 7 hours.

To record neurosporene productivity under optimized growth conditions, pathway variants were incubated overnight in 5 ml LB miller, followed by inoculating a 50 mL shake flask culture using 2xM9 media supplemented with 0.4% glucose and 10 mM arabinose. The culture was grown for 10 hours at 37˚C with 300 RPM orbital shaking. The neurosporene productivity was measured using 10 ml of the final culture as stated above.

Models and Computation

The RBS Calculator

The RBS Calculator v1.1 was employed to calculate the ribosome's binding free energy to bacterial mRNA sequences, and to predict the translation initiation rate of a protein coding sequence on a proportional scale that ranges from 0.1 to 100,000 or more. The thermodynamic model uses a 5-term Gibbs free energy model to quantify the strengths of the molecular interactions between the 30S ribosomal pre-initiation complex and the mRNA region surrounding a start codon. The free energy model is:

$$\Delta G_{total} = \Delta G_{mRNA:rRNA} + \Delta G_{spacing} + \Delta G_{start} + \Delta G_{standby} - \Delta G_{mRNA} \quad (1)$$

Using statistical thermodynamics and assuming chemical equilibrium between the pool of free 30S ribosomes and mRNAs inside the cell, the total Gibbs free energy change is related to a protein coding sequence's translation initiation rate, *r*, according to:

$$r \propto \exp(-\beta \Delta G_{total}) \quad (2)$$

This relationship has been previously validated on 132 mRNA sequences where the ΔG_{total} varied from −10 to 16 kcal/mol, resulting in well-predicted translation rates that varied by over 100,000-fold(Salis et al, 2009). The apparent Boltzmann constant, *β*, has been measured as 0.45 ± 0.05 mol/kcal, which was confirmed in a second study (Hao et al, 2011). In practice, we use a proportional constant of 2500 to generate a proportional scale where physiological common translation initiation rates vary between 1 and 100,000 au.

In the initial state, the mRNA exists in a structured conformation, where its free energy of folding is ΔG_{mRNA} (ΔG_{mRNA} is negative). After assembly of the 30S ribosomal subunit, the last nine nucleotides of its 16S rRNA have hybridized to the mRNA while all non-clashing mRNA structures are allowed to fold. The free energy of folding for this mRNA-rRNA complex is $\Delta G_{mRNA:rRNA}$ ($\Delta G_{mRNA:rRNA}$ is negative). mRNA structures that impede 16S rRNA hybridization or overlap with the ribosome footprint remain unfolded in the final state. These Gibbs free energies are calculated using a semi-empirical free energy model of RNA and RNA-RNA interactions(Mathews et al, 1999; Xia et al, 1998) and the minimization algorithms available in the Vienna RNA suite, version 1.8.5(Gruber et al, 2008).

Three additional interactions will alter the translation initiation rate. The $tRNA^{fMET}$ anti-codon loop hybridizes to the start codon (ΔG_{start} is most negative for AUG and GUG). The 30S ribosomal subunit prefers a five nucleotide distance between the 16S rRNA binding site and the start codon; non-optimal distances cause conformational distortion and lead to an energetic binding penalty. This relationship between the ribosome's distortion penalty ($\Delta G_{spacing} > 0$) and nucleotide distance was systematically measured. Finally, the 5' UTR binds to the ribosomal platform with a free energy penalty $\Delta G_{standby}$.

There are key differences between the first version of the RBS Calculator (v1.0)(Salis et al, 2009), and version v1.1 (Salis, 2011). The algorithm's use of free energy minimization was modified to more accurately determine the 16S rRNA binding site and its aligned spacing, particularly on mRNAs with non-canonical Shine-Dalgarno sequences, and to accurately determine the unfolding free energies of mRNA structures located within a protein coding sequence. For the purpose of this work, a ribosome binding site (RBS) sequence is defined as the 35 nucleotides located before the start codon of a protein coding sequence within a mRNA transcript. However, the presence of long, highly structured 5' UTRs can further alter the translation initiation rate of a protein coding sequence by manipulating its $\Delta G_{standby}$. The ribosome's rules for binding to long, highly structured 5' UTRs has been characterized (Espah Borujeni et al), and will be incorporated into a future version of the RBS Calculator (v2.0).

The RBS Library Calculator

The objective of the RBS Library Calculator is to identify the smallest RBS library that uniformly varies a selected protein's expression level across a targeted range to efficiently identify optimal protein expression levels and quantify expression-activity relationships. The RBS Library Calculator designs degenerate ribosome binding site (RBS) sequences that satisfy the following mini-max criteria: first, the RBS sequence variants in the library shall express a targeted protein to maximize coverage, C, of the translation rate space between a user-selected minimum (r_{min}) and maximum rate (r_{max}); second, the number of RBS variants in the library, $N_{variants}$, shall be minimized. The allowable range of translation rates is between 0.10 au and over 5,000,000 au though the feasible minimum and maximum rates will also depend on the selected protein coding sequence. These criteria are quantified by the following objective function:

$$\boldsymbol{F = 10C - 0.02N_{variants}} \tag{3}$$

The coverage of an RBS library is determined by first converting the translation rate space into a $\log_{10}$ scale and discretizing it into equal width bins. For this work, the bin width W is called the search resolution as it ultimately defines how many RBS variants will be present in the optimized RBS library. The total number of bins is determined by the user-selected maximum and minimum translation rates and the search resolution W, while the RBS library coverage C is determined by the ratio between filled bins and total bins, according to the following equations:

$$\boldsymbol{B_{total} = \left\lceil \frac{(r_{max}/r_{min})}{W} \right\rceil \quad C = \frac{B_{filled}}{B_{total}}} \tag{4}$$

For example, there will be a total of 17 bins when using a search resolution W of 0.30 and a translation rate space between 1.0 au to 100,000 au. A bin at position y in translation rate space will be filled when at least one RBS variant in the library has a predicted translation initiation rate that falls within the range $[y / 10^{W},\ y\ 10^{W}]$. An RBS library's coverage is one when all translation rate bins are filled by at least one RBS variant. The objective function F has a maximum value of 1 - 0.02 B_{total}, which is achieved when all bins are filled by a single RBS variant, yielding the most compact RBS library that expresses a protein with uniformly increasing translation rates.

The solution to the RBS Library Calculator optimization problem is a list of near-optimal degenerate ribosome binding site sequences. A degenerate RBS is a 35 nucleotide sequence that uses the 16 letter IUPAC code to indicate whether one or more nucleotides shall be randomly incorporated at a particular sequence position. The alphabet defines the inclusion of either single nucleotides (A, G, C, U/T), double nucleotides (W, S, M, K, Y, B), triple nucleotides (D, H, V), or all four nucleotides (N) in each sequence position. $N_{variants}$ is determined by the number of sequence combinations according to these degeneracies.

Chemical synthesis of degenerate DNA sequences creates a mixture of DNA sequence variants, which are then incorporated into a natural or synthetic genetic system, either plasmid- or chromosomally-encoded. Chemical synthesis of the degenerate DNA oligonucleotides may introduce non-random bias in nucleotide frequency, due to differences in amidite substrate binding affinities. The concentrations of manually mixed precursors can be adjusted to eliminate this bias.

Several properties of the RBS Library Calculator's mini-max optimization problem have influenced the selection of an appropriate optimization algorithm. First, the number of possible degenerate RBS sequences is very large (16^{35}), though many of these sequences will yield the same objective function. Further, the relationship between a degenerate RBS sequence and its library coverage is highly non-linear and discontinuous. The addition of degeneracy to some nucleotide positions will greatly increase library coverage, whereas modifying other nucleotide positions has no effect on coverage. The nucleotide positions that affect the library coverage will typically include portions of the Shine-Dalgarno sequence, but also other positions that modulate the energetics of mRNA structures. The locations of mRNA structures will depend on the selected protein coding sequence, which will significantly influence the optimal degenerate RBS sequence. Consequently, an evolutionary (stochastic) optimization algorithm was chosen to rapidly sample diverse sequence solutions, and use mixing (recombination) to identify nucleotide positions that are most important to maximizing library coverage.

A genetic algorithm is employed to identify an optimal degenerate RBS sequence that maximizes the objective function, *F*. The procedure performs iterative rounds of *in silico* mutation, recombination, and selection on a population of degenerate RBS sequences to generate a new population with improved fitness (**Figure 1B**). First, a mutation operator is defined according to the following frequencies: (i) 40%, two degenerate sequences are recombined at a randomly selected junction; 15%, the degeneracy of a randomly selected nucleotide is increased; 15%, the degeneracy of a randomly selected degenerate nucleotide is decreased; 15%, a non-degenerate nucleotide is mutated to another non-degenerate nucleotide; 10%, the degenerate sequence is not modified (designated elites); or 5%, a new degenerate sequence is randomly generated. Second, one or two degenerate sequences in the population are randomly selected with probabilities proportional to their evaluated objective functions, a randomly selected mutation operator is performed on these degenerate sequences, and the results are carried forward into the new population. This process is repeated until the objective function for the most fit sequence has reached the maximum value, the maximum objective function has not changed for a user-selected number of iterations, or when the total number of iterations has reached a user-selected maximum. The top five degenerate RBS sequences in the population are then returned, including the predicted translation initiation rates for each variant in the RBS library.

The genetic algorithm typically requires 50 to 100 iterations to identify optimal degenerate RBS sequences, starting from a population of randomly generated, non-degenerate RBS sequences. During the optimization procedure, the most common mutational trajectory is the broad expansion of sequence degeneracy towards maximizing coverage of the translation rate space, followed by targeted reduction of degeneracy to eliminate RBS variants with similar translation rates. The number of iterations is substantially reduced when a rationally designed RBS sequence is used as an initial condition, particularly when the selected maximum translation rate is over 10,000 au.

Kinetic Model Formulation, Transformation, and Identification

Mass action kinetics was utilized to formulate an ordinary differential equation model to quantify the rates of production and consumption of the 24 metabolite, free enzyme, and bound enzyme species in the pathway's reaction network. A derivation is found in the **Supplementary Information**. The reaction network includes 10 reversible reactions catalyzed by Idi, IspA, CrtE, CrtB, and CrtI enzymes, including reversible binding of substrate to enzyme and reversible unbinding of product from enzyme (**Supplementary Figure 4**). IspA, CrtE, CrtB, and CrtI catalyze multiple reactions. These reactions convert intracellular isopentenyl diphosphate (IPP) and Dimethylallyl diphosphate (DMAPP) to neurosporeneid. An additional

five mole balances on intracellular enzyme were derived. There are 48 unknown kinetic parameters.

De-dimensionalization of the model was carried out by transforming all metabolite and enzyme concentrations into ratios, compared to the concentrations in a reference pathway variant. For example, the forward v_{f1} and reverse v_{r1} reaction rates for the binding of IPP to *idi* enzyme were multiplied and divided by the reference pathway's concentrations for IPP and free *idi* enzyme, yielding:

$$v_{f1} = \underbrace{\left(k_1 * [IPP]_{ref} * [idi]_{ref}^{total}\right)}_{apparent\ kinetic\ parameter} * \underbrace{\frac{[IPP]}{[IPP]_{ref}}}_{\substack{metabolite \\ concentration \\ ratio}} * \underbrace{\frac{[idi]^{free}}{[idi]_{ref}^{total}}}_{\substack{enzyme \\ concentration \\ ratio}} \qquad v_{r1} = \underbrace{\left(k_{-1} * [CM1]_{ref}\right)}_{apparent\ kinetic\ parameter} * \underbrace{\frac{[CM1]}{[CM1]_{ref}}}_{\substack{enzyme \\ concentration \\ ratio}} \tag{5}$$

As a result, metabolite and enzyme concentration ratios are compared across pathway variants using dimensionless units. Accordingly, the total enzyme concentration ratios for each pathway variant were determined by comparing a pathway variant's translation rates to the reference pathway's translation rates. As an example, the *crtE* concentration ratio is:

$$\underbrace{\frac{[CrtE]^{total}}{[CrtE]_{ref}^{total}}}_{\substack{enzyme \\ concentration \\ ratio}} = \underbrace{\frac{translation\ initiation\ rate\ of\ crtE\ in\ a\ pathway\ variant}{translation\ initiation\ rate\ of\ crtE\ in\ the\ reference\ pathway}}_{\substack{translation\ initiation\ rate \\ ratio}} \tag{6}$$

The choice of the reference pathway variant will alter the apparent kinetic parameter values, but it will not alter the solution to the ODEs; increases in the apparent kinetic parameters are compensated by decreases in the enzyme concentration ratios. The reference pathway (#53) has predicted translation initiation rates of 72268, 20496, and 203462 au for *crtE*, *crtB*, and *crtI*, respectively.

Numerical integration of the transformed kinetic model is carried out using a stiff solver (ode23s, MATLAB) over a 7 hour simulated time period to correspond to experimental conditions. The inputs into the kinetic model are the kinetic parameter values and the total enzyme concentration ratios. The resulting neurosporene production fluxes r_p are related to measured neurosporene productivities by comparison to the reference pathway according to:

$$\underbrace{\frac{r_{p,i}}{r_{p,ref}}}_{\substack{simulated\ production \\ flux\ ratio}} = \underbrace{\frac{predicted\ neurosporene\ productivity\ of\ the\ i^{th}\ pathway\ variant}{measured\ neurosporene\ productivity\ of\ the\ reference\ pathway}}_{\substack{pathway\ productivity \\ ratio}} \tag{7}$$

The reference pathway has a neurosporene productivity of 196 ug/gDCW/hour when grown in LB media (non-optimized growth conditions). Each pathway variant will have a different neurosporene production flux and predicted neurosporene productivity as a result of the

different total enzyme concentrations, controlled by the *crtEBI* translation rates according to Equation 6. The kinetic parameters remain constant for all pathway variants.

Model reduction and identification were carried out to reduce the number of model degrees of freedom and to determine the kinetic parameter values that best reproduced the measured neurosporene productivities for the 73 pathway variants designed using *Search* mode. From the 48 unknown kinetic parameters, 10 non-independent parameters were eliminated, and an additional 5 were constrained using available biochemical data (**Supplementary Information**). A genetic algorithm was employed to identify the model's kinetic parameter values that best predicted the neurosporene productivities of the 72 non-reference pathway. On average, the resulting model predicts the neurosporene productivities to within 32% of the measurements (**Supplementary Figure 5**). We then performed inverse model reduction to determine the 48 kinetic parameter values that define the identified kinetic model (**Supplementary Table 8**). Model identification can be performed on the non-reduced model, though it would result in greater variability in best-fit kinetic parameters, longer optimization convergence times, and a requirement for more characterized pathway variants to achieve the same predictive error.

Acknowledgements

We thank B. Pfleger and J. Torella for valuable discussion; H. Wang and G. Church (Harvard University) for the gift of strains EcNR2 and EcHW2f; and the researchers who use the interactive website for their valuable feedback. This research was supported by the Office of Naval Research (N00014-13-1-0074), an NSF Career Award (CBET-1253641), DARPA CLIO (N66001-12-C-4017), a DARPA Young Faculty Award to H.M.S., and start-up funds provided by the Penn State Institute for the Energy and the Environment. M.G. and M.E. were supported by a NSF Research Experience for Undergraduate students. Computational resources provided by an Amazon AWS Research Grant.

Contributions

I.F. and H.M.S. designed the study, developed the algorithm, analyzed results, and wrote the manuscript. I.F., M.K., J.C., M.E., and M.G. conducted the experiments.

The authors declare competing financial interests: H.M.S. is a founder of De Novo DNA.

References

Ajikumar PK, Xiao W-H, Tyo KEJ, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science (New York, NY)* **330:** 70-74

Alper H, Miyaoku K, Stephanopoulos G (2006) Characterization of lycopene-overproducing E. coli strains in high cell density fermentations. *Applied microbiology and biotechnology* **72:** 968-974

Babiskin AH, Smolke CD (2011) A synthetic library of RNA control modules for predictable tuning of gene expression in yeast. *Molecular systems biology* **7**

Blount BA, Weenink T, Vasylechko S, Ellis T (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PloS one* **7:** e33279

Bonnet J, Subsoontorn P, Endy D (2012) Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proceedings of the National Academy of Sciences* **109:** 8884-8889

Cho SW, Kim S, Kim JM, Kim J-S (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology*

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339:** 819-823

Contador CA, Rizk ML, Asenjo JA, Liao JC (2009) Ensemble modeling for strain development of l-lysine-producing Escherichia coli. *Metabolic Engineering* **11:** 221-233

Du J, Yuan Y, Si T, Lian J, Zhao H (2012) Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic acids research* **40:** e142

Espah Borujeni A, Channarasappa AS, Salis HM Translation Rate is Controlled by Coupled Trade-offs Between Site Accessibility, Selective RNA Unfolding, and Sliding at Upstream Standby Sites. *Nucleic Acids Research* **accepted**

Esvelt KM, Wang HH (2013) Genome-scale engineering for systems and synthetic biology. *Molecular systems biology* **9**

Fell DA (1992) Metabolic control analysis: a survey of its theoretical and experimental development. *Biochemical Journal* **286:** 313

Fendt S-M, Buescher JM, Rudroff F, Picotti P, Zamboni N, Sauer U (2010) Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Molecular systems biology* **6**

Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6:** 343-345

Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. *Molecular systems biology* **7**

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL (2008) The Vienna RNA websuite. *Nucleic acids research* **36:** W70-74

Hao Y, Zhang ZJ, Erickson DW, Huang M, Huang Y, Li J, Hwa T, Shi H (2011) Quantifying the sequence–function relation in gene silencing by bacterial small RNAs. *Proceedings of the National Academy of Sciences* **108:** 12473-12478

Jaschke PR, Saer RG, Noll S, Beatty JT (2011) Modification of the genome of Rhodobacter sphaeroides and construction of synthetic operons. *Methods Enzymol* **497:** 519-538

Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* **31:** 233-239

Kholodenko BN, Westerhoff HV (1993) Metabolic channelling and control of the flux. *FEBS letters* **320:** 71-74

Lee ME, Aswani A, Han AS, Tomlin CJ, Dueber JE (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic acids research*

Li G-W, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484:** 538-541

Lou C, Liu X, Ni M, Huang Y, Huang Q, Huang L, Jiang L, Lu D, Wang M, Liu C (2010) Synthesizing a novel genetic sequential logic circuit: a push-on push-off switch. *Molecular systems biology* **6**

Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Molecular Microbiology* **52:** 1243-1253

Makino T, Skretas G, Kang T-H, Georgiou G (2011) Comprehensive engineering of Escherichia coli for enhanced expression of IgG antibodies. *Metabolic engineering* **13:** 241-251

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* **339:** 823-826

Maresca M, Lin VG, Guo N, Yang Y (2013) Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome research* **23:** 539-546

Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology* **288:** 911-940

Medina C, Camacho EM, Flores A, Mesa-Pereira B, Santero E (2011) Improved expression systems for regulated expression in Salmonella infecting eukaryotic cells. *PloS one* **6:** e23055

Mendez-Perez D, Gunasekaran S, Orler VJ, Pfleger BF (2012) A translation-coupling DNA cassette for monitoring protein translation in Escherichia coli. *Metabolic engineering* **14:** 298-305

Menzella HG (2011) Comparison of two codon optimization strategies to enhance recombinant protein production in Escherichia coli. *Microbial cell factories* **10:** 15

Moon TS, Lou C, Tamsir A, Stanton BC, Voigt Ca (2012) Genetic programs constructed from layered logic gates in single cells. *Nature* **491:** 249-253

Mutalik VK, Guimaraes JC, Cambray G, Lam C, Christoffersen MJ, Mai Q-A, Tran AB, Paull M, Keasling JD, Arkin AP (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature methods* **10:** 354-360

Na D, Lee S, Lee D (2010) Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC systems biology* **4:** 71

Paddon CJ, Westfall PJ, Pitera DJ, Benjamin K, Fisher K, McPhee D, Leavell MD, Tai a, Main a, Eng D, Polichuk DR, Teoh KH, Reed DW, Treynor T, Lenihan J, Fleck M, Bajad S, Dang G, Dengrove D, Diola D et al (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496:** 528-532

Pfleger BF, Pitera DJ, Smolke CD, Keasling JD (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nature biotechnology* **24:** 1027-1032

Plotkin JB, Kudla G (2010) Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12:** 32-42

Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS one* **5:** e9202

Quan J, Saaem I, Tang N, Ma S, Negre N, Gong H, White KP, Tian J (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nature biotechnology* **29:** 449-452

Ravasi P, Peiru S, Gramajo H, Menzella HG (2012) Design and testing of a synthetic biology framework for genetic engineering of Corynebacterium glutamicum. *Microbial cell factories* **11:** 1-11

Salis HM (2011) The ribosome binding site calculator. *Methods Enzymol* **498:** 19-42

Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology* **27:** 946-950

Sandoval NR, Kim JY, Glebes TY, Reeder PJ, Aucoin HR, Warner JR, Gill RT (2012) Strategy for directing combinatorial genome engineering in Escherichia coli. *Proceedings of the National Academy of Sciences* **109:** 10540-10545

Santos CNS, Xiao W, Stephanopoulos G (2012) Rational, combinatorial, and genomic approaches for engineering L-tyrosine production in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **109:** 13538-13543

Sharan SK, Thomason LC, Kuznetsov SG, Court DL (2009) Recombineering: a homologous recombination-based method of genetic engineering. *Nature protocols* **4:** 206-223

Sleight SC, Bartley BA, Lieviant JA, Sauro HM (2010) Designing and engineering evolutionary robust genetic circuits. *Journal of biological engineering* **4:** 12

Torella JP, Boehm CR, Lienert F, Chen J-H, Way JC, Silver PA (2013) Rapid construction of insulated genetic circuits via synthetic sequence-guided isothermal assembly. *Nucleic acids research***:** gkt860

Tran LM, Rizk ML, Liao JC (2008) Ensemble modeling of metabolic networks. *Biophysical journal* **95:** 5606-5617

Tseng H-C, Prather KL (2012) Controlled biosynthesis of odd-chain fuels and chemicals via engineered modular metabolic pathways. *Proceedings of the National Academy of Sciences* **109:** 17925-17930

Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. *Nature reviews Genetics* **11:** 636-646

Wang HH, Isaacs FJ, Carr Pa, Sun ZZ, Xu G, Forest CR, Church GM (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460:** 894-898

Wang HH, Kim H, Cong L, Jeong J, Bang D, Church GM (2012) Genome-scale promoter engineering by coselection MAGE. *Nature methods* **9:** 591-593

Westerhoff HV, Palsson BO (2004) The evolution of molecular biology into systems biology. *Nature biotechnology* **22:** 1249-1252

Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37:** 14719-14735

Xu P, Gu Q, Wang W, Wong L, Bower AG, Collins CH, Koffas MA (2013) Modular optimization of multi-gene pathways for fatty acids production in E. coli. *Nature communications* **4:** 1409

Yadav VG, De Mey M, Giaw Lim C, Kumaran Ajikumar P, Stephanopoulos G (2012) The future of metabolic engineering and synthetic biology: towards a systematic practice. *Metabolic engineering* **14:** 233-241

Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, Boldt J, Khandurina J, Trawick JD, Osterhout RE, Stephen R, Estadilla J, Teisan S, Schreyer HB, Andrae S, Yang TH, Lee SY, Burk MJ, Van Dien S (2011) Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nature chemical biology* **7:** 445-452

Zelcbuch L, Antonovsky N, Bar-Even A, Levin-Karp A, Barenholz U, Dayagi M, Liebermeister W, Flamholz A, Noor E, Amram S (2013) Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic acids research* **41:** e98-e98

Zhang F, Ouellet M, Batth TS, Adams PD, Petzold CJ, Mukhopadhyay A, Keasling JD (2012) Enhancing fatty acid production by the expression of the regulatory transcription factor FadR. *Metabolic engineering* **14:** 653-660

Zhao J, Li Q, Sun T, Zhu X, Xu H, Tang J, Zhang X, Ma Y (2013) Engineering central metabolic modules of Escherichia coli for improving β-carotene production. *Metabolic engineering* **17:** 42-50

Zouridis H, Hatzimanikatis V (2007) A model for protein translation: polysome self-organization leads to maximum protein synthesis rates. *Biophysical journal* **92:** 717-730