bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title: Efficient search, mapping, and optimization of multi-protein genetic systems in diverse 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

Developing predictive models of multi-protein genetic systems to understand and optimize their behavior remains a combinatorial challenge, particularly when measurement throughput is limited. We developed a computational approach to build predictive models and identify optimal sequences and expression levels, while circumventing combinatorial explosion. Maximally informative genetic system variants were first designed by the RBS Library Calculator, an algorithm to design sequences for efficiently searching a multi-protein expression space across a >10,000-fold range with tailored search parameters and well-predicted translation rates. We validated the algorithm's predictions by characterizing 646 genetic system variants, encoded in plasmids and genomes, expressed in six gram-positive and gram-negative bacterial hosts. We then combined the search algorithm with system-level kinetic modeling, requiring the construction and characterization of 73 variants to build a sequence-expression-activity map (SEAMAP) for a biosynthesis pathway. Using model predictions, we designed and characterized 47 additional pathway variants to navigate its activity space, find optimal expression regions with desired activity response curves, and relieve rate-limiting steps in metabolism. Creating sequence-expression-activity maps accelerates the optimization of many-protein systems, and allows previous measurements to quantitatively inform future designs.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Main Text

A microbe's ability to sense its environment, process signals, perform decision-making, and manufacture chemical products is ultimately controlled by its DNA sequence (Du et al, 2013; Du et al, 2012; Lee et al, 2013; Moon et al, 2012; Sandoval et al, 2012; Santos et al, 2012; Tseng & Prather, 2012; Xu et al, 2013; Yim et al, 2011; Zhao et al, 2013). The genetic part sequences controlling protein expression levels directly affect an organism's behavior by modulating binding occupancies, rates of catalysis, and the competition for shared resources. Finding a quantitative relationship between DNA sequence and host behavior has been a central goal towards understanding evolution and adaptation, treating human disease, and the engineering of organisms for biotechnology applications (de Vos et al, 2013; Lerman et al, 2013; Quandt et al, 2013; Strohman, 2002; Wessely et al, 2011).

Recent advances in DNA synthesis, assembly, and mutagenesis have greatly accelerated the construction and modification of large synthetic genetic systems. Combinatorial assembly methods enable the simultaneous introduction and modification of genetic parts to create many genetic system variants (de Raad et al, 2013; Dharmadi et al, 2013; Engler et al, 2009; Gibson et al, 2009; Guye et al, 2013; Sleight & Sauro, 2013; Torella et al, 2014). The development of multiplex genome engineering provides the ability to simultaneously introduce DNA mutations into several genomic loci (Esvelt & Wang, 2013; Wang et al, 2009; Wang et al, 2012). Further, Cas9-dependent and TALE-dependent nicking, cleavage, and mutagenesis has expanded site-directed genome engineering to diverse organisms (Bassett et al, 2013; Chang et al, 2013; Lo et al, 2013; Mali et al, 2013; Miller et al, 2010; Ran et al, 2013). Although techniques are readily available to construct or modify large genetic systems of interest, we currently cannot predict the DNA sequences that will achieve an optimal behavior, particularly when the actions of multiple proteins are responsible for a system's function.

This design challenge could be solved by creating a quantitative link between a genetic system's sequence, protein expression levels, and behavior in order to predict the effects of DNA mutations, map the phenotypic space accessible by natural evolution, and optimize non-natural DNA sequences towards a desired genetic system performance. These relationships, called sequence-expression-activity maps (SEAMAPs), can be formulated by combining sequence-dependent models predicting changes in protein expression levels (Borujeni et al, 2013; Brewster et al, 2012; Johnson et al, 2012; Kilpinen et al, 2013; Rhodius & Mutalik, 2010; Salis, 2011) with system-level models describing genetic system function (Fendt et al, 2010; Hyeon & Thirumalai, 2011; Smallbone et al, 2013; Sneppen et al, 2010; Wessely et al, 2011). Several types of models can be combined, utilizing thermodynamics, kinetics, mass transfer, and

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dynamical systems theory, to describe the multi-scale physical interactions affecting system function (Hyeon & Thirumalai, 2011). Validation of SEAMAP predictions across the range of possible behaviors critically tests our knowledge of the system's interactions, the modeling assumptions, and provides a systematic approach to optimizing genetic system function towards the most desired behavior (Smolke & Silver, 2011; Yadav et al, 2012).

However, the development and parameterization of such sequence-expression-activity models has been stymied by several forms of combinatorial explosion (Alper & Stephanopoulos, 2004; Bailey, 1999; Zelcbuch et al, 2013). First, the number of regulatory genetic part sequences that differentially control protein expression is astronomical; a single 30 nucleotide genetic part has 10^{18} possible sequences, while five of these genetic parts have more possible sequences than atoms in the Universe. Second, while protein expression can be modulated over a >100,000-fold range, there are relatively few genetic part sequences that express an extremely high or low amount of protein; finding these sequences using anecdotal rules or random mutagenesis is unreliable. Third, when multiple proteins synergistically work together to achieve the best possible behavior, the probability of finding genetic part sequences with optimal protein expression levels decreases combinatorially. Fourth, the differences in gene expression machinery across organisms will lead to host-dependent SEAMAPs. New approaches to developing SEAMAPs will be needed to circumvent combinatorial explosion, particularly as larger genetic systems with many proteins are targeted for engineering. Notably, while system-level models describing protein interactions have several unknown parameters, model reduction and rule-based simulations can significantly reduce the number of equations, transitions, co-dependent variables, and insensitive constants (Apgar et al, 2010; Conzelmann et al, 2008; Sneddon et al, 2010; Tran et al, 2008).

In this work, we demonstrate that biophysical modeling and computational design can be combined to efficiently create predictive sequence-expression-activity models for multi-protein genetic systems, while circumventing combinatorial explosion (**Figure 1A**). First, we employ predictive biophysical models to map the relationship between sequence and expression, and to develop an automated search algorithm that rationally designs the smallest number of genetic system variants whose protein expression levels cover the largest portion of the multi-protein expression space. This design process is tailored for each genetic system with the goal of maximizing the observable changes in system behavior across the entire physiologically possible range in a selected host organism. Second, we characterize a small number of genetic system variants and use overall activity measurements to parameterize a reduced, system-level mechanistic model that predicts how changes in protein expression control system function.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The resulting SEAMAP for the genetic system is repeatedly used to correctly design and optimize its function with different targeted behaviors.

We extensively validate our search algorithm's ability to design maximally informative genetic system variants using 646 genetic system variants, encoded on both plasmids and genomes, and characterized in diverse gram-negative and gram-positive bacteria. We demonstrate how combinatorial explosion can be overcome by showing that a small number of genetic system variants can be designed to map a multi-dimensional expression space with tailored search parameters and well-predicted expression levels. We then carry out our sequence-expression-activity mapping approach on an example 3-enzyme carotenoid biosynthesis pathway, characterizing 73 pathway variants to build a SEAMAP that predicts the possible physiological behaviors. We then test the model's predictions by correctly designing 19 additional pathway variants to access intermediate activities (interpolation); 28 additional pathway variants to access higher activities, including optimal pathway variants (extrapolation); and transcriptionally regulated pathway variants with desired activity response curves. We also use SEAMAP predictions to understand the relationship between DNA mutations and the pathway's evolutionary landscape. Finally, we compare three types of system-level models (mechanistic, geometric, and statistical) to analyze their ability to design genetic systems with targeted activities, and to provide re-usable information.

Results

Efficient Searching of the Sequence-Expression Space

The first step to mapping a genetic system's sequence, expression, and activity relationship is to design sequences to efficiently search its expression space. We developed an automated search algorithm, called the RBS Library Calculator, to design the smallest synthetic ribosome binding site (RBS) library that systematically increases a protein's expression level across a selected range on a >10,000-fold proportional scale. To do this, we use a predictive biophysical model to map mRNA sequence to translation initiation rate, employing a previously developed, and recently expanded, statistical thermodynamic model of the bacterial ribosome's interactions. We combine this sequence-dependent model with a genetic algorithm to perform optimization. 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**). When a genetic system is designed, the protein coding sequences' expression levels will be proportional to their translation initiation rates.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

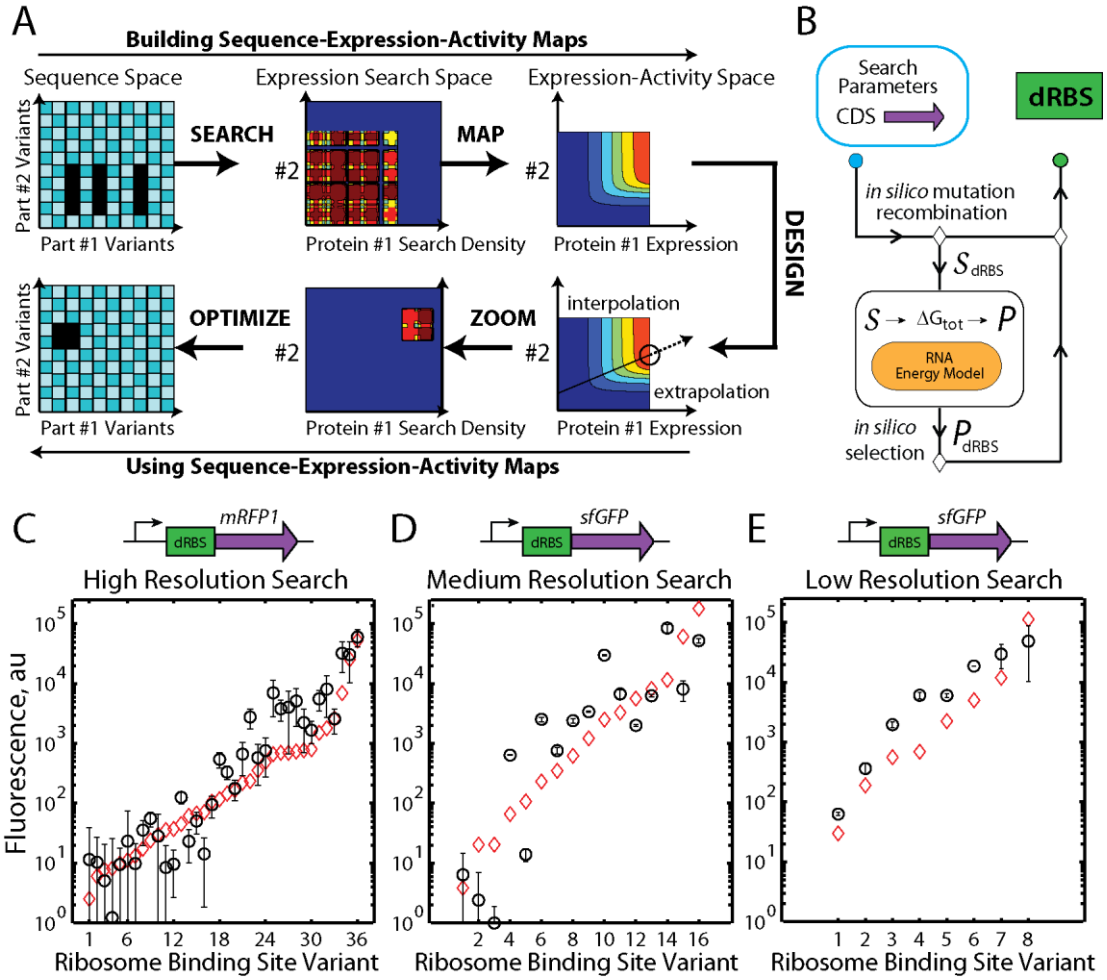

Figure 1: Building and using sequence-expression-activity maps (SEAMAPs). (A) The RBS Library Calculator designs a synthetic RBS library to efficiently search a multi-dimensional protein expression space. A kinetic mechanistic model maps the relationship between protein expression levels and genetic system activity, using a minimal number of measurements for parameterization. The SEAMAP's predictions are repeatedly used for different design objectives. The RBS Library Calculator designs a new RBS library to zoom onto a region of targeted protein expression levels for optimized genetic system performance. (B) The RBS Library Calculator combines a biophysical model of translation with a genetic algorithm to identify the smallest degenerate RBS sequence (dRBS) with maximal search coverage for an inputted protein coding sequence (CDS). The biophysical model calculates the ribosome's binding free energy ΔG_{tot} for an inputted mRNA sequence *S*, which is then related to its translation initiation rate and protein expression level *P*. (C, D, E) Fluorescence measurements show that optimized RBS libraries in *E. coli* DH10B searched 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. All p values < 0.001). Data averages and standard deviations from 6 measurements.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The algorithm has several modes to navigate multi-protein expression spaces. These modes enable one to control the search range, search resolution, and sequence design constraints according to the application and design objectives. In *Search* mode, a synthetic RBS library is optimized to cover the widest possible expression space using a desired resolution. In *Genome Editing* mode, the synthetic RBS library is designed towards introducing the fewest, consecutive mutations into the genome. Finally, in *Zoom* mode, the translation rate range is narrowed, and the search resolution is increased, to design an RBS library to target optimal expression levels.

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. 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 to achieve steady-state conditions. 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 84,000-fold range 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 (778-fold dynamic range) with a high 99% search coverage (**Figure 1E**).

The biophysical model of bacterial translation accurately predicted the translation initiation rates from these 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 with an apparent β of 0.45 mol/kcal. 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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that contains several outliers at low expression levels (average $\Delta\Delta G_{total}$ = 3.71 kcal/mol, R^2 = 0.72).

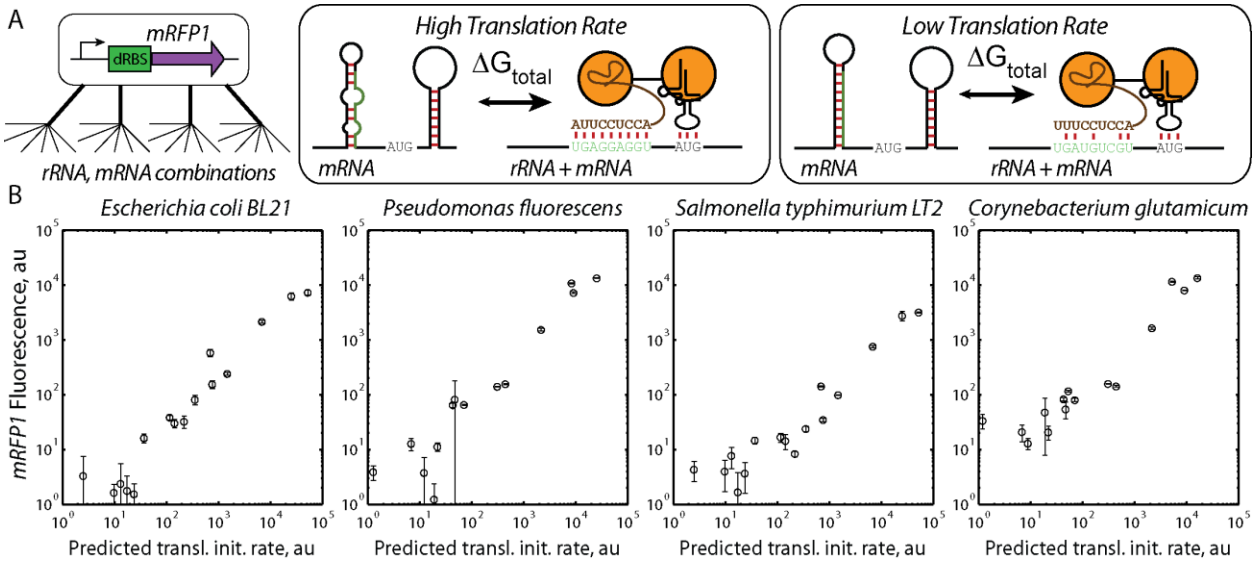

Figure 2: Searching expression spaces in diverse bacterial hosts. (A) Differences in bacterial 16S rRNA sequences lead to different mRNA translation rates. To overcome combinatorial explosion, the biophysical model predicts how changes in mRNA and rRNA sequences control translation rate. (B) A 14- to 16-variant optimized RBS library controlling *mRFP1* expression was characterized in *E. coli* BL21, *P. fluorescens*, *S. typhimurium* LT2, and *C. glutamicum*. Differences in 5' UTR and 3' 16S rRNA sequences are shown. (B) The biophysical model accurately predicted translation initiation rates across a >1000-fold range in *E. coli* BL21 ($\Delta\Delta G_{total}$ is 1.18 kcal/mol, R^2 is 0.93), *P. fluorescens* ($\Delta\Delta G_{total}$ is 1.63 kcal/mol, R^2 is 0.90), *S. typhimurium* LT2 ($\Delta\Delta G_{total}$ is 1.83 kcal/mol, R^2 is 0.89), and *C. glutamicum* ($\Delta\Delta G_{total}$ is 1.81 kcal/mol, R^2 is 0.88). All p values < 10^{-6}. Data averages and standard deviations from 3 measurements.

Navigation of Expression Spaces in Diverse Bacterial Species

The ribosome's interactions with mRNA can change between organisms, leading to varying translation rates. By accounting for known differences, we investigated whether the biophysical model can accurately predict translation rates in diverse bacterial hosts. Using the model to predict expression differences between organisms would enable the development of host-independent SEAMAPs, and the engineering of genetic systems in one organism for their eventual use in another. We selected four bacterial hosts currently used for biotechnology applications (**Supplementary Table 2**): *E. coli* BL21 for over-expression of recombinant proteins; *P. fluorescens* for production of biopolymers and soil decontamination; *S. typhimurium* LT2 for secretion of large proteins, including spider silk; and *C. glutamicum* for production of enzymes and amino acids (Becker et al, 2011; Monti et al, 2005; Widmaier et al, 2009). To predict translation initiation rates, we included promoter-dependent upstream sequences in the 5' UTR and selected the appropriate 3' 16S rRNA sequence for each host; 5'- ACCUCCUUU-3' for the

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

gram-positive *C. glutamicum* and 5'-ACCUCCUUA-3' for the remaining gram-negative species (**Figure 2A**).

We designed a 16-variant optimized RBS library to vary mRFP1 expression across a 14,000-fold range, introduced expression cassettes into broad host vectors with host-specific promoters, and measured fluorescence during long-time cultures in host-specific media (**Methods**). Overall, fluorescence measurements varied between 1051- and 10900-fold, depending on the host, and show that relative translation initiation rates were correctly predicted to within 2.1-fold (average error $\Delta\Delta G_{total}$ of 1.61 kcal/mol, Pearson R^2 is 0.89) (**Figure 2B**). Translation rate predictions were more accurate in gram-negative *E. coli* BL21 ($\Delta\Delta G_{total}$ of 1.18 kcal/mol, Pearson R^2 is 0.93), compared to gram-positive *C. glutamicum* ($\Delta\Delta G_{total}$ of 1.81 kcal/mol, Pearson R^2 is 0.88), though the difference is a single outlier (**Supplementary Table 1**). Interestingly, the apparent Boltzmann constant used to convert calculated binding free energies into predicted translation rates did not significantly vary between bacterial hosts (apparent β was 0.42 ± 0.02 mol/kcal). Consequently, these observations suggest that the free energies of *in vivo* RNA-RNA interactions remain the same regardless of the host organism, including the effects of molecular crowding on binding events (Tan et al, 2013).

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

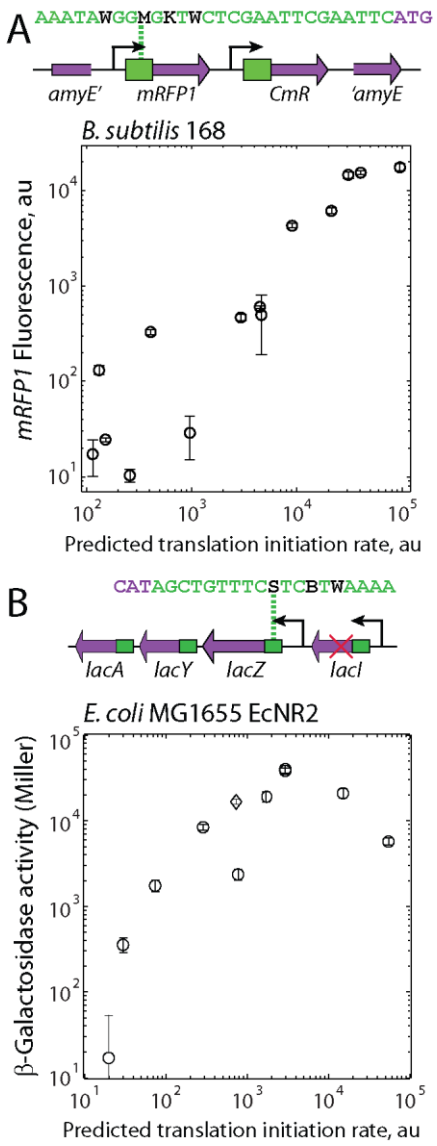

Figure 3: Searching genomic expression spaces. (A) Two RBS libraries were optimized to control the expression of a genomic single-copy of *mRFP1*, incorporated into the *amyE* locus of *B. subtilis*. Fluorescence measurements from 14 clones were compared to their predicted translation initiation rates (Pearson R^2 is 0.81, p value is $2x10^{-6}$). 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, p value is 0.02). The expression space was searched with 84% coverage. Data averages and standard deviations from 4 measurements.

Efficient search 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). We developed the *Genome*

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Editing mode to identify the minimal number of adjacent genomic RBS mutations that systematically 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 used to perform genome mutagenesis. We then evaluated our ability to search expression spaces by using optimized RBS libraries to modify the genomes of gram-positive and gram-negative bacteria.

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. Translation rate predictions use 5'-ACCUCCUUU-3' 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 3A**).

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 3B**). We first conducted three rounds of MAGE mutagenesis to introduce an in-frame stop codon into the *lacI* repressor coding sequence (**Supplementary Tables 3**). Using the algorithm's *Genome Editing* mode, we then designed a 12-variant degenerate oligonucleotide to target the *lacZ* RBS sequence and uniformly increase its translation initiation rate from 20 to 55,000 au. 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 3B**). 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).

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, suggesting that there is a critical point where translation initiation may no longer be the rate-limiting step in protein expression, potentially due to ribosomal pausing during translation elongation, or protein misfolding. Significant changes in specific growth did not occur (**Supplementary Table 1**). To further increase expression, one could replace the existing, natural genes in the genome with newly

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

designed protein coding sequences optimized for maximum expression control, which motivates the design of synthetic genomes (Lajoie et al, 2013).

Efficient search in multi-dimensional expression spaces

Complex genetic systems express multiple proteins to carry out their function. Building multi-protein SEAMAPs are particularly difficult, as it requires searching a larger, multi-dimensional expression space. We next evaluated *Search* mode's ability to efficiently explore a 3-dimensional expression space, compared to several types of randomly generated RBS libraries. Optimized RBS libraries were designed to encode 8 RBS variants with predicted translation initiation rates across a 5000-fold range (**Supplementary Table 4**). They 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. Separately, we constructed random RBS libraries by selecting six nucleotides of the Shine-Dalgarno sequence and randomly incorporating all possible choices to create 4096-variants with widely different predicted translation initiation rates (**Supplementary Table 5**). We employ combinatorial DNA assembly to construct bacterial operon variants encoding *cfp*, *mRFP1*, and *gfpmut3b* reporter proteins (Gibson et al, 2009), generating either 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, 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. For each case, 500 strains with operon variants were randomly selected, individually cultured, and their CFP, mRFP1, and GFPmut3b fluorescences were quantified by 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 well with the computationally predicted search coverage of 14.7% using Monte Carlo sampling of the predicted translation initiation rates (**Supplementary Methods**). Using the same method, 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. Overall, 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**).

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

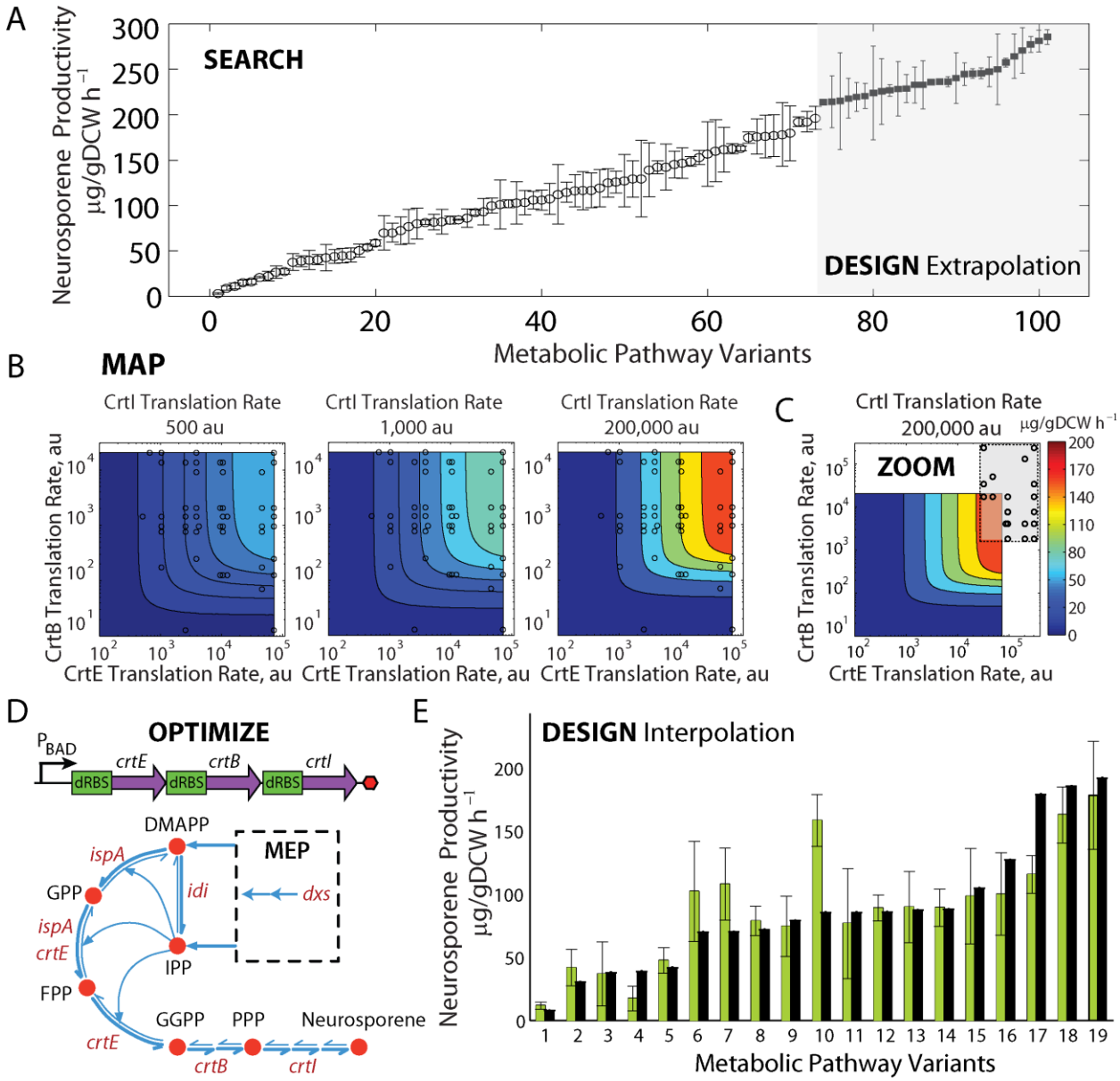

Figure 4: Sequence-expression-activity mapping of a multi-enzyme pathway. (A) Characterization of two libraries of neurosporene biosynthesis pathway variants, using the RBS Library Calculator in *Search* mode (left) or in *Zoom* mode (right). Averages and standard deviations from at least 3 measurements 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) To design pathways with higher activities, a translation rate region (gray box) is targeted using the RBS Library Calculator in *Zoom* mode. Translation rate predictions from selected pathway variants are shown (circles). (D) A schematic of the bacterial operon encoding *crtEBI*, and the corresponding reactions, genes, and metabolites in the biosynthesis pathway. Cofactors are not shown. (E) To evaluate the design of pathways with intermediate activities, 19 additional crtEBI pathway variants were characterized, and the predicted neurosporene productivities (black bars) were compared to the measured productivities (green bars). Data averages and standard deviations from 2 measurements.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mapping the Sequence-Expression-Activity Space of a Multi-enzyme Pathway

To demonstrate our approach, we created a SEAMAP for a multi-enzyme biosynthesis pathway, and then repeatedly used it to optimize the pathway's sequences and expression levels for different design objectives. The RBS Library Calculator in *Search* mode was employed to systematically vary *crtEBI* enzyme expression levels originating from a carotenoid biosynthesis pathway in *Rhodobacter 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 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 remained the rate-limiting step throughout the mapped space.

To formulate an expression-activity relationship for the pathway, we developed a mechanistic, kinetic model to describe the system-level behavior (**Methods**). We listed the 24 elementary chemical reactions that are responsible for enzymatic conversion of isoprenoid precursors (DMAPP and IPP) to neurosporene. All reactions are reversible, including enzymes' binding to substrates and the release of products (**Figure 4D** and **Supplementary Figure 4**). We developed a kinetic model of the reaction network, deriving a system of differential equations with 48 unknown parameters. Mole balances on each enzyme and flux constraints reduced the system of equations to having 33 unknown parameters (**Supplementary Methods**). We then 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 convert simulated reaction fluxes to measurable productivities. Changing a pathway variant's translation rates proportionally control the simulated enzyme concentrations, which alters the pathway's predicted neurosporene productivity. Finally, we employed model identification to determine a unique set of kinetic model parameters that reproduced the measured neurosporene productivities for the 72 non-reference pathway

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variants, across ten independent and randomly initialized optimization runs (**Supplementary Table 8**). The resulting kinetic model maps *crtEBI* translation rates to neurosporene productivities across a 10,000-fold, 3-dimensional translation rate space (**Figure 4B**).

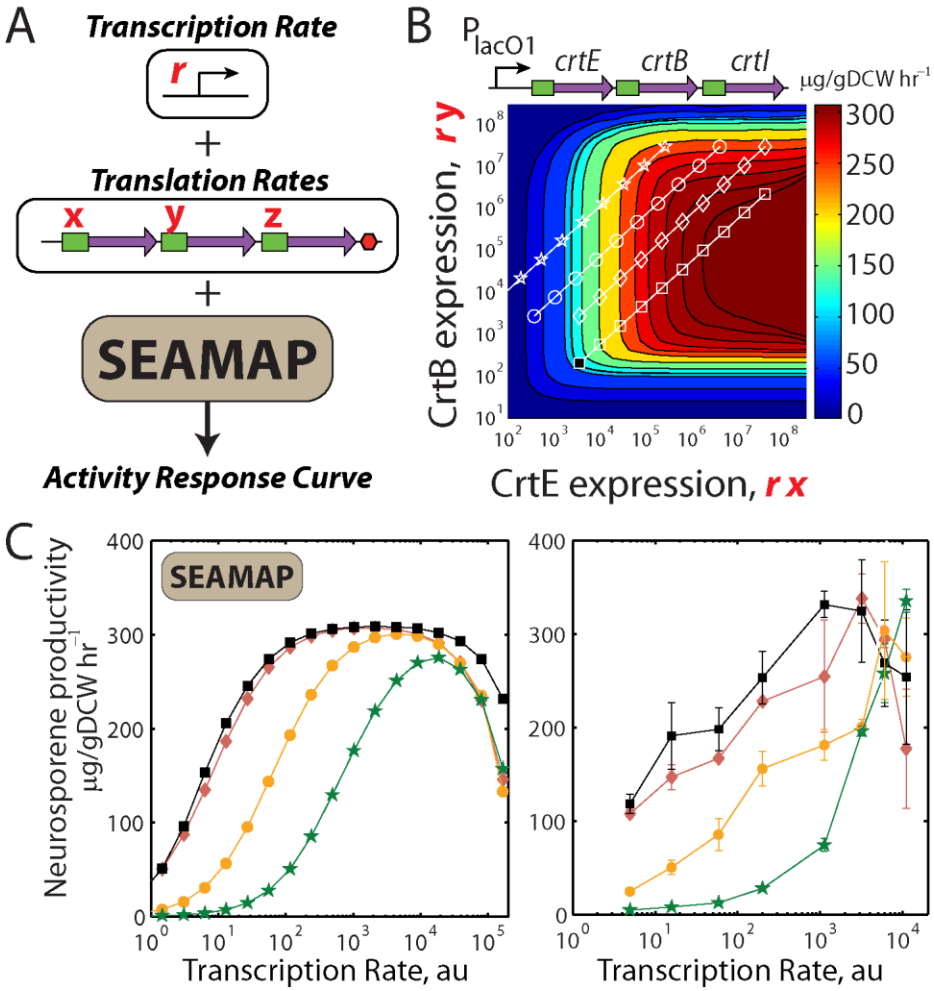

Figure 5: Using SEAMAPs to design multi-enzyme pathways with desired activity response curves. (A) A promoter's transcription rate *r* and the *crtEBI* translation rates (*x*, *y*, *z*) are inputted into SEAMAP's kinetic model to determine the pathway's productivity. (B) A slice of the CrtEB expression-activity space is shown, where CrtI expression is 200,000 au. The effects of transcriptional regulation for four pathway variants are shown as diagonal lines at their respective translation rates. The productivity of a reference pathway variant in one condition (black square) was characterized to determine its location in expression-activity space, which provides orientation for all other locations. (C, left) The effect of the P_{lacO1} promoter's transcription rate on the pathway variants' productivities are calculated. The location of the global maxima depends on the promoter's transcription rate and the mRNA translation rates. (C, right) The productivities of the four pathway variants are measured as transcription rate is increased via IPTG induction. Changes in translation rate cause the global maxima to appear at lower transcription rates, consistent with model calculations. Data averages and standard deviations from 2 measurements.

Design and Optimization of Multi-enzyme Pathways using SEAMAPs

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We tested the SEAMAP's predictions by using it to design *crtEBI* pathway variants according to different design objectives, beginning with pathways that exhibit intermediate activities (interpolation). 19 additional pathway variants were characterized, and the pathway's predicted productivities were compared to measured productivities. The biophysical model predicts the *crtEBI* translation rates from sequenced RBSs, and the kinetic model uses these translation rates to calculate each pathway 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.

Second, we tested the SEAMAP's ability to design improved pathways by identifying an expression region with higher activities beyond the existing observations (extrapolation), and employing the RBS Library Calculator in *Zoom* mode to target this region. We designed 8-variant RBS libraries with translation rate ranges of 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**). Importantly, the SEAMAP predicts sequence-dependent changes to pathway productivity, but not the effects of media formulation or growth conditions. For example, the best pathway variant's productivity was further increased to 441 µg/gDCW/hr when optimized media and aeration conditions were employed (Alper et al, 2006) (**Supplementary Figure 6**).

Third, we tested the SEAMAP's ability to predict a pathway's activity response curve when utilizing a tunable promoter to vary the *crtEBI* operon's transcription rate (**Figure 5A**). Changes in the transcription rate will proportionally vary all enzyme expression levels in the *crtEBI* operon, causing diagonal shifts in expression space that will lead to productivity changes. We modified the most productive *crtEBI* pathway variant, replacing its P_{BAD} promoter with the P_{lacO1} promoter, and measured a productivity of 119 µg/gDCW/hr in supplemented M9 media without IPTG addition. This measurement is used as a reference point, identifying where the modified pathway variant is located in the sequence-expression-activity space (**Figure 5B**, black square). The same promoter replacement was performed for three additional pathway variants with distinctly different *crtEBI* translation rates (**Supplementary Table 12**). Under these growth conditions, the change in promoter resulted in an overall increase in enzyme expression levels.

Model calculations are then combined to show how changes in the promoter's transcription rate and the pathway variants' translation rates will affect their productivities (**Figure 5B**). We use a model to first relate IPTG concentration to P_{lacO1} transcription rate (**Supplementary Figure**

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7), followed by multiplication with the translation rates to determine enzyme expression levels. These enzyme expression levels are substituted into the SEAMAP to predict the pathways' productivities. According to the model, the pathway variants' productivities will increase with promoter induction (**Figure 5C**, left), though variants with optimal translation rates will have higher productivities, and at lower transcription rates. Notably, the model predicts that excess enzyme expression will lower a pathway's productivity, due to sequestration of intermediate metabolites as enzyme-substrate complex.

We then characterized the four pathway variants' productivities with increasing IPTG induction (**Supplementary Table 12**). Though the pathway variants were expressed by the same promoter, their activity responses varied greatly and were highly consistent with model calculations (**Figure 5C**, right). Pathway variants with optimal translation rates had higher productivities, and achieved maximum activity at a lower transcription rate. However, additional increases in transcription lowered their productivities, due to excess enzyme expression levels. The SEAMAP clearly shows how changing the operon's transcription and translation rates can exhibit this non-linear behavior. Consequently, one can use the SEAMAP to guide the selection of a regulated promoter to dynamically control a pathway's. Regulated promoters can often serve as sensors for cellular stress, and they may be used to implement feedback control over a pathway's enzyme expression levels to maintain maximal activities. The use of dynamic regulation has been shown to significantly improve a pathway's productivity (Dahl et al, 2013; Zhang et al, 2012).

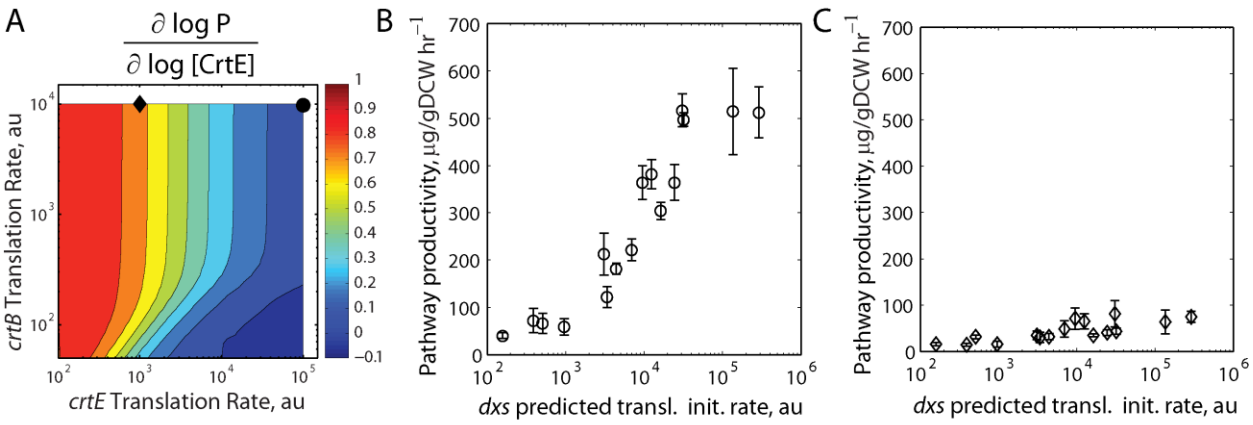

Figure 6: Increasing precursor biosynthesis for optimally balanced vs. imbalanced pathways. (A) The relationship between *crtEB* translation rates and CrtE's flux control coefficient is calculated using SEAMAP predictions. The *crtI* translation rate is 100,000 au. According to their FCCs, increasing precursor biosynthesis is predicted to improve the optimally balanced pathway variant (black circle) more than the imbalanced pathway variant (black diamond). An optimized RBS library is integrated to control genomic *dxs* translation initiation rate and systematically vary precursor biosynthesis, followed by productivity measurements using either (B) an optimally balanced pathway or (C) an imbalanced pathway variant. Predicted *crtEBI* translation initiation rates are (305000 au, 17120 au, 886364 au) for the optimally balanced pathway variant, and (1046 au, 20496 au, 200300 au) for the imbalanced pathway variant.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Expanding Search for Optimally Balanced Pathways

Analysis of the *crtEBI* pathway's SEAMAP reveals why metabolic optimization efforts have been generally laborious. First, each enzyme has the potential to be a rate-limiting step in the pathway (Bailey, 1999). Distributed control over the pathway's flux requires that all enzyme expression levels must be tuned to 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, although pathway optimization efforts aim to continually increase pathway productivities, it remains unclear when an engineered pathway has reached its maximum productivity, and thus pathway engineering efforts have no criteria for stopping. Optimization continues until a better variant cannot be found. Next, we use SEAMAP predictions to calculate a pathway variant's optimality, and to determine when a pathway variant has become optimally balanced. We then demonstrate that reaching this optimality condition is the appropriate time to redirect metabolic engineering efforts and increase precursor biosynthesis rates.

We define pathway optimality using flux control coefficients (FCCs) (Fell, 1992; Kholodenko & Westerhoff, 1993) and use the SEAMAP predictions to calculate the FCCs for the *crtEBI* pathway. FCCs quantify how differential fold-changes in enzyme expression control a pathway's overall productivity, and vary depending on the enzymes' expression levels (**Figure 6A; Supplementary Figure 8**). High FCC regions indicate where increasing an enzyme's expression will increase pathway's productivity, while low FCC regions show where increasing expression does not lead to a significant improvement in productivity. Negative FCC regions show where excess enzyme expression causes metabolite sequestration or growth toxicity.

A pathway is *balanced* when its enzymes' FCCs are equal; differential fold-increases in enzyme expression all have the same effect on pathway activity. Further, a pathway is *optimally balanced* when its enzymes' FCCs are zero at the global maxima in activity space; increasing enzyme expression has a minimal, or negative, impact on pathway productivity. Once a pathway is optimally balanced, it has shifted activity control over to the upstream metabolic module responsible for precursor biosynthesis.

We next investigated whether a pathway variant's FCCs provide the stopping criterion that indicates the need for further metabolic engineering of upstream pathways. We selected an optimally balanced pathway variant where all of its enzymes have approximately zero FCCs, and an imbalanced pathway variant where its enzymes have very different FCCs (0.65 for CrtE, near 0 for CrtB and CrtI) (**Figure 6A**). We then employed the RBS Library Calculator's *Genome Editing* mode to optimize a 16-variant RBS library controlling genomic *dxs* expression, the enzyme that controls the first committed step to isoprenoid biosynthesis. Using co-selection MAGE

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mutagenesis (Wang et al, 2012), 16 genome variants were constructed and verified. The RBS library varied *dxs* translation from 110 to 291000 au. The increase in *dxs* expression significantly improved the optimally balanced pathway's productivity up to 517 ug/gDCW/hr (**Figure 6B; Supplementary Table 14**), while only increasing the imbalanced pathway's productivity up to 81 ug/gDCW/hr (**Figure 6C**). Thus, a pathway variant's flux control coefficients provided the quantitative criteria for indicating when to cease pathway engineering efforts and redirect towards improving precursor biosynthesis rates. These results suggest an iterative metabolic engineering strategy where upstream pathways are additionally optimized, applying the optimality criteria in a successive fashion.

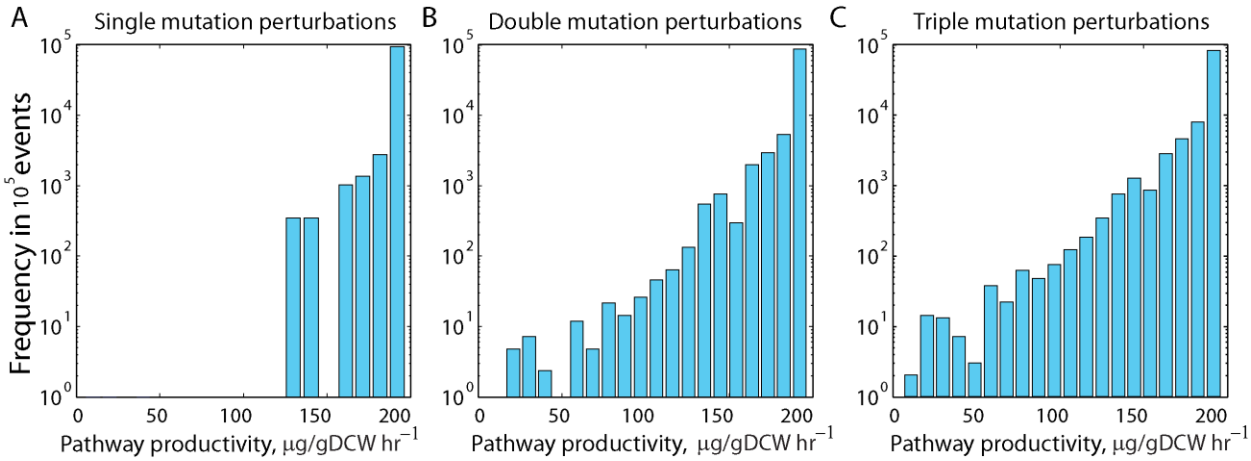

Figure 7: Predicting the evolutionary landscape of a multi-enzyme pathway. Histograms show that random mutations will more likely decrease a pathway's productivity. Either (A) one, (B) two, or (C) three nucleotide mutations are randomly introduced into the 35 nucleotide-long ribosome binding site sequences of *crtEBI*. Changes in enzyme expression levels and pathway productivities are predicted using the SEAMAP.

Discussion

A key challenge to successfully engineering cellular organisms has been the combinatorial vastness of their genetic instruction space, and the complex relationship between genotype and phenotype. We present a new approach to overcoming this design challenge by combining a biophysical model of gene expression with a systems-level mechanistic model to quantitatively connect a genetic system's sequence, protein expression levels, and behavior. We illustrate how to efficiently build sequence-expression-activity maps (SEAMAPs), performing the fewest number of characterization experiments, by using an automated search algorithm to uniformly explore a multi-dimensional expression space (**Figure 1**). Both the biophysical model and automated search algorithm are highly versatile; they can be used in diverse gram-negative and gram-positive bacteria, and to modify both plasmids and genomes (**Figures 2 and 3**). Using the search algorithm, we built a SEAMAP for a multi-enzyme pathway and demonstrated how it can be used to design pathway variants with targeted productivities

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**Figure 4**) and tailored activity response curves (**Figure 5**), while quantitatively guiding further metabolic engineering efforts to increase precursor biosynthesis rates (**Figure 6**). Altogether, creating a SEAMAP for a genetic system provides a coherent and predictive model that can be repeatedly used to optimize non-natural sequences and achieve complex design objectives.

More broadly, the study of evolutionary dynamics and different approaches to genetic optimization can both be understood as time-iterated DNA mutations to navigate an organism's sequence-expression-activity space. Evolution acts on a slower time-scale, and only selects for population members whose activities have improved their overall fitness; in contrast, optimizing genetic systems can be conducted independent of fitness evaluations, and can be directed towards rare DNA sequences. Both processes operate on the same sequence-expression-activity landscape. Using the multi-enzyme pathway's SEAMAP, we show how evolution could shape pathway productivity, due to random mutation. Starting from the optimally balanced pathway shown in **Figure 6**, the effects of single, double, and triple RBS mutations are calculated, showing that pathway productivity decreases in almost all cases (**Figure 7**). From an evolutionary perspective, if the multi-enzyme pathway is essential to cell growth, then these mutations will never proliferate in the population. However, for pathways involved in manufacturing chemical products, pathway productivity is more likely to be inversely coupled to cell growth, which will lead to mutation enrichment. Thus, building SEAMAPs allows one to visualize a genetic system's evolutionary landscape, and to optimize genetic system sequences to become insensitive to evolutionary dynamics.

Overall, our computational design approach combines principles from both Systems and Synthetic Biology to build predictive models (Kitano, 2002). We formulate models using physical principles that can be re-used across different systems and scales (Hyeon & Thirumalai, 2011), and their unknown parameters are parameterized using rationally designed genetic system variants with synthesized sequences. The number of unknown parameters can scale proportionally with the number of proteins in a genetic system, particularly in biosynthesis pathways and engineered genetic circuits where proteins typically interact with a small number of partners. Specifically, modeling each reversible reaction in a multi-enzyme pathway requires two kinetic parameters, and modeling a transcription factor's binding occupancy in a genetic circuit requires one thermodynamic parameter. Our search algorithm greatly improves the efficiency of model building by designing the minimal number of genetic system variants that maximally span the multi-dimensional expression-activity space, eliminating redundant measurements and increasing the information content of data-sets. As a result, the parameterized model can accurately represent the genetic system's non-linear behavior over its sequence-expression-activity space, while reducing characterization effort. These aspects of our approach become more important as larger genetic systems are modified and characterized,

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for example, using chip-based oligonucleotide synthesis and next-generation sequencing (Goodman et al, 2013; Kosuri et al, 2013).

When mechanistic information is unavailable, there are black-box approaches to building expression-activity relationships that utilize geometry, informatics, and statistics. Using our pathway variant data-set as an example, we illustrate the advantages and limitations of two alternative non-mechanistic system-level models. First, we use computational geometry to decompose the four-dimensional expression-activity space into Voronoi polygons, and to calculate unknown activities using linear interpolation between adjacent Voronoi cells (**Supplementary Information**). Using this approach, a geometric SEAMAP could use intermediate enzyme expression levels to predict pathway activities with 15% error (interpolation) (**Supplementary Figure 9, Supplementary Figure 10**). However, to design pathway variants with higher enzyme expression levels and activities (extrapolation), we would need to simplify the expression-activity relationship outside the model's convex hull, for example, by assuming linearity and independence. Based on the mechanistic modeling and FCC calculations, we know such assumptions are incorrect; the expression-activity relationship is non-linear and all enzyme expression levels co-dependently control pathway flux. Consequently, geometric models can be used to formulate accurate expression-activity models, although their ability to design genetic systems outside the previously characterized expression space is limited.

Second, we employ a statistical linear regression model to relate expression to activity, using an Exterior Derivative Estimator to determine best-fit coefficients (Aswani et al, 2011) that is similar to a recent approach (Lee et al, 2013). This statistical model assigns RBS variants as categories that are present or absent for each gene, and finds coefficients that linearly relate individual RBS variants' effects on expression to measured pathway productivities. The linear regression model could predict the pathway variants' activities with a 46% error (**Supplementary Information**). Using a previous approach (Lee et al, 2013), when the productivity data is log-transformed, a statistical model with the same training size predicted the pathways' activities with a 10% error. To design new pathway variants, the category-based statistical model predicts pathway productivities when utilizing RBS variants that were characterized in the existing training set. However, it cannot predict pathway productivities when utilizing newly designed RBS variants, which creates gaps in the expression-activity relationship. This becomes important when, for example, the optimal expression levels to achieve a desired design objective fall within a gap in the characterized expression space, or are located outside the initially selected expression range; a newly designed RBS sequence would be needed to achieve optimality.

Regardless of the modeling approach, it is important to critically test whether parameterization using randomly generated expression-activity data also leads to an accurate representation. A

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

well-defined model should not equally represent both real and random data. We carried out a test of this null hypothesis and found that the kinetic, computational geometry, and linear regression models all had high prediction errors (70% to 110%), except when log-transforming the random productivity data (15%). The variable transformation allowed the random data to fit the statistical model with similar error, compared to the real data (**Supplementary Figure 11**). These comparisons provide an important baseline to error measurements.

Our approach to designing genetic systems also departs from recent efforts to characterize genetic part toolboxes to control expression (Mutalik et al, 2013). A part-centric approach to statically tuning expression requires a tremendous amount of characterization to ensure that parts are modular, orthogonal, insensitive to surrounding sequence changes, and similarly functional in diverse hosts (Nielsen et al, 2013). Further, to be useful, a part-centric toolbox must both be wide and deep. Many genetic parts with similar expression levels must be available, together with many genetic parts that span the entire expression level range. Genetic parts must not have long repetitive sequences to minimize rates of homologous recombination and evolutionary navigation of the sequence-expression-activity space (Lovett, 2004; Sleight et al, 2010). By combining sequence-dependent biophysical models with optimization, we can generate an unlimited number of non-repetitive genetic parts that span the entire expression range, and in diverse bacterial hosts. In another distinction, a toolbox is a static list of DNA sequences, and cannot incorporate additional design criteria *ex post facto* without additional characterization to ensure similar function. However, as demonstrated, our validated model can be repeatedly re-used with different design objectives to optimize a genetic system's sequence towards a desired behavior.

As Synthetic Biology matures into an engineering discipline, additional effort will be needed to ensure continuity between design approaches, where numbers have a physical meaning and the origins of design rules have a molecular basis. In that regard, our computational design approach is seamlessly augmented by an improved understanding of translation initiation (Borujeni et al, 2013), and the development of new mechanistic models that use sequence information to predict changes in DNA bending and looping, transcription factor and nucleosome binding, transcription initiation and elongation rate, transcriptional termination efficiency, RNA folding and stability, polarity, translation elongation and coupling, regulation by cis- and trans- RNAs, and other interactions that affect protein expression levels (Brewster et al, 2012; Chen et al, 2013; Geggier & Vologodskii, 2010; Johnson et al, 2012; Kilpinen et al, 2013; Rhodius & Mutalik, 2010; Rodrigo et al, 2013). The integration of models, to relate sequence to expression and expression to activity, should be a central goal of Synthetic Biology as it will expand our ability to navigate a genetic system's behavior space, and optimize every nucleotide to achieve a desired sensing, signaling, and metabolic performance.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A software implementation of the RBS Library Calculator is available at http://www.salis.psu.edu/software, online since 2011. As of February 2014, 350 unaffiliated researchers have designed 3945 optimized RBS libraries for diverse biotechnology applications.

Materials and Methods

Strains and Plasmid Construction

All strains and plasmids are listed in **Supplementary Table 2**.

To construct plasmid-based RBS libraries in *Escherichia coli* strain DH10B, protein coding 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. Selected plasmids from the pIF1 series were transformed into *E. coli* BL21 and *S. typhimurium* for expression characterization. mRFP1 expression cassettes from the same plasmids were sub-cloned into a modified pSEVA351 vector, replacing J23100 with a Ptac promoter (Genbank accession JX560335, CmR, OriT replication origin), and transformed into *P. fluorescens* and *C. glutamicum* B-2784 by electroporation.

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 was inserted between the BamHI and EcoRI sites by ligation, and the constructs were verified by sequencing. The integration vector was integrated into the *amyE* genomic locus of *Bacillus subtilis 168* using the standard protocol, selected on 5 µg/mL chloramphenicol, and the integration verified by iodide starch plate assay.

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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 vary genomic *dxs* expression, co-selection MAGE was performed on an *E. coli* EcNR2 strain whose *lacZ* contains two early stop codons, performing 12 mutagenesis cycles using 1 µM of an oligonucleotide mixture to introduce the *dxs* RBS library and 10 nM of an oligonucleotide to restore *lacZ* expression (**Supplementary Table 2**). Culturing and selection took place over a 36 hour growth period in M9 minimal media supplemented with 0.4% lactose at 30 ˚C and shaking at 250 RPM. 40 colonies were selected, and 16 unique RBS variants were verified by sequencing PCR amplicons.

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 agar 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

Escherichia coli strains and *Pseudomonas fluorescens* were cultured in LB broth Miller (10g tryptone, 5g yeast extract, 10g NaCl) or M9 minimal media (6g Na_2HPO_4, 3g KH_2PO_4, 1g NH_4Cl, 0.5g NaCl, 0.24g $MgSO_4$, 0.011g $CaCl_2$), as indicated. *Salmonella typhimurium* LT2 and

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Corynebacterium glutamicum B-2784 were cultured in LB Lennox broth (10g tryptone, 5g yeast extract, 5g NaCl) and Brain Heart Infusion broth (6g brain heart infusion, 6g peptic digest of animal tissue, 14.5g digested gelatin, 3g glucose, 5g NaCl, 2.5g Na_2HPO_4, 7.4 pH), respectively.

To record fluorescence measurements from RBS variants controlling reporter expression, transformed strains and a wild-type *E. coli* 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 average auto-fluorescence. Growth temperatures for *P. fluorescens*, and *C. glutamicum* were 30°C. The growth temperature for *S. typhimurium* LT2 was 37°C.

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.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. To record neurosporene productivity from pathway variants using IPTG-inducible promoters, cultures were grown overnight and then diluted into 50 mL 2xM9 media supplemented with 2% glucuose , grown at 30˚C with 250 RPM shaking, and induced with increasing IPTG concentrations. Pathway productivity was recorded after 22 hours of growth.

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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

100,000-fold(Salis et al, 2009). The apparent Boltzmann constant, β, has been measured as 0.45 $\pm$ 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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (ODE) 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:

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

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:

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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/hr 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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We thank A.E. Borujeni, B. Pfleger, J. Torella, and A. Khodayari for valuable discussion; H. Wang and G. Church (Harvard University) for the gift of strains EcNR2 and EcHW2f; A. Demirci and T. Wood (Penn State) for the kind gift of strains *E. coli* BL21, *P. fluorescens*, and *C. glutamicum*; 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

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

Alper H, Stephanopoulos G (2004) Metabolic engineering challenges in the post-genomic era. *Chemical engineering science* **59:** 5009-5017

Apgar JF, Witmer DK, White FM, Tidor B (2010) Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems* **6:** 1890-1900

Aswani A, Bickel P, Tomlin C (2011) Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics* **39:** 48-81

Bailey JE (1999) Lessons from metabolic engineering for functional genomics and drug discovery. *Nature biotechnology* **17:** 616-618

Bassett AR, Tibbit C, Ponting CP, Liu J-L (2013) Highly Efficient Targeted Mutagenesis of< i> Drosophila</i> with the CRISPR/Cas9 System. *Cell reports* **4:** 220-228

Becker J, Zelder O, Häfner S, Schröder H, Wittmann C (2011) From zero to hero—Design-based systems metabolic engineering of< i> Corynebacterium glutamicum</i> for l-lysine production. *Metabolic Engineering* **13:** 159-168

Borujeni AE, Channarasappa AS, Salis HM (2013) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic acids research*

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Brewster RC, Jones DL, Phillips R (2012) Tuning promoter strength through RNA polymerase binding site design in Escherichia coli. *PLoS computational biology* **8:** e1002811

Chang N, Sun C, Gao L, Zhu D, Xu X, Zhu X, Xiong J-W, Xi JJ (2013) Genome editing with RNA-guided Cas9 nuclease in Zebrafish embryos. *Cell research* **23:** 465-472

Chen Y-J, Liu P, Nielsen AA, Brophy JA, Clancy K, Peterson T, Voigt CA (2013) Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature methods* **10:** 659-664

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

Conzelmann H, Fey D, Gilles E (2008) Exact model reduction of combinatorial reaction networks. *BMC systems biology* **2:** 78

Dahl RH, Zhang F, Alonso-Gutierrez J, Baidoo E, Batth TS, Redding-Johanson AM, Petzold CJ, Mukhopadhyay A, Lee TS, Adams PD (2013) Engineering dynamic pathway regulation using stress-response promoters. *Nature biotechnology*

de Raad M, Kooijmans SA, Teunissen EA, Mastrobattista E (2013) A Solid-Phase Platform for Combinatorial and Scarless Multipart Gene Assembly. *ACS synthetic biology*

de Vos MG, Poelwijk FJ, Tans SJ (2013) Optimality in evolution: new insights from synthetic biology. *Current opinion in biotechnology*

Dharmadi Y, Patel K, Shapland E, Hollis D, Slaby T, Klinkner N, Dean J, Chandran SS (2013) High-throughput, cost-effective verification of structural DNA assembly. *Nucleic acids research*: gkt1088

Du J, Bai W, Song H, Yuan Y-J (2013) Combinational expression of sorbose/sorbosone dehydrogenases and cofactor pyrroloquinoline quinone increases 2-keto-L-gulonic acid production in Ketogulonigenium vulgare-Bacillus cereus consortium. *Metabolic engineering*: 1-7

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

Engler C, Gruetzner R, Kandzia R, Marillonnet S (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PloS one* **4:** e5553

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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**

Geggier S, Vologodskii A (2010) Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences* **107:** 15421-15426

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

Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-Terminal codon bias in bacterial genes. *Science* **342:** 475-479

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

Guye P, Li Y, Wroblewska L, Duportet X, Weiss R (2013) Rapid, modular and reliable construction of complex mammalian gene circuits. *Nucleic acids research* **41:** e156-e156

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

Hyeon C, Thirumalai D (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nature communications* **2:** 487

Johnson S, Lindén M, Phillips R (2012) Sequence dependence of transcription factor-mediated DNA looping. *Nucleic acids research* **40:** 7728-7738

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

Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342:** 744-747

Kitano H (2002) Computational systems biology. *Nature* **420:** 206-210

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences* **110:** 14024-14029

Lajoie MJ, Rovner AJ, Goodman DB, Aerni H-R, Haimovich AD, Kuznetsov G, Mercer JA, Wang HH, Carr PA, Mosberg JA (2013) Genomically recoded organisms expand biological functions. *Science* **342:** 357-360

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* **41:** 10668-10678

Lerman JA, Chang RL, Hyduke DR, Palsson BØ (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology* **9**

Lo T-W, Pickle CS, Lin S, Ralston EJ, Gurling M, Schartner CM, Bian Q, Doudna JA, Meyer BJ (2013) Precise and heritable genome editing in evolutionarily diverse nematodes using TALENs and CRISPR/Cas9 to engineer insertions and deletions. *Genetics* **195:** 331-348

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

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

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

Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ (2010) A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29:** 143-148

Monti MR, Smania AM, Fabro G, Alvarez ME, Argaraña CE (2005) Engineering Pseudomonas fluorescens for biodegradation of 2, 4-dinitrotoluene. *Applied and environmental microbiology* **71:** 8864-8872

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

Nielsen AA, Segall-Shapiro TH, Voigt CA (2013) Advances in genetic circuit design: novel biochemistries, deep part mining, and precision gene expression. *Current opinion in chemical biology* **17:** 878-892

Quandt EM, Deatherage DE, Ellington AD, Georgiou G, Barrick JE (2013) Recursive genomewide recombination and sequencing reveals a key refinement step in the evolution of a metabolic innovation in Escherichia coli. *Proceedings of the National Academy of Sciences***:** 201314561

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ran F, Hsu PD, Lin C-Y, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154:** 1380-1389

Rhodius VA, Mutalik VK (2010) Predicting strength and function for promoters of the Escherichia coli alternative sigma factor, σE. *Proceedings of the National Academy of Sciences* **107:** 2854-2859

Rodrigo G, Landrain TE, Shen S, Jaramillo A (2013) A new frontier in synthetic biology: automated design of small RNA devices in bacteria. *Trends in Genetics* **29:** 529-536

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

Sleight SC, Sauro HM (2013) Randomized BioBrick Assembly: A novel DNA assembly method for randomizing and optimizing genetic circuits and metabolic pathways. *ACS synthetic biology* **2:** 506-518

Smallbone K, Messiha HL, Carroll KM, Winder CL, Malys N, Dunn WB, Murabito E, Swainston N, Dada JO, Khan F (2013) A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. *FEBS letters* **587:** 2832-2841

Smolke CD, Silver PA (2011) Informing biological design by integration of systems and synthetic biology. *Cell* **144:** 855-859

Sneddon MW, Faeder JR, Emonet T (2010) Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nature methods* **8:** 177-183

Sneppen K, Krishna S, Semsey S (2010) Simplified models of biological networks. *Annual review of biophysics* **39:** 43-59

Strohman R (2002) Maneuvering in the complex path from genotype to phenotype. *Science* **296:** 701-703

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tan C, Saurabh S, Bruchez MP, Schwartz R, LeDuc P (2013) Molecular crowding shapes gene expression in synthetic cellular nanosystems. *Nature nanotechnology* **8:** 602-608

Torella JP, Boehm CR, Lienert F, Chen J-H, Way JC, Silver PA (2014) Rapid construction of insulated genetic circuits via synthetic sequence-guided isothermal assembly. *Nucleic acids research* **42:** 681-689

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

Wessely F, Bartl M, Guthke R, Li P, Schuster S, Kaleta C (2011) Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Molecular systems biology* **7**

Widmaier DM, Tullman-Ercek D, Mirsky EA, Hill R, Govindarajan S, Minshull J, Voigt CA (2009) Engineering the Salmonella type III secretion system to export spider silk monomers. *Molecular Systems Biology* **5**

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

bioRxiv preprint doi: https://doi.org/10.1101/001008; this version posted March 3, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zhang F, Carothers JM, Keasling JD (2012) Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nature biotechnology* **30:** 354-359

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