SEQUENCE-TO-FUNCTION MODELS FOR EFFICIENT OPTIMIZATION OF METABOLIC PATHWAYS AND GENETIC CIRCUITS

A Dissertation in
Chemical Engineering
by
Iman Farasat

© 2015

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

August 2015
The dissertation of Iman Farasat was reviewed and approved* by the following:

Howard M. Salis  
Assistant Professor of Chemical Engineering and Biological Engineering  
Dissertation Advisor  
Chair of Committee

Costas D. Maranas  
Donald B. Broughton Professor of Chemical Engineering

Ali Borhan  
Professor of Chemical Engineering

Timothy C. Meredith  
Assistant Professor of Biochemistry and Molecular Biology

Janna K. Maranas  
Professor of Chemical Engineering  
Chair of Graduate Program

*Signatures are on file in the Graduate School
ABSTRACT

A central challenge in creating a biorenewable economy is the economically competitive production of valuable chemicals from renewable feedstocks. The latest advances in metabolic engineering and synthetic biology have yielded engineered microbes that manufacture hydrocarbon fuels, biodegradable plastics, and drugs. This has mainly been achieved by either manipulating a natural host’s metabolic network or by expressing heterologous enzymes that convert one of the host’s metabolites to the product of interest.

Recent attempts in synthetic biology and metabolic engineering have been focused on identifying well-defined and beneficial mutations to extend an organism’s metabolic capabilities. Technical developments in DNA synthesis and sequencing have enabled the rapid assembly of synthetic metabolic modules, direct genome mutagenesis, and the construction of synthetic bacterial genomes. These advances provide almost complete control over a microorganism’s DNA sequence, enabling the insertion of heterologous proteins and the variation of an organism’s protein expression levels. However, the space for possible DNA mutations is extremely large and cannot be brute-forced to optimize target genetic systems; for instance, the number of possible ways to mutate 100 base pairs of DNA is larger than the number of atoms on earth. Most mutations have neutral or negative effects on the target genetic systems’ output, and the beneficial mutations must be fished out by rational predictive design or found by exhaustive experimental study. The absence of a quantitative and predictive theory that relates a genetic system’s DNA sequence to its output phenotypic behavior has hampered our ability to design and optimize large genetic systems.

Here, we employ biophysical modeling to build quantitative maps that relate a genetic system’s DNA sequence with the expression rate of its genetic elements and final phenotypic activity. We use these maps to optimize three types of genetic systems: a heterologous metabolic
pathway, a genetic circuit controller, and a CRISPR/(d)Cas9 system. For each system, we first perform a minimal number of experiments to systematically develop a biophysical sequence-expression-activity map that we have called SEAMAP, then use it to optimize the genetic system at the DNA sequence level. We show that designing experiments based on the governing biophysics of the genetic systems substantially reduces the time and effort needed to optimize the genetic systems. For instance, we varied the protein production rate by over 1000-fold in five different microbial hosts and for five different proteins by performing 8 to 36 experiments in each case. As a comparison, a similar-fold change in the final activity of a protein-containing system may be achieved by conducting 100 to 100,000 experiments using generic DNA libraries (e.g. promoter and RBS libraries that alter expression of the protein) and well-established protein engineering approaches (e.g. Directed Evolution that alters activity of the protein).

In the first study (Chapter 1 and Appendix A), we perform biophysical guided optimization of a carotenoid biosynthesis metabolic pathway. We express three heterologous enzymes from R. sphaeroides on a bacterial operon to convert IPP, an E. coli metabolite, to neurosporene, a yellow-brown pigment. To optimize the expression rate of these enzymes, we first perform 73 systematic experiments to build a quantitative map that links the RBS sequence of each enzyme with its intracellular expression levels and the neurosporene productivity of the mutants. We then use the map to design the DNA sequence of the metabolic pathway for various applications. This includes navigating the pathway’s transcription and translation design space, identifying metabolic rate-limiting steps, examining the evolutionary robustness of genetic systems during long-term growth, quantifying optimality conditions for maximizing productivity, and designing pathway variants with maximal productivity. We show that the SEAMAP enables us to systematically vary productivity of the metabolic pathway over 100-fold.
The second case study (Chapter 2 and Appendix B) involves design and optimization of an analog genetic circuit controller: a signal amplifier circuit. The circuit is constructed by transmitting an input transcriptional signal through the cascade of two transcription factors (TFs). We first develop a thermodynamic-based map to quantify the effect of changes in genetic context and environmental conditions on transcriptional regulation in bacteria. We examine the map by performing 847 experiments in diverse microbial hosts and genetic contexts. Using the map, we also design a set of experiments to systematically measure *in vivo* binding energy of six TetR-homolog TFs to their DNA operator sites on an absolute energy scale (kcal/mol). We then employ the biophysical map to design the DNA sequence for a family of signal amplifying genetic circuits called genetic OpAmps that expand the dynamic range of cell sensors.

In the last case study (Chapter 3 and Appendix C), we focused on developing a SEAMAP to optimize activity and specificity of the revolutionary CRISPR/Cas9 system. This simple genetic system requires two basic elements, the Cas9 protein and a guide RNA, and can be programmed with any target sequence to perform high-throughput genome editing as well as high-throughput transcriptional regulation in diverse organisms. A major challenge in employing this system is its high rate of off-target activity, which may result in many unexpected indels or in misregulation. Our analysis of the available high throughput measurements shows that Cas9 activity and specificity are anti-correlated. Here, we employ statistical thermodynamics and enzyme kinetics to build a SEAMAP that relates the guide RNA and Cas9 RBS sequences with the expression of the total guide RNA-loaded Cas9 complex and its final phenotypic activity on any DNA locus in a target genome. We first combine the high-throughput measurements from six studies with a small set of systematic *in vivo* experiments to parameterize the SEAMAP. We then employ the map to solve the max Cas9 activity-max Cas9 specificity problem by considering the effect of Cas9 and guide RNA expression rate, mutations in guide RNA sequence, growth
condition, and genetic context on the final Cas9 activity at any genome locus. We finally use the map to explain the results of two recent high-throughput studies: a genome-wide analysis of targeting a specific DNA sequence in the λ-phage genome and an HIV treatment that involves excising the HIV cassette from the genome of HEK293T human cell line.
TABLE OF CONTENTS

List of Figures...............................................................................................................................x
List of Tables ..................................................................................................................................xiii
Achievements .................................................................................................................................xv
Acknowledgements....................................................................................................................xvi

Chapter 1 Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria .................................................................1
   Introduction...............................................................................................................................1
   Results......................................................................................................................................4
       Efficient Searching of the Sequence-Expression Space.........................................................4
       Navigation of Expression Spaces in Diverse Bacterial Species ...........................................8
       Efficient search in gram-positive and gram-negative bacterial genomes..........................10
       Efficient search in multi-dimensional expression spaces....................................................12
       Mapping the Sequence-Expression-Activity Space of a Multi-enzyme Pathway..............15
       Design and Optimization of Multi-enzyme Pathways using SEAMAPs............................18
       The Expanding Search for Optimally Balanced Pathways..................................................21
   Discussion...............................................................................................................................23
   Materials and Methods...........................................................................................................28
       Strains and Plasmid Construction .........................................................................................28
       Growth and Measurements.................................................................................................31
   Models and Computation.........................................................................................................33
       The RBS Calculator .............................................................................................................33
       The RBS Library Calculator ...............................................................................................35
       Kinetic Model Formulation, Transformation, and Identification.........................................38

Chapter 2 Whole-system genetic circuit design using biophysical measurements, modeling, and dimensionless units ......................................................41
   Introduction...............................................................................................................................41
   Results......................................................................................................................................44
       Measuring in vivo TF binding energies .................................................................................44
       Building a sequence-expression-activity map ....................................................................47
       Dimensionless design space of genetic circuits (Pt number) ...............................................51
       De novo design of genetic opAmps in dimensionless design space ....................................54
   Discussion...............................................................................................................................56
   Materials and Methods...........................................................................................................59
       Genetic circuit constructs ....................................................................................................59
       Growth and Measurements .................................................................................................61
       Estimating the measurements on absolute scales ...............................................................61
       Measuring In vivo binding energy of transcription factors ................................................62

Chapter 3 A Biophysical Model of CRISPR/Cas9 On- and Off-Target Activity for Rational Design of Genome Editing and Gene Regulation.................................65
### Table of Contents

**Introduction** ........................................................................................................ 65  

**Results** ............................................................................................................... 70  
  Model parameterization using *in vitro* measurements .................................. 70  
  Quantifying the Effects of Supercoiling on Adjacent DNA sites ................. 74  
  Binding initiation from non-canonical PAM sequences ............................. 77  
  The effects of crRNA:DNA mismatches ................................................. 78  
  The effect of host and experimental design factors on (d)Cas9 activity ..... 82  
  Genome-wide calculation of (d)Cas9 binding activity .............................. 85  
  Example: removing latent HIV provirus from genomes .......................... 89  

**Discussion and Conclusion** .............................................................................. 90  

**Materials and Methods** .................................................................................. 95  
  A Mechanistic Model of CRISPR/(d)Cas9 Binding and Cleavage .......... 95  
  Modeling the Expression and Formation of the Cas9:crRNA Complex ...... 96  
  Modeling Cas9's Random Walk to Determine its Search Rate .............. 99  
  A Free Energy Model for DNA Target Binding, R-Loop Formation, and  
  Cleavage ........................................................................................................ 101  
  Modeling the Genome-wide Occupancies at On-target and Off-target DNA  
  Sites ............................................................................................................... 105  
  The Effects of Supercoiling on Adjacent DNA sites ................................ 108  
  Additional Model Considerations and Assumptions ................................. 109  
  Model Simulation and Parameterization Approach ................................. 110  
  Growth and Measurements .......................................................................... 111  

**Reference** ......................................................................................................... 113  

**Appendix A Supplementary Information for Chapter 1** ................................ 130  
  Appendix A-Figures ......................................................................................... 130  
  Appendix A-Tables ......................................................................................... 139  
  Appendix A-Methods .................................................................................... 143  
  Appendix A-Method 1: The expected number of MAGE cycles to insert  
  optimized RBS libraries into a genome .................................................. 143  
  Appendix A-Method 2: Monte-Carlo estimation to determine the minimal  
  number of measurements ......................................................................... 145  
  Appendix A-Method 3: An elementary mass action kinetic model of the  
  carotenoid biosynthesis pathway's reactions ........................................ 146  
  Appendix A-Method 4: Using Computational geometry for building a non- 
  mechanistic model of the Carotenoid pathway .................................... 155  
  Appendix A-Method 6: Using statistical model for creating a non-mechanistic  
  model of the Carotenoid pathway .......................................................... 156  
  Appendix A-Method 7: Performance of kinetic, geometry, and statistical  
  modeling for building a SEAMAP ............................................................ 158  

**Appendix B Supplementary Information for Chapter 2** ................................ 162  
  Appendix B-Figures ......................................................................................... 162  
  Appendix B-Tables ......................................................................................... 168  
  Appendix B-Methods .................................................................................... 169
Appendix B-Method 1: Developing a statistical thermodynamic model of genetic circuits containing multiple TF binding sites ........................................ 169
Appendix B-Method 2: Characterizing genetic circuits in diverse genetic context and environmental conditions ........................................ 172
Appendix B-Method 3: Designing genetic circuits in Pt design space ........ 174

Appendix C Supplementary Information for Chapter 3 ................................ 178
Appendix C-Figures .................................................................................. 178
Appendix C-Tables .................................................................................. 182

Appendix D- Optimizing Furfural biodetoxification pathway .................... 183

Appendix E-Performance of the RBS Library Calculator for optimizing expression rate of different example genetic systems ........................................ 186

Appendix F- Rational Design of a Synthetic Entner-Doudoroff Pathway for Improved and Controllable NADPH Regeneration ........................................ 189

Introduction ................................................................................................. 189
Results ........................................................................................................... 194
   Rational Design and Construction of a Synthetic Entner-Doudoroff Pathway ...... 194
   Characterization of the synthetic Entner-Doudoroff pathway in ED1.0 .......... 196
   Efficient Search for Improved ED Pathway Variants in a 5-dimensional
   Expression Space ....................................................................................... 199
   Improving Terpenoid Biosynthesis using a Synthetic Entner-Doudoroff
   Pathway ................................................................................................... 206

4. Discussion .................................................................................................. 209
Materials and Methods ................................................................................ 213
   Design of synthetic operons and codon optimization .................................. 214
   Strain and plasmid construction .................................................................. 215
   Combining the RBS library Calculator and MAGE for pathway optimization ...... 216
   Chromosomal integration of Entner-Doudoroff pathway variants ......... 217
   Measurement of intracellular NADPH and NADP⁺ levels ......................... 218
   Quantifying NADPH levels with a modified fluorescent reporter ............... 219
   Measurement of NADPH-dependent carotenoid biosynthesis .................. 220

Supplementary Information ............................................................................. 222
Content ........................................................................................................... 222
Supplementary Materials and Methods .......................................................... 222
LIST OF FIGURES

Figure 1-1 Building and using sequence-expression-activity maps (SEAMAPs). ................6
Figure 1-2 Searching expression spaces in diverse bacterial hosts.................................8
Figure 1-3: Searching genomic expression spaces.......................................................10
Figure 1-4: Sequence-expression-activity mapping of a multi-enzyme pathway..............14
Figure 1-5: Using SEAMAPs to design multi-enzyme pathways with desired activity response curves........................................................................................................17
Figure 1-6: Increasing precursor biosynthesis for optimally balanced vs. imbalanced pathways...............................................................................................................20
Figure 1-7: Predicting the evolutionary landscape of a multi-enzyme pathway...............23
Figure 2-1: Measuring binding energy of transcription factors to their cognate DNA........44
Figure 2-2: Developing a sequence-expression-activity map for a signal inverter circuit containing a TF that targets multiple DNA loci.........................................................47
Figure 2-3: Extensive characterization of the genetic circuits in different hosts and growth conditions........................................................................................................51
Figure 2-4: Design of genetic opAmp circuits..................................................................53
Figure 2-5: Designing several example analog and digital genetic circuits in Pt dimensionless design space........................................................................................................56
Figure 3-1: The multi-step mechanism responsible for Cas9-mediated DNA site cleavage. ..69
Figure 3-2: In vitro validation of the Cas9 mechanistic model........................................72
Figure 3-3: In vivo validation of the Cas9 mechanistic model........................................73
Figure 3-4: The effect of mismatches between a crRNA and a target on its Cas9 binding affinity....................................................................................................................80
Figure 3-5: Sensitivity analysis of the Cas9 model’s parameters........................................82
Figure 3-6: Sensitivity analysis of the dCas9 model’s parameters......................................84
Figure 3-7: Genome-wide calculation of (d)Cas9 binding affinity to λ-phage plasmid.......85
Figure 3-8: Genome engineering for targeting HIV sequences........................................88
Appendix A-Figure 1: Sampling 3-dimensional CFP, mRFP1, GFPmut3b production rate space using optimized RBS libraries.................................................................130
Appendix A-Figure 2: Sampling 3-dimensional CFP, mRFP1, GFPmut3b production rate space using optimized RBS libraries.................................................................131
Appendix A-Figure 3: Estimating the number of measurements to find optimal protein expression levels in multi-protein genetic systems using random or optimized RBS libraries.........................................................................................132
Appendix A-Figure 4: The proposed elementary reaction model for the Carotenoid biosynthesis pathway. ..............................................................................................................133
Appendix A-Figure 5: The kinetic model's error distribution for neurosporene content predictions..................................................................................................................133
Appendix A-Figure 6: Neurosporene content measurements using optimized growth conditions......................................................................................................................134
Appendix A-Figure 7: Input-output transfer function of P_{lacO1} promoter. ......................135
Appendix A-Figure 8: flux control coefficients for the enzymes  
\textit{crtE},  
\textit{crtB}, and  
\textit{crtI} ..........136
Appendix A-Figure 9: Using computational geometry to search for optimal translation rates regions. ................................................................................................................137
Appendix A-Figure 10: Effect of training set size on model predictions. ...............................138
Appendix A-Figure 11: Null-hypothesis rejection test ............................................................139
Appendix B-Figure 1: Effect of additional TF binding sites on performance of circuits containing LacI and a TetR-homolog TF.................................................................................................162
Appendix B-Figure 2: Model parameterization for calculating RNA polymerase binding energy..........................................................................................................................163
Appendix B-Figure 3: Model parameterization for calculating TF binding energy. ..............164
Appendix B-Figure 4: Protein expression rate at different growth rate..................................165
Appendix B-Figure 5: Characterization of PhlF circuits in reverse orientation .................165
Appendix B-Figure 6: Characterization of PhlF circuits in multiple industrial microbial hosts ........................................................................................................................................166
Appendix B-Figure 7: Effect of cross activity on performance of genetic opAmp circuit. .........166
Appendix B-Figure 8: Thermodynamic states for binding of a TF and RNA polymerase to a promoter in presence of one additional TF binding site..................................................167
Appendix C-Figure 1: Measured versus predicted DNA cleavage in the *in vitro* experiments ................................................................. 178

Appendix C-Figure 2: Sensitivity analysis of individual kinetic parameters. ................................................................. 178

Appendix C-Figure 3: Measured versus predicted dCas9 repression activity in the *in vivo* experiments ................................................................. 179

Appendix C-Figure 4: Quantifying the effect of crRNA:target mismatches using currently available RNA:RNA and DNA:DNA energy parameters ................................................................. 180

Appendix C-Figure 5: Comparison between the *in vitro* and *in vivo* measurements and predictions by Cas9 model containing the second proposed dimerization mechanism. ... 181

Appendix D-Figure 1: Toxicity effect of Furfural and HMF on cellular growth. ................................................................. 183

Appendix D-Figure 2: Furfural and HMF biodetoxification pathway ................................................................. 184

Appendix D-Figure 3: Furan consumption rate for different HmfD variants. ................................................................. 185

Appendix E-Figure 1: Characterization result for a subset of genetic variants designed by the RBS Library Calculator ................................................................. 186

Appendix E-Figure 2: Combinatorial optimization for the expression rate of genetic elements involved in three multi-protein genetic circuits. ................................................................. 186

Appendix E-Figure 3: Measuring the translation rate capacity of a codon-optimized gene... 187

Appendix E-Figure 4: Recombinant Vaccine Development ................................................................. 188

Appendix F-Figure 1: EMP versus ED Pathways ................................................................. 193

Appendix F-Figure 2: Design and construction of ED1.0 and combinatorial ED variants. ... 194

Appendix F-Figure 3: Characterization of ED1.0 ................................................................. 199

Appendix F-Figure 4: Uniform Sampling of the 5-dimensional Expression Space. .... 200

Appendix F-Figure 5: Characterization of ED-expressing genome variants ................................................................. 201

Appendix F-Figure 6: The Effects of Changing ED Enzyme Expression Levels on NADPH Regeneration Rates ................................................................. 205

Appendix F-Figure 7: Improving Terpenoid Biosynthesis using the Synthetic Entner-Doudoroff Pathway ................................................................. 207
LIST OF TABLES

Table 2-1: Binding energy of RNA polymerase and 6 TetR-homolog transcription factors to their cognate promoters.................................................................45

Table 3-1: A summary of all studies used to estimate the Cas9 model's parameters........67

Table 3-2: Parameter values used in this study.................................................................72

Table 3-3: Cas9:crRNA binding energies to canonical and non-canonical PAM sites ....76

Appendix A-Table 1: Characterization data for all RBS library variants controlling protein expression. ........................................................................................................139

Appendix A-Table 2: Strains and plasmids used in the study. ..............................................140

Appendix A-Table 3: Designed MAGE oligonucleotides for genome engineering. ..........140

Appendix A-Table 4: The optimized RBS libraries for sampling the translation rate space of the 3-color CFP, mRFP1, and GFPmut3b operon. ........................................140

Appendix A-Table 5: Random RBS libraries for sampling translation rate space of the designed triple-color operon. ........................................................................140

Appendix A-Table 6: The optimized RBS libraries for searching the enzyme expression space of the Carotenoid biosynthesis pathway......................................................140

Appendix A-Table 7: Characterization data for the 73 Carotenoid pathways generated using Search mode.................................................................140

Appendix A-Table 8: The carotenoid pathway's kinetic parameters determined by model identification. ........................................................................................................140

Appendix A-Table 9: Characterization of 19 additional carotenoid pathway variants to test the kinetic model's predictions. .........................................................141

Appendix A-Table 10: The optimized RBS libraries for targeting CrtE, CrtB, and CrtI expression levels using Zoom mode.........................................................141

Appendix A-Table 11: Characterization of 28 Carotenoid pathway variants generated using Zoom mode..........................................................................................141

Appendix A-Table 12: Characterization of pathways with novel input-output transfer functions.................................................................................................141

Appendix A-Table 13: The designed RBS library for varying expression of dxs enzymes. ...141

Appendix A-Table 14: Characterization of balanced and imbalanced pathways. ...............141
Appendix A-Table 15: The random RBS libraries used to calculate the coverages of example 3-, 4-, and 6-protein expression spaces .......................................................... 141

Appendix A-Table 16: The optimized RBS libraries used to calculate the coverages of example 3-, 4-, and 6-protein expression spaces using Search mode ........................................ 142

Appendix A-Table 17: The optimized RBS libraries used to calculate the coverages of example 3-, 4-, and 6-protein expression spaces using Zoom mode ...................................... 142

Appendix A-Table 18: Comparison between kinetic, geometry, and statistical model for building a SEAMAP ........................................................................................................ 142

Appendix B-Table 1: Strains used in the study .................................................................. 168

Appendix B-Table 2: Characterization data for all genetic circuits containing LacI and a TetR-homolog TF ........................................................................................................ 168

Appendix B-Table 3: Characterization data for all genetic opAmp circuits ...................... 168

Appendix C-Table 1: Parameters used in (d)Cas9 activity analysis .................................. 182

Appendix C-Table 2: Parameters used in genome-wide calculation for λ-phage plasmid ...... 182

Appendix C-Table 3: Parameters used in HIV calculation .................................................. 182
ACHIEVEMENTS

The main achievement of my PhD was to develop a protein expression optimization framework, the RBS Library Calculator, to design efficient combinatorial libraries that incrementally vary protein production rate by over 10000-fold. This framework facilitates altering the expression of every protein in a genetic system with a minimal number of experiments. At this time, we have employed this algorithm to optimize the expression rate of 21 proteins of different classes: 10 enzymes in three metabolic pathways, T7 RNA polymerase and 4 transcription factors in three analog genetic circuits, and 6 individual proteins in diverse Gram-negative and Gram-positive microbes (Chapter 1 and Appendix D-F). The RBS Library Calculator has been licensed by several biotech companies and has been used by over 1000 researchers.

I also developed three biophysics-based algorithms to simplify the engineering of microbes: the Cas9 Calculator to estimate CRISPR/Cas9 binding affinity to any genome locus, Pathway Map Calculator to systematically develop a kinetic model of a heterologous multi-enzyme pathway with a minimal number of experiments, and TF Binding Affinity Calculator to systematically measure in vivo binding affinity of a TF to any target DNA and quantify its binding affinity. In addition, I was involved in improving the accuracy of the RBS Calculator program and in developing the Operon Calculator algorithm to automate de novo design of bacterial operons.
ACKNOWLEDGEMENTS

The PhD track helped me to learn many life skills that are essential for reaching my future career and personal goals. During this journey, many people shared their wonderful experiences and thoughts with me, and finishing the journey could have been extremely difficult or impossible without their guidance. First, I am very thankful for Dr. Howard M. Salis. His continuous support was beyond the technical aspects of my PhD. The extraordinary help of my doctoral committee, Dr. Ali Borhan, Dr. Costas D. Maranas, and Dr. Timothy C. Meredith, was an essential actuator for me. Many ideas that I used here were the result of direct communication with my committee members or based on my understanding from the materials that they taught me in their graduate courses. I was very lucky that I had many wonderful lab mates who treated me like family and helped me to successfully pass the difficulties in my journey: Dr. Jason Collens, Dr. Manish Kushwaha, Amin Espah Borujeni, Tian Tian, Chiam Yu Ng, Long Chen, Grace Vezeau, Sean Halper, and Alexander Reis. Last, but not least, I would like to acknowledge the support that I have received from many friends, which enabled me to continue focusing on my goals: Dr. Ali R. Zomorrodi, Dr. Banafsheh Keshavarzi, Dr. Ehsan Espah Borujeni, Dr. Samira Khalili, Mahsa Hadidi, Seyed Pouria Motevalian, Ali Khodayari, Anupam Chowdhury, Dan Ye, Pouria Fattahi, Mona Hatami, Negar Hashemian, and many more.
Chapter 1 Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria

A modified version of this chapter was published in Molecular System biology journal[1].

Introduction

A microbe's ability to sense its environment, process signals, perform decision-making, and manufacture chemical products is ultimately controlled by its DNA sequence [2,3,4,5,6,7,8,9,10,11]. 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 [12,13,14,15,16].

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 [17,18,19,20,21,22,23]. The development of multiplex genome engineering provides the ability to simultaneously introduce DNA mutations into several genomic loci [24,25,26]. Further, Cas9-dependent and TALE-dependent nicking, cleavage, and mutagenesis has expanded site-directed genome engineering to diverse organisms [27,28,29,30,31,32]. 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 [33,34,35,36,37,38] with system-level models describing genetic system function [15,39,40,41,42]. Several types of models can be combined, utilizing thermodynamics, kinetics, mass transfer, and dynamical systems theory, to describe the multi-scale physical interactions affecting system function [41]. 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 [43,44].

However, the development and parameterization of such sequence-expression-activity models has been stymied by several forms of combinatorial explosion [45,46,47]. 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 [48,49,50,51].

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 1-1A). 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. 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 1-1B). When a genetic system is designed, the protein coding sequences' expression levels will be proportional to their translation initiation rates.

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 Appendix A-Table 1.
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 $\Delta G_{\text{tot}}$ 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. p-value<0.01). Data averages and standard deviations from 6 measurements.

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 1-1C), 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 1-1D). 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 1-1E). The biophysical model of bacterial translation accurately predicted the translation initiation rates from these 60 RBS variants with an average error $\Delta \Delta G_{\text{total}}$ of 1.74 kcal/mol, which is equivalent to predicting the measured translation initiation rate to within 2.2-fold with an apparent $\beta$ of 0.45 mol/kcal. The biophysical model's predictions were particularly accurate for the high and low resolution libraries (average $\Delta \Delta G_{\text{total}} = 1.05 \text{ kcal/mol}, R^2 = 0.88$; and average $\Delta \Delta G_{\text{total}} = 0.46 \text{ kcal/mol}, R^2 = 0.89$, respectively) in contrast to the medium resolution library that contains several outliers at low expression levels (average $\Delta \Delta G_{\text{total}} = 3.71 \text{ kcal/mol}, R^2 = 0.72$).
Figure 1-2 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 ($\Delta\Delta G_{\text{total}}$ is 1.18 kcal/mol, $R^2$ is 0.93), *P. fluorescens* ($\Delta\Delta G_{\text{total}}$ is 1.63 kcal/mol, $R^2$ is 0.90), *S. typhimurium* LT2 ($\Delta\Delta G_{\text{total}}$ is 1.83 kcal/mol, $R^2$ is 0.89), and *C. glutamicum* ($\Delta\Delta G_{\text{total}}$ is 1.81 kcal/mol, $R^2$ is 0.88). The p-value < 0.01. 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 (Appendix A-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 [52,53,54]. 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 gram-positive *C. glutamicum* and 5’-ACCUCCUUA-3’ for the remaining gram-negative species (Figure 1-2A).

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 $\Delta \Delta G_{\text{total}}$ of 1.61 kcal/mol, Pearson $R^2$ is 0.89) (Figure 1-2B). Translation rate predictions were more accurate in gram-negative *E. coli* BL21 ($\Delta \Delta G_{\text{total}}$ of 1.18 kcal/mol, Pearson $R^2$ is 0.93), compared to gram-positive *C. glutamicum* ($\Delta \Delta G_{\text{total}}$ of 1.81 kcal/mol, Pearson $R^2$ is 0.88), though the difference is a single outlier (Appendix A-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 $\beta$ 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 [55].
Figure 1-3: Searching genomic expression spaces. (A) Two RBS libraries were optimized to control the expression of a genomic single-copy of \textit{mRFP1}, incorporated into the \textit{amyE} locus of \textit{B. subtilis}. Fluorescence measurements from 14 clones were compared to their predicted translation initiation rates (Pearson $R^2$ is 0.81, p value is $2 \times 10^{-6}$). 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 \textit{lacZ} expression. Predicted translation initiation rates are compared to measured \textit{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 [24,26,28,56,57,58,59]. We developed the \textit{Genome 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’-ACCUCUCCUUU-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 $\Delta \Delta G_{\text{total}} = 1.77$ kcal/mol) (Figure 1-3A).

Second, we employed MAGE mutagenesis on the E. coli MG1655-derived EcNR2 genome [26], targeting its lacI-lacZYA locus and controlling lacZ protein expression levels (Figure 1-3B). We first conducted three rounds of MAGE mutagenesis to introduce an in-frame stop codon into the lacI repressor coding sequence (Appendix A-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 1-3B). 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 (Appendix A-Table 1). To further increase expression, one could replace the existing, natural genes in the genome with newly designed protein coding sequences optimized for maximum expression control, which motivates the design of synthetic genomes [60].

**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 (Appendix A-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 (Appendix A-Table 5). We employ combinatorial DNA assembly to construct bacterial operon variants encoding cfp, mRFP1, and gfpmut3b reporter proteins [17], 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 (Appendix A-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% (Appendix A-Figure 2), which agrees well with the computationally predicted search coverage of 14.7% using Monte Carlo sampling of the predicted translation initiation rates (Appendix A-Methods). Using the same method, similar search coverages of 16.7% and 19.1% are computationally predicted for a 4096-variant library NNNGGANN [61] and a 23328-variant library DDRRRRDDDD [26], 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 (Appendix A-Figure 3).
Figure 1-4: 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.
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 \(\text{crtEBI}\) enzyme expression levels originating from a carotenoid biosynthesis pathway in \(\text{Rhodobacter sphaeroides}\). Three 16-variant optimized RBS libraries were designed to vary \(\text{crtE}, \text{crtB},\) and \(\text{crtI}\) from 445 to 72000 au, 3 to 20000 au, and 97 to 203000 au, respectively (Appendix A-Table 6). 3-part combinatorial DNA assembly onto a ColE1 vector resulted in up to 4096 clonal pathway variants, transcribed by the arabinose-induced \(\text{P}_{\text{BAD}}\) promoter. 73 clones containing unique pathway variants were randomly selected, sequenced, transformed into \(\text{E. coli} \, \text{MG1655}\)-derived EcHW2f strain (Appendix A-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 \(\mu\text{g/gDCW/hr}\) (Figure 1-4A and Appendix A-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 1-4B), which eliminates redundant measurements and thus maximizes the measurements' information content. As \(\text{crtEBI}\) 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 1-4D and Appendix A-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 (Appendix A-Methods). We then use an ensemble modeling approach [50,62] 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 variants, across ten independent and randomly initialized optimization runs (Appendix A-Table 8). The resulting kinetic model maps *crtEBl* translation rates to neurosporene productivities across a 10,000-fold, 3-dimensional translation rate space (Figure 1-4B).
Figure 1-5: 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

We tested the SEAMAP’s predictions by using it to design \textit{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 \textit{crtEBI} translation rates from sequenced RBSs, and the kinetic model uses these translation rates to calculate each pathway variant’s productivity (Figure 1- 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) (Appendix A-Table 9). Overall, kinetic model predictions were more accurate at higher \textit{crtEBI} translation rates (Appendix A-Figure 1- 5). In general, a high \textit{crtE} translation rate was necessary for high biosynthesis rates, while low \textit{crtB} and high \textit{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 \textit{crtE}, 1800 to 232000 au for \textit{crtB}, and 26000 to 1347000 au for \textit{crtI} (Figure 1- 4C and Appendix A-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 1- 4A) (Appendix A-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 [63] (Appendix A-Figure 6).
Third, we tested the SEAMAP’s ability to predict a pathway’s activity response curve when utilizing a regulated promoter to vary the \textit{crtEBI} operon's transcription rate (Figure 1-5A). Changes in the transcription rate will proportionally vary all enzyme expression levels in the \textit{crtEBI} operon, causing diagonal shifts in expression space that will lead to productivity changes. We modified the most productive \textit{crtEBI} pathway variant, replacing its \text{P}_{\text{BAD}} promoter with the \text{P}_{\text{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 1-5B, black square). The same promoter replacement was performed for three additional pathway variants with distinctly different \textit{crtEBI} translation rates (Appendix A-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 1-5B). We use a model to first relate IPTG concentration to \text{P}_{\text{lacO1}} transcription rate (Appendix A-Figure 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 1-5C, 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 (Appendix A-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 1-5C, 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 [64,65].

Figure 1-6: 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.
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 [46]. 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 [39]. 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) [66,67] 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 1-6A; Appendix A-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 criteria that indicates the need for further metabolic engineering efforts in 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 1-6A). 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 mutagenesis [25], 16 genome variants were constructed and verified. The RBS library varied dxs translation from 110 to 291000 au. These dxs expression changes significantly increased the optimally balanced pathway's productivity up to 517 ug/gDCW/hr (Figure 1-6B; Appendix A-Table 14), while only increasing the imbalanced pathway's productivity up to 81 ug/gDCW/hr (Figure 1-6C). Consequently, a pathway engineering effort should cease when an optimally balanced pathway variant has been found. Metabolic engineering of the upstream precursor biosynthesis pathway will then increase the downstream pathway's productivity. These results suggest an iterative pathway engineering approach where additional upstream pathways are successively optimized, applying the optimality criteria in a recursive fashion.
Figure 1-7: 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 1-1). 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 1-2 and 1-3). 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 (Figure 1-4) and tailored activity response curves (Figure 1-5), while quantitatively guiding further
metabolic engineering efforts to increase precursor biosynthesis rates (Figure 1-6). 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 1-6, the effects of single, double, and triple RBS mutations are calculated, showing that pathway productivity decreases in almost all cases (Figure 1-7). 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 [68]. We formulate models using physical principles that can be re-used across different systems and scales [41], 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, for example, using chip-based oligonucleotide synthesis and next-generation sequencing [69,70].

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 (Appendix A-Methods). Using this approach, a geometric SEAMAP could use intermediate enzyme expression levels to predict pathway activities with 15% error (interpolation) (Appendix A-Figure 9, Appendix A-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 [71] that is similar to a recent approach [2]. 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 (Appendix A-Methods). Using a previous approach [2], 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 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 (Appendix A-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 [61]. 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 [72]. 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 [73,74]. 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 [36], 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 [33,34,35,38,75,76,77]. 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.

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 Appendix A-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 [17] 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 sigma70 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 *mRFP* 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 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 (Appendix A-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 DpnI digested, gel purified, and joined together into the pFTV vector using a 3-part CBAR assembly reaction [17], 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 Na₂HPO₄, 3g KH₂PO₄, 1g NH₄Cl, 0.5g NaCl, 0.24g MgSO₄, 0.011g CaCl₂), as indicated. *Salmonella typhimurium* LT2 and *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 Na₂HPO₄, 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 OD₆₀₀ 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*, *C. glutamicum*, and *S. typhimurium* were 30°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\textsubscript{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 \textit{lacZ} activities using Miller assays, \textit{E. coli} EcNR2 genome variants containing \textit{lacI}
knockouts and \textit{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 µg/ml chloramphenicol. Cultures were then
diluted into fresh supplemented LB Miller media and cultured at 30°C to an OD\textsubscript{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\textsubscript{420} - 1.75 OD\textsubscript{550}) / OD\textsubscript{600} during the times when the product synthesis rate was constant.

To measure neurosporene productivities, pathway variants were incubated for 16hrs at 30°C, 250
RPM orbital shaking in 5 ml culture tubes, then washed with PBS, dissolved in fresh LB miller
(50 µg/ml chloramphenicol, and 10mM arabinose), and grown for another 7 hours. Cells were
centrifuged (Allegra X15R at 4750 RPM) for 5 minutes, washed with 1 ml ddH2O, and dissolved
in 1 ml acetone. The samples were incubated at 55 °C for 20 minutes with intermittent vortexing,
centrifuged for 5 minutes, and the supernatants transferred to fresh tubes. Absorbance was
measured at 470 nm using NanoDrop 2000c spectrophotometer and converted to µg
Neurosporene (x 3.43 µg/nm absorbance). The remaining pellet was heated at 60 °C for 48 hrs to
determine dry cell weight. Neurosporene content was calculated by normalizing Neurosporene
production by dry cell weight. Neurosporene productivity was determined by dividing by 7 hours.
To record neurosporene productivity under optimized growth conditions, pathway variants were incubated overnight in 5 ml LB miller, followed by inoculating a 50 mL shake flask culture using 2xM9 media supplemented with 0.4% glucose and 10 mM arabinose. The culture was grown for 10 hours at 37˚C with 300 RPM orbital shaking. The neurosporene productivity was measured using 10 ml of the final culture as stated above. To record neurosporene productivity from pathway variants using IPTG-inducible promoters, cultures were grown overnight and then diluted into 50 mL M9 media supplemented with 2% glucose, 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:

\[ \Delta G_{\text{total}} = \Delta G_{\text{mRNA:rRNA}} + \Delta G_{\text{spacing}} + \Delta G_{\text{start}} + \Delta G_{\text{standby}} - \Delta G_{\text{mRNA}} \]  

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:

\[ r \propto \exp(-\beta \Delta G_{\text{total}}) \]
This relationship has been previously validated on 132 mRNA sequences where the $\Delta G_{\text{total}}$ varied from $-10$ to 16 kcal/mol, resulting in well-predicted translation rates that varied by over 100,000-fold[78]. The apparent Boltzmann constant, $\beta$, has been measured as $0.45 \pm 0.05$ mol/kcal, which was confirmed in a second study [79]. 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 $\Delta G_{\text{mRNA}}$ ($\Delta G_{\text{mRNA}}$ 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 $\Delta G_{\text{mRNA},\text{rRNA}}$ ($\Delta G_{\text{mRNA},\text{rRNA}}$ 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[80,81] and the minimization algorithms available in the Vienna RNA suite, version 1.8.5[82].

Three additional interactions will alter the translation initiation rate. The tRNA$^{\text{fMET}}$ anti-codon loop hybridizes to the start codon ($\Delta G_{\text{start}}$ 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 ($\Delta G_{\text{spacing}} > 0$) and nucleotide distance was systematically measured. Finally, the 5' UTR binds to the ribosomal platform with a free energy penalty $\Delta G_{\text{standby}}$.

There are key differences between the first version of the RBS Calculator (v1.0)[78], and version v1.1 [37]. 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 $\Delta G_{\text{standby}}$. The ribosome's rules for binding to long, highly structured 5' UTRs has been characterized [83], and will be incorporated into a future version of the RBS Calculator (v2.0).

The RBS Library Calculator

The objective of the RBS Library Calculator is to identify the smallest RBS library that uniformly varies a selected protein's expression level across a targeted range to efficiently identify optimal protein expression levels and quantify expression-activity relationships. The RBS Library Calculator designs degenerate ribosome binding site (RBS) sequences that satisfy the following mini-max criteria: first, the RBS sequence variants in the library shall express a targeted protein to maximize coverage, $C$, of the translation rate space between a user-selected minimum ($r_{\text{min}}$) and maximum rate ($r_{\text{max}}$); second, the number of RBS variants in the library, $N_{\text{variants}}$, 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:

$$F = 10C - 0.02N_{\text{variants}}$$  \hspace{1cm} (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:

$$B_{total} = \left\lfloor \frac{r_{max}/r_{min}}{W} \right\rfloor \quad C = \frac{B_{filled}}{B_{total}} \quad (4)$$

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 $[y/10^W, y10^W]$. 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 $1 - 0.02 \times B_{total}$, which is achieved when all bins are filled by a single RBS variant, yielding the most compact RBS library that expresses a protein with uniformly increasing translation rates.

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

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

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

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

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

**Kinetic Model Formulation, Transformation, and Identification**

Mass action kinetics was utilized to formulate an ordinary differential equation (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 Appendix A-Methods. 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 (Appendix A-Figure 4). IspA, CrtE, CrtB, and CrtI catalyze multiple reactions. These reactions convert intracellular isopentenyl diphosphate (IPP) and
Dimethylallyl diphosphate (DMAPP) to neurosporenone. 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:

$$v_{f1} = \left( k_1 \times \frac{[IPP]_{ref} \times [idi]_{total}}{\text{apparent kinetic parameter}} \right) \times \frac{[IPP]}{[IPP]_{ref}} \times \frac{[idi]_{free}}{[idi]_{ref}}, \quad v_{r1} = \left( k_{-1} \times \frac{[CM1]_{ref}}{\text{apparent kinetic parameter}} \right) \times \frac{[CM1]}{[CM1]_{ref}}$$

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:

$$\frac{[CrtE]_{total}}{[CrtE]_{ref}} = \frac{\text{translation initiation rate of } crtE \text{ in a pathway variant}}{\text{translation initiation rate of } crtE \text{ in the reference pathway}} \times \frac{\text{translation initiation rate ratio}}{\text{enzyme concentration ratio}}$$

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:

$$\frac{r_{p,i}}{r_{p,\text{ref}}} = \frac{\text{predicted neurosporene productivity of the } i^{th} \text{ pathway variant}}{\text{measured neurosporene productivity of the reference pathway}}$$

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 (Appendix A-Methods). 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 (Appendix A-Figure 5). We then performed inverse model reduction to determine the 48 kinetic parameter values that define the identified kinetic model (Appendix A-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.
Chapter 2 Whole-system genetic circuit design using biophysical measurements, modeling, and dimensionless units

Introduction

With new advances in synthetic biology, researchers can now vary different gene expression steps such as replication[84], transcription[85], mRNA degradation[86], translation[87], and protein degradation[88]. This has provided numerous genetic parts for constructing genetic circuits that reprogram biological cells for sophisticated applications such as controlling metabolic networks[64,65], sensing cellular disorders[89], and curing diseases[90]-[91]. A key challenge for rationally constructing novel genetic circuits is the absence of a systematic design methodology. Current design approaches rely on the assembly of multiple well-characterized genetic parts and devices with the assumption that each component’s transfer function remains unchanged[61,92]. However, this design strategy underestimates the multi-layered nature of regulatory interactions between the constructed genetic systems and the remaining context. For instance, decoy sites for the regulators[93], retroactivity of genetic parts[93,94,95], and growth conditions[95,96] change the dynamics of the designed circuits, altering their expected output signal by over 10-fold. Combinatorial optimization and directed evolution[92,97] can debug small defective circuits, however, the number of required characterizations increases astronomically with the complexity of the circuit. This has hindered researchers from building larger genetic circuits[95].

To build a system-wide design methodology, we apply biophysics to enumerate all major factors controlling a genetic circuit’s performance. This allows us to construct a sequence-expression-activity map (SEAMAP) that predicts how genetic context and experimental conditions affect the expression and activity of transcription factors (TFs), the largest
characterized family of regulatory elements, with 5 to 10% genome occupancy[98,99]. A key ingredient for building these maps is the accurate measurement of TF binding energy to their regulons[100,101,102]. The estimated energies by current in vitro approaches[103,104,105] may provide correct predictions[101] or result in large unexpected error[102,106], likely due to differences in in vitro and in vivo conditions such as crowding, diffusion, time-scale of equilibrium, and ionic strength[102,107]. Several in vivo approaches[108,109] estimate the binding energies on relative scales, or use in vitro assumptions to estimate the absolute energies[109]. These measured energies are usually functions of the experimental methods; for example, using three different approaches, Zeigler et al.[108] have reported 1.32 au (flow cytometry), 3.00 au (ChIP-seq), and 4.87 au (a combination of both methods) for binding energy of Cbf1 to its cognate binding site. This inconsistency restricts our ability to use biophysical design methods for many natural and synthetic circuits when a transcription factor targets multiple DNA loci. We address this problem by developing a high-throughput assay that measures in vivo TF binding energy on an absolute scale.

During the design of genetic circuits, the choice of circuit-harboring host, circuit components, and growth conditions affects circuit output and determines design success or failure[95]. These choices create a co-dependent combinatorial design space that must be explored to obtain circuits with correct functionality[93,95]. In other science and engineering disciplines[110], when many co-dependent parameters exist in a process, non-dimensionalization is an effective approach to understand the process and forward engineer de novo systems, for instance Reynolds number in fluid dynamics and Elasticity in economics. To reduce design complexity of genetic circuits, we define a dimensionless number (Pt) that encapsulates the effect of all the parameters controlling TF partitioning among its binding sites. This allows us to create a dimensionless design space where activity of each transcriptional regulatory element is quantified by a dimensionless value. By combining this design space with sequence-expression-
activity mapping, we develop an efficient design framework to minimize the number of unknown independent parameters that must be adjusted when aiming to build genetic systems at DNA sequence level.

Here, we first measure binding energy of 6 TetR-homolog transcription factors to their cognate operator DNA (Figure 2-1). This allows us to develop a biophysical map that estimates performance of the circuits composed of these TFs (Figure 2-2). The map accounts for 16 parameters divided into three distinct categories: genetic elements, host-dependent parameters, and growth conditions. The genetic element parameters include TF binding energy, number of TF specific sites, the TF's RBS, circuit orientation, RNA polymerase binding energy, circuit copy number, rate of open complex formation on the output promoters, TF's transcription rate, binding cooperativity of TF to its operator, TF's mRNA degradation rate, and TF’s protein degradation rate. The host-dependent parameters include 16S rRNA sequence, number of available RNA polymerases, and genome size. The growth condition parameters include media and temperature. We extensively examine the map by performing 847 experiments in 6 gram-positive and gram-negative bacterial hosts, resulting in over 1000-fold change in the circuit output (Figure 2-3). We then use the developed design framework to build a family of genetic opAmp circuits that amplify a transcriptional signal up to 20-fold (Figure 2-4).
Figure 2-1: Measuring binding energy of transcription factors to their cognate DNA. (A) The developed high throughput assay includes three steps: construction of genetic circuits that contain multiple TF binding sites, characterization of the circuits by flowcytometry, and model parameterization to calculate the TF binding energy ($\Delta G_R$). (B) Effect of auxiliary sites on the activity of 5 additional TFs. YFP: yellow fluorescent protein; N: number of TF binding sites.

Results

Measuring in vivo TF binding energies

When a TF is used to repress a promoter, addition of TF’s decoy sites sequesters a fraction of TF molecules, and increases binding probability of RNA polymerase to the promoter. The change in the expression rate of this promoter is a function of RNA polymerase binding affinity to the promoter and TF’s binding energy to its operator sites. The change in the expression rate can be assayed in a high-throughput fashion using luminescence and fluorescence assays, and can be analyzed using a nonlinear model identification method (Methods) to estimate
the binding energies. For a heterologous TF, this could be done by inserting up to 4 additional TF binding sites on the same expression vector, and quantifying the change in mRNA transcription using a transcriptional fusion assay while titrating the rate of TF expression using an IPTG-inducible promoter (Figure 2-1B; Appendix B-Figure 1).

Table 2-1: Binding energy of RNA polymerase and 6 TetR-homolog transcription factors to their cognate promoters; \( r_{\text{max}} \) is the rate of open complex formation by RNA polymerase on each cognate promoter that limits the maximum transcription rate from the promoter; \( \Delta G_R \): TF binding energy; \( \Delta G_{\text{RNAP}} \): RNA polymerase binding energy; \( \delta_p \): degradation rate of TF proteins; \( \delta_m \): TF’s mRNA degradation rate.

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>( \Delta G_R ) (kcal/mol)</th>
<th>( \Delta G_{\text{RNAP}} ) (kcal/mol)</th>
<th>( r_{\text{max}} ) (mRNA/(s . promoter))</th>
<th>( (\delta_m \delta_p)^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhlF</td>
<td>-9.2 ( \pm ) 0.5</td>
<td>-4.7 ( \pm ) 0.3</td>
<td>2.84</td>
<td>11395</td>
</tr>
<tr>
<td>BM3R1</td>
<td>-7.6 ( \pm ) 0.2</td>
<td>-4.7 ( \pm ) 0.3</td>
<td>0.13</td>
<td>271053</td>
</tr>
<tr>
<td>AmtR</td>
<td>-8.9 ( \pm ) 1</td>
<td>-4.8 ( \pm ) 0.4</td>
<td>1.42</td>
<td>43839</td>
</tr>
<tr>
<td>LmrA</td>
<td>-8.1 ( \pm ) 0.4</td>
<td>-4.6 ( \pm ) 0.4</td>
<td>6.47</td>
<td>138009</td>
</tr>
<tr>
<td>SrpR</td>
<td>-7.4 ( \pm ) 0.3</td>
<td>-4.4 ( \pm ) 0.3</td>
<td>2.92</td>
<td>8074</td>
</tr>
<tr>
<td>HlyIIR</td>
<td>-6.4 ( \pm ) 0.3</td>
<td>-4.7 ( \pm ) 0.4</td>
<td>2.41</td>
<td>50462</td>
</tr>
</tbody>
</table>

We applied our approach to measure the \textit{in vivo} binding free energies of 6 TetR-homolog TFs (Figure 2-1B; Appendix B-Figure 1). The characterized TFs were PhlF, BM3R1, AmtR, LmrA, SrpR, and HlyIIR, from \textit{P. fluorescens}, \textit{B. megaterium}, \textit{C. glutamicum}, \textit{B. subtilis}, \textit{P. putida}, and \textit{B. cereus}[85]. The HlyIIR belonged to RutR subfamily that binds as a single dimer to its operator DNA[111]. The remaining TFs belonged to AcrR subfamily that forms a pair of dimers on its operator site[112]. Each dimer binds to one DNA strand, and the dimers likely have minimal energetic contacts[113,114]. We also systematically performed a 50% perturbation in the copy
number of the circuit, number of available RNA polymerase, and the apparent degradation rate of the TFs to estimate the maximum effect of these parameters on the binding energies within the perturbation (Appendix B-Figure 2 and 3). The results of model identifications for these TFs are reported in Table 2-1.

The calculated parameters were consistent with the known biophysics of the genetic circuits. The target promoters contained identical upstream element (UP), and similar -10 and -35 motifs[115], with no more than two nucleotide differences from the consensus motifs[115] in total. Therefore, we expected to observe similar RNA polymerase binding energy ($\Delta G_{\text{RNAP}}$) for these promoters. The calculated apparent degradation rates ($\delta_m \delta_p$) of some characterized TFs were lower than expected values for typical endogenous macromolecules (e.g. LacI[116]). These circuits contained heterologous elements, and the difference might be as a result of lower mRNA stability, protein stability, dimerization rate, and protein folding that could not be dissected by the current assay. The lower stability is common for heterologous proteins[117]. The presence of the auxiliary sites had the largest and smallest effect on the functionality of PhlF and HlyIIR constructs respectively. Consequently, the binding affinity of PhlF must be stronger than HlyIIR (Table 2-1).
Figure 2-2: Developing a sequence-expression-activity map for a signal inverter circuit containing a TF that targets multiple DNA loci. The TF target sites could be promoters or repetitive sequences that control dosage-response of the promoters. (A) Interactions among the circuit’s genetic parts and its host’s gene expression machineries. (B) The developed biophysical model of the genetic circuit. (C) Systematic modifications in PhlF circuit. The developed map correctly predicted the effect of several changes including the number of PhlF targets, copy number of the circuit, ribosome binding site (RBS) of the PhlF, growth media, temperature, and the input signal. To consider the effect of these changes, \( N, c, \Delta G_{\text{tl}}, \mu, \mu, \text{and IPTG concentration} \) (using a predictive model[118] to estimate \( r_{\text{in}} \)) were adjusted respectively; \( \beta: (k_B T)^{-1} (1.6 \text{ kcal/mol}) \); \( N: \text{number of specific TF binding sites; } d: \text{number of TF dimers on an operator DNA; } R: \text{concentration of free TF dimer; } R_c: \text{intracellular concentration of TF dimers; } c: \text{circuit copy number; } \Delta G_{\text{tl}}: \text{the binding energy of ribosome to TF’s RBS; } r_{\text{aRNA}}: \text{rate of antisense RNA production; } \phi: \text{cooperativity of RNA polymerase and replication fork; } \delta_p: \text{degradation rate of the TF’s mRNA; } \delta_p: \text{degradation rate of TF protein; } \Delta G_{\text{DNA}}: \text{binding energy of TF to its cognate DNA; } \Delta G_{\text{RNAP}}: \text{binding energy of RNA polymerase to the target promoter; } P_{\text{RNAP}}: \text{binding affinity of a free RNA polymerase to the promoter; } r_{\text{max}}: \text{rate of open complex formation on the promoter by RNA polymerase; } L_G: \text{genome size; } \mu: \text{specific growth rate; } r_{\text{out}}: \text{the output transcription rate; } \text{RNAP: RNA polymerase.} \)

Building a sequence-expression-activity map

We developed a sequence-expression-activity map (SEAMAP) to link input signal and DNA sequence of a repression circuit, containing a TF with multiple target sites, to its output transcription rate. These target sites could be located inside promoters to control their
transcription rates\cite{100,109}, or within intergenic regions to adjust intracellular concentration of the TF\cite{93}. Many factors control genetic parts’ outputs with a diverse effect, for instance from 0.75- to 100000-fold change\cite{78,87,94,96,100,109,119,120,121,122,123,124,125}. Several studies quantified the effect of individual factors\cite{96,100,109,119,122,123,125}. Here, we used transcription rate of an IPTG-inducible promoter as an input transcriptional signal to analyze the effect change in multiple factors. This allowed us to develop a system-wide model that accounts for the effect all these factors on the circuit output.

The binding competition between available RNA polymerase and cognate TFs has the largest effect on transcription rate of a promoter\cite{101,109} (Figure 2-2). RNA polymerase performs a random 3D walk to find the promoter and transcribe its adjacent DNA code\cite{126}. Although multiple intermediates have been identified\cite{127,128}, two major rate-limiting steps determine the rate of transcription from a promoter: binding of RNA polymerase holoenzyme to form a closed complex with the promoter, and creation of a transcription bubble (open complex formation) that initiates the transcription process. We quantify these steps using $\Delta G_{RNAP}$ and $r_{max}$ respectively. Binding of TFs to the promoter or its adjacent sequence changes the transcription rate by altering the binding rate of RNA polymerase\cite{101,109}. TetR-homolog TFs often bind as a dimer\cite{129}, two dimers\cite{112,114,129}, or higher-order dimers\cite{130}. The maximum number of binding dimers is typically a function of the operator size: usually 1 dimer when the operator is about 15 to 17 nucleotides\cite{129} (e.g. TetR), and 4 dimers for operators as large as 55 nucleotides\cite{130} (e.g. EthR). The intracellular concentration of free TF dimers (R), binding affinity to its cognate sites ($\Delta G_R$), and the number of bound dimers at each operator site (d) dictate the binding rate of a TF to any DNA site. Of the defined parameters, R is the only unknown variable, and the remaining parameters are the characteristics of the target promoter’s DNA sequence.
The cooperativity of many genetic factors determines the intracellular concentration of free dimer TF (R). A change in the concentration could be as a result of partitioning the TF molecules among its binding sites, or a modification at any step of gene expression, from DNA replication to folding (Figure 2-2). We used a previously proposed model of IPTG-inducible promoters[118] to estimate the transcription rate of TF's mRNA as a function of its inducer's concentration (IPTG). The model accounted for interactions among the inducer, LacI tetramers, and the TF-expressing promoter. To predict the TF translation rate, we used our developed model of translation[78,87,125] that accounted for multiple interactions between 3' 16s rRNA and the ribosome binding site (RBS) of TF's mRNA (ΔGₜ). Previous studies[119,131] had demonstrated reduction in transcription rate (to about 75%, even lower especially during time-lapse assays) when RNA polymerase moves toward replications fork counter-currently. We modified the TF's transcription rate according to the direction, φ=1 for co-current, and φ=0.75 for counter-current moves with respect to the replication fork. Changes in growth rate could alter the overall TF synthesis rate[132]. We analyzed the available measurements[96,132] at different growth conditions, and observed a good consistency between specific growth rate and protein production rate (µ₀.7; R²=0.98) (Appendix B-Figure 4).

We enumerated the major thermodynamic states during transcription and translation processes to build a sequence-expression-activity map of the repression genetic circuit in two steps. In the first step, we estimated the total number of active TF dimers by calculating its transcription and translation rates at different growth conditions (µ) and expression cassette orientations (φ). We then performed a mass balance on the number of dimers to estimate their partitioning among TF’s specific and non-specific binding sites, and calculated the number available dimers that can readily bind with the target promoter. This enabled us to calculate the intracellular concentration of free TF dimers (R) from the DNA sequence and growth condition. In the second step, we counted the number of possible ways that the TF and RNA polymerase can
bind to a pool of target promoters (copy number = c) (Appendix B-Methods). This enabled us to develop a biophysical model for estimating the change in the output transcription rate from changes in circuit copy number and free TF concentration (Figure 2-2B). Both steps together allowed us to predict the phenotype of the genetic circuit from its genotype.

We examined the developed sequence-expression-activity map by investigating the effect of changes in the analyzed factors. We first systematically altered the output transcription rate of PhlF circuits by modifying circuit copy number, PhlF transcription and translation rates, growth media and temperature. In all cases, the map correctly predicted the circuit's output (Figure 2-2C). We then studied the effect of the additional factors by altering multiple host and circuit-dependent parameters simultaneously; we re-characterized the PhlF genetic systems in E. coli pir+ (Appendix B-Table 1) in the reverse orientation (Appendix B-Figure 5), as well as in 4 additional gram-negative and gram-positive hosts (Appendix B-Figure 6) that are commonly used for industrial chemical and therapeutic drug production [52,133,134]: E. coli BL21, S. typhimurium LT2, P. fluorescens, and C. glutamicum (Appendix B-Table 1). Changing the genetic context simultaneously altered multiple factors; for instance, moving the circuit to C. glutamicum changed the number of TF non-specific sites as well as its translation rate. We also designed synthetic ribosome binding sites (RBSs) for LmrA and AmtR circuits. In total, we performed 847 measurements for the genetic circuits composed of two cascaded TFs (LacI and a TetR-homolog TF). The developed sequence-expression-activity map correctly predicted the effect of these genotypic and environmental changes on the output transcription rate (Figure 2-3A) when the inducer’s concentration (IPTG) varies.
Figure 2-3: Extensive characterization of the genetic circuits in different hosts and growth conditions. (A) Measurements (flu) vs. predictions (au) by the developed sequence-expression-activity map. (B) Pt number accounts for partitioning the TFs among their binding sites, and can encapsulate the effect of many genetic factors as a single dimensionless parameter. The circuit measurements are color coded: PhlF: black, LmrA: red, AmtR: green, SrpR: cyan, HlyIIR: blue, BM3R1: meganta. The shaded area shows the effect of two-fold shift in the x-axis; $\beta$: $(k_B T)^{-1} \ (1.6 \ kcal/mol)$; $c$: circuit copy number; $R_t$: total number of TF dimers; $L_G$: genome size; $N$: number of TF specific binding site; $\Delta G_R$: TF binding energy; flu: fluorescence unit; au: arbitrary unit

**Dimensionless design space of genetic circuits (Pt number)**

When a circuit’s genetic part is swapped or altered, the corresponding genetic parameters changes. In order to preserve circuit output, additional modifications to the circuit DNA or growth condition is required. However, altering one factor causes a chain of modifications in the others which could dampen or uplift the intended adjustment; for instance, changing a promoter sequence to increase $\Delta G_{RNAP}$ may simultaneously alter $\Delta G_R$ as well.

To reduce the cooperativity among the unknown parameters of the genetic parts during circuit design, we defined a dimensionless number, called Pt after Mark Ptashne. The Pt number encapsulates several design factors, including number of specific and non-specific TF binding
sites, dosage of target promoter, TF binding energy to the target, and the number of active TF dimers. This number quantifies the binding affinity of a TF dimer molecule to its target promoter with respect to binding affinity to all specific and non-specific targets (Figure 2-3B), and is an analog of the Damköhler dimensionless number in transport phenomena[135]. To our knowledge, Pt is the first introduced dimensionless number in synthetic biology. The Pt number allows us to transform the cooperative design space of a genetic circuit into a space, called Pt design space, such that activity of each transcriptional regulatory element is represented by a dimensionless number. This minimizes the number of unknown parameters that must be adjusted independently. As a demonstration, we converted the design space of the 847 characterized genetic systems into a one-dimension Pt space to estimate their output transcription rate solely as a function of Pt. The results demonstrated Pt as a single parameter that controls performance of simple repression circuits (Figure 2-3B).

Using the dimensionless transformation, we developed a systematic framework for constructing genetic circuits in two fold, circuit design and DNA compilation. For designing a circuit, we first identify a set of design criteria in the dimensionless design space that results in correct circuit output. This enables us to design genetic circuits, independent of genetic context and environmental conditions, as a set of dimensionless numbers (Appendix B-Methods). We then use our developed sequence-expression-activity map to compile the circuit’s DNA sequences according to the calculated criteria.
Figure 2-4: Design of genetic opAmp circuits. (A) A genetic opAmp is composed of two cascaded TFs to (1) intensify an input transcriptional signal and (2) increase its dynamic range (two criteria). The output dynamic range is a function of Pt values for the individual genetic parts. (B) Pt design space of genetic opAmps. The blue area shows the Pt values for the genetic parts that could result in over 50% increase in the dynamic range of an input signal, generated by an IPTG-inducible promoter. The circles demonstrate the constructed circuits. The colored lines show the change in the Pt number of both genetic parts when the input transcription rate changes. (C) Comparison between the measurements and predictions by SEAMAP for the circuits. As expected, SA, SP, LP, and SL circuits amplify the transcriptional signal up to 20 fold. The numbers under the orange half circles show the translation rate of each transcription factor (protein/ (s. mRNA)). $r_{in}$, $r_1$, $r_{out}$ are the transcription rate of the input, internal, and output promoters respectively. Pt*$_1$: Pt$_1$ X min (1, $r_{out,max} / r_{in,max}$); Pt*$_2$: Pt$_2$ X $r_{1,max} / r_{in,max}$; SL: SrpR-LmrA; LP: LmrA-PhlF; SP: SrpR-PhlF; SA: SrpR-AmtR; SB: SrpR-BM3R1; LB: LmrA-BM3R1; RBS: ribosome binding site; YFP: yellow fluorescent protein.
De novo design of genetic opAmps in dimensionless design space

A large fraction of natural genetic sensors have a low output dynamic range[136,137,138] that may not be sufficient to trigger correct outputs for many synthetic circuits[85,92,139]. This mismatched output-input transfer function is a challenge in design of complex genetic circuits, and an effective solution to this problem is using genetic opAmp (operational amplifier) circuits[140]. These circuits intensify the signal and increase its dynamic range simultaneously.

We designed and created a family of genetic opAmp circuits by pair-wise assembly of two consecutive signal inverter circuits to amplify a transcriptional input signal (Figure 2-4A). The dynamic range of the output signal is controlled by transfer function of each inverter circuit while the maximum transcription rate of output promoters ($r_{\text{out, max}}$) determines the maximum theoretical intensity of the output signal. This maximum intensity may not be available for many circuits (Appendix B-Figure 7); however Rail-to-Rail opAmps could reach the theoretical value, where cross- and retro-activities are at their minimal level. In total, the choices of the genetic parts and their internal elements such as RBS and promoter, genetic context such as host and expression vector, and growth condition such as media and temperature determines the circuit's functionality, and many configurations result in signal attenuation.

According to our developed map, 25 unknown parameters must be adjusted during the design of the opAmp circuit including $\Delta G_{R,TF1}$, $\Delta G_{R,TF2}$, $\Delta G_{RNAP,TF1}$, $\Delta G_{RNAP,TF1}$, $r_{\text{max,1}}$, $r_{\text{max,2}}$, $\Delta G_{\text{tl,TF1}}$, $\Delta G_{\text{tl,TF2}}$, $\phi_{TF1}$, $\phi_{TF2}$, $d_{TF1}$, $d_{TF2}$, $\delta_{m,TF1}$, $\delta_{m,TF2}$, $\delta_{p,TF1}$, $\delta_{p,TF2}$, $r_{\text{asRNA,TF1}}$, $r_{\text{asRNA,TF2}}$, $N_{TF1}$, $N_{TF2}$, the target host (16S rRNA sequence), $L_{G}$, $P$, $\mu$, and $c$. Several design parameters are cooperative and any attempt to change the corresponding DNA sequence of one parameter may affect the others such as $\Delta G_{R,TF1}$ and $\Delta G_{RNAP,TF1}$, $\Delta G_{R,TF2}$ and $\Delta G_{RNAP,TF2}$, $\Delta G_{\text{tl,TF1}}$ and 16S rRNA, and $\Delta G_{\text{tl,TF2}}$ and 16S rRNA. Imposing additional criteria may reduce the number of unknowns, for instance to 21 when the circuit-harboring host and growth condition are pre-set. Further reduction
may be possible by imposing more assumptions however may result in completely defective circuit, for example using BM3R1 and its cognate promoter as the second genetic part (Figure 2-4). In addition, the circuit may need to be re-designed when the context changes.

To simplify the forward engineering of the Rail-to-Rail genetic opAmp, we transformed the multi-dimensional design space into a Pt design space. This reduced the number of design parameters to two dimensionless unknowns representing the regulatory state of the genetic parts (Figure 2-4A). We identified the range of Pt values that resulted in amplification of an input signal generated by an IPTG-inducible promoter (Figure 2-4B). We then constructed four genetic opAmp circuits using combinations of different TFs that have Pt values varying across the amplification area (Figure 2-4B). All four circuits successfully amplified the input signal up to 20-fold. We also designed two circuits that did not follow the predicted safe-design area. Characterization of these circuits showed about 5-fold reduction in the signal (Figure 2-4C). These results are consistent with our prediction, and demonstrate the ability of the Pt number to simplify design of genetic circuits with multiple co-dependent parameters.

In total, the circuits contained three cascaded TFs (LacI and two TetR-homolog TFs) that pump up the change in the intracellular concentration of the inducer (IPTG) as the output transcription rate. We used the developed sequence-expression-activity map (SEAMAP) to simulate the outcomes of the individually designed circuit configurations. The results confirmed that the map can correctly predict the functionality of these circuits (Figure 2-4C).
Figure 2-5: Designing several example analog and digital genetic circuits in Pt dimensionless design space. We used Pt number to transform design space of four digital (NOT, NOR, NAND, and AND gates) and two analog (toggle switch and analog-digital convertor) circuits. In these circuits, the input signals are transmitted through one or multiple independent and cascaded transcription factors. The transformed Pt-design space for NOT- and NOR-gates reveals that Pt₁ and Pt₁+Pt₂ numbers (assigning a Pt value for each input signal) must pass through the shaded regions respectively to maximize input-output fold-change by the circuits. For the toggle switch, the red area shows the presence of multiple states that can be targeted to switch ON and OFF the output transcription rate. For the remaining circuits, successful designs should have maximum change in output circuit functionality, for instance by following the white arrows. The slope of the arrow can be simply adjusted by altering translation rate of each transcription factor. Pt₁ and Pt₂ values can be proportionally altered by changing the input transcriptional signals. Pt₁, max and Pt₂, max are the maximum Pt values for TF1 and TF2 respectively. OTR: the fold-change ratio in the output transcription rates, equal to the maximum of r₁/r₂ and r₂/r₁.

Discussion

We created a sequence-expression-activity map (SEAMAP) that predicts the effect of changes in DNA sequence on outputs of TF-containing genetic circuits. To build the map, we identified the growth-dependency of the interactions among circuits’ genetic parts and their host's gene expression machineries. This allowed us to create a system-wide model that quantified the effect of several parameters including circuit copy number (c), binding energy of RNA
polymerase ($\Delta G_{\text{RNAP}}$) and transcription factors ($\Delta G_{R}$) to their cognate promoters, rate of transcriptional bubble formation ($r_{\text{max}}$), cooperativity of TF dimers ($d$), number of TF specific (N) and non-specific ($L_{G}$) binding sites, binding energy of ribosome ($\Delta G_{\text{t}}$) to its genetic parts, orientation of expression cassette with respect to replication fork ($\phi$), change in 16S rRNA, rate of antisense RNA production ($r_{\text{asRNA}}$), degradation of TF’s mRNA ($\delta_{m}$) and TF’s protein ($\delta_{p}$), number of available RNA polymerase (P), and growth condition (i.e. media and temperature) ($\mu$). Changes in the governing factors could have over 1000-fold variation on genetic systems’ performance (Figure 2-3; Appendix B-Figure 5). Building a SEAMAP allows us to consider the effect of these environmental and genetic context parameters during design of a genetic circuit’s DNA sequence.

A key limitation in building SEAMAPs for TF-based genetic circuits was the absence of reliable protein-DNA binding affinity measurements which we addressed by developing a high-throughput in vivo method. This approach only required characterization of a simple repression cassette in the presence of multiple TF target sites, followed by a model parameterization (Figure 2-1; Methods) that determines RNA polymerase and TF binding energies to the promoter (Table 2-1). We examined the calculated in vivo energies by performing 847 experiments for several genetic systems at different growth conditions in 6 gram-positive and gram-negative hosts including E. coli DH10B, E. coli pir+, E. coli BL21, S. typhimurium LT2, P. fluorescens, and C. glutamicum (Appendix B-Table 1). The biophysical model correctly predicted the activity of these genetic systems (Figure 2-3A; Appendix B-Methods). Accurate measurement of in vivo energies is key for precise and reliable gene expression prediction.

We then used the SEAMAP to develop a systematic design framework for building genetic circuits. The choices of genetic parts and their internal elements, DNA context, and environmental conditions provide a complex combinatorial design space where change in one parameter may cooperatively alter functionality of the others. The developed framework dissects
design of circuit from its DNA sequence in two steps. The first step is transforming the cooperative design space to a dimensionless space such that activity of each transcriptional regulatory element is determined by a dimensionless number called Pt. The Pt number encapsulates the effect of all the identified parameters controlling TF binding to DNA loci. This allows us to determine circuit outcomes as a function of Pt values for all the regulatory reactions, independent of genetic contexts and environmental conditions, and identify a set of safe-design criteria for each genetic part as a function of its Pt values (Appendix B-Methods). The second step is employing the previously described sequence-expression-activity map that can translate these Pt design criteria to DNA sequence of each genetic part for a desired context.

We utilized the framework to design several analog and digital circuits (Figure 2-4 and 5; Appendix B-Methods). The dimensionless transformation reduced the number of parameters in design space of a NOT-gate, NOR-gate, NAND-gate, AND-gate, analog-digital convertor and toggle switch from 16, 16, 25, 35, 25, and 25 dimensional unknowns to 1, 2, 2, 3, 2, and 2 dimensionless unknowns respectively. As a demonstration, we forward engineered a family of genetic opAmp circuits. These circuits transmitted an input transcriptional signal through two cascaded TFs to amplify the signal up to 20-fold. The transformation reduced the number of unknown parameters from 25 to 2. In this new design space, we identified a set of safe-design criteria for successful amplification. These biophysical design criteria successfully distinguished defective circuits from signal amplifiers (Figure 2-4).

In total we constructed genetic circuits composed of one, two, or three transcription factors, and examined them in different genetic and environmental contexts. In all cases the developed SEAMAP correctly predicted the circuit outputs from the changes in one or more of the input signals (inducer concentration), genetic parts, and growth conditions. Similarly, we recently demonstrated the power of such biophysics-driven sequence-expression-activity map in simplifying the forward engineering of metabolic pathways[87]. Together, both works
demonstrate the capability of biophysical modeling to design and predict outcomes of complex genetic systems at their DNA level using available biophysical knowledge.

Currently computer-aided design (CAD) programs greatly help researchers to search, copy-paste, and assemble independently characterized genetic parts[61,141,142]. However, these parts are usually characterized under a specific and narrow range of environmental conditions and genetic contexts. As we have shown here, the transfer function of genetic parts varies when any system parameter changes, resulting in lower design reliability of such CAD-based programs especially when aiming to build larger genetic systems, or using them in different hosts. In contrast, designing in dimensionless space enables us to code a genetic system as a set of independent dimensionless numbers that can be later translated on demand to DNA sequence for desired genetic contexts and growth conditions using sequence-expression-activity maps.

**Materials and Methods**

**Genetic circuit constructs**

The cloning was started with pRiboJ-PhlF, pRiboJ-AmtR, pRiboJ-LmrA, pSarJ-BM3R1, pSccJ-SrpR, pVtmoJ-HlyIIR plasmids[85]. The additional TF binding sites were inserted on the each plasmid using digestion/ligation. The insert DNA for each of BM3R1, SrpR, and HlyIIR circuits was designed as a 500 nucleotide DNA block containing 4 TF target sites separated by about 80 nucleotide randomized DNA and a restriction enzyme, and was purchased from Integrated DNA Technologies (Coralville, Iowa) as a gBlock. A multi-cloning site was inserted on the plasmids to add EcoRI, AvrII, NheI, SacI, AflIII, SpeI, and XbaI binding sites using XhoI and SphI restriction enzymes. The DNA inserts were digested by XhoI and one of EcoRI, AvrII, SpeI, and SacI to create small DNA sequences containing one, two, three, and four TF binding
sites. These sequences were ligated to a gel-purified vector containing the genetic circuits digested by XhoI and one of EcoRI, AvrII, and SacI enzymes. The vector contained p15A replication origin, Amp\textsuperscript{R} antibiotic resistance gene, LacI, a gene expressing the proper TF, and YFP driven by a TF-targeted promoter. The insert DNA for AmtR, PhlF, and LmrA containing one or two binding sites was created using PCR assembly, and ligated to a gel-purified vector digested by XhoI and SphI restriction enzymes. The circuits with 3 or 4 auxiliary binding sites were created by a follow up digestion/ligation using XbaI and SphI restriction enzymes.

The effect of change in genetic context was studied by modifying the circuit sequence, the vector, and harbored hosts. RBS libraries were designed for AmtR, PhlF, and LmrA genes using The RBS library Calculator\cite{87}, and degenerate oligomers were purchased from Integrated DNA Technologies. Plasmid libraries were constructed using Gibson DNA assembly\cite{17}, and selected colonies were sequence-confirmed. The PhlF circuits containing no, 2, and 4 additional binding sites were PCR extended, and cloned in a reverse direction on a pBac vector containing R6K and F-origin, and Kan\textsuperscript{R} antibiotic resistance gene. These circuits were then transformed into both \textit{E. coli} DH10B and \textit{E. coli} pir\textsuperscript{+} that over-expresses \pi-protein as a part of R6k replication origin. Selected PhlF plasmids were transformed into \textit{S. typhimurium} and \textit{E. coli} BL21, also sub-cloned into a pSEVA351 vector (Genbank JX560335), and electroporated into \textit{C. glutamicum} and \textit{P. fluorescens}.

Genetic opAmps were created by assembling individual circuits. The RBS sequence of each TF was altered by PCR extension of its DNA sequence with a proper RBS sequence. The PCR product was re-inserted into the original vector using Gibson DNA assembly\cite{17}. The new circuits were then individually PCR-amplified and assembled by a second round of Gibson DNA assembly. Isolated colonies were selected and sequence-confirmed.
Growth and Measurements

Each construct and a proper negative control (an empty vector) were individually grown overnight at 200 R.P.M in a 96 deep well plate containing appropriate rich growth media and antibiotic. 5 µl of each well was transformed into a selected media in a 96-well microplate. The plate was incubated using a TECAN plate reader. To maintain steady state growth at log growth-phase, two subsequent serial dilutions were performed when OD$_{600}$ reached to about 0.2. The cultures from the third plate were transformed into a microplate containing Phosphate Buffered Saline (PBS) and 2 mg/ml Kanamycin, and fluorescence distribution of at least 20,000 cells was recorded by a LSR-II Fortessa flow cytometer (BD biosciences). The fluorescence content for each well was determined by subtracting an average auto-fluorescence from the arithmetic mean of the recorded distribution. A detailed characterization protocol was reported by Farasat et al.[87].

Estimating the measurements on absolute scales

The previously reported translation rate for LacZ was 0.31 protein / (s. mRNA)[143], the equivalent of 732 au by our model of translation[87]. Their ratio provides a conversion factor to estimate our model’s output on an absolute scale. Therefore, we approximated the translation rate of the TF genes (protein / (s. mRNA)) using Eq. 1 where $\Delta G_{t}$ was the binding energy of ribosome to the mRNA, calculated by our model:

$$\text{translation rate} = 1.1e^{-0.45\Delta G_{t}}$$

(1)

To calculate the transcription rate of the IPTG-inducible promoter on an absolute scale, we used characterization data of E. coli’s native lac promoter[143]. At growth rate of 0.8 doubling/hr and in the presence of 0.5 mM IPTG, the reported synthesis rate of the native
promoter was about 0.3 mRNA / s [143]. At this growth rate, transcription rate of this native promoter was about half of a commonly used synthetic promoter, PlacUV5[144], and this synthetic promoter had a similar activity with Ptac promoter[145]. Therefore, we reasoned that the transcription rate for a single copy of the utilized IPTG-inducible promoter must be about 0.6 mRNA / s at 0.5 mM IPTG. This allowed us to convert the measured YFP production rates to mRNA synthesis rate for the promoter.

We used the calculated YFP conversion factor to quantify the output transcription rate ($r_{\text{max}}$) of TF-targeted promoters on their absolute scales. This required imposing further assumptions about the system; the presence of a RiboJ[139] upstream of YFP coding section may alter its mRNA stability in control construct by modifying accessibility of RNaseE to the mRNA. Besides, this additional sequence may also slightly alter translation rate of YFP by changing mRNA secondary or tertiary structure. These factors may proportionally change all the reported $r_{\text{max}}$ values (Table 1) but unlikely, alter the order of magnitude of these values. Although, we discounted the effect of these factors during our calculations, the estimated $r_{\text{max}}$ values allowed us to correctly predict the output transcription rate of genetic opAmp circuits.

**Measuring In vivo binding energy of transcription factors**

The developed approach included three steps: construction, characterization, and model parameterization. In the first step, a simple-repression genetic circuit was constructed where a TF targets its cognate promoter driving YFP. The transcription rate of the TF was controlled by an IPTG-inducible promoter. In theory, one auxiliary site is sufficient to estimate the energies while additional circuits could reduce experimental noise and improve accuracy of the method. Up to 8 additional sites for PhlF and up to 4 additional sites for the remaining circuits were placed outside expression cassettes, and separated by about 25 to 80 nucleotide randomized DNA. In the second
step, IPTG concentration was varied from 0 to 1 mM, and the output fluorescence production rate of the constructed circuits was measured using flow cytometry. In the final step, a model parameterization was applied to calculate TF and RNA polymerase binding energies. The model contained 5 unknowns: rate of open complex formation ($r_{\text{max}}$), rate of antisense RNA production ($r_{\text{asRNA}}$), TF binding energy ($\Delta G_R$), RNA polymerase binding energy ($\Delta G_{\text{RNAP}}$), and apparent degradation rate of TF ($\delta_m\delta_p^{-1}$). To estimate $r_{\text{max}}$, the measured YFP production rate at low IPTG concentration were averaged, placed in the thermodynamic model, to approximate the transcription rate when TF concentration approaches to zero. The remaining parameters were varied from 0 to 0.5% of maximum input transcription rate, -12 to 0 kcal/mol, -12 to 0 kcal/mol, and 10000 to 1000000 s$^{-2}$ respectively, the output transcription rate for each circuit was calculated using the developed thermodynamic model, and the values were identified that minimized error of prediction by the model.

The energy values were then corrected to avoid artificial error amplitude for measurements with low fluorescence level, close to background level. During the correction, a range for each energy value was identified that resulted in errors within 10% of the calculated minimal model error. The range was later screened visually to find the best energy values that represent the measurements. Two types of error could reduce the accuracy of the calculated energies. The first type was due to the utilized nonlinear parameterization algorithm. To compensate for this error, standard deviation of all energy values that corresponded to model errors within 10% of the minima, were calculated. The second error type was due to uncertainty in model parameters that was quantified by performing a 50% perturbation in the number of available RNA polymerases, ($\delta_m\delta_p^{-1}$), and circuit copy number. For each perturbation, each energy value was varied from -12 to 0 kcal/mol, and maximum and minimum prediction error was calculated to create two max and min datasets (Appendix B-Figure 2 and 3). These datasets were screened to identify the binding energies with least prediction errors, and calculated
standard deviation of all the local optimal energies as a measure for parameter uncertainty. The reported standard deviations in Table 2-1 are the largest values between these two calculations.
Chapter 3 A Biophysical Model of CRISPR/Cas9 On- and Off-Target Activity for Rational Design of Genome Editing and Gene Regulation

Introduction

The RNA-mediated Cas9 adaptive immune system (CRISPR type II) has revolutionized genome engineering by enabling the precision cutting of DNA that can be customized to target any sequence [57,146,147,148,149,150], while being functional in a broad range of prokaryotes and eukaryotes, including bacteria, yeast, flies, fish, plants, worms, monkeys, mice, rats, rabbits, frogs, and human cell lines [28,147,151,152,153,154,155,156,157,158,159,160,161]. By forcing the host to repair these precision DNA cuts, the CRISPR/Cas9 system allows recombinant DNA to be inserted at desired genome locations, and therefore can be used for performing high-throughput gene knockouts, loss-of-function screening, artificial immunization, removal of latent genome-encoded viruses, and site-specific gene therapy applications [162,163,164,165]. A nuclease-deficient version of Cas9, called dCas9, retains its RNA-guided DNA binding activity and has been used as a transcription factor to tightly control gene expression levels and rewire a host's transcriptional regulatory network [166]. Multiple dCas9-based repression and activation devices, including within layered genetic circuits, have been developed in bacteria, yeast, and mammalian cells; these genetic circuits can regulate a targeted promoter's transcription rate by up to 1000-fold [149,167,168,169,170]. In principle, the expression of multiple guide RNAs, working with dCas9, enables the regulation of many promoters simultaneously, and provides an almost limitless source of programmable transcription factors.

Based on recent observations, the CRISPR/Cas9/dCas9 system is highly versatile, but has imperfect specificity and activity under a wide range of environmental and genotypic conditions [168,171,172], motivating a study of its mechanisms and the development of a model to
rationally design its guide RNAs [164]. One major challenge has been binding to off-target DNA sites, resulting in off-target mis-cutting of genomic DNA by Cas9 or gene mis-regulation by dCas9 [171,173,174,175,176]. Several strategies have been shown to reduce Cas9 off-target behavior by manipulating its cleavage activity [27,176,177,178,179,180,181,182]. For example, two guide RNAs expressed together with a partially nuclease-deficient Cas9 nickase have been used to make two single-strand cuts at adjacent locations, increasing the rate of on-target repair by homologous recombination [27]. Further, fusing dCas9 to the FokI nuclease increased the specificity of its nuclease activity to a 20 bp recognition sequence [182]. These strategies address off-target cutting, but not off-target binding and gene regulation. A system-wide understanding of how guide RNAs work together with Cas9/dCas9 to control off- and on-targeting binding would enable the rational design of guide RNAs, and other controllable factors, to improve Cas9/dCas9 specificity and activity. In particular, when engineering dCas9-based genetic circuits, it will be desirable to modulate dCas9’s ability to regulate gene expression through the introduction of guide RNA mismatches [152]. However, the quantitative relationship between guide RNA sequence and dCas9’s binding affinity is currently unknown.

In this section, we develop a comprehensive, mechanistic model of CRISPR/Cas9 that predicts how experimental conditions and guide RNA sequences (crRNAs) control target site selection and cleavage activity. To initially parameterize this model, we analyze the large amount of structural, biochemical, and next-generation sequencing data that has recently measured several aspects of CRISPR/Cas9’s function with different crRNAs under varied experimental conditions [148,172,176,178,180,181,183,184,185,186]. We formulate a single system-wide model that explains how these disparate observations can originate from the same CRISPR/Cas9 mechanism of function. We also present quantitative criteria for designing guide RNA sequences with targeted binding and cleavage activities. By accounting for several important factors beyond the guide RNA sequence, our design rules are a significant improvement over existing, and
somewhat contradictory, sequence design rules whose outcomes have also depended on the
selected experimental conditions [152,164,176,180,184].

Table 3-1, we summarize the several types of experimental measurements used to
parameterize and validate the model, including the number of degrees of freedom and the number
of data-points in each experimental data-set. To identify a narrow range of best-fit parameter
values, model parameterization was performed by using either a simple simplex method
(fminsearch) or a Levenberg-Marquardt method (lsqnonlin) in MATLAB to minimize the sum of
squared relative errors, followed by a parameter sensitivity analysis and visual comparisons to
more precisely identify best-fit model parameters.

Table 3-1: A summary of all studies used to estimate the Cas9 model's parameters

<table>
<thead>
<tr>
<th>Parameter Used in Description</th>
<th>Parameter Used in Description</th>
<th>Reference</th>
<th>Assay</th>
<th>Data used</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔG_{PAM}</td>
<td>Table 3-3</td>
<td>Current study</td>
<td>Flow cytometry</td>
<td>Multiple on-target sequences</td>
</tr>
<tr>
<td>k_d, k_c, k_f, k_{dr}, k_i</td>
<td>Fig. 2</td>
<td>Jiang et. al.[147] Sternberg et. al. [181] Sternberg et. al. [181] Szczelkun et. al.[186]</td>
<td>Deep sequencing Phosphoimaging DNA gel single-tethered DNA curtains Single-molecule DNA supercoiling</td>
<td>Randomized PAM λ2 target on plasmid &amp; synthesized DNA Single molecule and bulk assays</td>
</tr>
<tr>
<td>ΔG_{supercoiling}</td>
<td>Fig. 2</td>
<td>Depew and Wang[187] Wang and Peck[188]</td>
<td>Distribution of topoisomers</td>
<td>--</td>
</tr>
<tr>
<td>Δσ</td>
<td>Fig. 3</td>
<td>Current study</td>
<td>Flow cytometry</td>
<td>Multiple on-target sequences</td>
</tr>
<tr>
<td>ΔG_{exchange} (dataset I)</td>
<td>Fig. 4</td>
<td>Pattanayak et.al.[176]</td>
<td>Deep sequencing</td>
<td>CLTA1, CLTA2, CLTA3</td>
</tr>
<tr>
<td>ΔG_{exchange} (dataset II)</td>
<td>Fig. 4</td>
<td>Hsu et. al.[180]</td>
<td>Deep sequencing</td>
<td>Emx1.1, Emx1.2, Emx1.3, Emx1.6, Emx1.10, Emx1.11, Emx1.12, Emx1.13, Emx1.14, mx1.15, Emx1.16, Emx1.17, Emx1.18, Emx1.19, Emx1.20 gRNA1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mali et. al.[183] Pattanayak et.</td>
<td>Deep sequencing</td>
<td>CLTA1, CLTA2, CLTA3</td>
</tr>
</tbody>
</table>
To develop this model, we employed statistical thermodynamics and the law of mass action to formulate a five-step mechanism that accounts for concentration-dependent, cell volume-dependent, host genome-dependent, and crRNA-dependent changes to Cas9 complex formation, diffusion, target specificity, and target activity (Figure 1). Kinetic and thermodynamic constants were estimated by analyzing six studies of Cas9/dCas9 function (Table 3-1). We validated this model using in vitro Cas9-dependent cleavage rate data (Figure 2), obtained by Sternberg et al. [181], together with new data collected in this study, measuring in vivo dCas9-dependent transcriptional repression in synthetic genetic circuits within bacterial cells (Figure 3). Further, to predict how a guide RNA controls target specificity, we used deep sequencing data [147,176,180,183] to compile a position-dependent, nearest neighbor binding model that accounts for canonical and non-canonical PAM recognition sites, R-loop formation, and mismatches with DNA target sites (Figure 4). By performing a sensitivity analysis on the model, we show the optimal experimental conditions to maximize on-target Cas9 cleavage activity and minimize Cas9/dCas9 off-target binding (Figures 5, 6). Finally, we applied the model to predict Cas9 cleavage off-target activity, and the resulting locations of undesired genomic indels, when using crRNAs to remove latent HIV viral DNA segments from a human cell line [91]. As a key conclusion, we present the optimal experimental conditions to minimize off-target cleavage activity, and calculate the necessary deep sequencing coverage to correctly identify 99% of genomic indels formed (Figure 7).
Figure 3-1: The multi-step mechanism responsible for Cas9-mediated DNA site cleavage. (A) Each crRNA strand is expressed with rate \( r_{crRNA} \). The active crRNA is formed by either hybridization of an expressed tracrRNA with an expressed precrRNA or by direct expression of a synthetic guide RNA (sgRNA). The Cas9 endonuclease is expressed with rate \( r_{Cas9} \). (B) Cas9 binds to the mature crRNA with a forward kinetic association constant \( k_b \). Cas9 may also form dimers with the Cas9:crRNA complex with a forward kinetic association constant \( k_{di} \). After loading the crRNA, the structure of the Cas9:crRNA undergoes an isomerization with forward kinetic constant \( k_I \) to create a DNA channel. \( N_{crRNA} \), \( N_{Cas9} \), \( N_{intermediate} \), and \( N_{Cas9:crRNA} \) are their numbers of molecules. (C) The resulting active complex performs a 3D random walk with molar flow rate \( r_{RW} \). The probability that it binds to a DNA site is determined by the site sequence, including the presence of a protospacer adjacent motif (PAM), the number of same-sequence DNA sites (\( N_{target, j} \)), and their binding free energy (\( \Delta G_{target, j} \)). (D) The formation of a stable Cas9:crRNA:DNA complex occurs in several steps: Cas9:crRNA recognizes the PAM site, unwinds the DNA duplex, and sequentially replaces DNA:DNA base pairings with RNA:DNA base pairings in an exchange reaction to form a DNA:RNA:DNA complex, called an R-loop. The DNA target site's binding free energy to Cas9:crRNA (\( \Delta G_{target} \)) sums together its PAM interaction energy (\( \Delta G_{PAM} \)), the energy needed to unwind the supercoiled DNA (\( \Delta \Delta G_{supercoiling} \)), and the crRNA:DNA exchange energy during R-loop formation (\( \Delta \Delta G_{exchange} \)). During these steps, the Cas9:crRNA:DNA complex may dissociate with first order kinetic constant \( k_d \) or it may be cleave the bound DNA site with pseudo first order kinetic constant \( k_C \). After cleavage, the Cas9:crRNA:DNA complex remains bound to the cleaved DNA, and is considered a no-turnover enzyme. Additional model parameters include the DNA replication rate (\( \mu \)) and the degradation or dilution rates of Cas9 (\( \delta_{Cas9} \)), crRNA (\( \delta_{crRNA} \)), and Cas9:crRNA complex (\( \delta_{Cas9:crRNA} \)).
Results

Model parameterization using in vitro measurements

We utilized the in vitro measurements obtained by Sternberg et. al. [181] to determine the kinetic parameter values that quantify Cas9:crRNA complex formation, pre-cleavage dissociation, and Cas9-dependent cleavage. In this study, the binding locations and cleavage rates of Cas9 were measured in solution and across a plasmid DNA to characterize the multi-step process by which Cas9 finds DNA targets, initiates R-loop formation, and cleaves DNA sites. Here, we utilized the authors' dynamic measurements of DNA site cleavage at different concentrations of Cas9 and crRNA, using either an on-target site on plasmid DNA (Figure 2B in [181]) or an on-target site on a double-stranded DNA fragment (Extended Data Figure 5 in [181]). By assuming Cas9's characteristic length of $\lambda_{\text{Cas9}} \approx 150 \, \text{Å}$ [185,189], we estimated the number of microstates that Cas9 can perform a random walk to them, about $3 \times 10^{15}$ microstates in a 10 µl reaction solution. Based on the buffer's diffusivity of 45 µM$^2$/s [190], we approximated that the diffusive flow rate of a Cas9:crRNA molecule to any defined microstate in the solution was about $4.05 \times 10^{-10}$ 1/sec during Sternberg et. al.'s experiments (Equation 5). The presence of the high copy plasmid DNA in the solution (25 nM plasmid DNA, $1.5 \times 10^{11}$ individual plasmid molecules) increased the collision rate between a Cas9:crRNA and a plasmid DNA molecule. Using these numbers, we estimated that the Cas9:crRNA molecule's random walk rate to any plasmid molecule was 61 1/sec.

We then determined the kinetic parameter values controlling Cas9:crRNA association ($k_a$), isomerization ($k_i$), and dimerization ($k_d$), pre-cleavage dissociation ($k_d$), and cleavage activity ($k_c$) by calculating the rate of cleavage ($r_c$) across a range of Cas9 and crRNA concentrations, mirroring the experimental conditions, and comparing to experimental cleavage
measurements (56 experiments; $R^2=0.97$; **Appendix C-Figure 1**) using the 25 nM of plasmid DNA as template [181]. The model solution was evaluated for an initial 10 minute time period, followed by *in silico* addition of the DNA substrate and an additional 30 minute time period. The best-fit kinetic parameter values were then determined through optimization to minimize the relative error between calculated and measured cleavage rates (**Methods**). Based on our analysis (**Appendix C-Figure 2**), we could uniquely parameterize $k_r$, $k_i$, $k_{di}$, and the ratio $k_c/k_d$ (**Table 3-2**). Surprisingly, the rate of cleavage was found to be less than the rate of pre-cleavage dissociation ($k_c/k_d \ll 1$), suggesting that (d)Cas9 must engage in multiple aborted rounds of binding and R-loop formation before successfully cleaving the DNA site. Using the best-fit parameter values, the model was able to accurately capture the experimentally observed time-dependent cleavage rates while varying the Cas9 and crRNA concentrations (**Figure 2A**).

As expected, when the Cas9 concentration is limiting, the calculated amount of cleaved DNA is almost equal to the Cas9 concentration because Cas9 does not turn-over after cleavage takes place. However, when the same cleavage reaction was conducted using short relaxed DNA fragments instead of negatively supercoiled plasmid DNA [181], the total cleavage activity dropped about fivefold while the time-dependent rate of DNA cleavage increased (**Figure 2B**). Sternberg et. al. hypothesized that the lower total cleavage activity was a result of partial enzymatic activity of Cas9 but left the improved cleavage rate observation unexplained. According to our calculations, the Cas9 binding probability to the correct target in the short DNA fragment was close to 1, but this would overshoot the improvements in the cleavage rates. Interestingly, the unexplained difference in DNA cleavage rates was consistent with the effects of negatively supercoiled DNA on R-loop formation, suggesting that changes in the DNA templates' superhelical densities may be responsible. To test both Cas9 partial activity and supercoiling effect hypotheses, we first reduced Cas9 concentration in our *in silico* calculations to about 20% of the reported concentrations, then compared the apparent drop in time-dependent cleavage rate
on relaxed DNA to the calculated difference in cleavage as a result of the free energy penalty \( \Delta \Delta G_{\text{supercoiling}} \) (Equation 12). The apparent drop in cleavage rate was equivalent to a free energy penalty of 0.72 k_B T during R-loop formation on a 20 nucleotide DNA site. The energy difference is remarkably similar to the supercoiling energy, calculated by equation 12 when the initial superhelical density varied from -0.06 (supercoiled DNA) to 0 (relaxed DNA). This correctly explains how increasing Cas9 concentration affects cleavage activity on relaxed DNA sites (Figure 2C). The results demonstrated that the \( \Delta \Delta G_{\text{supercoiling}} \) may be correctly estimated by this equation for relaxed, plasmid, or genomic DNA sites with varying superhelical densities (\( \sigma_I \)).

**Figure 3-2: In vitro validation of the Cas9 mechanistic model.** Recording the time-dependent cleavage was started by adding target DNA to equimolar mixtures of Cas9 and crRNA after 10 minutes pre-incubation. (A) Cleavage rate predictions (line) and measurements (circle) for 25 nM supercoiled target DNA. (B) Comparison between predicted (black) and measured (green) total cleavage rate for non-supercoil target DNA after a long time cleavage reaction. (C) Cleavage predictions (line) and measurements (circle) for non-supercoiled target DNA. Data points from Sternberg et. al. [181].

**Table 3-2: Parameter values used in this study**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Calculated/ Used Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_{NS} )</td>
<td>-0.06</td>
<td>kcal/mol</td>
</tr>
<tr>
<td>( k_d )</td>
<td>0.08</td>
<td>s(^{-1})</td>
</tr>
<tr>
<td>( k_c )</td>
<td>0.022</td>
<td>hr(^{-1})</td>
</tr>
</tbody>
</table>
The ratio between \( k_c \) and \( k_d \) determines steady-state cleavage rate from each target (most variation in the cleavage rate), and not their absolute values \((k_c/k_d=0.0016)\). To estimate the individual kinetic parameters, we used the residence time of \((d)\text{Cas9}\) on a relaxed DNA, shown to be 130 seconds[186]. This was the equivalent of 27.7 \( \text{hr}^{-1} \) \( \text{Cas9:crRNA} \) dissociation rate. We employed equation 11 to inversely calculate \( k_d \) (13.5 \( \text{hr}^{-1} \)), then used the \( k_c/k_d \) ratio to estimate \( k_c \) (0.022 \( \text{hr}^{-1} \)). The parameter values used in this study are reported in Table 3-2.

The table below summarizes the kinetic parameters used in the study:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_d )</td>
<td>13.5</td>
<td>( \text{hr}^{-1} )</td>
</tr>
<tr>
<td>( \Delta \sigma )</td>
<td>0.0065</td>
<td>[]</td>
</tr>
<tr>
<td>( k_i )</td>
<td>1</td>
<td>( \text{s}^{-1} )</td>
</tr>
<tr>
<td>( k_f )</td>
<td>0.08</td>
<td>( \text{s}^{-1} )</td>
</tr>
<tr>
<td>( k_{di} )</td>
<td>0.035</td>
<td>( \text{s}^{-1} )</td>
</tr>
<tr>
<td>( \Delta G_{\text{PAM,ref (CGGTA)}} )</td>
<td>-9.9</td>
<td>kcal/mol</td>
</tr>
<tr>
<td>( \Delta G_{\text{single-mismatch}} )</td>
<td>0.78</td>
<td>kcal/mol</td>
</tr>
</tbody>
</table>

**Figure 3-3: In vivo validation of the Cas9 mechanistic model.** (A) Addition of target sites dilutes \( \text{Cas9:crRNA} \), and increases the rate of RNA polymerase binding to the promoter. (B and C) Prediction and experimental data for the presence of (stars) 0, (circles) 1, (diamonds) 2, (squares) 4, or (triangles) 8 additional on-target sites. Data points and bars represent the mean and standard deviation of 2 measurements, performed in this study.
Quantifying the Effects of Supercoiling on Adjacent DNA sites

We used a novel expression system to quantify the supercoiling coupling effect for the presence of multiple adjacent Cas9 targets. In this system, the expression of a crRNA that targets a YFP-driven promoter enables (d)Cas9 to act as a transcriptional regulator and repress the promoter. The presence of the additional target results in higher YFP production by sequestering a fraction of the dCas9:crRNA complex. The difference in fluorescence production is proportional to the binding affinity of dCas9:crRNA with the sequence, which can be assayed in a high-throughput fashion to measure the dCas9:crRNA binding affinity with any sequence of interest. A main source of this change is change in $\Delta \Delta G_{\text{supercoiling}}$, that would enable us to inversely calculate the change in superhelical density of a target ($\Delta \sigma$) for additional of each auxiliary site.

We first introduced a heterologous cassette of YFP, driven by a synthetic $\sigma^{70}$ promoter in *E. coli*, using a high-copy plasmid. We expressed a precrRNA that suppressed the promoter using dCas9:crRNA, and measured fluorescence production by maintaining steady-state log-phase growth in a rich media. The resulting low YFP level demonstrated that Cas9:crRNA successfully suppressed the promoter activity. Expressing the reporter protein on a low copy plasmid reduced the activity below detection level (Data are not shown).

We then inserted 1, 2, 4, or 8 additional identical targets within a short distance (60 to 80 nucleotide randomized DNA between each two targets) on the same plasmid (*Figure 3A*) to quantify the effect of multiple R-loop interference, and observed over 100-fold increase in YFP production. Over-expression of the precrRNA by an IPTG-inducible promoter reduced the fluorescence production up to 21-fold (*Figure 3B*) for all constructs, while in the absence of the precrRNA (a negative control), the maximum YFP production reached to 14200 flu. This demonstrates even leaky transcription of the guide RNA was sufficient for strong suppression of the promoter activity.
Each R-loop opens at least 20 nucleotides, and over 20% of the targeted region must be melted for formation of all R-loops independently (opening over 180 nucleotides out of about 900 nucleotides for the construct with 8 auxiliary sites). This could generate a large torsional stress on the DNA that enhances Cas9 dissociation from the DNA, and inhibits the formation of additional R-loops within a short distance. As a result, the binding sites must have cooperativity.

We used our developed biophysical model to estimate the effect of cooperativity between the binding sites. We simplified Eq. 14 & 15 for the \textit{in vivo} condition and parameterized the model with our experimental data. The number of non-specific sites was double \textit{E. coli}'s genome size (N=9.2x10^6). The promoter's degradation rate (δ_{target}) was set to the measured specific growth rate (μ=0.72 hr^{-1}). We approximated the crRNA production as a linear function of precrRNA expression rate that was quantified at different IPTG concentrations using a YFP transcriptional fusion assay. The remaining unknown parameters were N_{Cas9}, ΔG_{targets}, Δσ, and σ_{promoter}. For the other model parameters (k_I, k_d, k_{di}, k_f), we used the estimated values from the \textit{in vitro} measurements. We then employed a nonlinear least square algorithm (Methods) to identify the unknown parameters (6 unknowns, 55 measurements; R^2=0.97; Appendix C-Figure 3). The resulting model correctly calculated the fluorescence measurements (Figure 3B). The computed ΔG_{targets}, Δσ, and σ_{promoter} were -10.7 kcal/mol, 0.0065, and -0.1 respectively (Table 3-2). Since the crRNA and target DNA were full complementary, the effect of ΔΔG_{exchange} was small, and the ΔG_{target} was mainly determined by ΔG_{PAM}. Based on the calculated σ_{promoter}, the positive driving force due to high supercoiling density around the promoter resulted in -0.77 kcal/mol beneficial binding energy. Consequently, the ΔG_{PAM} for CGGTA PAM, as our reference PAM sequence in the following calculations, was about -9.9 kcal/mol. The calculated large penalty change in superhelical density (Δσ=0.0065) suggests a strong R-loop inhibition for adjacent DNA sites even in the presence of 60 to 80 DNA gap between the targets. As a comparison with other cellular
processes that involve melting the strands of DNA, increasing superhelical density of a target DNA inhibits RNA polymerase binding to nearby DNA sequence, within 250 bp [191,192].

The maximum dynamic range of the targeted promoter varied with the number of auxiliary targets. In the presence of two additional targets, the maximum observed repression in the YFP production increased 21-fold (Figure 3B). Increasing the number of on-targets to 9 reduced the dynamic range to about 6-fold. According to the model predictions, the dynamic range can be recovered by simultaneously increasing expression of crRNA and Cas9, as has been observed in vivo[149].

In total, the presence of 9 on-target sites on the high copy plasmid generated over 1000 copies of the target in *E. coli*, and the activity of the YFP promoter still remained under 40% of the negative control. Cas9 is a relatively small protein machinery that can diffuse very quickly into the cell (random walk of about 1 second per Cas9 complex to any location in *E. coli*). This property enables Cas9 to efficiently search for its target DNA and serve hundreds of targets simultaneously without substantial reduction in its efficiency.

**Table 3-3:** Cas9:crRNA binding energies to canonical and non-canonical PAM sites (kcal/mol). The energies are average values of all combinations in the first and fifth positions. N.B: no statistically significant binding

<table>
<thead>
<tr>
<th>2nd position</th>
<th>4th position</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>AAA</td>
</tr>
<tr>
<td>AAT</td>
<td>-6.6</td>
</tr>
<tr>
<td>AAC</td>
<td>N.B.</td>
</tr>
<tr>
<td>AAG</td>
<td>-6.6</td>
</tr>
<tr>
<td>TAA</td>
<td>N.B.</td>
</tr>
<tr>
<td>TAT</td>
<td>-6.8</td>
</tr>
<tr>
<td>TAC</td>
<td>N.B.</td>
</tr>
</tbody>
</table>
Binding initiation from non-canonical PAM sequences

Genome-wide distribution and accessibility of PAM sequences dictate affinity of Cas9 for initiating an R-loop on a target[181]. Recent studies have demonstrated a reduced rate of R-loop formation from non-canonical PAM sequences, the source for a considerable fraction of Cas9 off-target activity[147,178,184]. To calculate the binding energy to different PAMs on an energy scale (ΔG_{PAM}), we collected data from a recent in vivo study[147] that quantified cleavage rate from randomized PAMs, proximal to a constant protospacer for S.pyogenes' CRISPR/Cas9. We defined CGGTA a reference PAM sequence, as we measured its binding energy previously (ΔG_{PAM,ref} =-9.9 kcal/mol). We then employed Eq. 14 and 15 at steady state condition for each characterized PAM to compute the change in its ΔG_{target} with respect to a target, containing the reference PAM. Because all protospacers were identical, ΔΔG_{exchange} remained almost unaltered, and the relative change in ΔG_{PAM} was mainly due to the change in ΔG_{target}. By adding the reference to the calculated energies, we estimated ΔG_{PAM} on an absolute energy scale (kcal/mol).

A subsequent analysis revealed that the effect of the first and fifth positions was small, and a core of 3 nucleotides (position 2 to 4) accounted for most binding affinity. We calculated the core’s binding energy by averaging the energies for all combinations in the first and fifth
positions. To remove false bindings, we only averaged energies for sequences that resulted in at least 0.1% cleavage. We also excluded any core sequence with less than 6 remaining combinations, and considered as no binding (Table 3). The average coefficient of variability for the reported energies was 9%.

As expected, NGG (N: A, T, C, G) had the highest binding affinity to Cas9. Lower binding rates occurred from NAG and NGW (W: A, T), about 20- and 70-fold less than NGG. The presence of a positive gap (NNGGN), 2 positive gaps (NNNGG), or a negative gap (GGNNN) between the PAM and the protospacer reduced the affinity by about 20-, 80-, or 70-fold. Overall, Cas9 had high binding affinity to A, G-rich sequences. A recent study also demonstrated a high rate of binding from NGA sequence[193].

The effects of crRNA:DNA mismatches

Mismatches between a crRNA and its target DNA usually delay formation of a stable Cas9:crRNA:DNA R-loop[178,181,186]. To quantify the effect, we used available high-throughput studies that examined the binding of multiple crRNAs to degenerate target libraries. These experiments either measured cleavage rate by Cas9 or repression/activation rate by dCas9 as a transcription factor, and the activity was determined by next generation sequencing[176,180,183].

For each pair of tested degenerate libraries and crRNAs, we related ΔG_{target} with the measured Cas9:crRNA activity. For cleavage activity[176,180], this was done by simplifying Eq. 14 and 15 for the experimental condition. For the transcriptional activation data[183], we employed a simple activation model of transcription[194] to estimate the binding energy. We used the crRNA’s full complementary sequence as a reference to calculate the relative change in ΔG_{target} for any combination of target:crRNA mismatches. For further analysis, we only selected
the sequences with at least 50X read coverage that contained a PAM identical to the reference's PAM. This eliminated the effect of PAM modification (change in $\Delta G_{\text{PAM}} \approx 0$) and supercoiling ($\Delta \Delta G_{\text{supercoiling}} \approx 0$), and the calculated change in $\Delta G_{\text{target}}$, with respect to the reference, was equal to the change in $\Delta \Delta G_{\text{exchange}}$.

A subsequent analysis showed that a single mismatch could vary $\Delta G_{\text{exchange}}$ up to about 4 kcal/mol (an 800-fold change in the Cas9 binding rate), and the effect is mismatch- and position-dependent. Interestingly, in the presence of CA sequence in a protospacer, proximal to PAM (Table 3-1; CLTA4 crRNA), the binding affinity of a few commonly accepted PAM sequences (NGG) dropped up to 5-fold. Insertion or deletion of one nucleotide between CA and the PAM recovered the binding affinity (beneficial mismatches). We did not observe a dependency to other combinations which suggested that the presence of CA next to the PAMs may destabilize duplex unzipping through an unknown mechanism. For PAM proximal mismatches, the measurements displayed compatible changes in the exchange energy while we observed a discrepancy for some PAM distal mismatches within the measurements. Overall, the calculated $\Delta \Delta G_{\text{exchange}}$ values by Pattanayak et. al.[176] (in vitro) measurements were small while the apparent Cas9:crRNA binding was more sensitive to mismatches in Hsu et. al.[180] measurements (in vivo). As an example, the average penalty energy for presence of any mismatch in the characterized protospacers was 0.14 and 0.78 kcal/mol for Pattanayak et. al.’s and Hsu et. al.’s measurements respectively. The difference could be as a result of variations in their characterization protocols.

For further analysis, we created two datasets, one containing 3671 measurements from the in vitro study, and one containing all analyzed sequences (5979 measurements) (Table 3-1).
Figure 3-4: The effect of mismatches between a crRNA and a target on its Cas9 binding affinity. The left (3671 measurements) and right (5979 measurements) graphs were created based on the dataset I (in vitro) and dataset II (both in vitro and in vivo) measurements respectively (Table 3-1). (A) Positional weights for mismatches at different locations of a target. The weights are normalized to their maximum values. (B) The developed NN mismatch energy model for hybridization of a target with a crRNA. The white boxes show the unknown energy parameters, about 14% (dataset II) to 20% (dataset I) of the energy parameters (C) Predictions versus measurements for ΔΔG$_{exchange}$; Pearson correlation of 0.86 and 0.78 for dataset I and dataset II respectively.

We then developed a comprehensive model that accounts for all possible base pairings and mismatches between the crRNA and the possible DNA sites. Mismatch position-dependency
is a result of asymmetric binding of crRNA across its target DNA for CRISPR type I[195], and likely for Cas9 (CRISPR type II). However, the available RNA/DNA nearest-neighbor (NN) energy models presume uniform base-pairing accessibility[196,197,198]. Also, only a few mismatch energy parameters for RNA/DNA heteroduplexes had been quantified, and using them may result in a poor predictive power (Appendix C-Figure 4). Hence, we relaxed the energy values for NN mismatches to obtain by a model identification (Methods). Because the mismatch effect is position dependent, we also defined a weight for each position along the protospacer. We parameterized the developed energy model using the calculated apparent ΔΔG \text{exchange}. For each characterized sequence, we computed ΔΔG \text{exchange} as a position-weighted summation of the local transient energies according to the NN model. For each dataset, we utilized a nonlinear least-square algorithm to identify a set of weights (Figure 4A) and energy parameters (Figure 4B) with the least calculated error (Methods). We also reduced the number of unknown energies by assuming symmetric penalty energy when both nucleotides were mismatched.

The developed NN model correctly determined the apparent energies (Figure 4C) that indicates the NN energy model’s capability to estimate the effect of mismatches for novel Cas9 targets. However, after model identification, about 14% (dataset II) to 20% (dataset I) of NN energy parameters remained unknown for the resulting models, mostly accounting for two consecutive mismatches. These combinations did not exist in any of our tested sequences. Also, depended on the source of the training data, the confidence interval for some calculated NN energies varied over 2 kcal/mol, likely a result of inaccuracy in some measurements or employing different characterization protocols. A more comprehensive systematic measurement can increase the accuracy of the NN model and reduce the confidence interval for each parameter. The calculated weights showed a high energy dependency at PAM proximal regions (the first 8 nucleotides), accounting for about 67% (dataset II) to 81% (dataset I) of ΔΔG \text{exchange} variation.
In a recent *in vivo* study [172], 87% of sequences with high binding affinities to a Cas9:crRNA complex have at most 1 mismatch within the first 8 nucleotides.

![Figure 3-5: Sensitivity analysis of the Cas9 model’s parameters.](image)

(A) The effect of change in each design parameter on the time-course cleavage of an on-target (blue), an off-target with a single mismatch (dark green), and off-target with double mismatch (red), and an off-target with three mismatches (light blue). (B) The effect of systematic change in each design parameter on the steady state rate of target cleavage (blue) and target specificity (black). Each condition’s fitness (light green) is calculated as the product of the cleavage and specificity. Mismatch (mm).

The effect of host and experimental design factors on (d)Cas9 activity

The rate of Cas9:crRNA binding to a target is a function of many design factors including the target's binding energy, Cas9 and crRNA concentration, time, the host's genome size and volume, and the total number of on- and off-targets. To examine each factor's effect, we used human cell as a model organism, and simulated the time-course cleavage of four hypothetical sequences: an on-target, an off-target with a single mismatch, an off-target with two mismatches, and an off-target that contained three mismatches. The total number of potential target sites was double the genome size (6.4x10^9 positions). We set the apparent number of Cas9 and crRNA expressed during a doubling time, total assay duration, Cas9's *in vivo* diffusion coefficient,
doubling time, and the on-target's binding energy to 1000, 1000, 100 hr, 11.25x10^{-12} \text{m}^2/\text{s}, 20 hr
and -9.9 \text{kcal/mol} \text{ respectively (Appendix C-Table 1). Here, for simplicity, we assumed 1) equal
accessibility to all the targets 2) a constant penalty energy for presence of each mismatch
(\Delta \Delta G_{\text{single-mismatch}}=0.78 \text{kcal/mol}) 3) nucleus volume is constant and about 8% of the entire
cell[199]. 4) No other specific target sequence exists in the cell. These assumptions enabled us to
estimate the Cas9:crRNA binding rate to the hypothetical targets.

We used the developed mechanistic model to simulate the effect of the change in each
design parameter on the time-dependent cleavage rate from each target (Figure 5A). Increase in
the on-target binding energy, reduced its Cas9 binding affinity and cleavage rate. The change
could be performed by adding mismatches to 5’ crRNA respectively[148,152,176,179,183].
Decrease in Cas9 and crRNA expression reduced the Cas9 activity from all targets but lowered
the off-target more than the on-target cleavage rate. Selecting a larger host increased the number
of non-specific binding events and reduced the probability of binding to the on-target. Similarly,
reducing the doubling time lowered the number of correct binding events. Overall, any attempt to
increase target specificity reduced cleavage rate from all targets in a non-linear fashion. The
result demonstrated that Cas9 activity and specificity have a pareto optimality, and cleavage rate
from all potential targets must be considered simultaneously to find an optimal experimental
condition with the minimal off-target and the maximal on-target activity.

We then employed a multi-objective optimization approach by quantifying target
specificity as the relative cleavage rate from the on-target in comparison with the total cleavage
rate from the on- and off-targets. To find an optimal value for each design factor, we varied every
model's parameter, and calculated each condition’s fitness as a product of the on-target’s cleavage
rate and specificity at steady state to identify their optimal values with maximal fitness (Figure
5B). This enabled us to calculate the quantitative change required for each design parameter that
maximizes the fitness. For any application, the design factors must be adjusted to obtain maximal productivity and the mechanistic model can be used to quantify the changes.

We also used the model to study the effect of the change in each design factor on activity of a promoter, targeted by dCas9 (Figure 6). Similar to Cas9, reducing the target binding energy and genome size increased the probability of binding to the targeted promoter, and lowered the promoter's activity at steady state. Also, increase in the expression rate of Cas9 and crRNA elevated the number of active complex molecules, and reduced transcription rate of the promoter. The reduction occurs almost linearly (in log-log scale) in a large range in comparison with typical transcription factors[100]. This is simply because Cas9 binding energy to the target DNA is much larger than many characterized transcription factors. As an example, the binding energy of LacI to its common operators (O1, O2, and O3) is -9.2, -8.3, and -5.8 kcal/mol while the Cas9 binding

---

**Figure 3-6: Sensitivity analysis of the dCas9 model's parameters.** (A) Effect of change in each design parameter on the time-course activity of an on-target (blue), an off-target with a single mismatch (dark green), and off-target with double mismatches (red), and an off-target with three mismatches (light blue). (B) The effect of the change in each design parameter on the steady state promoter activity (blue) and dCas9 specificity on the target (black); Mismatch (mm)
energy is about -10 kcal/mol. This behavior can be also observed for binding of transcription factors to synthetic operators with strong binding affinities[100]. Overall, the developed thermodynamic model could enable us to design novel synthetic genetic circuits containing dCas9 as a transcription factor.

**Figure 3-7: Genome-wide calculation of (d)Cas9 binding affinity to λ-phage plasmid.** (A) the target binding energy, (B) the PAM binding energy, (C) the crRNA:DNA exchange penalty energy for any locus in λ-phage plasmid. (D,F) the binding probability of Cas9:crRNA to any plasmid locus initially and after 10 minutes incubation. (E) the time-dependent binding rate to λ2 target (black solid line), to any off-target site (OSt, dash line), the major off-target site (OS1, green line), and an off-target with binding energy of zero (OS2, inset). *: λ2 target; x: OS1 target; blue: plus strand; red: minus strand.

**Genome-wide calculation of (d)Cas9 binding activity**

After forming a Cas9:crRNA molecule, the active complex performs a random 3D walk to examine any accessible DNA target for a stable binding. The Cas9:crRNA residence time at each individual non-target site is usually small[181]. However, trapping in many short-lived
binding events delays Cas9:crRNA arrival to the correct target, and reduces its occupancy at any
given time. To estimate the delay, the ensemble of all governing equations for all the possible
target sites must be considered and solved simultaneously. This allows us to simulate the
movement of any Cas9:crRNA molecule on the available DNA sites and capture their time-
dependent movement trajectories to any desired DNA site.

Here, as a demonstration, we calculated the time-dependent binding rate of a crRNA-
loaded Cas9 to any possible DNA sites in λ-phage plasmid, and compared it with experimental
measurements by Sternberg et. al.[181] that quantified the binding affinity of a Cas9: crRNA to
all λ-phage loci. In this experiment, 10 nM Cas9 was pre-incubated for about 10 minutes with 100
nM λ2-crRNA that mainly targets a specific sequence (called λ2 target) on λ-phage plasmid
(Figure 7). The binding reaction was started by adding the formed active complex to 0.1 nM λ-
phage DNA substrate, and the binding distribution was visualized after 10 minutes using total
internal reflection fluorescence microscopy.

We first analyzed the genetic system and quantified the Cas9 binding affinity to each
target site: the size of λ-phage plasmid was 46,500 bp and contained about 97,000 potential Cas9
targets. For each target, we used our parameterized model to estimate its Cas9 binding energy.
This included using the location-dependent weights (Figure 4) and a constant penalty energy for
each mismatch (ΔΔG_{single-mismatch}=0.78 kcal/mol) (Table 3- 2) to estimate DNA:crRNA exchange
energy at each target (Figure 7). Also, PAM binding energy at each site was determined. 5670
and 34363 canonical and non-canonical PAM sequences (Table 3- 3) were located on its both
DNA strands and had strong Cas9 binding affinity. The PAM binding energy for the remaining
DNA sites was insignificant and set to zero (Figure 7). This calculation enabled us to estimate
the binding energy of the Cas9:crRNA to any DNA site. The binding energy varied about 25
kcal/mol among these sites: the correct target site had the strongest binding affinity while only
about 4000 out of the 97000 targets had negative binding energy showing that the penalty energy

86
of exchanging RNA:DNA for the remaining sites significantly inhibits Cas9 binding and R-loop formation.

We then quantified the binding probability of any free active complex to all target sites. We created the binding profile using equation 7 that estimates binding probability at any given site. The profile revealed that the short-lived off-target bindings accounted for about 6% of total Cas9 residence time. We used equation 14 and 15 to calculate the time-dependent binding rate to λ2 target sequence (Appendix C-Table 2). The total number of governing ODEs in the system was about 97000. Numerical integration of all these equations simultaneously could cause a large calculation error due to solution instability and equation stiffness. Instead, we simplified the system by creating two pools of targets: one pool that contained the λ2 target, and one pool that contained the all the off-targets (OSt target). For each pool, we assigned an apparent binding affinity which was the summation of the binding affinities (i.e. $\exp(-\Delta G_{\text{target}}/k_B T)/N$) for the sequences in the pool. We then integrated equation 14 and 15 for this simplified pseudo system to calculate time-dependent Cas9 binding rate to the targets (Figure 7). Similarly, to track the time-dependent binding from two example off-target sequences, we modified and extended the pseudo system by creating two additional pools and assigning a binding affinity score to each pool: one for an off-target with the strongest binding affinity (called OS1 target) and one for an off-target with binding energy of about zero (called OS2 target) (Figure 7). Consistent with the experimental results by Sternberg et. al., our calculations demonstrated that Cas9:crRNA hops among all these off-target sites, identifies the λ2 target, and initiates a stable binding by forming an R-loop within 2 minutes of the reaction start. This expected delay can be controlled by designing the genetic system (referred to Figure 2). It is worthwhile to mention that the binding is temporary, and only in about 0.2% of these bindings, a conformation change in Cas9:crRNA occurs that leads to a stable permanent binding to the sequence.
Our simulation also demonstrated that given sufficient time to an active complex or sufficiently high concentration of active complex, it eventually finds any target sequence in the system and initiates an R-loop. As a comparison, elevating Cas9:crRNA concentration in an *in vitro* system[176] significantly increased the rate of Cas9 binding and cleavage from all targets. At any given time point Cas9 arrival time on any sequence is proportional to its binding affinity and sequence occupancy. For instance, initially, the binding rates for λ2, OSt, OS1, and OS2 targets were $1.3 \times 10^5$ 1/sec, $8 \times 10^3$ 1/sec, $5.7 \times 10^2$ 1/sec, and $9 \times 10^{-3}$ 1/sec respectively in a 100 µl reaction. In this case, the total number of sites from each type was about $6 \times 10^6$. Within few minutes of initiating the reaction, the λ2 target became mainly occupied. As a result, the total partition function value reduced from 161.6 to 10.6 after occupation of the λ2 target. This changed the partitioning of Cas9:crRNA among the remaining sites and increased their binding rates (Figure 7).

**Figure 3-8: Genome engineering for targeting HIV sequences.** The yellow, blue, and green colors show the default model, 10-fold lower and 10-fold higher expression rate of Cas9 and crRNA, respectively. (A) Cleavage rate (%) from all Cas9 targets that contain no more than 5
mismatches in comparison with LTR-B crRNA. (B) The cumulative fraction of indels at different cleavage (%) cut-off. (C) The sequencing read coverage required to detect cumulative fraction of indels. (D) The time-course cleavage (solid line) and specificity (dash line) of the selected target.

Example: removing latent HIV provirus from genomes

Recently, CRISPR/Cas9 has been examined as an efficient DNA cleavage machinery for curing HIV from infected genomes[91]. However, over 670,000 off-target indels, mostly called non-specific or statistically insignificant, were identified after 14 days treatment with 2 crRNAs using 15-30X deep sequencing coverage. Here, we used the developed Cas9 model for reasoning the high rate of observed indels. We performed a bioinformatic search[200] to identify all sequences that contained 5 or fewer mismatches inside human genome in comparison with the selected targets, and identified about 6000 potential off-target sequences. A fraction of the off-targets contained mismatches that their penalty energy did not exist in the currently developed NN energy model. Hence, we utilized the weights calculated by dataset I but used an average maximal penalty energy for a mismatch at any position (2.3 kcal/mol) instead of the NN energies. Using the calculated weights, this value corresponds to an average change of 0.78 kcal/mol in ΔΔG_{exchange} for the presence of a single mismatch between a crRNA and its target. The other model’s parameters were adjusted according to the reported values in Appendix C-Table 3. We also assumed that each cleavage would result in an indel. For further analysis, we selected one of the designed crRNAs (LTR-B), and simulated the time-dependent cleavage from its identified targets (3106 DNA sites).

We examined the effect of change in Cas9:crRNA production rate on the number of off-target indels by changing Cas9 and crRNA expression 10-fold lower (0.1X) and 10-fold higher (10X) than the default model (1X). In all cases, most off-target sequences had a low cleavage probability (Figure 8A); 99%, 95%, and 80% of the sequences had less than 1% cleavage rate for
0.1X, 1X, and 10X expression respectively. However, the total number of off-target cleavage events was significant, and accounted for most indels (Figure 8B). These events may result in cell heterogeneity that cannot be simply assayed by low coverage deep sequencing. We then calculated the number of sequencing reads at each locus, required to detect the on- and off-target indels with 99% confidence level. The number of reads is inversely correlated with each locus's cleavage rate (Figure 8C). For instance, to detect an indel with 1% cleavage rate, at least 127 reads must be obtained for 99% confidence level. The number of sequencing reads to statically detect 99% of the indel events for 0.1X, 1X, and 10X is in the order of $10^5$, $10^4$, and $10^3$ respectively. A lower number of sequencing reads increases the number of false negative events and may statistically reject many low rate off-target cleavage events while the actual number of off-targets is larger.

**Discussion and Conclusion**

We have developed the first mechanistic model of CRISPR/Cas9-mediated cleavage. The model contains 5 kinetic steps including Cas9 and crRNA expression, Cas9:crRNA active complex formation, random walk of the active complex to all potential targets, stabilizing target binding, and cleavage reaction (Figure 1). We used statistical thermodynamics and the law of mass action to build a time-dependent kinetic model that was later examined in vitro and in vivo using *S. pyogenes*’ CRISPR machinery (Figures 2 & 3, Table 3-1). According to the biophysics, Cas9:crRNA goes through multiple cycles of binding and unbinding from the correct targets before cleaving the DNA. The developed mechanistic model reveals a series of design factors that control in vivo cleavage rate including the expression rate of Cas9 and crRNA, Cas9:crRNA binding energy to each on- and off-target, and a host's properties such as genome size, doubling
time, and volume. Change in genotype or growth condition can alter these factors. For instance, target binding energy can be altered by adding 5’ mismatches to crRNA[152,179].

Consistent with our analysis in vivo and in vitro, Cas9 may dimerize at its high concentrations. A mobility gel shift assay has also confirmed the existence of additional larger protein complex molecule at high Cas9 concentration[181]. This could explain why a fraction of Cas9 molecules are inactive at high concentration, and do not cleave their target DNA, while at low concentrations almost all complexes actively cleave their target DNA (Figure 2). The exact dimerization mechanism is unknown. Here, we assumed that the dimerization occurs between a wild-type Cas9 and a crRNA-loaded Cas9 that is well supported by our in vitro and in vivo analysis. Other Cas9 dimerization or polymerization mechanism may also exist that are consistent with our data. For instance, assuming the dimerized complex can efficiently cleave its target DNA, provides an additional reaction to our proposed mechanism. In this case, a slightly weaker correlation may be observed between the in vitro and in vivo data, and the model prediction (Appendix C-Figure 5). Regardless, our biophysical analysis as well as indirect in vitro and in vivo experiments (ex. protein gel and presence of inactive complex) strongly support the dimerization of Cas9 and/or Cas9 complex.

We employed the model to study the effect of each design factor on activity and specificity of (d)Cas9. A sensitivity analysis of these parameters (Figures 5 and 6) indicated a toxicity effect at high concentration of Cas9 and crRNA- a higher cleavage rate occurred even from sequences that were multiple mismatches away from the correct target[176]. DNA targets with different superhelical densities have different Cas9 cleavage rates (Figure 2). Larger genomes contain more specific and non-specific off-target sites. As a result, a high rate of off-target binding may be observed in eukaryotes, but is unlikely in bacteria[149,184] (using a correct design strategy[152]). Since, Cas9 performs a random 3D walk, larger volume reduces diffusion to all on- and off- targets, and delays PAM recognition step. Similarly, increasing its doubling
time elevates the probability of PAM recognition and increases Cas9 activity from all targets. The design factors can be tuned to find an optimal phenotype with a controllable on-target activity and maximal specificity.

Depended on the experimental condition and input substrates, each cleavage intermediate step may limit the final Cas9 cleavage rate. The estimated rate of complex formation is sufficiently large ($k_i \approx 1 \text{s}^{-1}$), and may not become a major rate-limiting step during *in vivo* characterization. In contrast, for cleaving short DNA fragments, *in vitro* complex formation rate might be comparable or even slower than target binding step, and could also control the time-dependent rate of cleavage. For instance, a faster *in vitro* cleavage rate has been observed for pre-incubated Cas9, crRNA in comparison with not pre-incubated mixtures[148]. Change in Cas9 recognition signal stem loop could reduce the complex formation rate/activity that results in a lower cleavage rate[176,178], as has been observed for other CRISPR proteins as well[201]. For a correct target, RNA/DNA strand displacement usually occurs fast[186], and PAM recognition and cleavage reaction are the rate-limiting steps of the remaining reactions. The residence time of Cas9:crRNA on a target is mainly determined by the exchange energy, for instance, the residence time of Cas9:CrRNA on a sequence containing even 12 non-target adjacent PAMs is still under 1% of a correct target DNA with similar concentration[181]. When the number of potential targets is small (e.g. the *in vitro* experiment; Figure 2), the rate of PAM recognition is large, providing enough timing for Cas9 to visibly go through a cycle of binding and unbinding before completing the cleavage reaction. In contrast, in the presence of many targets (e.g. bacterial and mammalian genomes), the rate of PAM recognition drops substantially, and this step may also become a major rate-limiting step. The presence of multiple mismatches between a target and crRNA inhibits the cleavage reaction. However, Cas9 can still bind with a lower rate and reduce accessibility of the target (e.g. a promoter)[152]. As a comparison, in a ChIP-seq
experiment[172], only 8.7% of off-target sites with high binding affinity had a significantly larger cleavage rate than the examined negative control.

The state of the art in Cas9 recombineering is designing a crRNA, that is complementary to the target sequence, but is multiple mismatches away, for instance 3 nucleotides[164], from the rest of the genome. The mismatch-allowance cutoff plays a dominant role in this design. As the cutoff drops the number of ignored mismatch-containing sequences elevates astronomically. A seed of 8 nucleotides accounts for about 67 to 81% of a target's binding energy, and a mismatch in the seed reduces the average binding affinity about 14-fold while even 8 PAM-distal mismatches can be tolerated[152]. The rate of cleavage at each filtered off-target site might be low, while the expected number of total off-target indels is large, and may account for most cleavage events (Figure 8). The potential Cas9 off-target sites are randomly distributed across the host's genome, and could result in cellular heterogeneity (Figure 8). For instance, during a genome-wide cleavage study[184], the number of detected off-target sites with indel rate of above 1%, between 0.1% and 1%, and between 0.01% and 0.1% were 0, 5, and 170 sequences respectively. Deep sequencing of target DNA is a common high-throughput approach to count the number of cleaved and uncut sites, and measure cleavage rate [91,176,180]. Theoretically, to measure a cleavage rate between 1% and 99% (1% precision), at least 127 high-quality reads must be obtained for 99% confidence level. In many studies, the average read coverage for each locus is low, for instance about 100 or less[91,176,180,183]. The number of reads can be linearly increased with a high expense. This hinders next generation sequencing assays to detect most off-targets with low cleavage rate (ex. 0.1% or below), and may fail to uncover the resulting heterogeneity.

By analyzing the available in vivo and in vitro measurements with our mechanistic model (Table 3- 1), we quantified the effect of three main sources of Cas9/dCas9 off-target activity. The first source is the presence of non-canonical PAM sequences. The most commonly observed
PAM for *S. pyogenes’* CRISPR/Cas9 is NGG[163]. We calculated the Cas9 binding energy to other PAMs, and showed that a lower binding rate can occur from non-canonical PAMs such as NAG and NGW motifs (**Table 3**). We also computed the effect positive and negative gaps between a protospacer and a PAM, and demonstrated that it could reduce Cas9/dCas9 binding affinity. The third source of off-target activity is the presence of many genome loci with multiple mismatches away from the correct target. We developed an energy model that accounts for all combinations of mismatches between a crRNA and all target sequences (**Figure 4**). By incorporating these models into the CRISPR array design, one can estimate the probability of cleavage at any genome locus. These models enable us to solve the max activity-max specificity optimization problem at DNA level, and design crRNAs with minimal cross activity.

Exposure time is also a determinant factor in the number of off-target indels. According to our *in vitro* analysis, cleavage occurs only at about 0.2% of the binding events to the correct target, and a lower rate for mismatch-containing sequences. The probability of most off-target indel events is small when measured within a host's two- or three-doubling time, and even may go below detection level. However, as has been demonstrated for CRISPR type I[202], large exposure time by CRISPR/Cas9 primes off-target cleavage from degenerate targets that contain multiple mismatches in comparison with the correct target. This might be a reason for current discrepancy among proposed Cas9's off-target patterns[152,172,179], also suggests that the CRISPR/Cas9 machinery must be turned off after usage for vital applications such as gene therapy.

As been demonstrated here and in other studies[149,152] dCas9 can successfully identify its target promoters and suppress their activities (or activate for dCas9-activator fusions[152,183]). The strong suppression reduces the dynamic range of the promoter activity (**Figure 3**). Simultaneous change in expression of dCas9 and crRNA can improve the dynamic range of the target promoters[149] while due to cross-activity, this strategy cannot be used for
adjusting the activity of multiple co-expressed crRNAs. Instead, we demonstrated that the addition of auxiliary sites can increase the dynamic range for a constant level of dCas9 production. A larger dynamic range may be achieved by addition of mismatches between crRNA and a target to reduce Cas9 binding affinity[152]. The developed mismatch energy model can be also used here to quantitatively alter the dCas9 binding affinity and design crRNA sequences with diverse activity.

Presence of CRISPR/Cas9 machinery increases cellular autonomous mutation rate as a function of Cas9 and crRNA expression rates. A low rate of cross-activity might be acceptable for many purposes such as high-throughput genome engineering and functionality screening[164] but may become challenging for critical applications such as gene therapy. In addition, design of dCas9-based genetic systems requires a combinatorial DNA search to find the number, type, and position of mismatches that provide the desire dCas9 binding affinity to each target sequence. Biophysical modeling would enable us to estimate binding affinity to all on- and off-target sequences, and quantitatively adjust the governing design factors at DNA sequence level, for a robust control on the activity of the targets while reducing unexpected modifications.

Materials and Methods

A Mechanistic Model of CRISPR/(d)Cas9 Binding and Cleavage

The DNA specificity and activity of Cas9-mediated cleavage is dictated by a 5-step mechanism that includes the expression of Cas9 and crRNA, the formation of active Cas9:crRNA complex, a random intracellular walk to search for DNA target sites, the formation of a Cas9:crRNA:DNA complex (an R-loop) at DNA sites, and finally DNA site cleavage (Figure 1).
We developed a dynamical mechanistic model that incorporates all known biomolecular interactions and processes that control the rates of these steps. When using Cas9, the model predicts the cleavage rates for all plasmid- and genome-encoded DNA sites inside the cell. When using nuclease-deficient dCas9, the model predicts the binding occupancy of dCas9:crRNA to the same DNA sites. The mechanistic model accounts for how several factors control all the DNA sites' cleavage rates, including changing Cas9 and crRNA expression levels, different crRNA protoscaler (guide) sequences, different DNA site sequences, both canonical and non-canonical PAM recognition DNA site sequences, and the effects of DNA site supercoiling. The model also explicitly accounts for the host's specifications, including its genome sequence, genome length, cell size, and growth rate. Moreover, the model allows for the expression of multiple crRNA guide strands, and it will determine how the competitive binding of crRNAs to Cas9 will also affect the DNA sites' cleavage rates. In the following, we describe the biophysical interactions controlling Cas9-mediated cleavage and the corresponding equations used to quantitatively model the resulting dynamics.

**Modeling the Expression and Formation of the Cas9:crRNA Complex**

Mature crRNA guide strands can be expressed in two ways: transcription of a single chimeric synthetic guide RNA (sgRNA) that contains the 5' target recognition region, followed by a conserved Cas9-binding hairpin [185]; or transcription of a precrRNA array and a tracrRNA that form an RNA duplex that is subsequently processed by RNAse III into a mature crRNA [146,203]. As a key difference, the precrRNA can contain multiple target recognition sequences, each separated by a repetitive spacer sequence. The tracrRNA binds to these repetitive spacers and forms a double-stranded complex with precrRNA, becoming a target for RNAse III cleavage [146,203,204]. The resulting RNAse processing can generate multiple mature crRNAs from a
single precrRNA. Cas9 may bind with the tracrRNA before landing on the precrRNA, and facilitate the tracrRNA:precrRNA hybridization [203]. After the mature crRNA is loaded into Cas9, an unidentified RNA exonuclease trims its 5' end, leaving a target recognition sequence of about 20 nucleotides [189,204]. When not bound to a crRNA, wild-type Cas9 remains in a structural conformation that inhibits its cleavage activity [205]. During the crRNA loading process, Cas9 undergoes a rotational shift that exposes a DNA binding channel, yielding an active Cas9:crRNA complex.

In our model, we first introduce the production rates of mature crRNA guide strands (r_{crRNA}) and Cas9 proteins (r_{Cas9}) as zero order reactions. These production rates can be varied by altering the DNA copy numbers or transcription rates of the precrRNA, sgRNA, or Cas9 as well as the translation rate of Cas9's mRNA [149,193]. We then employ mass action kinetics to describe the irreversible formation of an intermediate Cas9:crRNA complex, followed by an irreversible isomerization reaction that produces an active Cas9:crRNA complex. The rate of intermediate complex formation is quantified using a second order kinetic constant k_f and the isomerization reaction's rate is quantified using a first order kinetic constant k_i. As first order reactions, the crRNA, Cas9, and intermediate Cas9:crRNA complex degrade or become diluted at a rate quantified by the kinetic constants δ_{crRNA}, δ_{Cas9}, and δ_{Cas9:crRNA}. Finally, the rate of target binding for each active Cas9:crRNA complex is designated r_{binding}, and will be derived below.

Several in vitro and in vivo studies have measured the dynamics of sgRNA, precrRNA, and Cas9 complex formation under different experimental conditions [146,149,176,181,203,205]. At low Cas9 concentrations, almost all Cas9 proteins form active Cas9:crRNA complexes and cleave their target DNA [181]. However, when the Cas9 concentration is greatly increased, there appears to be a fraction of inactive Cas9:crRNA complex that does not bind or cleave its target site [181]. Because many CRISPR and/or DNA binding proteins can form homodimers [206,207,208,209], we speculated that it is possible for Cas9 to dimerize with an intermediate
Cas9:crRNA complex. We correspondingly introduced a dimerization reaction quantified by a second order kinetic constant $k_d$. Consistent with this hypothesis, at high Cas9 concentrations, a mobility gel shift assay identified at least two Cas9:crRNA:DNA complexes with different molecular weights [181].

The resulting differential equations (Equations 1 to 4) describe the dynamics of Cas9 and crRNA expression and active complex formation in terms of their molecular counts, assuming that the cell has a constant volume. For our first biophysical model of the CRISPR/Cas9 system, we have ignored the effects of stochastic gene expression as well as the effects of discrete cellular division. Here, the kinetics of dimerization ($k_d$) are assumed to be slower than the kinetics of active complex formation ($k_f$). In addition, to account for the production of multiple crRNA guide strands with different sequences, we expanded the system of differential equations by an index $i$ to describe their production, active complex formation, and rate of target binding. We assumed that all expressed crRNA guide strands bind equally well to Cas9, and form active complexes at the same rate, with the same kinetic parameters ($k_f$ and $k_i$). However, through competitive binding, the fraction of Cas9 bound to each crRNA guide strand will depend on the crRNAs' differing expression levels. The rates of Cas9-dependent cleavage will also differ across different crRNA guide strand sequences (index $i$) as well as different DNA site sequences (index $j$), designated by $r_{C[i,j]}$.

$$\frac{dN_{crRNA,i}}{dt} = r_{crRNA,i} - \delta_{crRNA,i} N_{crRNA,i} - k_f N_{Cas9} N_{crRNA,i}$$

(1)

$$\frac{dN_{Cas9}}{dt} = r_{Cas9} - \delta_{Cas9} N_{Cas9} - k_f N_{Cas9} \sum_j N_{crRNA,j} - k_{di} N_{Cas9} \sum_j N_{intermediate,j}$$

(2)
Modeling Cas9's Random Walk to Determine its Search Rate

Once formed, active Cas9:crRNA complexes do not undergo facilitated diffusion or hopping, but instead engage in three-dimensional molecular diffusion to search for DNA sites [181]. The rate of diffusion is governed by the diffusivity of the Cas9:crRNA complex ($D$), and also several host-specific factors, including the volume of the compartment ($V$) and the characteristic length between sites of production and binding ($\lambda$). Here, we assume that the cellular compartment is well-mixed such that the rate of net molar flow is zero, though the time required for a Cas9 protein to find a target DNA site depends on the rate of molecular diffusion. Accordingly, the rate of molecular diffusion for active Cas9:crRNA complexes using the $i^{th}$ crRNA guide strand ($r_{RW,i}$) will be proportional to its concentration [210]:

$$r_{RW,i} = \frac{6D\lambda N_{Cas9:crRNA,i}}{V}$$

Equation (5) is the molar flow rate, or contact rate, between active Cas9:crRNA complexes and all possible DNA sites inside the cell. We then use the sequences of the crRNA guide strand and the DNA site to calculate the probability that, once contact has been made, the active Cas9:crRNA complex binds to the DNA site and forms a stable Cas9:crRNA:DNA complex, called an R-loop. The rate of binding of the $i^{th}$ Cas9:crRNA complex to the $j^{th}$ DNA site is simply the product of the contact rate and the binding probability ($P_{[i,j]}$):
To calculate this binding probability, we assume that the pool of active Cas9:crRNA complexes have reached chemical equilibrium with the pool of both on-target and off-target DNA sites. This assumption is valid because the number of potential DNA sites is always much larger than the number of Cas9:crRNA complexes. In addition, when the Cas9:crRNA levels have reached steady-state conditions, the system will become ergodic. Accordingly, we derive a partition function in terms of the $i^{\text{th}}$ active Cas9:crRNA complex's binding free energy to the $j^{\text{th}}$ DNA site sequence ($\Delta G_{\text{target},[i,j]}$) as well as the number of accessible DNA sites with the $j^{\text{th}}$ sequence ($N_{\text{target},j}$). Here, our reference state is a DNA sequence that binds non-specifically to Cas9:crRNA with a zero binding free energy. As the total number of non-specific DNA binding sites, we use twice of the host's genome length $N$. The binding probability will follow a Boltzmann distribution, and we may use both the reference state and partition function as normalization factors to calculate the probability that the $i^{\text{th}}$ Cas9:crRNA complex binds successfully to the $j^{\text{th}}$ DNA site:

$$P_{[i,j]} = \frac{N_{\text{target},j} \exp \left( - \frac{\Delta G_{\text{target}[i,j]}}{k_B T} \right)}{1 + \sum_{n,m} \frac{N_{\text{target},m}}{N} \exp \left( - \frac{\Delta G_{\text{target}[n,m]}}{k_B T} \right)}$$

Together, Equations 6 and 7 provide a systematic approach for comparing the rates of binding for different crRNA sequences and for determining how one Cas9:crRNA complex affects the binding of other Cas9:crRNA complexes when multiple crRNAs are expressed inside the same cell. Our next step was to develop a sequence-dependent free energy model to calculate and predict these binding rates for any crRNA guide strand sequence.
A Free Energy Model for DNA Target Binding, R-Loop Formation, and Cleavage

The binding free energy of an active Cas9:crRNA complex to a particular DNA site controls its binding occupancy, and ultimately, its cleavage rate. Several interactions control the magnitude of this binding free energy, including the presence of a protospacer adjacent motif (PAM) site, the rate of R-loop formation during a multi-step exchange reaction, and the effects of supercoiling at the DNA site. Here, we employed thermodynamics to quantify the energetics of these interactions and developed a multi-term free energy model that calculates $\Delta G_{\text{target}[i,j]}$ for different crRNA guide strand sequences, DNA site sequences, canonical and non-canonical PAM sequences, and varying amounts of DNA site supercoiling. Altogether, the free energy model sums together the strengths of these interactions, according to:

$$\Delta G_{\text{target}[i,j]} = \Delta G_{\text{PAM},j} + \Delta \Delta G_{\text{exchange}[i,j]} + \Delta \Delta G_{\text{supercoiling},j} \tag{8}$$

Next, we describe the mechanism of R-loop formation and how these interactions’ free energies are quantified.

After contacting a DNA site, a Cas9:crRNA complex recognizes and binds to the PAM sequence [178,211]. The canonical PAM site for the Cas9 from *Streptococcus pyogenes* is NGG, though additional non-canonical sequences have also been recognized [147,148,152,180]. The Cas9:crRNA complex then pulls apart the double-stranded DNA upstream of the PAM sequence, which is an energetically intensive process. Cas9 does not hydrolyze an energy-providing cofactor, such as ATP or GTP. Instead, its only significant source of external energy input originates from the binding interactions between the Cas9 protein and the PAM recognition sequence [181], which we designate as $\Delta G_{\text{PAM}}$. As we show below, the most canonical PAM recognition sequence has an apparent $\Delta G_{\text{PAM}}$ of about -9.5 kcal/mol, which is sufficient to pull apart between four and eight DNA base pairings depending on sequence composition. Non-
canonical PAM sequences have less energetically favorable interactions with Cas9, but can still support R-loop formation and cleavage [147,180,186,193].

The Cas9:crRNA complex continues to pull apart double-stranded DNA by performing an exchange reaction, allowing the crRNA guide strand to form RNA:DNA base pairings with its complementary DNA strand [181,186,212]. In step-wise transitions, each DNA base pair is pulled apart, and the corresponding nucleotide from the crRNA binds to form a Watson-Crick base pair, resulting in the formation of a DNA:Cas9:crRNA:DNA sandwich, called an R-loop. R-loop formation is directional and sequential, beginning at the PAM site, and proceeding upstream. Before the R-loop is completed, strand displacement can stall and reverse, resulting in Cas9:crRNA dissociation, whenever the DNA:DNA complex becomes more stable than the DNA:Cas9:crRNA:DNA complex. We designated this difference in stability as \( \Delta \Delta G_{\text{exchange}} \); if \( \Delta \Delta G_{\text{exchange}} \) becomes positive and large, the R-loop cannot successfully form. To investigate whether Cas9 plays a role in target specificity, we then developed and parameterized two versions of a free energy model to calculate \( \Delta \Delta G_{\text{exchange}} \) for a given crRNA and DNA site sequence, where the first model incorporates only nucleic acid interactions, while the second model accounts for both nucleic acid and Cas9-dependent interactions.

In the first model version, when the crRNA and DNA site are fully complementary, \( \Delta \Delta G_{\text{exchange}} \) is governed by the difference in free energy between the RNA:DNA duplex and its corresponding DNA:DNA duplex. Interestingly, this difference is free energy is sequence-dependent; for example, the binding free energy of the dinucleotide base pair rAC:dGT is 1.0 kcal/mol more stable than dAC:dGT, while the binding free energy of rCG:dCG is 1.6 kcal/mol less stable than dCG:dCG [197]. These nearest-neighbor free energies are designated as \( \Delta G_{\text{RNA:DNA}} \) and \( \Delta G_{\text{DNA:DNA}} \), and may be calculated using previously developed free energy models that have been parameterized using calorimetry measurements [196,197,198,213]. Second, because of the sequential nature of R-loop formation, when the crRNA has non-complementary
bases with the DNA site, the effect of the resulting mismatches will depend on their distance from the PAM site. For simplicity, we introduce a position-dependent multiplicative weight $d_k$ that modulates the impact of these free energy differences. $k$ is location and varies from 0 to the crRNA guide strand's length; the value of $d_1$ will be larger than $d_{20}$.

Therefore, our first approach for calculating $\Delta \Delta G_{\text{exchange}}$ compares the thermodynamic stability of the $i^{\text{th}}$ crRNA:DNA complex to the stability of the $j^{\text{th}}$ DNA:DNA duplex, using the following expression:

$$\Delta \Delta G_{\text{exchange}}[ij] = \sum_k d_k \left[ \Delta G^{\text{RNA:DNA}}_{k,k+1} - \Delta G^{\text{DNA:DNA}}_{k,k+1} \right]$$

(9)

where the summations proceed over the lengths of the crRNA:DNA and DNA:DNA sequences. Equation (9) has 21 unknown $d_k$ parameter values and uses dinucleotide free energies that were previously parameterized in the absence of Cas9 [196,197,198,213].

However, it is possible that the Cas9 protein alters the stability of the R-loop in a sequence-specific fashion. To investigate this possibility, our second approach to calculating $\Delta \Delta G_{\text{exchange}}$ is to formulate an entirely empirical nearest-neighbor model, which enumerates all possible dinucleotide RNA:DNA duplexes and mismatches together with the distance-dependent coefficients, resulting in 277 unknown parameters. In the result section below, we determined these parameter values using thousands of experimental measurements of off-target and on-target Cas9 activity. Once parameterized, the following expression is used to calculate $\Delta \Delta G_{\text{exchange}}$ for any crRNA and DNA site sequence:

$$\Delta \Delta G_{\text{exchange}}[ij] = \sum_k d_k \Delta \Delta G^{\text{Cas9:crRNA:DNA}}_{k,k+1}$$

(10)

where the summation proceeds over the length of the crRNA:DNA sequence. In the results section, we systematically compared the accuracy of these two models to quantitatively determine Cas9's effect on DNA site specificity.
Next, we incorporated the effects of DNA site supercoiling into the model of Cas9:crRNA's binding energetics. Negative supercoiling, the untwisting of helical DNA, increases the stability of an R-loop by lowering the stability of the competing DNA:DNA complex [186]; however, there is a free energy input to form supercoiled DNA. When relaxed B-form helical DNA of length \( n \) is (un)twisted by \( 10\sigma \) turns, the change in free energy will be \( \Delta G_{\text{supercoiling}} = 10n\sigma^2k_bT \), where \( \sigma \) is the superhelical density, \( k_b \) is the Boltzmann constant, and \( T \) is temperature [214]. Due to the activity of topoisomerases and gyrases inside cells, the superhelical density of bacterial and human genomic DNA varies between \( \sigma = -0.02 \) and \( -0.1 \), depending on the location's distance from the origin of replication and its proximity to highly transcribed genes [215,216]. If a DNA site has already been negatively supercoiled by the host’s native enzymes, then a free energy input is not needed to stabilize the R-loop. However, if the DNA site is relaxed or positively supercoiled, then the additional free energy needed to untwist it will increase the dissociation rate of the Cas9:crRNA complex as it forms the R-loop. Accordingly, the dissociation kinetic constant of the Cas9:crRNA complex will depend on the degree of DNA site supercoiling according to

\[
k_{d,j} = k_d^* \exp\left( \frac{\Delta\Delta G_{\text{supercoiling},j}}{k_bT} \right)
\]

where we determine the free energy input needed to untwist the DNA site by comparing the superhelical density at the target site (\( \sigma_j \)) with an average density for all nonspecific binding sites (\( \sigma_{\text{NS}} \)) using the expression:

\[
\Delta\Delta G_{\text{supercoiling},j} = 10nk_bT\left( \sigma_{\text{NS}}^2 - \sigma_j^2 \right)
\]

For instance, an average superhelical density of -0.06 (\( \sigma_{\text{NS}} \)) has been previously reported for \textit{E. coli} genome [214,217].
After the R-loop has formed, the DNA:Cas9:crRNA:DNA complex has the ability to cut the DNA strands, one at a time, typically at the third nucleotide upstream of the PAM site [148]. As measured by a time-course cleavage assay, an appreciable amount of nicked DNA accumulates before double-stranded DNA breaks are observed, indicating that Cas9’s endonuclease reaction is a slow, rate-limiting step. Unlike most enzymes, after Cas9 has doubly cut its DNA site, the Cas9:crRNA complex remains stably bound to the DNA site and does not have the ability to cleave DNA at another site [181]. This absence of turnover causes Cas9 to become a limiting reactant. However, before Cas9 has doubly cut its DNA site, optical trap pulling experiments have shown that the formation of the R-loop is reversible and that the DNA:Cas9:crRNA:DNA complex can dissociate [186]. In light of these two competing pathways, we derived an expression for the cleavage rate of the $i^{th}$ Cas9:crRNA complex bound to the $j^{th}$ DNA site:

$$r_{C[i,j]} = \frac{k_c r_{\text{binding}[ij]}}{k_c + k_{d,j}}$$

(13)

where the rate of cleavage is controlled by a first-order kinetic constant $k_c$, and the effects of DNA supercoiling on the dissociation kinetic constant, $k_d$, are determined using Equation 11.

**Modeling the Genome-wide Occupancies at On-target and Off-target DNA Sites**

Finally, the total numbers of free, bound, and cut DNA sites over time are determined by deriving two differential equations quantifying all sources of production and consumption. In Equation 14, we show how the rates of Cas9:crRNA binding and cleavage control the numbers of bound and uncut DNA:Cas9:crRNA:DNA complexes per cell ($N_{\text{complex}}$). Here, the rate of cell
division, designated as $\mu$, quantifies how fast the cell replicates its DNA and distributes its chromosomes to daughter cells. The Cas9:crRNA complexes bound to those chromosomes are distributed to daughter cells at the same time.

$$\frac{dN_{\text{complex}[i,j]}}{dt} = r_{\text{binding}[i,j]} - r_{\text{C}[i,j]} - \mu N_{\text{complex}[i,j]}$$  \hspace{1cm} (14)$$

In Equation 15, we sum together the rates of Cas9-dependent cleavage, potentially arising from multiple crRNA guide strands, to determine the total rate of cleavage at the $j^{th}$ type of DNA site. Initially, the host organism begins with $N_{\text{total,j}}$ copies of these DNA sites. For chromosomally encoded DNA sites, $N_{\text{total,j}}$ will vary between 0 and 2, depending on their distance from the chromosome's origin of replication and whether the site is located within accessible euchromatin or inaccessible heterochromatin. For plasmid-encoded DNA sites, $N_{\text{total,j}}$ can be in the hundreds. DNA sites with the same sequence are treated as the same DNA site with a higher copy number.

After Cas9 binds and cleaves a DNA site, we assume that the double-stranded break is repaired either via homologous recombination or non-homologous end-joining. After the DNA site has been repaired, it is no longer available to be cleaved again until a new copy have been replicated. Once a newly available DNA site has been replicated, it is also distributed to daughter cells during division. Therefore, the net production rate of available DNA sites is $\mu(N_{\text{total,j}} - N_{\text{target,j}})$. Together, the rate of DNA replication and Cas9-dependent cleavage determines the total number of cut and uncut DNA sites within the organism, according to:

$$\frac{dN_{\text{target,j}}}{dt} = \mu(N_{\text{total,j}} - N_{\text{target,j}}) - \sum_i r_{\text{C}[i,j]}$$  \hspace{1cm} (15)$$

Altogether, for a genetic system that expresses $\eta$ crRNAs in a host with $\zeta$ available DNA sites, the formally complete biophysical model of CRISPR/Cas9 consists of $3\eta + \zeta (\eta+1) + 1$
ordinary differential equations, which is a large number. However, with further analysis that distinguishes between DNA sites with either high or low Cas9-dependent cleavage rates (on-target or off-target, respectively), we greatly reduce the number of differential equations and more easily calculate the fraction of DNA sites that are free, bound, or cut over time. First, we note that, for on-target DNA sites, the differential equations are highly coupled; as Cas9 binds and cleaves these sites, the number of available DNA sites decreases, and the rate of Cas9-dependent cleavage correspondingly decreases until the system reaches a steady-state. However, when a site's cleavage rate is very small, we can effectively ignore this coupling, and separately calculate the numbers of cleaved on-target and off-target DNA sites. Accordingly, we use the following approximate expression for determining the steady-state number of cleaved off-target DNA sites ($N_{\text{off-target,}j}$) with the $j^{th}$ sequence,

$$
\frac{N_{\text{off-target,}j}}{N_{\text{total,}j}} = \sum_i \frac{r_{\text{C,}i,j}}{\mu N_{\text{total,}j}} \iff \sum_i r_{\text{C,}i,j} << \mu N_{\text{total,}j}
$$

(16)

where the rates of Cas9-dependent cleavage implicitly depend on $N_{\text{off-target,}j}$. After substitution of all expressions, solving for $N_{\text{off-target,}j}$ requires the solution of a system of multivariate quadratic polynomials, which can be efficiently computed using either an iterative hybrid Krylov method [218] or a sequential root finder, for example, using Dixon Resultants [219]. With this approximation, the model now contains $3\eta + \zeta_{\text{on-target}}(\eta+1) + 1$ differential equations, which is much fewer. Importantly, even though their individual cleavage rates are small, a genome has such a large number of off-target sites that a significant fraction of Cas9 will bind to one of them. Therefore, while we can assume steady-state conditions for individual off-target DNA sites, we cannot ignore their presence.
The Effects of Supercoiling on Adjacent DNA sites

When a Cas9:crRNA complex binds to a DNA site, the formation of the R-loop will result in positive supercoiling of the surrounding DNA sites, due to conservation of the DNA linking number in the absence of topoisomerase or gyrase activity [220]. Positive supercoiling of DNA will alter the affinities of DNA-binding proteins, such as RNA polymerase [124] or other Cas9:crRNA complexes. These longer-range effects become important when crRNAs are designed to bind to several nearby on-target DNA sites, for example, when targeting two different DNA sites with a chimeric dCas9-FokI fusion [182], when inserting recombinant DNA between two nicked or doubly cleaved DNA sites, or when using dCas9 to regulate the transcription rate of a promoter using multi-input logic. Whenever multiple on-target DNA sites are adjacently located, we therefore modified the free energy model for $\Delta G_{\text{target}}$ to incorporate the site-to-site effects of supercoiling.

Consider multiple DNA sites located within a short segment of DNA surrounded by a type of fixed end, for example, between two active promoters, DNA replication origins, or other sites where DNA-binding proteins constrain DNA topology. When Cas9:crRNA binds to one of these DNA sites, the unwinding of the DNA site during R-loop formation increases the superhelical density of the remaining DNA segment by an amount $\Delta \sigma$ (more positive), which depends on the lengths of the DNA site and the DNA segment. With the increase in supercoiling from $\sigma_j$ to $\sigma_j + \Delta \sigma$ (from negative to less negative), Cas9:crRNA will require an additional free energy input to bind to the remaining DNA sites within the segment and form an R-loop, according to Equation (12). As more DNA sites are bound by Cas9:crRNA, we assume that the linking number is conserved, yielding an increase in superhelical density from $\sigma_j$ to $\sigma_j + c \Delta \sigma$ for $c$ bound DNA sites. Eventually, the free energy needed to stabilize the R-loop will become sufficiently large to prevent Cas9:crRNA from binding additional DNA sites within this DNA
segment. According to our calculations below, $\Delta \sigma$ is about 0.0065. To calculate these binding probabilities, we modified the partition function in Equation (7), accounting for the combinations of states where Cas9:crRNA has bound $c$ adjacent DNA sites with their corresponding supercoiling-dependent energy penalties.

**Additional Model Considerations and Assumptions**

There are additional factors, not included within this model, that can affect Cas9's ability to recognize and bind crRNAs as well as cleave DNA sites. Outside of the crRNA guide sequence, the tracrRNA and sgRNA form four stem loop structures that are responsible for recognizing and binding to Cas9 [185]. While the third and fourth stem loops are not essential for recognition, truncation of these structures did reduce the stability of the Cas9:crRNA complex. In another study, it was observed that truncated sgRNAs resulted in lower cleavage rates at both on-target and off-target DNA sites, which suggests that there were either fewer active Cas9:crRNA complexes or that active complexes had lower intrinsic cleavage activities [176]. Here, the biophysical model assumes that the tracrRNA and sgRNA fold into the wild-type structure. Further, while Cas9 can bind well to both single- and double-stranded DNA, its cleavage rate is significantly reduced when bound to single-stranded DNA or a truncated double-stranded DNA site [181]. The current biophysical model only considers double-stranded DNA sites within long contiguous DNA, such as plasmids and genomes.

Overall, the developed mechanistic model can estimate the probability of binding and cleavage for any Cas9 target DNA. In addition to degradation rate of all the involved molecules ($\delta_i$), the final model’s parameters are $k_f$, $k_{dis}$, $k_i$ (complex formation step), $\Delta G_{PAM}$, $\Delta \Delta G_{exchange}$, $\Delta G_{supercoiling}$, $k_d$, $c$ (stabilizing target binding), and $k_c$ (cleavage step). The input parameters are the exposure time ($t$) and the production rate of Cas9 ($r_{Cas9}$) and crRNA ($r_{crRNA}$). For a system
containing 1 type of crRNA and N on- and off-targets, the concentrations of free Cas9, crRNA, intermediate complex, free active Cas9:crRNA complex and the targets are unknown, and can be calculated by solving N+4 ordinary differential equations (Eq. 1-4, 14-15) simultaneously. In the following sections, we have used multiple in vivo and in vitro measurements to estimate the model parameters in different conditions. A summary of the studies and the utilized data is provided in Table 3-1.

Model Simulation and Parameterization Approach

Differential equations were numerically integrated using a variable-order, adaptive time-stepping stiff numerical solver (ode15s) in MATLAB. For comparison to experimental measurements, the relative errors between model solution and experimental measurements were calculated over the measurements' time interval or after a steady-state condition was reached. In Table 3-1, we summarize the several types of experimental measurements used to parameterize and validate the model, including the number of degrees of freedom and the number of data-points in each experimental data-set. To identify a narrow range of best-fit parameter values, model parameterization was performed by using either a simple simplex method (fminsearch) or a Levenberg-Marquardt method (lsqnonlin) in MATLAB to minimize the sum of squared relative errors, followed by a parameter sensitivity analysis and visual comparisons to more precisely

Strains and Plasmids

To validate model predictions, we constructed three plasmids that employ dCas9 to transcriptionally regulate expression of a reporter protein. The first plasmid expresses the YFP fluorescent protein reporter on a R6K vector using a KanR antibiotic marker. The YFP expression cassette contains a σ70 promoter (J23100), a synthetic ribosome binding site designed by the RBS Calculator[1,37,78,125,221], a codon-optimized YFP coding sequence, and an efficient
transcriptional terminator [77]. A primary crRNA binding site is located within the promoter region with the sequence (5' -- TATCGTTAAGGTTACTAGAG -- 3'). Where noted, between one to eight auxiliary crRNA binding sites with the same sequence as the primary crRNA binding site, each separated by 80 nucleotides of randomized DNA, were inserted downstream of the transcriptional terminator. To insert auxiliary binding sites, gBLOCK DNA fragments (Integrated DNA Technologies) were synthesized and assembled with a digested vector fragment using T4 ligation. The second plasmid constitutively expresses Cas9 and tracrRNA on a p15A vector using an AmpR antibiotic marker. Plasmid constructed was performed by PCR-amplifying the Cas9 and tracrRNA expression cassettes from the pdCas9 plasmid [152] and assembling with a PCR-amplified p15A vector fragment using Gibson's method[17]. The third plasmid expresses the precrRNA using an IPTG-inducible P_{tac} promoter on a ColE1 vector using a CmR antibiotic marker, and was constructed by PCR-amplifying the precrRNA cassette from pdCas9 and assembling it with a PCR-amplified ColE1 fragment using Gibson's method. The precrRNA contains two BsaI sites flanking the protospacer region, which were utilized to insert new crRNA guide sequences into the precrRNA with digestion and ligation of annealed oligonucleotides. Cloned plasmids were verified by sequencing. The three plasmids were electroporated together into E. coli pir116, and selected on triple antibiotic agar plates.

**Growth and Measurements**

Transformed strains were grown overnight at 37°C and 200 RPM in LB Miller supplemented with 10 μg/ml chloramphenicol, kanamycin, and ampicillin (Sigma-Aldrich). 5 μl of cultures were diluted into fresh selective media in a 96-well microplate, incubated, and shaken at 37°C in a TECAN M1000 spectrophotometer. Serial dilutions were performed twice to maintain cells in the exponential phase of growth for a 12 hour period. 10 μl samples were
extracted after the second and third serial dilutions and added to 200 μl PBS supplemented with 2 mg/mL kanamycin for halting growth. Single-cell YFP fluorescence from at least 20,000 cells were recorded by an BD Fortessa flow cytometer. The average YFP expression level was determined by taking the average of the fluorescence distribution and subtracting the average auto-fluorescence of *E. coli* pir116.
Reference


Appendix A Supplementary Information for Chapter 1

Appendix A-Figures

Appendix A-Figure 1: Sampling 3-dimensional CFP, mRFP1, GFPmut3b production rate space using optimized RBS libraries. (A) Three optimized RBS libraries were designed using the RBS Library Calculator (search resolution=0.35) to contain 8 variants that efficiently search the translation rate space of a 3-color operon encoding CFP, mRFP1, and GFPmut3b (Appendix A-Table 4). A combinatorial library of three colored operon was created from these RBS libraries and transformed into *E. coli*. (B) Five hundred colonies were randomly selected and characterized for their fluorescence production.
Appendix A-Figure 2: Sampling 3-dimensional CFP, mRFP1, GFPmut3b production rate space using optimized RBS libraries. (A) A library of triple-color operon variants was created by combinatorial assembly of three random RBS libraries for CFP, mRFP1, and GFPmut3b each contained NNNNNNN within Shine-Dalgarno region (Appendix A-Table 5). (B) Five hundred variants were characterized using flow cytometry.
Appendix A-Figure 3: Estimating the number of measurements to find optimal protein expression levels in multi-protein genetic systems using random or optimized RBS libraries (A) The number of measurements needed to search for optimal protein expression levels of a genetic system (CrtE, CrtB, CrtI, IspA, Idi, and Dxs) were determined using either optimized RBS libraries (squares) or random RBS libraries (diamonds). First, the entire translation rate space between 1 to 100,000 au was targeted. Random (Appendix A-Table 15) and optimized (Appendix A-Table 16) RBS libraries were created and their abilities to cover the translation rate space were compared for three metabolic pathways with (left) three, (middle) four, or (right) six proteins. In all cases, optimized RBS libraries improve coverage and reduce the number of measurements. (B) Zoom mode was employed to design optimized RBS libraries to target a narrow translation rate range between 30,000 and 300,000 au (Appendix A-Table 17). High coverage of the targeted translation space were achieved using a small number of measurements.
Appendix A-Figure 4: The proposed elementary reaction model for the Carotenoid biosynthesis pathway. CM1 to CM14 indicates the intermediate enzyme-metabolite complexes. IPP, isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; GGPP, geranyl geranyl diphosphate. Cofactors are omitted.

Appendix A-Figure 5: The kinetic model's error distribution for neurosporene content predictions. The model's predictions have higher accuracy at higher neurosporene productivity.
production rate. \( r_p \) is the predicted pathway productivity. \( r_a \) is the actual, measured pathway productivity. The average error of prediction for all characterized pathways is 32%. There are two outliers responsible for the most error.

**Appendix A-Figure 6: Neurosporene content measurements using optimized growth conditions.** The Carotenoid pathway variant with the highest neurosporene content was cultivated with high aeration during exponential growth phase for either 7 hrs in LB miller media or 10 hrs in 2xM9 media supplemented with 0.4% glucose. Their neurosporene productivities were measured by hot acetone extraction and dry cell weight measurements. Ten milliliter of each culture was heated for 48 hrs at 60 °C and the dry cell weights were measured (A) immediately after baking; or (B) after allowing the samples to equilibrate at room temperature for 4 hrs. The difference in dry cell weight measurements yielded an average difference of 27% in dry cell weight, which resulted in a 28% change in the neurosporene productivity. All carotenoid content and pathway productivity measurements were obtained using the (B) approach, allowing samples to equilibrate after heating.
Appendix A-Figure 7: Input-output transfer function of $P_{lacO1}$ promoter. A ColE1 based plasmid was constructed to express $mRF1$ fluorescence reporter under control of $P_{lacO1}$ promoter, and transformed into the plasmid cured *E. coli* Echw2f. IPTG concentration was varied from 0mM to 1mM, and expression of *mRFP1* fluorescence reporter from $P_{lacO1}$ promoter was measured by flow cytometry during exponential phase in supplemented M9 media. A model of the $P_{lacO1}$ promoter's response to induction was constructed by fitting the experimental data to a hill equation model \[118\], which is \[ r(au) = \frac{9100}{1 + \frac{1450}{(IPTG(mM))^2}}. \]
Appendix A-Figure 8: flux control coefficients for the enzymes *crtE*, *crtB*, and *crtI*.

Using the kinetic model predictions, flux control coefficients for the enzymes *crtE*, *crtB*, and *crtI* in the neurosporene biosynthesis pathway are evaluated across the 3-dimensional translation rate space. Numerical finite differences are utilized to calculate the partial derivatives at 50x50x50 logarithmically distributed grid points. Contour maps show flux control coefficient values. High FCCs indicate where increasing an enzyme's expression will increase pathway's productivity, while low FCCs show regions where increasing expression does not lead to a significant improvement in productivity. Negative FCCs show regions where excess enzyme expression lowers the pathway's productivity, due to metabolite sequestration.
Appendix A-Figure 9: Using computational geometry to search for optimal translation rates regions. (A) The CrtEBI expression space is decomposed into Voronoi polygons, and the pathway productivities within the convex hull of the space are computed by interpolation. The white areas are outside the convex hull. (B) 8 out of 19 carotenoid pathway variants from the test data-set (Figure 4E) have *crtEBI* translation rates within the convex hull. Their measured productivities (green) are compared to productivities determined according to geometric interpolation (black). (C) Zoom mode can be similarly applied to non-mechanistic models to identify translation rate regions with improved performances.
Appendix A-Figure 10: Effect of training set size on model predictions. (A) Error of interpolation (B) percentage of the test set pathway variants that could be evaluated by the models. To have a reliable prediction from kinetic model, the training set size should be larger than the number of unknown parameters in the model (dashed line).
Appendix A-Figure 11: Null-hypothesis rejection test. (A) The distribution of the randomly generated productivities, which are used for training. Error of prediction for (B) kinetic (C) Geometry (D) statistical models trained with actual measurement training set (diamond) or randomly generated training set (circle). (E) The model failed to reject the null hypothesis while the other models successfully falsified the null hypothesis.

Appendix A-Tables

All Supplementary Tables are Located in the Supplementary Data.

Appendix A-Table 1: Characterization data for all RBS library variants controlling protein expression. The measured fluorences for optimized RBS library variants controlling expression of mRFP1, sfGFP, and LacZ. Translation initiation rates are predicted using RBS Calculator v1.1.
Appendix A-Table 2: Strains and plasmids used in the study. This table contains the DNA content of the studied and created plasmids and strains.

Appendix A-Table 3: Designed MAGE oligonucleotides for genome engineering. Two chemically modified 90-mer oligonucleotides were synthesized to introduce a stop codon into lacI and an optimized RBS library to uniformly control lacZ expression. The 5' phosphothiorate (*) and 2' fluoro uracil (/i2FU/) modifications are denoted. LacZ_oligo_m1_v1(**) oligonucleotide was, used as described in Wang et al, to create LacZ mutant. LacZ_restore oligonucleotide was design to recover lacZ activity during Co-selection MAGE.

Appendix A-Table 4: The optimized RBS libraries for sampling the translation rate space of the 3-color CFP, mRFP1, and GFPmut3b operon. The RBS Library Calculator in Search mode was used to design three 8-variant RBS libraries with a search resolution of 0.35. Translation initiation rates are predicted using RBS Calculator v1.1.

Appendix A-Table 5: Random RBS libraries for sampling translation rate space of the designed triple-color operon. The table contains the random RBS libraries controlling expression of CFP, mRFP1, and GFPmut3b, and their predicted translation rates. Translation initiation rates are predicted using RBS Calculator v1.1.

Appendix A-Table 6: The optimized RBS libraries for searching the enzyme expression space of the Carotenoid biosynthesis pathway. The RBS Library Calculator in Search mode was used to design 16-variant RBS libraries controlling CrtE, CrtB, and CrtI expression. Translation initiation rates are predicted using RBS Calculator v1.1. When multiple in-frame start codons are present, each capable of producing a full length or extended protein, we select the start codon with the highest predicted translation initiation rate. In particular, there is an extra in-frame GTG start codon in the RBS controlling CrtE expression with a calculated translation initiation rate between 445 and 2585 au. There are internal start codons in two RBS variants controlling CrtI expression; their calculated translation initiation rates are 189 au and 558 au. In addition, two of the 73 pathway variants were observed to have single point mutations in the CrtE RBS, and their translation initiation rates were calculated and used instead of the originally designed RBS sequence.

Appendix A-Table 7: Characterization data for the 73 Carotenoid pathways generated using Search mode. Their translation rate predictions and neurosporene content measurements are shown. Translation initiation rates are predicted using RBS Calculator v1.1. The carotenoid pathways' predicted productivities were determined using either the kinetic model or the computational geometry model.

Appendix A-Table 8: The carotenoid pathway's kinetic parameters determined by model identification. The calculated best-fit 48 kinetic parameters for all 24 elementary reactions of the proposed Carotenoid pathway’s kinetic model are shown. Sensitivity analysis
revealed that a 2-fold increase or decrease in a single parameter increased the prediction's error to 35%. However, a 10-fold increase or decrease in any two parameters (all pair-wise combinations) increased the prediction's error to 46%. In addition, multiplying all kinetic parameters by a positive factor does not alter the model's predictions. Therefore, model identification resulted in identifying best-fit kinetic parameters that are sensitive to 2-parameter variations, but insensitive to 1-parameter or time-scale changes. As a result, the best-fit kinetic parameters are not unique, but are located within a relatively small manifold within the large kinetic parameter space.

Appendix A-Table 9: Characterization of 19 additional carotenoid pathway variants to test the kinetic model's predictions. The predicted translation initiation rates and measured carotenoid productivities are shown. Translation initiation rates are predicted using RBS Calculator v1.1. The kinetic model predicts the carotenoid pathway's productivity at selected \textit{crtEBI} translation rates. The pathway's productivities were predicted with an error of 25%. Stars (*) denote where the in-frame GTG start codon was determined to have a higher predicted translation initiation rate than the ATG start codon. Any pathway variants that is outside the convex hull formed the training data cannot be analyzed by the geometry model (NaN).

Appendix A-Table 10: The optimized RBS libraries for targeting CrtE, CrtB, and CrtI expression levels using Zoom mode. The translation rate ranges were 32,000 to 305,000 au for CrtE, 1800 to 232,000 au for CrtB, and 26,500 to 1,347,000 au for CrtI. Translation initiation rates are predicted using RBS Calculator v1.1.

Appendix A-Table 11: Characterization of 28 Carotenoid pathway variants generated using Zoom mode. These variants target the high neurosporene content region in the translation rate space of CrtE, CrtB, and CrtI.

Appendix A-Table 12: Characterization of pathways with novel input-output transfer functions. Four pathway variants were designed by SEAMAP to have linear-linear or log-linear behaviors.

Appendix A-Table 13: The designed RBS library for varying expression of dxs enzymes. The optimized RBS library for altering translation rate of dxs enzyme generated by Genome Editing mode of the RBS Library Calculator was shown.

Appendix A-Table 14: Characterization of balanced and imbalanced pathways. The measured neurosporene productivities of the characterized balanced and imbalanced pathways at 16 different dxs production rates were shown.

Appendix A-Table 15: The random RBS libraries used to calculate the coverages of example 3-, 4-, and 6-protein expression spaces. The random libraries controlling \textit{crtE}, \textit{crtB}, \textit{crtI}, \textit{dxs}, \textit{idi}, \textit{ispA} are shown. The translation rates were predicted by RBS Calculator v1.1.
Appendix A-Table 16: The optimized RBS libraries used to calculate the coverages of example 3-, 4-, and 6-protein expression spaces using Search mode. Six optimized RBS libraries were designed to cover the translation rate space of *crtE*, *crtB*, *crtI*, *dxs*, *idi*, and *ispA*. The translation rates were predicted by RBS Calculator v1.1.

Appendix A-Table 17: The optimized RBS libraries used to calculate the coverages of example 3-, 4-, and 6-protein expression spaces using Zoom mode. Optimized RBS libraries for *crtE*, *crtB*, *crtI*, *dxs*, *idi*, and *ispA* enzymes were designed to cover a narrow translation rate range between 30,000 and 300,000 au. The translation rates were predicted by RBS Calculator v1.1.

Appendix A-Table 18: Comparison between kinetic, geometry, and statistical model for building a SEAMAP. Statistical model 1 and model 2 use formula S1 and S2 for error calculation respectively. The Run time is the order of magnitude of time required for evaluating a pathway variant using MATLAB software on windows 7 32-bit, 2 Quad CPU 2.83 GHz. * They could not provide information for internal gaps in expression space.
Appendix A-Methods

Appendix A-Method 1: The expected number of MAGE cycles to insert optimized RBS libraries into a genome

A general formula is derived for determining the number of MAGE mutagenesis cycles $N$ needed to introduce the optimized RBS libraries into $K$ targeted protein coding sequences within the genome. Multiplexed genome engineering (MAGE) uses mutagenic oligonucleotides and oligo-mediated allelic recombination to incorporate targeted mutations into an engineered host strain.

The efficiency of incorporation is determined by the number of mutagenesis cycles, and depends on the oligonucleotide's allelic replacement efficiency. Several factors affect the allelic replacement efficiency (AR). First, targeting the lagging strand of the genome is 10- to 100-fold more efficient than the leading strand[222]. Second, chemically modified oligonucleotides improve AR by increasing the oligonucleotide's intracellular stability and preventing the mismatch DNA repair system (mutS) from repairing mutations. AR is generally improved by knocking out mutS completely[223]. Third, the kinetics of oligonucleotide hybridization are important as there is a 1-second window for allelic recombination[25]. Inhibitory secondary structures at the 5’ and 3’ oligonucleotide ends prevent hybridization.

We design oligonucleotides with 5’ phosphorothioate modifications, internal 2’ fluoro bases, and minimal 5’ and 3’ secondary structures to have allelic replacement efficiencies up to 25% in a mutS knockout and up to 6% in a mutS+ strain. Increasing the number of consecutive mutations $M$ in a mutS knockout will reduce allelic replacement efficiency according to the mismatch identity[26,224], following the empirical relationship $AR_{K=1} = 0.26e^{-0.135*(M-1)}$ as determined by Wang et. al[26].
The chance of incorporating a single RBS variant into a single site within the genome in one mutagenesis cycle is a Bernoulli trial with probability $AR_{K=1}$. The chance of incorporation increases with additional mutagenesis cycles according to a cumulative geometric distribution, where the number of cycles $N$ determines the total fraction of mutated genome $F$ according to:

$$F_1 = AR_{K=1} + (1 - AR_{K=1})AR_{locus} + (1 - AR_{K=2})AR_{K=1} + \cdots + (1 - AR_{K=1})^{N-1}AR_{K=1}$$

Simplification of this equation leads to:

$$F_1 = 1 - (1 - AR_{K=1})^N$$

where $(1 - AR_{K=1})^N$ is the probability of having a genome remain unmodified after $N$ cycles. For example, the introduction of the stop codon in lacI (a C -> T replacement) required only 3 mutagenesis rounds to modify 50% of the genome population (an observed $AR_{K=1}$ of 20%). Introduction of the RBS library into lacZ required twenty rounds to modify over 90% of the genome population.

The targeting of multiple RBS sequences and protein coding sequences are largely independent if distantly located (>500 kbp), though the incorporation of mutations at nearby genes may be correlated[25,225]. Assuming independence and a constant overall oligonucleotide concentration, the fraction of mutated genomes will decrease as additional genes are targeted, according to the relationship:

$$F_K = [1 - (1 - AR_{K=1}/K)^N]^K$$

where the chance of incorporating mutations into two sites is simply $F_1 \times F_1$. The number of mutagenesis cycles $N$ to ensure at least $F$ fraction of the genome has been modified is then:

$$N = \frac{\log \left( 1 - F \frac{1}{K} \right)}{\log \left( 1 - \frac{AR_{K=1}}{K} \right)}$$

When designing mutagenic oligonucleotides, we select up to 7 consecutive mutation variants to ensure a sufficiently high allelic replacement efficiency to maximize genome
modification and minimize the number of cycles. According to this design specification, 50% of
the genome population will contain at least one modification in each of its $K$ targets with
$N = -20K \log_{10} \left( 1 - 0.5^{1/k} \right)$ cycles. To reduce sequencing costs and increase the fraction of
modified genome to 90%, the number of mutagenesis cycles is increased to
$N = -20K \log_{10} \left( 1 - 0.9^{1/k} \right)$. Increasing the number of targets while maintaining reasonably
low mutagenesis cycles requires the use of coselective MAGE [25].

Appendix A-Method 2: Monte-Carlo estimation to determine the minimal number of measurements

Monte Carlo sampling could estimate the number of measurements for a library of multi-
protein genetic systems that will be sufficient to explore its N-dimensional translation rate space.
For each protein in the genetic system, an RBS library is used to vary its translation rate and
expression level. The translation rates of RBS sequences will depend on the protein coding
sequences in the genetic system. Optimized RBS libraries are designed by the RBS Library
Calculator in Search Mode to uniformly vary the translation rates of the specified coding
sequences between 1 and 100,000 au with a medium search resolution (0.30), yielding libraries
with 16 variants. Random RBS libraries contained 6 random nucleotides within their Shine-
Dalgarno sequences (ie: NNNNNN or NNNGGANNN) or 9 degenerate consensus nucleotides
(DDRRRRRRD), creating libraries with 4096, or 23328 variants. In all cases, the translation
rates of the RBS library variants are predicted using the RBS Calculator v1.1. Each RBS library
presents a pool of choices. Library construction is assumed to generate a homogeneous mixture of
RBS variants where all potential choices have equal weight.

For each measurement simulation, one RBS variant is randomly selected from each of the
N RBS library pools. The measurement's position in the N-dimensional translation rate space is
determined, and its corresponding bin is classified as covered. The number of choices increases combinatorially as the number of proteins in the genetic system increases. Random selection is repeated for M measurements, and the total coverage of the N-dimensional translation rate space is calculated by extension of Equation (4). This process is repeated in 100 independent runs to determine the average and standard deviation of coverage. Monte Carlo sampling is performed identically for optimized or random RBS libraries, and for genetic systems with 2 or more proteins.

Appendix A-Method 3: An elementary mass action kinetic model of the carotenoid biosynthesis pathway’s reactions

The development of the kinetic model has three parts. First, we derive a mechanistic, mass action model of the pathway to relate changes in enzyme expression level to changes in pathway productivity. Second, we carry out model reduction and dimensional analysis to determine a pathway's productivity in comparison to a reference pathway variant. The dimensional analysis eliminates the need to determine the concentrations of the intracellular metabolites or enzymes, while retaining the ability to predict changes in pathway productivities. Third, we perform model identification to determine the unknown variables’ best-fit values that reproduce the experimentally measured pathway productivities. Validation of the kinetic model is performed on an additional set of characterized pathway variants.

Part I. Kinetic model formulation We developed a kinetic model of 10 intracellular reactions catalyzed by Idi, IspA, CrtE, CrtB, and CrtI enzymes that includes (i) reversible binding of substrate to enzyme; and (ii) reversible unbinding of product from enzyme. IspA, CrtE, CrtB, and CrtI catalyze multiple reactions. These reactions convert intracellular isopentenyl diphosphate (IPP) and Dimethylallyl diphosphate (DMAPP) to the neurosporene carotenoid.
There are a total of 24 metabolites, free enzyme, and enzyme complexes in the kinetic model (Appendix A-Figure 4).

Using mass action kinetics, we derive 24 ordinary differential equations describing the rates of production and consumption of all chemical species within the reaction network. We assume isochoric conditions. There are accordingly 48 unknown kinetic parameters quantifying the rates of the first and second order reactions involved in forward and reverse reactions. In addition, five mole balances on the amounts of free and bound enzyme are derived. The amounts of free and bound enzyme must sum to a constant amount. The metabolic dynamics of all 73 pathway variants are governed by the same set of kinetic parameter values and mole balances.

All reactions in the kinetic model are potentially reversible. The net increases in species concentrations (C_i) are determined by the forward (v_{f,j}) and reverse (v_{r,j}) rates, or the net reaction rates V_{j}^{net}, of the N_i participating reactions that consume or produce metabolite i, according to:

\[
\frac{dC_i}{dt} = F_{\text{in}}^i + \sum_{j=1}^{N_i} v_{f,j} - v_{r,j} = F_{\text{in}}^i + \sum_{j=1}^{N_i} V_{j}^{net}, \quad i = 1 \text{ to } 24
\]

The final reaction’s net rate determines the pathway’s overall neurosporene productivity.

The rates of the forward and reverse reactions are determined by either first order \( (k_i[S_j]) \text{ or } k_{-i}[S_j] \) or second order \( (k_i[S_j][S_k]) \text{ or } k_{-i}[S_j][S_k] \) mass action kinetics where \( k_i \) and \( k_{-i} \) are the kinetic parameters for the forward and reverse reactions, respectively. \( F_{\text{in}}^i \) is the input flux for metabolite i which is zero for all metabolites except IPP and DMAPP. \( [S_j] \) refers to the concentration of the j\text{th} chemical species, which may be a metabolite, a free enzyme, or a bound enzyme. The net reaction rate is the sum of the forward and reverse reaction rates:

\[ V_{j}^{net} = v_{f,j} - v_{r,j} \]

To illustrate, the first reaction describes the binding of IPP to the enzyme \( idi \). Its forward and reverse reaction rates are calculated according to:
\[ v_{f,1} = k_1 [idi][IPP] \text{ and } v_{r,1} = k_{-1} [CM1] \]

where [\(idi\)] is the concentration of free \(idi\) enzyme, \([CM1]\) is the concentration of bound \(idi\) enzyme, and \([IPP]\) is the concentration of metabolite isopentenyl diphosphate.

Next, we perform a dimensional analysis to transform unknown metabolite and enzyme concentrations into metabolite and enzyme concentration ratios, using a reference pathway variant for comparison. We multiply and divide the forward and reverse reaction rates by the metabolite and enzyme concentrations within the reference pathway variant, which do not alter the calculated answer. We then re-arrange the variables to yield an apparent kinetic parameter, a metabolite concentration ratio, and an enzyme concentration ratio. As a result, the concentration ratios are bounded and appropriately scaled. To illustrate, we carry out this transformation for the first reaction’s forward and reverse rates, yielding:

\[
v_{f1} = \left( \frac{k_1 \cdot [IPP]_{ref} \cdot [idi]_{total}}{[IPP]_{ref}} \right) \cdot \frac{[IPP]}{[idi]_{total}} \cdot \frac{[idi]_{free}}{[CM1]_{ref}}
\]

\[
v_{r1} = \left( \frac{k_{-1} \cdot [CM1]_{ref}}{[CM1]_{ref}} \right) \cdot \frac{[CM1]}{[CM1]_{ref}}
\]

Importantly, the choice of the reference pathway variant will alter the apparent kinetic parameter values, but it will not alter the solution to the ordinary differential equations; increases in the apparent kinetic parameters are compensated by decreases in the enzyme concentration ratios. We selected pathway variant #53 as the reference pathway variant. The total enzyme concentration ratios for each pathway variant are determined by comparing that pathway variant’s translation rates to the translation rates of the reference pathway. As an example, for \(crtE\), the equation is:
The reference pathway has predicted translation initiation rates of 72268, 20496, and 203462 au for \textit{crtE}, \textit{crtB}, and \textit{crtI}, respectively.

To determine a pathway's metabolite concentrations and productivity, we input the enzymes' concentrations and kinetic parameters into the ordinary differential equations and utilize a stiff numerical integrator (MATLAB's ode23s) to determine a numerical solution. The time period of the numerical solution is the same as the experimental growth conditions; 7 hours post-induction for all pathway productivity predictions shown in Figure 2. The numerical solution provides the time-dependent metabolite concentrations, including the production flux of neurosporene. We denote the neurosporene production flux as \( r_p \). Each pathway variant will have a different neurosporene production flux as a result of the different total enzyme concentrations, controlled by the \textit{crtEBI} translation rates.

We then use the reference pathway to determine how changes in the model's calculated neurosporene production flux will control the pathway's neurosporene productivity. By comparing the ratio in neurosporene production fluxes between reference and non-reference pathway variants, we predict the non-reference pathway's neurosporene productivity according to the equation:

\[
\frac{r_{pi}}{r_{pref}} = \frac{\text{predicted neurosporene productivity of the } i^{th} \text{ pathway variant}}{\text{measured neurosporene productivity of the reference pathway}}
\]

The neurosporene productivity of the reference pathway is 196 µg/gDCW/hour. The reference pathway's neurosporene production flux is calculated using the same kinetic parameters and mole balances as the non-reference pathway variants.
The kinetic parameters control the pathway's metabolic dynamics and its overall productivity; however, the mass action kinetic parameters for each enzyme in the reaction network (idi, ispA, CrtE, CrtB, and CrtI) are not known, and are difficult to measure. We use model identification to determine the best-fit kinetic parameters that reproduced the measured productivities for the 73 characterized pathway variants.

Prior to carrying out model identification, we employed model reduction and dimensional analysis to eliminate non-independent variables within the kinetic model, and to bound all kinetic parameters, metabolite concentrations, and enzyme concentrations. This procedure reduces the overall number of degrees of freedom within the kinetic model, while enabling the optimization procedure to rapidly converge to a single set of parameter values that satisfies all observations.

**Part II. Model reduction and dimensional analysis** We use the ensemble modeling approach[50,62] to reduce and de-dimensionalize the proposed kinetic model and to convert all unknown kinetic parameters and concentrations into bounded variables. This transformation eliminates time-scale variation and enables the model identification process to rapidly converge to best-fit kinetic parameters with low predictive errors.

First, we eliminate non-independent, redundant kinetic parameters by quantifying the net rate of a reaction in terms of a reaction reversibility. A reaction reversibility $R_i$ of 1 indicates that the $i^{th}$ reaction proceeds only in the forward direction; similarly, a $R_i$ of -1 indicates that the reaction proceeds only in the reverse direction, according to the equations:

$$R_i = \frac{v_{f,i} - v_{r,i}}{v_{f,i} + v_{r,i}} = -\frac{v_{i}^{net}}{v_{f,i} + v_{r,i}} = \frac{C}{v_{f,i} + v_{r,i}}$$

where we note how the net reaction rate and reaction reversibility are directly related. Due to carbon balancing under steady-state conditions, the net reaction rate on a per-carbon basis would become a constant $C$ for all reactions. Therefore, for the reference pathway, we assume
that its metabolic dynamics have reached a steady-state condition such that the $C$ becomes a constant. We do not assume steady-state conditions for non-reference pathway variants.

By substituting first or second order rate laws into the forward and reverse reaction rates, we can relate the kinetic parameters and metabolite concentrations to the reaction reversibilities, and eliminate an unknown variable. To illustrate, model reduction can express both the forward and reverse reaction rates using fewer unknown variables:

$$v_{f,i} = \frac{1}{2} C \left(1 - \frac{1}{R_{fi}}\right) \text{ and } v_{r,i} = -\frac{1}{2} C \left(1 + \frac{1}{R_{ri}}\right)$$

where the $C$ constant is a known quantity as it is directly related to the pathway’s production rate of neurosporene, which is experimentally measured. Consequently, for each pair of reactions in the reference pathway's kinetic model, we eliminate a non-independent variable.

Overall, there are 24 chemical species and 24 elementary reactions in the reaction network. Without model reduction and dimensional analysis, there are 48 unknown kinetic parameters quantifying the forward and reverse rates. Model identification can be performed on the non-reduced model, but 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. Alternatively, applying model reduction reduces the number of unknown variables to 24 unknown reaction reversibilities and 14 unknown bound enzyme concentration ratios, yielding 38 unknown variables. One additional unknown parameter defines the flux ratio between FPP and GPP entering the system, totaling 39 unknown parameters. These 39 unknown parameters are bounded and well-scaled; they vary between -1 and 1, or between 0 and 1.

Previous characterization of the pathway’s enzymes enables a further reduction in the number of unknown variables [226,227,228]. Bennet et. al. [227] found that 83% of studied
metabolites in *E. coli* central metabolism have intracellular concentrations that are much higher than their enzymes’ corresponding Km values, indicating that these enzymes operate near their maximum velocity. We accordingly imposed two additional assumptions in the kinetic model to constrain IPP and geranyl diphosphate (GPP) concentrations to be 10 times higher than their Km values in *ispA* and *idi*. We also set the ratio between production rate of isopentenyl diphosphate and dimethylallyl diphosphate to be 6:1 [226]. These constraints simplified the upper portion of the reaction network that governs flux distribution between IPP, GPP, and FPP, and reduced the total number of unknown variables to 33.

*Part III. Model Identification to Identify the Values of the Unknown Variable Model*

Model identification is an optimization procedure that compares a model’s solution to a data-set of measured observations. A sufficiently large set of accurate observations will allow precise identification of all unknown variables within a model. Here, we have 72 non-reference pathway variants and one reference pathway variant with accurately measured neurosporene productivities, and a kinetic model with 33 unknown variables. The solution to the kinetic model is the neurosporene production fluxes for the 72 non-reference pathway variants and the one reference pathway variant. We then use reference pathway variant's neurosporene production flux and measured neurosporene productivity to predict the non-reference pathway variants' neurosporene productivities. The quality of the model identification is determined by comparing the model’s predicted neurosporene productivities to the experimentally measured productivities. A perfect kinetic model would have equivalent comparisons across all 72 non-reference pathway variants.

We use genetic algorithm optimization to identify the values of the 33 unknown variables that maximizes the quality of the kinetic model across the 72 non-reference pathway variants. We performed ten separate and independent optimization runs to identify the best-fit values to the unknown variables. We used a genetic algorithm population size of 50, and a convergence
tolerance of 1% variation. We then selected the set of parameter values with the lowest error. The resulting kinetic model predicts the non-reference pathway variants' neurosporene productivities to within 32% of the measured productivities (Appendix A-Figure 5). The same kinetic model is able to predict the productivities of the 19 additional pathway variants to within 24%; the measurements of these pathway variants were not used during the model identification process. Taking into account all constraints, we then performed inverse model reduction to determine the 48 kinetic parameter values that define the identified kinetic model (Appendix A-Table 8).

The 24 non-reduced ordinary differential equations

\[
\frac{d[IPP]}{dt} = F_{IPP} + k_{-1}[CM1] - k_{1}[idi][IPP] + k_{-4}[CM3] - k_{4}[CM2][IPP] + k_{-7}[CM5]
- k_{7}[CM4][IPP] + k_{-10}[CM7] - k_{10}[CM6][IPP] + k_{-13}[CM9]
- k_{13}[CM8][IPP]
\]

\[
\frac{d[DMAPP]}{dt} = F_{DMAPP} + k_{2}[CM1] - k_{-2}[idi][DMAPP] - k_{3}[ispA][DMAPP] + k_{-3}[CM2]
\]

\[
\frac{d[GPP]}{dt} = k_{5}[CM3] - (k_{-5} + k_{6})[ispA][GPP] + k_{-6}[CM4] - k_{9}[CrtE][GPP] + k_{-9}[CM6]
\]

\[
\frac{d[FPP]}{dt} = k_{11}[CM7] - (k_{-11} + k_{12})[CrtE][FPP] + k_{-12}[CM8] + k_{8}[CM5]
- k_{-8}[FPP][ispA]
\]

\[
\frac{d[GGPP]}{dt} = k_{14}[CM9] - k_{-14}[CrtE][GGPP] - k_{15}[CrtB][GGPP]^2 + k_{-15}[CM10]
\]

\[
\frac{d[PPPP]}{dt} = k_{16}[CM10] - (k_{-16} + k_{17})[CrtB][PPPP] + k_{-17}[CM11]
\]

\[
\frac{d[Phytoene]}{dt} = k_{18}[CM11] - k_{-18}[CrtB][Phytoene] - k_{19}[CrtI][Phytoene] + k_{-19}[CM12]
\]

\[
\frac{d[Phytofluene]}{dt} = k_{20}[CM12] - (k_{-20} + k_{21})[CrtI][Phytofluene] + k_{-21}[CM13]
\]
\[
\frac{d[\xi - \text{carotene}]}{dt} = k_{22}[CM13] - (k_{-22} + k_{23})[CrtI][\xi - \text{carotene}] + k_{-23}[CM14]
\]
\[
\frac{d[\text{Neurosporene}]}{dt} = k_{24}[CM14] - k_{-24}[CrtI][\text{Neurosporene}]
\]
\[
\frac{d[CM1]}{dt} = k_1[idi][IPP] - (k_{-1} + k_2)[CM1] + k_{-2}[idi][DMAPP]
\]
\[
\frac{d[CM2]}{dt} = k_3[ispA][DMAPP] - k_{-3}[CM2] + k_{-4}[CM3] - k_4[CM2][IPP]
\]
\[
\frac{d[CM3]}{dt} = k_4[CM2][IPP] - (k_{-4} + k_5)[CM3] + k_{-5}[ispA][GPP]
\]
\[
\frac{d[CM4]}{dt} = k_6[ispA][GPP] - k_{-6}[CM4] + k_{-7}[CM5] - k_7[CM4][IPP]
\]
\[
\frac{d[CM5]}{dt} = k_7[CM4][IPP] - (k_{-7} + k_8)[CM5] + k_{-8}[ispA][FPP]
\]
\[
\frac{d[CM6]}{dt} = k_9[CrtE][GPP] - k_{-9}[CM6] + k_{-10}[CM7] - k_{10}[CM6][IPP]
\]
\[
\frac{d[CM7]}{dt} = k_{10}[CM6][IPP] - (k_{-10} + k_{11})[CM7] + k_{-11}[CrtE][FPP]
\]
\[
\frac{d[CM8]}{dt} = k_{12}[CrtE][FPP] - k_{-12}[CM8] + k_{-13}[CM9] - k_{13}[CM8][IPP]
\]
\[
\frac{d[CM9]}{dt} = k_{13}[CM8][IPP] - (k_{-13} + k_{14})[CM9] + k_{-14}[CrtE][GGPP]
\]
\[
\frac{d[CM10]}{dt} = k_{15}[CrtB][GGPP]^2 - (k_{-15} + k_{16})[CM10]
\]
\[+ k_{-16}[\text{PrePhytoene diphosphate}][CrtB]
\]
\[
\frac{d[CM11]}{dt} = k_{17}[CrtB][\text{PrePhytoene diphosphate}] - (k_{-17} + k_{18})[CM11]
\]
\[+ k_{-18}[\text{Phytoene}][CrtB]
\]
\[
\frac{d[CM12]}{dt} = k_{19}[CrtI][\text{Phytoene}] - (k_{-19} + k_{20})[CM12] + k_{-20}[\text{Phytofluene}][CrtI]
\]
\[
\frac{d[CM13]}{dt} = k_{21}[CrtI][\text{Phytofluene}] - (k_{-21} + k_{22})[CM13] + k_{-22}[\xi - \text{carotene}][CrtI]
\]
\[
\frac{d[CM14]}{dt} = k_{23}[CrtI][\xi - \text{carotene}] - (k_{-23} + k_{24})[CM14] + k_{-24}[Neurosporene][CrtI]
\]

The mole balances summing together the free and bound forms of the five enzymes

\[
[idi] + [CM1] = [idi]_{\text{total}}
\]
\[
[ispA] + [CM2] + [CM3] + [CM4] + [CM5] = [ispA]_{\text{total}}
\]
\[
[CrtE] + [CM6] + [CM7] + [CM8] + [CM9] = [CrtE]_{\text{total}}
\]
\[
[CrtB] + [CM10] + [CM11] = [CrtB]_{\text{total}}
\]
\[
[CrtI] + [CM12] + [CM13] + [CM14] = [CrtI]_{\text{total}}
\]

Flux model constraints

The metabolic flux of IPP and DMAPP entering the reaction network has a ratio of about 6:1. This is enforced through a model constraint. \( F_{IPP} = \frac{6}{7}, F_{DMAPP} = \frac{1}{7} \)

Appendix A-Method 4: Using Computational geometry for building a non-mechanistic model of the Carotenoid pathway

There are many types of non-mechanistic models that can capture high-dimensional relationships between inputs (translation rates) and outputs (pathway productivities). As an example, we created a non-mechanistic model that uses computational geometry and linear interpolation to calculate a pathway variant’s productivity within a translation space region.

The computational geometric model first partitions the translation rate space into small Voronoi polygons [229]. The center of each polygon is a vertex in the 4-dimensional space, defined by the predicted translation rates and measured productivities of a pathway variant. There are 73 Voronoi polygons to represent the 73 pathway variants in the initially characterized data-
set. The interior volume of the translation rate space is bounded by the 73 polygon’s convex hull. The convex hull forms the border of the translation rate regions where interpolation may be used to calculate the pathway productivity of any point within the interior volume.

Next, we employed linear interpolation to calculate the productivity of a pathway whose enzymes have translation rates within the interior volume of the convex hull. The productivity of the pathway is the distance-weighted sum of the productivities from the surrounding centers of Voronoi polygons (Appendix A-Figure 9A). It is not possible to use linear interpolation to calculate a pathway’s productivity when its translation rates exist outside the convex hull; accordingly, these regions are shown as white areas. 56 out of initially characterized 73 pathway variants have translation rates that exist within the convex hull. Linear interpolation is able to reproduce the measured pathway variants’ productivities to within 32%. For the 19 additionally characterized pathway variants, only 8 have translation rates within the convex hull and linear interpolation is able to predict these pathways’ productivities to within 15% (Appendix A-Figure 9B).

Appendix A-Method 6: Using statistical model for creating a non-mechanistic model of the Carotenoid pathway

Recently Lee et al.[2] engineered Violacein pathway in *Saccharomyces cerevisiae* by developing a linear statistical model that predicts synthesis of the pathway’s intermediate and final products. The model’s coefficient of determination ($R^2$) for the log(Violacein) prediction was 0.64 and their model only required characterization data for less 3% of their entire combinatorial library (91 separate clones out of 3125 potential variants). The relatively good $R^2$ and simplicity of the model motivated us to examine the statistical model for constructing a SEAMAP of Neurosporene biosynthesis pathway.
To build a linear statistical framework, an independent categorical state $i$ is defined as any combination of RBS variants for every targeted gene (total number of states=$P$), and an unknown parameter ($\beta_i$) is assigned for each state to include its fitness in the model’s predictions ($Y$). Combinatorial space of characterized pathway variants is mapped by assigning a binary variable ($X_{j,i}$) for present or absent of categorical variant $i$ in characterized pathway variant $j$ (Total number of characterized pathway variants=$N$). Since the predictions have an arbitrary unit, an extra parameter ($t$) is added as a linear translation operator. This map could be interpreted as a mathematical transformation from gene expression space to a virtual high-dimensional simplex $A$ with $P$ vertices ($a_1=[\beta_1,0,0,\ldots,0]_{1xP}$, $a_2=[0,\beta_2,0,\ldots,0]_{1xP}$, ..., $a_P=[0,0,0,\ldots,\beta_P]_{1xP}$). The final linear regression model is presented as:

$$\log(Y_j) = t + X_{j,i}\beta_i \quad i = 1 \cdots P \quad j = 1 \cdots N$$

The common approach to identify the unknown parameters of a linear regression is least-square. However, according to Lee et al., this may result in poor fitting and unfavorite linear transformation. Instead, Exterior Derivative Estimator (EDE) method[71] could find better values for the $P+1$ model parameters by learning the intrinsic data constraints[2].

We developed a linear statistical model for Neurosporene biosynthesis pathway. According to Lee et al.[2], a relatively low number of pathway variants, between 1% and 2% (around 30 to 60 clones of their potential 3125 pathway variants), was sufficient to train an accurate predictive model for Violacein pathway. Hence, we chose 73 synthetic Neurosporene pathway variants as the training set (around 2% of total combinatorial library) and a set of additional 19 variants for model validation. The training set contained 44 independent RBS variants for CrtE, CrtB, and CrtI, resulted in a final linear regression model with 45 unknown parameters ($P=44$). We used EDE method to train the model. This category-based model can only provide information about a pathway variant if its RBS variants exist in the training set. One CrtI RBS variant (TIR=312 au) was present in two pathway variants (out of 19 variants) of the test set.
but did not exist in any of the 73 training variants thus, these data were excluded from error calculation. The average error of productivity prediction for the remaining test variants was 46%.

The proposed statistical model can only perform limited interpolation in discrete intervals of Neurosporene expression-activity space. This prevents the model from estimating the productivity of novel pathway variants (with uncharacterized RBS sequences). As a result, the model fails to identify highly productive pathways; the best computed productivity by the linear model is 145 µg/gDCW/hr while the highest measured productivity among the training set and all characterized variants are 196 and 286 µg/gDCW/hr respectively.

Appendix A-Method 7: Performance of kinetic, geometry, and statistical modeling for building a SEAMAP

Biophysical modeling of gene expression[37,78,125] could be coupled with a kinetic, statistical, or geometry modeling to quantitatively link microorganisms’ DNA sequences with their phenotypic behaviors. Availability and quality of characterization data, complexity of the pathway, and application of the map affects the competency of the choices. Here, we use Neurosporene biosynthesis pathway to compare these models. We analyze their interpolation and extrapolation predictive powers, the effect of data throughput on their predictions, and their ability to falsify incorrect data acquisition and mathematical transformation.

The first criterion for selecting a model is usually its interpolation and extrapolation power. We evaluated interpolation power of the trained kinetic, geometry, and statistical models by estimating pathway productivity for a set of 19 Neurosporene pathway variants. These models were able to provide information for 100%, 42%, and 89% of the test variants with average errors of 24.8%, 12.5%, and 46% respectively. We then tested their extrapolation power for additional 28 pathway variants with high enzyme expression levels. The kinetic model estimated the
productivity for the entire set with 17% error. However, both geometry and statistical models failed to reveal any information about the extrapolation set.

The smallest number of measurements to develop a reliable model is also an essential model criterion. To study the effect of data throughput, we varied the training set size by randomly selecting 5 to 73 pathway variants, retraining the kinetic, geometry, and statistical models, and calculating the average relative error for 19 test variants. Any pathway variant, which existed in spatial gaps of geometry or statistical models, was excluded from their error calculations. With any number of training data, the kinetic model evaluated the behavior of entire test set while geometry model estimated the productivity of the only variants that were inside the convex hull formed by the training set data (Appendix A-Figure 10A). The later model locally has a great predictive power and enables smooth and continuous predictions but requires large amount of data to cover a large region in expression-activity space. Unlike kinetic and geometry models, the linear statistical model is a discrete model and, it can only predict the phenotypic behavior of a pathway variant if all its RBS sequences exist as parts of training set pathways. The number of training measurements significantly influences predictions of this model (Appendix A-Figure 10B).

An incorrect mathematical transformation could hide intrinsic characteristics of data, and falsely show high predictive power for a simplified model. For instance, just by changing error definition from Formula S1 to Formula S2, one can artificially lower prediction error of an identical statistical model from 46% to 10%. Thus, any new model needs to be evaluated for false-positive rejection. Here, as a null hypothesis, we trained the kinetic, geometry, and statistical models with randomly generated data (instead of the actual experimental productivity measurements) and calculated error of prediction for the existing test set. We generated 73 random numbers (Appendix A-Figure 11A) utilizing a normal distribution random generator (with mean and standard deviation of actual log (productivity) for training set), and trained the
proposed models with randomly selected training sets that contained 10 to 73 generated variants. A good model that was trained by the actual measurements should provide small prediction error for the test set (Ea) while training the same model with the randomly generated data should result in a large calculation error (E_R). The E_R/Ea ratio determines the maximum error dynamic range that could be achieved by the model using the training and test sets. We define successful null-hypothesis rejection for a model if its dynamic range is at least 2-fold. When using Formula S1 for defining error, we observed successful null-hypothesis rejection for the kinetic, geometry, and statistical models (Appendix A-Figure 11BCD). Lee et al.[125] compared log(measurements) with log(predictions) by a statistical model for estimating Violacein production and achieved relatively high R². To study the effect of log transformation on results of Neurosporene statistical model, we used Formula S2 for error calculation of the model. As a result, error of prediction from identical statistical model substantially reduced (10% error). The calculated error for the null hypothesis also dropped to 15.5% and resulted in model failure for rejecting the null-hypothesis (Appendix A-Figure 11E). Log-transformation is commonly used for linearizing data. However, this mathematical transformation could be misleading by artificially lowering the prediction error for quantities that vary only one or two orders of magnitudes.

\[ \text{Error 1(\%) = 100 \times \sum_{i=1}^{N} \text{abs} \left( \frac{\text{Measurement}_i - \text{Prediction}_i}{\text{Measurement}_i} \right) } \]  

Formula S1

\[ \text{Error 2(\%) = 100 \times \sum_{i=1}^{N} \text{abs} \left( \frac{\text{log(Measurement}_i) - \text{log(Prediction}_i)}{\text{log(Measurement}_i)} \right) } \]  

Formula S2

A summary of strengths and weaknesses of kinetic, geometry, and statistical modeling for building a SEAMAP has been provided in Appendix A-Table 18. Success rate of statistical models is a function of data richness; when high-throughput data is available, statistical models might be the simplest (not an accurate way) to develop a SEAMAP. Geometry models perform precise interpolation but require massive characterizations for large pathways. These models could be used locally to calculate a quantity such as flux control coefficients of a pathway.
variant's enzymes without acquiring information about the entire expression space. Kinetic modeling utilizes the pre-existing biochemistry information (ex. Km of each enzymes and stoichiometry of the involved reactions) and it only requires small number of characterizations to provide information across sequence-expression activity map.
Appendix B Supplementary Information for Chapter 2

Appendix B-Figures

Appendix B-Figure 1: Effect of additional TF binding sites on performance of circuits containing Lacl and a TetR-homolog TF; 1 (O), 2(◊), 3 (X), 4(○), 5(☆), and 9(□) TF target sites.
Appendix B-Figure 2: Model parameterization for calculating RNA polymerase binding energy. The blue line shows the relative error of model prediction for circuits containing 1, 2, 3, 4, and 5 TF targets. The black lines show the maximum and minimum changes in error when model parameters are perturbed 50% including number of RNA polymerases, apparent degradation rate of TF, and circuit copy number.
Appendix B-Figure 3: Model parameterization for calculating TF binding energy.
The blue line shows the relative error of model prediction for circuits containing 1, 2, 3, 4, and 5 TF targets. The black lines show the maximum and minimum changes in error when model parameters are perturbed 50% including number of RNA polymerases, apparent degradation rate of TF, and circuit copy number.
Appendix B-Figure 4: Protein expression rate at different growth rate. The data were collected from Klumpp et al.[132]. The measurements with doubling time of 30 minutes or longer were fitted to a power-law model ($R^2=0.98$).

Appendix B-Figure 5: Characterization of PhIF circuits in reverse orientation: 1 (O), 3(◊), and 5(X) TF target sites. Transcription rate of TF is reduced when the circuit orientation changes (pir+ strain). Correcting for cooperativity between replication fork and RNA polymerase has little/no effect on model predictions for simple circuits containing only one PhIF target. However, the functionality of circuits with multiple PhIF binding sites (3 or 5) deviates from uncorrected models (specially inside the box). The solid lines show model predictions, black for the circuit containing 1 PhIF target, red for the circuit with 3 PhIF targets, and green for the circuit containing 5 PhIF targets; with correction: $\varphi=0.75$, without correction: $\varphi=1$. 
Appendix B-Figure 6: Characterization of PhlF circuits in multiple industrial microbial hosts; 1 (O), 3(◊), 5(X), and 9(☆) TF target sites.

Appendix B-Figure 7: Effect of cross activity on performance of genetic opAMP circuit. The cross activity of AmtR and LmrA transcription factors[85] reduces the maximum output transcription rate of the LmrA-cognate promoter. The red diamonds: output transcription rate from LmrA-cognate promoter in the absence of LmrA gene; the black circles: output transcription rate of AmtR-LmrA genetic opAMP; the blue diamonds: output transcription rate by the IPTG-inducible promoter that controls AmtR expression level.
Appendix B-Figure 8: Thermodynamic states for binding of a TF and RNA polymerase to a promoter in presence of one additional TF binding site.
Appendix B-Tables

Appendix B-Table 1: Strains used in the study. The table contains all strains used to examine the constructed genetic circuits.

<table>
<thead>
<tr>
<th>Names</th>
<th>Descriptions</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> DH10B</td>
<td>F– mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara leu) 7697 galU galK rpsL nupG λ–</td>
<td>[230]</td>
</tr>
<tr>
<td><em>E. coli</em> pir+</td>
<td>Escherichia coli KmR M13ΔuidA::pir+</td>
<td>[231]</td>
</tr>
<tr>
<td><em>E. coli</em> BL21</td>
<td>Escherichia coli BL21</td>
<td>[232]</td>
</tr>
<tr>
<td><em>S. typhimurium</em></td>
<td>Salmonella typhimurium LT2</td>
<td>[134]</td>
</tr>
<tr>
<td><em>C. glutamicum</em></td>
<td>Corynebacterium glutamicumB-2784 (ATCC# 13032)</td>
<td>[233]</td>
</tr>
<tr>
<td><em>P. fluorescens</em></td>
<td>Pseudomonas fluorescens B2</td>
<td>[234]</td>
</tr>
</tbody>
</table>

Appendix B-Table 2: Characterization data for all genetic circuits containing LacI and a TetR-homolog TF. The characterization data for genetic circuits containing LacI and one TetR-homolog TF are located in two Excel worksheets (values and standard deviations). For each circuit, the table includes the TF translation rate, utilized expression vector, host, growth media and temperature, and the measured output fluorescence values at 11 IPTG concentrations. All translation rates were converted to an absolute scale (Protein/ (mRNA. s)). The “auxiliary sites” column shows the number of additional TF binding sites placed on the same expression vector.

Appendix B-Table 3: Characterization data for all genetic opAmp circuits. The characterization data for genetic circuits containing LacI and two TetR-homolog TFs are located in two Excel worksheets (values and standard deviations). The table includes the name and translation rate of each TF, as well as circuit output at 11 IPTG concentrations spanning between 0 and 1mM. All translation rates were converted to an absolute scale (Protein/ (mRNA. s)).
Appendix B-Methods

Appendix B-Method 1: Developing a statistical thermodynamic model of genetic circuits containing multiple TF binding sites.

Building a thermodynamic model of transcription requires two steps: 1) calculating the number of free TF dimers from total number of expressed TF molecules. 2) Enumerating all possible ways for binding of RNA polymerase to calculate its binding probability that is proportional to its output transcription rate. The number of available TF dimers (R) is a function of total expressed TF, the number of specific and non-specific binding sites, and TF binding energy for each DNA loci. TetR-homolog TFs usually bound as a pair of dimers or high order dimers. Binding of the additional dimers are involved in a cooperative process; binding of the first dimer is usually enthalpy-driven while the remaining is entropy- and/or enthalpy-driven[113]. Here, we assumed that the binding is a sequential process with minimal changes in the binding energy of higher-order dimers (d: number of dimers). This allowed us to only consider one binding energy for binding all the dimers to one DNA site (ΔGR), although the formulation could be simply expanded to include the cooperativity energy. We performed a mass balance on the total dimers (Rt) to estimate the number of bound and unbound molecules at thermodynamic equilibrium, in presence of N specific target sites and LG non-specific sites (Eq. 1).

\[ R + N \sum_{i=1}^{d} \left( \frac{e^{-\beta \Delta G_R}}{L_G} R \right)^i = R_t \]  

To estimate the rate of output transcription, we first developed a thermodynamic model of single-copy repression circuits containing multiple TF binding sites. In these circuits, the binding competition between TF and RNA polymerase dictates promoter occupancy by RNA polymerase while rate of open complex formation (rmax) determines the maximum output
transcription rate. We used the previously proposed thermodynamic states[109,235] for binding of RNA polymerase and TFs to build a simple repression model. The thermodynamic states for presence of two TF binding sites were shown in Appendix B-Figure 8. In this case, the binding energy of RNA polymerase (ΔG_{RNAP}), number of available RNA polymerase (P), intracellular concentration of free TF (R), and TF binding energy to the target sites (ΔG_{R}) determined the output transcription rate (r):

\[ r = \frac{e^{-βΔG_{RNAP}}P \left(1 + \sum_{i=1}^{d} \left( \frac{e^{-βΔG_{R}}}{L_G} \right)^i \right)}{1 + e^{-βΔG_{RNAP}}P \left(1 + \sum_{i=1}^{d} \left( \frac{e^{-βΔG_{R}}}{L_G} \right)^i \right)^2} \]  

(2)

Addition of TF target sites increases the number of thermodynamic states in the system. These new states alter the circuit dynamics, and must be incorporated in the model. For instance, when there are three TF binding sites, one near the promoter and two outside the expression cassette, the output transcription rate can be calculated by:

\[ r = \frac{1 + 2 \sum_{i=1}^{d} \left( \frac{e^{-βΔG_{R}}}{L_G} \right)^i + \left( \sum_{i=1}^{d} \left( \frac{e^{-βΔG_{R}}}{L_G} \right)^i \right)^2}{1 + 2 \sum_{i=1}^{d} \left( \frac{e^{-βΔG_{R}}}{L_G} \right)^i + \left( \sum_{i=1}^{d} \left( \frac{e^{-βΔG_{R}}}{L_G} \right)^i \right)^2} \]  

(3)

In presence of N target sites where N-1 sites are outside of the expression cassette, the model can be expanded to Eq. 4. In this formulation, in both numerator and denominator, the summation term inside the curly brackets represents the total number of binding states that the TF can bind with the N-1 auxiliary sites. By crossing out this term, we derived a simplified equation (Eq. 5).
When the circuit copy number \((c)\) increases, the number of possible ways that the TF and RNA polymerase can bind to the promoter changes. Recently, a comprehensive model has been proposed\([100]\) that consider many thermodynamic state of TF binding. However, this model is extremely non-linear, and could result in high error propagation in estimated parameters during our non-linear error minimization. Instead, we assumed that a TF dimer can independently bind with the promoter or the auxiliary sites. The independency assumption allowed us to enumerate total TF binding states as the multiplication of the number of binding states. The binding probability of RNA polymerase to the promoter is proportional to the fraction of all thermodynamic states that RNA polymerase can bind with the promoter. Using binomial expansion, we approximated the number of states for binding of RNA polymerase \((N_{RNA\text{-}promoter})\) and TF \((N_{TF\text{-}promoter})\) by Eq. 6 and Eq. 7 respectively.

\[
N_{RNA\text{-}promoter} = \left(1 + \frac{e^{-\beta \Delta G_{RNA\text{-}P}}}{L_G} \right)^c - 1 \tag{6}
\]

\[
N_{TF\text{-}promoter} = \left[1 + \sum_{i=1}^d \left(\frac{e^{-\beta \Delta G_{R}}}{L_G} \right)^i \right]^c - 1 \tag{7}
\]

In this case, the output transcription rate can be calculated as the fraction of \(r_{\text{max}}\) that RNA polymerase is bound to the promoter (Eq. 8):

\[
r = \frac{\left(\sum_{j=0}^{\binom{N-1}{j}} \left(\frac{\sum_{i=1}^d \left(\frac{e^{-\beta \Delta G_{R}}}{L_G} \right)^i}{1 + e^{-\beta \Delta G_{RNA\text{-}P}} L_G + \sum_{i=1}^d \left(\frac{e^{-\beta \Delta G_{R}}}{L_G} \right)^i} \right) \right)^c}{r_{\text{max}}} \tag{4}
\]

\[
r = \frac{\frac{e^{-\beta \Delta G_{RNA\text{-}P}}}{L_G}}{1 + \frac{e^{-\beta \Delta G_{RNA\text{-}P}}}{L_G} + \sum_{i=1}^d \left(\frac{e^{-\beta \Delta G_{R}}}{L_G} \right)^i} r_{\text{max}} \tag{5}
\]
For any prediction during model parameterization, we calculated the output transcription rate by utilizing Eq. 1 and 8. The lowest level of background was set as about 5% of auto fluorescence measurements, and any lower calculated transcription rate was adjusted to this level. For estimating the number of free dimers, we used MATLAB to calculate the largest root of the polynomial (Eq. 1) that was smaller than $R_t$. The remaining parameters are context-specific and are fixed or estimated during model parameterization.

\[ r' = \left( \frac{1+e^{-\beta \Delta G_{RNAP}}}{L_G} \right)^c - 1 \left[ \left( \frac{1+e^{-\beta \Delta G_{RNAP}}}{L_G} \right)^c + \left( 1+\sum_{i=1}^{d} \left( \frac{e^{-\beta \Delta G_{R_i}}}{L_G} \right)^c \right) \right]^{-1} \]

(8)

For any prediction during model parameterization, we calculated the output transcription rate by utilizing Eq. 1 and 8. The lowest level of background was set as about 5% of auto fluorescence measurements, and any lower calculated transcription rate was adjusted to this level. For estimating the number of free dimers, we used MATLAB to calculate the largest root of the polynomial (Eq. 1) that was smaller than $R_t$. The remaining parameters are context-specific and are fixed or estimated during model parameterization.

**Appendix B - Method 2: Characterizing genetic circuits in diverse genetic context and environmental conditions**

We re-characterized PhlF genetic circuits in different growth conditions as well as in 4 industrial gram positive and gram negative bacterial hosts, and re-evaluated predictions by the developed model. The *Salmonella* was characterized in minimal media while the remaining hosts were grown in rich media and at their optimum growth temperatures. In all cases, we kept the PhlF binding energy constant while other systems parameters were changed accordingly.

Specific growth rate was the only modified parameter when we examined different growth media (minimal or rich) and temperature (30 °C or 37 °C). This parameter allowed us to correctly predict effect of change in the growth media (Figure 2). Change in temperature has multiple effects on TF productions rate such as change in growth rate, transcription rate, translation elongation rate, and mRNA and protein degradation rates[236,237,238,239]. Here, we only considered the growth rate as a factor that alters total TF dimer concentration. The model showed a correct trend in the circuit out with a change in temperature (Figure 2).
To examine the effect of change in genetic context, we first re-characterized PhlF genetic circuits containing no, 2, and 4 PhlF auxiliary sites in *E. coli* BL21 (Appendix B-Figure 6). The new host’s genome was slightly smaller than *E. coli* DH10B; therefore, we altered the number of non-specific binding sites. We also changed specific growth rate in the model accordingly, since the circuit constructs grew faster in BL21 than DH10B.

We then characterized the circuits in *Salmonella*, and re-evaluated PhlF model parameters. The result revealed that although the PhlF and RNA polymerase binding energies remained unchanged, the apparent degradation rate \((\delta_m \delta_p)^{-1}\) must be reduced to about one-third. This difference could be as a result of change in PhlF transcription, mRNA degradation, protein folding, and protein degradation. A further characterization of the input signal in *E. coli* DH10B and *Salmonella* rejected the presence of a significant change in input PhlF transcription rate between these hosts. PhlF is a heterologous protein in Salmonella, and difference is likely due to lower mRNA or protein stability.

PhlF was originally purified from *P. fluorescens* that controls Phl operon[240]. We characterized the PhlF circuits (containing no, 2, and 4 auxiliary sites) inside the native host. The host’s genome was about 40% larger than *E. coli*’s genome that increased the number of non-specific binding sites. While the calculated PhlF binding energy did not change, \(P \times \exp (-\beta \Delta G_{RNAP})\) was changed about 5-fold. This might be due to the effect of larger number of free RNA polymerase, higher RNA polymerase binding affinity to its native promoter, or different rate of random walk by RNA polymerase to this promoter.

We finally transformed the PhlF circuits (containing no, 2, and 4 auxiliary sites) into *C. glutamicum*, and re-evaluated the model parameters. This host has a different 16S rRNA sequence than *E. coli* DH10B. We used our model of translation[78,87,125] to estimate the change in PhlF translation rate.
Appendix B-Method 3: Designing genetic circuits in Pt design space

The choice of genetic parts, circuit-harboring host, and growth conditions control dynamics of genetic circuits, and must be properly adjusted to develop genetic circuits with correct outputs. These choices create a complex combinatorial design space of unknown parameters with few numbers of correct solutions. Using Pt number, we can transform design space of genetic circuits into a new space with small number of independent unknowns. Here, we have used this design space to design 4 digital and 2 analog genetic circuits: NOT, NOR, NAND, AND, analog-digital convertor, and toggle switch (Figure 5).

We first transformed the design space of all unknown parameters to a dimensionless design space. We used Eq. 9 to relate output transcription rate of a promoter to the characteristics of its cognate TF, encapsulated as a Pt value. This allowed us to derive governing equations of these circuits as functions of Pt values for each TF and a dimensionless number that represents maximum change by each promoter with respect to input signal.

\[
\frac{r_{\text{max}}}{r} = 1 + 0.5Pt^2
\]  

(9)

For the NOT gate circuit, the governing equation is identical to Eq. 9. For the NOR gate, change in each input transcription rate proportionally alters Pt value of the TF. For simplicity, we have defined two Pt values each for one input signal. In this case, the apparent Pt value for this TF is the summation of both Pt values, and the output transcription rate becomes solely a function of this apparent value. Therefore, the NOR gate governing equations was simplified to:

\[
\frac{r_{\text{max}}}{r} = 1 + 0.5(Pt_1 + Pt_2)^2
\]  

(10)
The NAND gate is composed of two TFs that independently controlled two output promoters. The total output transcription rate is a summation of transcription rates from these two promoters. When the maximum theoretical transcription rates of these promoters are equal, the governing equations are simplified to Eq. 11:

\[ \frac{r}{r_{\text{max}}} = \frac{1}{1 + 0.5Pt_{1}^2} + \frac{1}{1 + 0.5Pt_{2}^2} \]  

(11)

The AND gate is constructed by assembling a NAND and a NOT gate. This gate contains three TFs, and its output can be calculated by solving Eq. 9 and Eq. 11 when the output transcription rates for all promoters are similar. In this case, the governing equations can be simplified to Eq. 12. The $Pt_{3,\text{max}}$ is the maximum $Pt_3$ value for the TF3.

\[ \frac{r_{\text{max}}}{r} = 1 + 0.5 \left[ Pt_{3,\text{max}} \left( \frac{1}{1 + 0.5Pt_{1}^2} + \frac{1}{1 + 0.5Pt_{2}^2} \right) \right]^2 \]  

(12)

The analog-digital convertor circuit contains two cascaded TFs. The output transcription rate of the first TF’s cognate promoter can be calculated by Eq. 9. Change in activity of this promoter alters intracellular level of TF2 and affects circuit output. Therefore, the $Pt$ number of the second TF must be properly adjusted to accommodate for the change (Eq. 13). The $Pt_{2,\text{max}}$ is the maximum $Pt_2$ value for the TF2.

\[ Pt_{2} = \frac{r_{a}}{r_{a,\text{max}}} \cdot Pt_{2,\text{max}} = \frac{Pt_{2,\text{max}}}{1 + 0.5Pt_{1}^2} \]  

(13)

\[ \frac{r_{\text{max}}}{r} = 1 + 0.5Pt_{2}^2 \]  

(14)
Similarly, the toggle switch contains two TFs. However, the transcription rate of each TF is regulated by the other TF. Therefore, the output transcription rate must be calculated by simultaneously solving the governing equations (Eq. 15-18). The $P_{t1,max}$ and $P_{t2,max}$ are the maximum $P_{t1}$ and $P_{t2}$ values for the TF1 and TF2 respectively.

$$\frac{r_a}{r_{a,max}} = \frac{1}{1+0.5P_{t1}^2}$$  \hfill (15)

$$\frac{r_b}{r_{b,max}} = \frac{1}{1+0.5P_{t2}^2}$$  \hfill (16)

$$P_{t1} = \frac{r_b}{r_{b,max}} P_{t1,max}$$  \hfill (17)

$$P_{t2} = \frac{r_a}{r_{a,max}} P_{t2,max}$$  \hfill (18)

We then identified the design criteria that resulted in correct circuit output. To maximize signal inversion by the NOT gate, the change in input transcription rate must change the $P_t$ number from a small value (lower than 1) to a large value (larger than 10). To obtain a safe-design criterion, we assumed that the maximum dynamic range of the promoter is 1000-fold, and further increase the TF expression keeps the system in the OFF state. Using Eq. 9, this provided a safe lower-limit of $P_t$ for OFF-state ($45 \leq P_t$). Also, we considered ON state when the output transcription rate is within 50% of $r_{max}$. This provided an upper-limit of $P_t$ value for ON-state ($P_t \leq 1.41$). Therefore, a safe-design criterion for NOT-gate circuit is designing circuit components to alter $P_t$ value passing the entire calculated rage when the input transcriptional signal changes. For instance, one can adjust translation rate of the TF appropriately to change $P_t$ within the calculated range.

For NOR gate, the apparent $P_t$ value is the summation of individually defined $P_t$ values. Therefore, the calculated $P_t$-range for the NOT gate can be applied to the individual $P_t$ values as well as their summations.
For the NAND gate, each TF must suppress its cognate promoter completely when the input transcription signal is at its maximum level. Therefore similar to NOT gate, the individual Pt values must vary across a range wider than the calculated Pt-range (starts before and finishes after the range).

Since, the AND gate is composed of a NAND and a NOT gate, the Pt-values for the first two TFs must follow the NOT-gate safe criterion. The maximum Pt value for the third TF must be large enough to completely invert the signal, for instance between 10 and 100.

For the analog-digital convertor circuits, similarly the Pt value of the first TF must follow the proposed criterion to maximize the fold change in the input transcriptional signal of the second TF. The $\text{Pt}_{2,\text{max}}$ is a constant unknown parameter and must be adjusted such that the second TF can completely suppression the output promoter. Therefore, as a safe-design criteria $\text{Pt}_{2,\text{max}}$ should be larger than 50.

For the toggle switch, the set of nonlinear equations may result in one or multiple thermodynamic states for the circuit outputs. The Pt values must be selected from the regions that have multiple states, and the output of these states spans in a large range, for instance, about 100-fold change.

Designing genetic circuits in Pt space reduces the design complexity; however, it may increase error of outcome prediction when determining thermodynamic state of each regulation solely as a function of Pt. Therefore, the selected design criteria must be selected conservatively to increase design success rate.
Appendix C Supplementary Information for Chapter 3

Appendix C-Figures

Appendix C-Figure 1: Measured versus predicted DNA cleavage in the *in vitro* experiments

Appendix C-Figure 2: Sensitivity analysis of individual kinetic parameters.
Appendix C-Figure 3: Measured versus predicted dCas9 repression activity in the *in vivo* experiments.
Appendix C-Figure 4: Quantifying the effect of crRNA:target mismatches using currently available RNA:RNA and DNA:DNA energy parameters. For each dataset (Table 1), a set of 21 positional weights were determined that minimized the error of model predictions. (A) Positional weights for mismatches at different locations of a target. (B) The calculated exchange energy for each base-pair as the difference between RNA:RNA and DNA:DNA energy parameters using available energy values. (C) Predictions versus measurements for $\Delta\Delta G_{\text{exchange}}$. Pearson correlation of 0.56 and 0.26 for dataset I and dataset II respectively.
Appendix C-Figure 5: Comparison between the in vitro and in vivo measurements and predictions by Cas9 model containing the second proposed dimerization mechanism. A reaction was added to the proposed Cas9 model for conversion of the dimerized molecule (Cas9-Cas9:crRNA) to active complex, assuming that it can bind and cleave the target DNA with a rate similar to Cas9:crRNA. In vitro model parameterization only altered the value of $k_{di}/k_1$ from 0.035 to 0.2 while the other parameters remained unchanged. (A) time-dependent Cleavage rate predictions (line) and measurements (circle) for the supercoiled target. (B) time-dependent predictions (line) and measurements (circle) for the non-supercoiled target. (C) Model prediction versus (black) and measured (green) total cleavage rate for non-supercoil DNA after about 30 minute incubation. (D) In vivo model validation. Model was re-parameterized using the new $k_{di}$ value. The calculated $\Delta\sigma$ and $\Delta G_{PAM}$ remained unchanged.
Appendix C-TABLES

### Appendix C-Table 1: Parameters used in (d)Cas9 activity analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_f$</td>
<td>-0.06</td>
<td>--</td>
</tr>
<tr>
<td>$N_{\text{Cas9}}$</td>
<td>1000</td>
<td>molecule</td>
</tr>
<tr>
<td>$N_{\text{crRNA}}$</td>
<td>1000</td>
<td>molecule</td>
</tr>
<tr>
<td>Doubling time</td>
<td>20</td>
<td>hr</td>
</tr>
<tr>
<td>$N$</td>
<td>$6.4\times10^9$</td>
<td>DNA sites</td>
</tr>
<tr>
<td>$\Delta G_{\text{target}}$</td>
<td>-9.9</td>
<td>kcal/mol</td>
</tr>
</tbody>
</table>

### Appendix C-Table 2: Parameters used in genome-wide calculation for $\lambda$-phage plasmid

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{\text{Cas9}}$</td>
<td>10</td>
<td>nM</td>
</tr>
<tr>
<td>$C_{\text{crRNA}}$</td>
<td>100</td>
<td>nM</td>
</tr>
<tr>
<td>$N$</td>
<td>97000</td>
<td>DNA sites</td>
</tr>
<tr>
<td>Reaction volume (V)</td>
<td>100</td>
<td>$\mu$L</td>
</tr>
</tbody>
</table>

### Appendix C-Table 3: Parameters used in HIV calculation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_f$</td>
<td>-0.06</td>
<td>--</td>
</tr>
<tr>
<td>$N_{\text{Cas9}}$</td>
<td>100</td>
<td>molecule</td>
</tr>
<tr>
<td>$N_{\text{crRNA}}$</td>
<td>100</td>
<td>molecule</td>
</tr>
<tr>
<td>Doubling time</td>
<td>36</td>
<td>hr</td>
</tr>
<tr>
<td>$N$</td>
<td>$6.4\times10^9$</td>
<td>DNA sites</td>
</tr>
</tbody>
</table>
Appendix D- Optimizing Furfural biodetoxification pathway

The economic viability of many bioproducts (e.g. biofuels) requires an inexpensive feedstock. Recently lignocellulose feedstock has received a lot of attention since it can be simply converted to cheap sugar feedstock using a pretreatment process[241]. However, during this process several byproducts are also being produced that have toxicity effect on microbial growth such as Furfural and hydroxy-methy furfural (HMF). Small amount of these toxins reduces \textit{E. coli} growth by 50% (Appendix D-Figure 1).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Toxicity effect of Furfural and HMF on cellular growth. \textit{E. coli} was grown in minimal M9 media supplemented with 0.4\% glucose. The dash line represents the doubling time in absence of Furfural and HMF.}
\end{figure}

Recently, Koopman et. al. [242] characterized a metabolic pathway for furan detoxidation in \textit{C. basilensis}. In this pathway, shown in (Appendix D-Figure 2), during eight enzymatic reactions, HMF and furfural are being converted to 2-oxoglutaric acid which is a TCA cycle.
intermediate. Consequently, this pathway detoxifies these toxins by producing an energy resource for microbes.

![Diagram of the Furfural and HMF biodetoxification pathway]

**Appendix D-Figure 2: Furfural and HMF biodetoxification pathway**

Here, we used our biophysical approach to optimize the expression of HmfD, as one of the first enzymes in the metabolic pathway. The work was a collaboration with Jason Collens. We expressed HmfABCDE enzymes in *P. putida* on a bacterial operon, driven by an IPTG-inducible promoter. We then employed the RBS Library Calculator to design an optimal RBS library for HmfD and created a library of 16 possible HmfD variants using Gibson DNA assembly [17]. Nine variants were selected at random and sequence verified. Each construct was grown overnight at 30C, 250 RPM, in LB media. The cultures were centrifuged, and re-suspended at a calculated OD600 of 0.5 in M9 media spiked with 0.5 g/L furoic acid. Furan concentrations and OD600’s were monitored on 3 hour intervals. The results demonstrated that the furan consumption varied over 30-fold (**Appendix D-Figure 3**), showing that small number of systematic experiments (nine experiments) is sufficient to get a significant change in the productivity of the metabolic pathway.
Appendix D-Figure 3: Furan consumption rate for different HmfD variants.
Appendix E-Performance of the RBS Library Calculator for optimizing expression rate of different example genetic systems

A) Five different proteins in *E. coli*

\[
\begin{align*}
\text{LacZ} & \quad \text{RFP} & \quad \text{sGFP} & \quad \text{CFP} & \quad \text{GFPmut3b} \\
\begin{array}{cc}
\text{Expression rate au} & \text{Variant} \\
10^0 & 10^1 & 10^2 & 10^3 & 10^4 & 10^5 & 10^6 & 10^7 & 10^8 & 10^9 & 10^{10} & 10^{11} & 10^{12} & 10^{13} & 10^{14} & 10^{15} & 10^{16} & 10^{17} & 10^{18} & 10^{19} & 10^{20} & 10^{21} \\
4 & 8 & 12 & 16 & 20 & 24 & 28 & 32 & 36 & 40 & 44 & 48 & 52 & 56 & 60 & 64 & 68 & 72 & 76 & 80 & 84 \\
\end{array}
\end{align*}
\]

B) RFP expression in five different gram-negative and gram-positive hosts

\[
\begin{align*}
\text{S. typhimurium} & \quad \text{C. glutamicum} & \quad \text{E. coli BL21} & \quad \text{P. fluorescens} & \quad \text{B. subtilis} \\
\begin{array}{cc}
\text{Expression rate au} & \text{Variant} \\
10^0 & 10^1 & 10^2 & 10^3 & 10^4 & 10^5 & 10^6 & 10^7 & 10^8 & 10^9 & 10^{10} & 10^{11} & 10^{12} & 10^{13} & 10^{14} & 10^{15} & 10^{16} & 10^{17} & 10^{18} & 10^{19} & 10^{20} & 10^{21} \\
4 & 8 & 12 & 16 & 20 & 24 & 28 & 32 & 36 & 40 & 44 & 48 & 52 & 56 & 60 & 64 & 68 & 72 & 76 & 80 & 84 \\
\end{array}
\end{align*}
\]

**Appendix E-Figure 1:** Characterization result for a subset of genetic variants designed by the RBS Library Calculator; au: arbitrary unit

**Appendix E-Figure 2:** Combinatorial optimization for the expression rate of genetic elements involved in three multi-protein genetic circuits. In each case, an RBS library was designed for each individual protein (i.e. PhiF in the Signal Inverter, AmtR and LmrA in the Signal Amplifier, and TetR and T7 RNA polymerase in the transcriptional Resource Generator). The combinatorial libraries were created by Gibson DNA assembly method. For the Signal Inverter and Signal Amplifier circuits, 11 and 10 different circuit variants were first sequence-confirmed then characterized by flowcytomtry. For the Resource Generator circuit, 96 colonies
were selected at random and characterized by flowcytometry. The experimental work for the Resource Generator circuit was solely performed by Manish Kushwaha; au: arbitrary unit; the Amplification is the fold change in the output in comparison with the fold change in the input signal that was created by a LacI-inducible promoter (Figure 2-2).

Appendix E-Figure 3: Measuring the translation rate capacity of a codon-optimized gene. The translation initiation rates of (A) gfpmut3b and (B) adhp::gfpmut3 fusion proteins were uniformly increased across a 10,000-fold scale using an optimized RBS library to identify the critical point where translation initiation is no longer the rate-limiting step in protein expression and folding. The expression of both proteins reached the same plateau (dashed line) at similar translation initiation rates. Data averages and standard deviations from 3 measurements.
Appendix E-Figure 4: Recombinant Vaccine Development. Vaccination is the most effective and widespread treatment for disease prevention. However, a vaccine’s protective immunity is often lost in time, leading to the need for re-vaccination (e.g. Tetanus shots). Modern vaccine development is hindered by trial-and-error optimization, where mixtures of antigens are varied in the search for a vaccine that maximizes the protective immunity duration. The antigen dosage level plays an important role in determining the duration of a vaccine’s protective immunity. With too little antigen, individual T-cells do not mount a sufficient response, resulting in poor T-cell expansion. With too much antigen, T-cell overexpansion results in poor differentiation to memory cells. Here, as a demonstration, we created a chimeric fluorescent protein containing an epitopes (GP33) of LCMV-GP model antigen fused to Cerulean (CFP) inside a S. Typhimurium vector.
Appendix F- Rational Design of a Synthetic Entner-Doudoroff Pathway for Improved and Controllable NADPH Regeneration

A modified version of this appendix was published in Metabolic Engineering journal[243]

Introduction

Most metabolic reactions that produce industrially important compounds depend on electron-carrying cofactors, such as NADH and NADPH. In particular, NADPH plays a vital role in the biosynthesis of drugs [244,245,246], chiral alcohols [247,248], fatty acids and biopolymers [249,250,251,252], while also being required for lipid biosynthesis, biomass formation, and cell replication [253,254]. As a result, the regeneration rate of NADPH is often the rate-limiting step for the over-production of desired chemicals, while maintaining robust cellular growth. Therefore, increasing NADPH regeneration rates can increase both pathway productivities and product yields [244,245,252,255,256,257,258,259]. Here, our objective is to develop a modular, drop-in pathway that rapidly regenerates NADPH, and provides control over redox supply levels, to increase the productivity of NADPH-dependent metabolic pathways.

In E. coli, the three major sources of NADPH regeneration are the pentose phosphate pathway (PPP), tricarboxylic acid (TCA) cycle, and the transhydrogenase system [260]. To increase NADPH regeneration rates, a common strategy has been to redirect carbon flux through PPP by the deletion of pgi or pfkA/pfkB [244,261,262]; and by the over-expression of glucose-6-phosphate dehydrogenase (zwf) or 6-phosphogluconate
dehydrogenase (gnd) [250,263,264]. Following these approaches, titers of leucocyanidin and thymidine, both limited by NADPH availability, were improved by up to 3.8-fold [244] and 4.85-fold [246], respectively. However, the resulting release of carbon dioxide within PPP [265] lowers product carbon yield and the growth defect caused by a pgp deletion limits productivity [265,266,267]. To overcome this challenge, it is possible to redirect carbon flux through the Entner-Doudoroff (ED) pathway, which regenerates NADPH without a concomitant carbon loss.

The Entner-Doudoroff (ED) pathway combines the enzymes glucose-6-phosphate dehydrogenase (zwf), 6-phosphogluconolactonase (pgl), 6-phosphogluconate dehydratase (edd), and 2-keto-3-deoxygluconate-6-phosphate (KDPG) aldolase (eda) to convert glucose 6-phosphate into two units of pyruvate, while generating equimolar amounts of ATP, NADH, and NADPH (Appendix F-Figure 1). In contrast, the well-known Embden-Meyerhof-Parnas (EMP) glycolysis pathway performs the same conversion, but produces two moles each of ATP and NADH. There are several additional, and important, differences between these otherwise substitutable glycolytic pathways. First, the lower amount of ATP synthesis causes the ED pathway to become highly exergonic, favoring catalysis in the forward direction [268]. As a result, the ED pathway has been shown to require 3.5-fold less enzyme to achieve the same EMP pathway flux, implying a similar reduction in the cost of assembling the catalytic machinery. Second, bacterial strains that rely on the ED pathway to perform glycolysis generally produce more NADPH than their anabolic demand [268,269,270]. To supplement ATP synthesis, like other facultative organisms, ED-dependent bacteria carry out aerobic respiration and catabolize additional non-glycolytic substrates [255,268]. Finally, when both the EMP
and ED pathways are available in the same organism, the ED pathway often fulfills an alternative role. For example, the conditionally expressed ED pathway in *E. coli* evolved to carry out gluconate metabolism [262,263,271,272,273,274,275].

Engineering the natural *E. coli* ED pathway may not enable tunable control over its NADPH regeneration rate, due to endogenous layers of transcriptional, translational, and allosteric regulation. Instead, a promising strategy is to heterologously express a highly active version of the pathway from a different organism [276,277]. We therefore selected the highly active ED pathway from *Zymomonas mobilis*. This organism relies solely on the ED pathway for glycolysis, has a high sugar uptake rate, has a high regeneration rate of ATP and NAD(P)H, and produces large amounts of ethanol that surpasses many yeast strains [278,279]. The high glycolytic flux of *Z. mobilis* has been attributed to the high turnover numbers, minimum allosteric control, and high expression levels of its ED enzymes [279,280,281]. Its glucose 6-phosphate dehydrogenase (*zwf*) enzyme is known to regenerate both NADH and NADPH, enabling autonomous redox balancing [270]. To the best of our knowledge, a complete *Z. mobilis* Entner-Doudoroff pathway has not yet been expressed in *E. coli*.

In this study, we designed, constructed, and systematically optimized a synthetic Entner-Doudoroff pathway as a drop-in module that significantly increases a bacterial host's NADPH regeneration rate. Using computational optimization and biophysical models, we rationally designed two operon sequences to heterologously express the four-enzyme ED pathway as well as phosphoglucone isomerase (*pgi*) to obtain maximum control over their expression levels (Appendix F-Figure 2). We constructed and assembled the resulting 8.9-kbp genetic system, and integrated it into the *E. coli*
MG1655-derived genome. We then efficiently explored the 5-dimensional expression space by employing the RBS Library Calculator to design optimized genome mutations [282] together with multiplex automated genome engineering (MAGE) mutagenesis to implement the genome mutations [26], generating libraries of $10^6$ ED pathway-genome variants. Using a NADPH-dependent fluorescent protein, we screened 624 ED pathway-genome variants for high NADPH regeneration rates, and then extensively characterized 22 re-integrated pathways by measuring in vivo NADPH regeneration rates and NADPH-dependent biosynthesis rates. As a result, an optimized ED pathway increased NADPH-dependent fluorescence by 25-fold and increased the production titer of an already optimized carotenoid biosynthesis pathway by 97%.
Appendix F-Figure 1: EMP versus ED Pathways. The major glycolytic pathways in *E. coli*, showing the Embden-Meyerhof-Parnas (EMP) and Entner-Doudoroff (ED) pathways. Native *E. coli* genes are shown in light blue color. The optimized heterologous genes in our synthetic bacterial operons are shown highlighted in yellow boxes.
Appendix F-Figure 2: Design and construction of ED1.0 and combinatorial ED variants. The synthetic ED-tetAR operons were designed using the Operon Calculator, with the \(Zm\)-zwf, \(Zm\)-pgi, \(Zm\)-edd, \(Zm\)-eda and \(Zm\)-pgl genes from Zymomonas mobilis ZM4 as input. \(Zm\)-zwf and \(Zm\)-pgi were grouped into the first operon and \(Zm\)-edd, \(Zm\)-eda and \(Zm\)-pgl were grouped into the second operon. Both operons were under the control of promoter \(P_{\text{tac}}\). This genetic system was integrated into the chromosome of \(E. coli\) EcNR2 strain between \(tonB\) and \(yciL\) locus, resulting in strain ED1.0. 16-variant RBS library (dRBS\(_i\), \(i = 1, 2, \ldots, 5\)) for each gene \(i\) was designed using RBS Library Calculator and introduced into strain ED1.0 using MAGE mutagenesis resulting in a large combinatorial population.

Results

Rational Design and Construction of a Synthetic Entner-Doudoroff Pathway

We selected five enzymes from \(Z. mobilis\) ZM4 for heterologous expression in \(E. coli\): glucose-6-phosphate dehydrogenase (ZMO0367/\(Zm\)-zwf), 6-phosphogluconolactonase (ZMO1478/\(Zm\)-pgl), 6-phosphogluconate dehydratase (ZMO0368/\(Zm\)-edd), 2-keto-3-deoxygluconate-6-phosphate (KDPG) aldolase (ZMO0997/\(Zm\)-eda), and phosphoglucose isomerase (ZMO1212/\(Zm\)-pgi). The first four enzymes constitute the ED pathway that converts glucose 6-phosphate to pyruvate and glyceraldehyde-3-phosphate, while the fifth reversibly interconverts fructose-6-phosphate to glucose-6-phosphate (Appendix F-
We included the enzyme \textit{Zm-pgi} to regulate metabolic flux at the major glycolysis branch point, glucose-6-phosphate. Throughout the paper, the operons consisting of the five \textit{Z. mobilis} enzymes are designated as the synthetic ED operons.

Protein expression levels are regulated by several genetic elements, including promoters, ribosome binding sites (RBSs), and protein coding sequences. In natural genetic systems, changes in transcription, translation, and mRNA stability collectively control a protein's expression level. As a key strategy to optimizing the ED pathway, we developed an optimization procedure, called the Operon Calculator, that designs bacterial operon sequences with the overall objective of concentrating expression control to the fewest number of short genetic parts, while eliminating undesired genetic elements that confound our ability to control protein expression. The Operon Calculator minimizes the number of undesired internal start codons, ribosome pause sequences, repetitive sequences, and restriction sites, while selecting 5' UTR sequences, synonymous codon sequences, and terminators for high translation rate capacities and termination efficiencies (Section 2.1). As a result, two synthetic bacterial operon sequences were designed grouping together the enzymes \textit{Zm-zwf} and \textit{Zm-pgi} into the first operon, and \textit{Zm-pgl}, \textit{Zm-eda} and \textit{Zm-edd} into the second operon (Appendix F-Figure 2, Supplementary Figure S1). We selected an IPTG-inducible \( P_{\text{tac}} \) promoter to transcribe both operons. We also designed the initial ribosome binding site sequences for all five enzyme coding sequences to have translation initiation rates of about 1,000 au on the RBS Calculator v1.1 proportional scale, which is a moderate translation rate. Importantly, these RBS sequences were designed by the RBS Library Calculator [282] so that a small
number of adjacent mutations could greatly vary the coding sequences’ translation rates. In addition, we chose tetAR as our selection marker for genome integration.

The synthetic ED operons were first constructed and inserted into a pre-constructed vector (pCN-L) along with tetAR operon resulting in a 11.8 kb plasmid (pCN-LEDT) by combining DNA synthesis, DNA assembly, and molecular cloning. We then employed PCR amplification and homologous recombination to integrate the ED-tetAR operons (8.9 kb) into the EcNR2 chromosome (Section 2.2). We refer to this strain, harboring the first version of our synthetic Entner-Doudoroff pathway, as ED1.0 (Appendix F-Figure 2).

Characterization of the synthetic Entner-Doudoroff pathway in ED1.0

We first characterized the activity of the synthetic Entner-Doudoroff pathway in ED1.0, compared to its parent strain EcNR2, by measuring the NADPH/NADP+ intracellular redox ratio and by measuring the in vivo NADPH regeneration rate. Using a glucose 6-phosphate dehydrogenase assay on cell extract (Section 2.5), we found that the ED1.0 strain has a NADPH/NADP+ redox ratio that is 1.87-fold higher than its parent EcNR2 strain after both are cultured to the exponential growth phase using M9 minimal media (two-tailed, two-sample t-test, p-value = 0.037) (Appendix F-Figure 3A). We then selected mBFP, a NADPH-dependent fluorescent protein reporter, as a large consumption sink for NADPH that also serves as an observable readout for its in vivo regeneration rate. mBFP is a short-chain dehydrogenase that actively oxidizes NADPH and proportionally emits fluorescence [283].

196
As an initial control to validate the mBFP assay, we over-expressed the transhydrogenase PntAB on an R6K origin vector within *E. coli* Pir116 together with an IPTG-inducible mBFP expression plasmid to measure its effect on mBFP fluorescence (Supplementary Materials and Methods). For a comparison, we also created an R6K-origin plasmid that did not express any enzymes, and co-transformed this control plasmid with the mBFP expression plasmid in the same strain. We continuously measured mBFP fluorescence after IPTG induction during long-time cultures maintained in the exponential phase of growth using M9 minimal media. After a lag period, mBFP fluorescence increased linearly in time over a long duration, indicating that NADPH availability was a rate-limiting step to fluorescence emission (Supplementary Figure S2A). We quantified NADPH regeneration rate by calculating the first-derivative (slope) of the time-course mBFP fluorescence per OD\textsubscript{600} over the linear regime, which we refer to as the mBFP fluorescence production rate (mBFP flu rate). Overexpression of PntAB together with induced mBFP expression, using 0.5 mM IPTG, increased the mBFP fluorescence production rate by 81%, compared to the control (93.6 ± 20.7 au/OD\textsubscript{600}.h vs 51.8 ± 28.9 au/OD\textsubscript{600}.h, two-tailed, two-sample t-test, p-value = 0.057) (Supplementary Figure S2B). Therefore, mBFP was found to be a quantitative reporter of the *in vivo* NADPH regeneration rate.

We then modified the mBFP expression plasmid to ensure that NADPH availability was the rate-limiting factor that controls mBFP fluorescence across a larger dynamic range. We first replaced the IPTG-inducible promoter controlling mBFP expression with a constitutive promoter. We then replaced its ribosome binding site sequence with a rationally designed one to substantially increase its translation rate. The
resulting plasmid (pCN-mBFP) constitutively expresses mBFP with a very high expression level.

We then employed the pCN-mBFP plasmid to measure the \textit{in vivo} NADPH regeneration rate of the ED1.0 strain, compared to its parent EcNR2 strain. Transformed strains were grown in long-time cultures maintained in the exponential growth phase using M9 minimal media. mBFP fluorescence was monitored continuously. We found that the mBFP fluorescence production rate for the parent strain EcNR2 was relatively constant regardless of the addition of IPTG. In contrast, the mBFP fluorescence production rate for the strain ED1.0 was substantially higher and was further increased when adding IPTG to induce the ED pathway's expression. The highest mBFP fluorescence production rate was observed at an intermediate IPTG concentration (25 µM) and there was no statistically significant increase in mBFP production rate when additional IPTG was added (p-values > 0.1 for all pair-wise two-sample t-tests between 25 µM and 0.5 mM IPTG) (\textbf{Appendix F-Figure 3B}). In the absence of IPTG, the observed level of mBFP production could be explained by transcriptional leakiness of the P\textsubscript{lac} promoter. Based on these results, the synthetic Entner-Doudoroff in strain ED1.0 is improving the NADPH regeneration rate by 4.8-fold when induced by 25 µM IPTG, compared to its parent strain. We expected that optimization of the ED pathway would be necessary to further increase its activity.
Appendix F-Figure 3: Characterization of ED1.0. (A) The NADPH/NADP⁺ redox ratios were measured for strains EcNR2 and ED1.0. Expression of the ED pathway was induced using 0.5 mM IPTG. ED1.0 had a significantly higher NADPH/NADP⁺ ratio than its parent strain (two-tailed, two-sample t-test, p-value = 0.037). Values and error bars represent the averages and s.d. of the ratios for four replicates. NADPH and NADP⁺ concentrations (μmol/ g DCW) for EcNR2 and ED1.0 are also reported (mean ± s.d. for n = 4). (B) The mBFP fluorescence production rates were measured for strains (blue bar) EcNR2 and (red bar) ED1.0. Increasing amounts of IPTG were added in separate experiments. The expression of the ED pathway in ED1.0 is controlled by an IPTG-inducible P_tac promoter. The pound sign (#) indicates a near-background mBFP production rate for EcNR2 strain at 0 mM IPTG. Values and error bars represent the means and s.d. of two replicates. *P < 0.1; **P < 0.05.

Efficient Search for Improved ED Pathway Variants in a 5-dimensional Expression Space

We initially designed the synthetic operons in ED1.0 to express all the ED pathway's enzymes with translation initiation rates of about 1,000 au on RBS Calculator proportional scale. The initial values were chosen to match the typical translation initiation rates controlling the expression of enzymes found in the native glycolysis, PPP, and ED pathways of *E. coli* MG1655. We anticipated that our initial translation rate guesses may not lead to the highest possible NADPH regeneration rate, and during the initial design of the synthetic operons, employed our RBS Library Calculator algorithm to design ribosome binding site sequences that could be easily mutagenized to provide large changes in translation initiation rate.
Specifically, the RBS Library Calculator in Genome Editing mode was applied to design 16-variant RBS libraries that systematically varied the translation rates for each enzyme coding sequence from about 10 to 900,000 au on the RBS Calculator proportional scale [282] (Supplementary Table S3). The resulting optimized RBS libraries contained a small number of nearby degenerate nucleotides that could be readily incorporated into the *E. coli* genome using a site-directed genome mutagenesis technique. A key advantage of this approach is the compactness of the resulting combinatorial RBS library and the broad coverage of the 5-dimensional expression space. Complete combinatorial incorporation of the five 16-variant RBS libraries will create a library of $16^5$ (1,048,576) variants that will uniformly sample a 5-sided hypercube with lengths that span at least a 10,000-fold change in expression (Appendix F-Figure 4).

Appendix F-Figure 4: Uniform Sampling of the 5-dimensional Expression Space. (A) For each of the enzyme, degenerate RBS (dRBS) for 16-variant RBS library was designed with RBS Library Calculator. By having just four degenerate nucleotides on the RBS sequence of *Zm-*
zwf, one can uniformly vary the predicted translation rate across a 4500-fold range. (B) (Black dots) The predicted translation initiation rates for all possible combinations of the five optimized RBS libraries, showing two-dimensional slices of the 5-dimensional space. (Red circles) The predicted translation rates for the 22 selected ED-expressing genome variants and (blue squares) the initial ED1.0 strain showing the sampling of the space after 40 cycles of MAGE genome engineering. Translation rates are predicted using RBS Calculator v1.1.

Appendix F-Figure 5: Characterization of ED-expressing genome variants. (A) The normalized mBFP fluorescence production rates of 387 genome variants and 22 control strains (including 7 ED 1.0 mBFP, 10 WT mBFP and 5 no BFP control strains) were measured and ranked. Normalized mBFP fluorescence production rates were calculated by dividing all measurements by the average mBFP fluorescence production rate for the EcNR2 strain, which was 28.2 ± 9.4 au/OD$_{600}$/h. Boxplots represent the ranking distribution of ED1.0 mBFP, WT mBFP and the no mBFP control strains. (B) The growth rates of each variant were measured and appear in the same order as in A. Bar colors are (blue) ED-expressing genome variants, (green) parent strain EcNR2, (red) strain ED1.0, and (grey) negative controls EcNR2 and ED1.0 without mBFP overexpression.

We then applied oligo-mediated allelic replacement, also known as MAGE [26], to incorporate these ribosome binding site mutations directly into the E. coli genome.
According to an allelic replacement efficiency calculation [225], we estimated that 40 MAGE cycles were required to generate 14% of genomes with at least 4 out of 5 RBS mutations and 2.3% of genomes with all 5 RBSs mutated. Over a span of 12 days, 40 MAGE cycles were conducted using mutagenic oligonucleotides that correspond to the optimized RBS library sequences. Bulk sequencing of genome pools after the 12th, 30th, and 40th cycles revealed that RBS sequences became increasingly mutated at specifically targeted nucleotide positions. We then characterized the pool of genome variants from the 40th cycle of MAGE.

We transformed the pool of ED-expressing genome variants with pCN-mBFP, our constitutively over-expressed mBFP plasmid, and then isolated 624 single colonies. We cultured them using M9 minimal media supplemented with 0.4% w/v glucose and 1 mM IPTG, and measured their mBFP fluorescence production rates. As controls, we also characterized the mBFP fluorescence production rates of strain ED1.0 and parent strain EcNR2 in each set of measurements. When varying their RBS sequences and enzyme expression levels, we expected that ED-expressing genome variants would have different NADPH regeneration rates and growth rates, due to changes in pathway flux and the accumulation of toxic intermediates (e.g. KDPG). 237 ED-expressing genome variants (38%) exhibited poor growth after induction of the ED pathway, indicating that some combinations of RBS sequences and enzyme expression levels resulted in imbalanced pathways. The remaining 387 ED-expressing genome variants displayed varied mBFP fluorescence production rates across a 710-fold range (between 0.045-fold to 32.12-fold), indicating that these optimized RBS sequence mutations greatly affected the ED pathway's enzyme expression levels, and correspondingly, its overall NADPH
regeneration rate (Appendix F-Figure 5). 336 of these variants had lower mBFP-linked NADPH regeneration rates than the average of ED1.0, suggesting that our initially selected translation rates of 1,000 au were a suitable initial condition. Interestingly, there was no observed correlation between the growth rate of a genome variant and its mBFP-linked NADPH regeneration rate ($R^2 = 0.04$).

Next, we selected 22 ED-expressing genome variants, including 16 high mBFP-producing variants and 6 low mBFP-producing variants, for sequencing, re-integration into fresh *E. coli* EcNR2 genomes, and further characterization. We found that all 22 variants contained unique combinations of RBS sequences (Appendix F-Figure 6). 7 variants had four out of five modified RBS sequences. 4 variants had three modified RBSs, 6 variants had two modified RBSs, and 5 variants had a single modified RBS. None of the selected genome variants had all five of their RBS sequences modified. The RBS controlling *Zm-pgl* was modified in 73% of the selected genome variants, while only 36% of genome variants had RBS modifications controlling *Zm-zwf* expression. Otherwise, 41%, 50%, and 62% of variants had RBS modifications controlling *Zm-eda*, *Zm-edd*, and *Zm-pgi* expression. In one variant (ED9), there was a spontaneous, non-designed mutation to the RBS controlling *Zm-pgi* expression. The translation initiation rates for all RBS modifications were calculated by the RBS Calculator's biophysical model (Appendix F-Figure 6). Notably, the ED3 genome variant contained the most highly translated RBS for *Zm-zwf*, but also a frame-shift mutation in the protein coding sequence that abrogated its expression, indicating that *Zm-zwf* over-expression was highly toxic.
We re-characterized the mBFP fluorescence production rates of the 22 selected ED-expressing genome variants and ranked them (Appendix F-Figure 6, Supplementary Table S4). Strain ED2 consistently ranked first among the selected variants with 25-fold higher mBFP fluorescence production rate, compared to the parent strain EcNR2. Notably, its translation rate profile is not significantly different from ED1.0, which ranked within the top three. The differences between ED2 and ED1.0 were a 3.7-fold higher translation rate for Zm-eda and a 10-fold lower translation rate for Zm-pgl, indicating that small changes in translation rate, and correspondingly enzyme expression level, can have a beneficial effect on NADPH regeneration rate. However, there remains ED-expressing genome variants in the larger pool with even higher mBFP fluorescence production rates and NADPH regeneration rates.

We examined the relationship between translation rates, enzyme expression levels, and NADPH regeneration rate for these 22 ED-expressing genome variants. Qualitatively, we note that higher enzyme expression levels did not always yield higher NADPH regeneration rates. However, we do not observe a quantitative pattern in contrast to our previous study, where we employed mass action kinetics to formulate such a quantitative relationship. Several factors could confound this analysis, including the activity of endogenous enzymes, competition for transcriptional or ribosomal resources, allosteric feedback control, insufficient sampling of the high-dimensional expression space, and variations in day-to-day measurements.
Appendix F-Figure 6: The Effects of Changing ED Enzyme Expression Levels on NADPH Regeneration Rates. (A) The normalized mBFP fluorescence production rates were measured and ranked for 22 selected ED-expressing genome variants that were re-integrated into the genome of parent strain EcNR2. Values and error bars represent the means and s.d. of 2-12 replicates. (B) The translation initiation rates controlling the expression of the five ED enzymes are shown, comparing the (red) 22 modified ED variants to (blue) ED1.0. Translation initiation rates were predicted using the RBS Calculator v1.1.
Improving Terpenoid Biosynthesis using a Synthetic Entner-Doudoroff Pathway

In bacteria, the methyl erythritol phosphate (MEP) pathway synthesizes the terpenoid precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), consuming equimolar amounts of glyceraldehyde-3-phosphate (G3P) and pyruvate [284]. In *E. coli*, the enzymes within the MEP pathway primarily use NADPH as the source of reducing equivalents [285] (Supplementary Table S5), and the reaction catalyzed by the first enzyme in the MEP pathway, *dxs*, is a key rate-limiting step to precursor biosynthesis [26,286,287,288]. Interestingly, the Entner-Doudoroff pathway synthesizes both the carbon and redox precursors to terpenoid biosynthesis [272,289] (Appendix F-Figure 7A). As a consequence, we expected that a highly active ED pathway would increase the rate of terpenoid biosynthesis, particularly when the activities of the MEP and downstream terpenoid biosynthesis pathways were optimized such that NADPH availability became a greater rate-limiting factor. To test this hypothesis, we took advantage of our previous work that systematically optimized a carotenoid biosynthesis pathway from *Rhodobacter sphaeroides*, expressed within an engineered strain that had a substantially higher MEP pathway flux [282].
Appendix F-Figure 7: Improving Terpenoid Biosynthesis using the Synthetic Entner-Doudoroff Pathway. (A) The metabolic reactions that convert glucose 6-phosphate into the carotenoid neurosporene via the Entner-Doudoroff, MEP, and carotenoid biosynthesis pathways. The enzymes \textit{crtEBI} from \textit{Rhodobacter sphaeroides} produce the brown pigment neurosporene. (B) Neurosporene production was measured when combining selected ED-expressing genome variants and control strains with expression of the \textit{crtEBI} operon. ED pathways were integrated into either EcNR2 or EcIF15. The strain EcIF15 has an engineered RBS substantially increasing \textit{dxs} expression. Values and error bars represent the means and s.d. of between 2 to 7 replicates.

First, we transformed selected ED-expressing genome variants with a \textit{crtEBI}-expressing plasmid, which produces the carotenoid neurosporene under control of an IPTG-inducible \textit{P}_{\text{lacO1}} promoter. Cultures were grown in 2X M9 minimal media with 0.4% w/v glucose and 0.2 mM IPTG in short (10 hour) batches, followed by hot acetone extraction, and measurement of their neurosporene content and dry cell mass. The strains ED13 and ED1.0 produced 79% and 43% higher neurosporene production titers, respectively, compared to the parent EcNR2 strain (Appendix F-Figure 7B). To compare the use of the ED pathway to alternative approaches to increasing NADPH regeneration, we knocked out expression of \textit{pgi} in the parent strain EcNR2 and found that it increased neurosporene production by 24% using the same growth conditions and
measurements. In contrast, the strains ED2, ED5, and ED11 produced about the same amount of neurosporene, compared to their parent EcNR2 strain (4% to 15% lower).

After these measurements, we questioned whether the relatively low neurosporene production rate was creating a sufficiently large burden on NADPH availability. If NADPH regeneration is to become a rate-limiting step in neurosporene production, it must be produced at a high rate with a high demand for NADPH. Next, we selected a previously engineered strain EcIF15 and transformed it with plasmid that expresses the neurosporene biosynthesis pathway now under control of an arabinose-inducible P_BAD promoter to enable orthogonal transcriptional control over both the ED and crtEBI pathways. EcIF15 was previously engineered to significantly over-express 1-deoxy-D-xylulose-5-phosphate synthase (dxs) and increase the biosynthesis of the precursors IPP and DMAPP [282]. As a result, with 10 mM arabinose induction, it produces 791% and 843% more neurosporene, compared to control strains E. coli MG1655 and EcNR2, respectively (Appendix F-Figure 7B). We also characterized the P_BAD-crtEBI pathway in EcIF15 strain expressing ED17’s ED-tetAR to confirm that the large titer change was not a result of swapping promoters (Supplementary Table S6).

We next introduced ED pathway variants into the EcIF15 strain and evaluated the synergistic effects of an increased NADPH regeneration rate together with an increased precursor biosynthesis rate. In contrast to the EcNR2 strain, the expression of any of the selected ED pathway variants (ED1.0, ED2, ED11, ED13, and ED17) in EcIF15 substantially increased neurosporene production. The variant ED11 yielded the highest improvement with a 1336% increase, compared to EcNR2, and a 97% increase, compared to the parent strain EcIF15 (Appendix F-Figure 7B). Interestingly, ED11 has
three higher translation rates controlling \textit{Zm-zwf}, \textit{Zm-edd} and \textit{Zm-eda} expression, compared to ED1.0 (\textbf{Appendix F-Figure 7B}). Based on these results, when the bottleneck in the MEP pathway was eliminated, expression of the synthetic ED pathway was able to produce more precursors and regenerate more NADPH, leading to large improvements in neurosporene production.

\textbf{4. Discussion}

In the field of metabolic engineering, increasing the availability of NADPH has been a significant challenge, driven by the need to supply greater amounts of reducing equivalents towards the over-production of a wide range of chemical products. To solve this challenge, previous efforts have deleted or over-expressed selected genes, such as oxidoreductases, transhydrogenases, and NAD kinases [247,250,251,252,263,264,290,291,292]. In a recent effort, the \textit{E. coli} NAD$^+$-dependent glyceraldehyde-3-phosphate dehydrogenase (GAPDH) encoded by \textit{gapA} gene was replaced with a \textit{Clostridium acetobutylicum} \textit{gapC} gene encoding for a NADP$^+$-dependent GAPDH [257]. Another recent study also showed that replacing the promoters of \textit{E. coli} \textit{edd-eda} operon and \textit{zwf} gene with a constitutive promoter and a strong promoter, respectively, increased intracellular NADPH/NADP$^+$- ratio [293].

In this study, we engineered a synthetic version of the Entner-Doudoroff (ED) pathway to rapidly regenerate NADPH. The pathway combines five enzymes from \textit{Zymomonas mobilis}, expressed together within two synthetic bacterial operons that were rationally designed to achieve maximum expression control. Starting from the first
version of the pathway, we then carried out systematic optimization of the enzymes' expression levels to improve the pathway's activity, first employing a NADPH-dependent fluorescent protein reporter to measure NADPH regeneration rates, followed by measuring the ED pathway's effect on an NADPH-dependent terpenoid biosynthesis pathway. By combining MAGE genome mutagenesis with our RBS Library Calculator algorithm, we introduced targeted genome modifications to greatly vary the ED pathway's individual enzyme expression levels and to efficiently search its 5-dimensional expression space. In principle, continued MAGE cycling will generate up to a million genome variants, though it was only necessary to characterize a much smaller number to identify ED pathway variants with greatly improved NADPH regeneration rates. As a result, one of our ED pathway variants exhibited a 25-fold higher NADPH regeneration rate, as measured by the fluorescent protein reporter, and another variant increased terpenoid biosynthesis by 97%. The synthetic ED pathway exists as a drop-in module; in principle, it can be transferred and expressed in many different bacterial hosts to increase their NADPH regeneration rate and enhance the production of NADPH-dependent products.

A novel and important aspect of our design approach was to commit, at an early stage, to integrating the ED pathway into the genome of our bacterial host before optimizing its expression levels. In comparison, pathway engineering efforts have traditionally relied upon multi-copy plasmids to over-express desired enzymes. Multi-copy plasmids can express more protein than expression cassettes inside genomes, but they require active selection (e.g. the addition of antibiotics) to maintain plasmid stability over long culture times, which is undesirable in industrial applications. When pathways
are plasmid-encoded, any optimization of their expression levels may be problematic when the final version of the pathway must eventually be genome-integrated for industrial applications. Instead, the first version of our genome-integrated ED pathway (ED1.0) was remarkably stable over the course of 40 cycles of MAGE mutagenesis and during 2-day cultures. Taking advantage of genome engineering techniques, this design choice enabled us to rapidly insert, optimize, copy, and re-insert large genetic modules within and across genomes, while improving the modules' stabilities and maintaining the same copy number.

With our genome-centric strategy, we needed to ensure that our operons' transcription rates, translation rates, and mRNA stabilities could be sufficiently high to express high levels of enzyme even though the DNA copy number is very low (1 to 2 copies per cell). Our Operon Calculator applies several design criteria to ensure that the operons' mRNAs have fewer undesirable genetic elements and that the desired protein coding sequences have the potential to be translated at extremely high rates. Among our 22 sequenced ED pathway variants, the translation rates for Zm-pgl, Zm-edd, and Zm-zwf achieved extremely high translation initiation rates (500,000+ au on the RBS Calculator v1.1 proportional scale). If the ED enzymes had exhibited low turnover numbers, then such high translation rates would have been necessary to achieve high pathway activities.

In a previous study, applying a kinetic modeling formalism enabled us to determine a quantitative relationship between sequence, enzyme expression level, and pathway activity [282]. However, while our optimization strategy yielded highly productive pathway variants, we did not observe such a clear relationship for the synthetic ED pathway in the EcNR2 parent strain. Previous reports have shown that
accumulation of KDPG is toxic [294,295] and will therefore impact the organism's growth rate [296], which could confound any expected relationship. Notably, amongst the ED variants that both over-produced neurosporene and emitted large amounts of mBFP fluorescence, ED11 has a consistent increase in Zm-eda, Zm-edd, and Zm-zwf expression that should result in an overall increase in pathway flux, while ED2 has increased Zm-eda and decreased Zm-pgl expression, which should minimize the accumulation of KDPG. It appears that increased neurosporene production in EcIF15-derived strains also relies on a consistent flux through the entire ED pathway, including production of the MEP precursors pyruvate and G3P, and not only on the NADPH regeneration rate.

Finally, together with previous work [5,282,286,297,298,299,300,301,302,303,304,305,306,307], this study highlights the common challenges of performing expression optimization on multi-enzyme pathways in high-dimensional expression spaces. For enzymes with low turnover numbers, large changes in enzyme expression are needed to observe a significant change in pathway activity. In contrast, when enzymes have higher turnover numbers, as in the case of the ED pathway, smaller expression level changes can have a significant impact. Multiple enzymes work together to control a pathway's overall activity. As we observed in this study, synergistic changes in the individual enzyme expression levels are needed to create a more active and balanced pathway. Expression imbalances can affect growth rate, due to the accumulation of toxic intermediates.

Further, to avoid exhaustively searching high-dimensional expression spaces, new modeling formalisms will be needed to convert large sets of sequences and measurements into accurately predicted optimal expression levels. These challenges will continue as the
field designs and optimizes longer pathways. New approaches will also be needed to combine pathway modules with predictably matched pathway fluxes, tailored to the organism's existing metabolic network. The success of these approaches together with advanced genome engineering techniques will accelerate the development of extremely large genetic systems, encoded within genomes, capable of radically redirecting carbon, energy, redox, uptake, and transport fluxes towards the production and secretion of desired products.

Materials and Methods

Chemicals were obtained from Sigma-Aldrich Co. (St. Louis, MO) and VWR International (Radnor, PA). Enzymes were purchased from New England Biolabs Inc. (Ipswich, MA). *E. coli* TOP10 strain (Invitrogen), Pir116 strain (TransforMax™ EC100D™ pir-116) and *E. coli* K12 ER2267 (LacIq) strain (NEB) were used for plasmid construction and propagation. Plasmid pQE-mBFP [283] was obtained from Dr. Geun-Joong Kim’s lab (Chonnam National University, South Korea). Tetracycline-resistance gene cassette tetAR was obtained from Dr John Roth’s Lab (UC Davis, CA). DNA fragment and oligonucleotide synthesis were performed by Integrated DNA Technologies Inc. (Coralville, IA) and GeneArt (Regensburg, Germany). Gene sequencing was performed by QuintaraBio (Boston, MA) and the Penn State Genomics Core. Unless stated otherwise, M9 minimal media is 1x M9 salt (Sigma-Aldrich Co.), 2 mM magnesium sulfate, 100 µM calcium chloride, supplemented with 0.4% w/v glucose, 0.34
g/L thiamine and 0.05 g/L leucine, adjusted to a pH of 7.4. For EcNR2 or EcIF15 derived strains, biotin was added to a final concentration of 250 µg/L.

**Design of synthetic operons and codon optimization**

Starting from amino acid sequences, gene and operon designs were carried out using a recently developed design procedure for synthetic operons (Operon Calculator v0.50) (Unpublished data). Synonymous codon selections for each coding sequence were initially determined by using weighted random choice from a custom codon usage table, followed by multi-objective optimization according to several design criteria. The custom codon usage table heavily weighted synonymous codons from the 164 most highly translated *E. coli* MG1655 protein coding sequences, as predicted by the RBS Calculator v2.0 [308], while eliminating rare codons. Additional synonymous codon mutations were made according to the following criteria: (i) 5' untranslated regions and the beginnings of each protein coding sequence were co-optimized to achieve high translation rate capacities *i.e.* at least a 100,000 translation initiation rate on the RBS Calculator v1.1 proportional scale, (ii) the translation initiation rates of internal start codons were minimized, (iii) Shine-Dalgarno-like ribosomal pause sequences [309] were removed, (iv) repetitive sequences, inverted repeat, and selected restriction enzyme recognition sequences were removed, and (v) 5' and 3' untranslated regions had minimum necessary lengths. Optimized ribosome binding sites and protein coding sequences for the five enzymes were assembled into two bacterial operons. Transcription of both operons is
initiated by an IPTG-inducible P_tac promoter, and terminated by BBa_B0021 and BBa_K780000 transcriptional terminators from the Registry of Standard Biological Parts.

**Strain and plasmid construction**

The parent strain for initial genome integrations and ED optimization is *E. coli* EcNR2, derived from *E. coli* MG1655 with modifications *bioA/bioB::λ-Red-bla ΔmutS::cat* [26]. The parent strains for re-integration and characterization of selected ED pathways are EcNR2 and EcIF15, which is a derivative of EcHW2f [26] with a mutated *dxs* ribosome binding site (*Supplementary Table S1*) [282]. The EcNR2 Δ*pgi* strain was constructed by inserting two consecutive stop codons into the *pgi* coding sequence with co-selection MAGE (*Supplementary Materials and Methods S1.1*). The synthetic operons described in section 2.1 were divided into three segments (ZmED1, ZmED2 and ZmED3) for gene synthesis. The first two segments were further divided into six gBlocks Gene Fragments each with lengths of about 500 bp. Both ZmED1 (2417 bp) and ZmED2 (2318 bp) were then assembled using Gibson Assembly [310]. The third segment (ZmED3) was synthesized by GeneArt (Life Technologies). Plasmid pCN-LED was constructed by assembling ZmED1, ZmED2 and ZmED3 into a separately constructed vector backbone (pCN-L) that contained an R6K origin, a chloramphenicol (Cm) resistance marker, and the *lacI* gene. Once assembled, pCN-LED contained the two ED bacterial operons flanked by 40 bp sequences and restriction sites to enable recombination into the *tonB/yciL* intergenic region within the *E. coli* MG1655 genome.
Additionally, the \textit{tetAR} operon was then amplified from \textit{Salmonella typhimurium} LT2 TT25401 strain and inserted into plasmid pCN-LED. After sequence verifying the resulting 11.8 kb plasmid pCN-LEDT (pCN-065), the recombination cassette (ED-tetAR) was amplified by PCR and integrated into the EcNR2 genome at the \textit{tonB/yciL} intergenic region using lambda red recombination approach (see section 2.4). In order to increase \textit{tetA} expression required for counter-selection with nickel salts or fusaric acid during co-selection MAGE \cite{311}, additional MAGE genome mutagenesis was employed to increase the translation initiation rate of the integrated \textit{tetA} marker from 228 to 48,372 au, employing the RBS Library Calculator to design RBS genomic mutations (\textit{Supplementary Materials and Methods S1.2.}). The resulting strain is henceforth denoted as ED1.0.

\textbf{Combining the RBS library Calculator and MAGE for pathway optimization}

For each of the five enzyme coding sequences, we used the RBS Library Calculator in Search mode to design 16-variant degenerate RBS libraries with translation initiation rates that spanned between 4,500 to 61,000-fold range, while constricting degenerate nucleotides to within a 9 bp region \cite{282} (\textit{Supplementary Table S3}). RBS libraries were encoded within 90 bp degenerate oligonucleotides with four phosphorothioated bases at the 5’ terminus (\textit{Supplementary Table S2}). RBS libraries were incorporated into the ED1.0 chromosome by performing 40 manual rounds of MAGE using a pool size of 80 (16 x 5) oligonucleotides adjusted to a final concentration...
of 1 to 2 µM. Daily MAGE rounds were carried out by first growing ED1.0 or previously mutagenized strains in 5 mL of SOC broth, supplemented with 0.4% w/v glucose, for 12 to 16 hours at 30°C and 250 RPM shaking, followed by dilution to an initial OD600 of 0.05 in 5 mL of SOC broth. The first MAGE round begins by incubating cells at 30 °C at 250 to 300 RPM until their OD600 reached 0.5 to 0.7, followed by heat shock at 42 °C for 15 min to induce λ-prophage (bet, gam, exo) genes expression, pelleting, and washing three times with cold sterile water to induce electrocompetency. 50 µL of the oligonucleotide mixture was then added to electrocompetent cells and electroporated at 1,800 V. To begin the second MAGE round, cells were recovered in pre-warmed SOC until their OD600 reached 0.4 to 0.6, followed by repetition of heat shock, recovery, and electroporation. Three to four MAGE rounds were performed daily. The resulting pool of ED-genome variants were then characterized using several assays.

Chromosomal integration of Entner-Doudoroff pathway variants

ED-tetAR linear DNA cassettes were PCR amplified from selected variants and re-integrated into the chromosomes of strains EcNR2 or EcIF15 via homologous recombination using the following approach. Overnight cultures of isolated colonies were grown in SOC broth supplemented with 0.4% w/v glucose for 12 to 16 hours at 30 °C with aeration. Overnight cultures were then diluted to an initial OD600 of 0.05 to 0.1 in 5 ml of SOC broth and were grown at 30 °C to OD600 of 0.4 to 0.6. Induction of the λ-Red recombination proteins were performed by shifting the cultures to 42 °C water bath with
shaking for 15 minutes. The cultures were subsequently chilled for 10 min, washed three
times with cold sterile distilled water and finally resuspended in 150 µL of distilled
water. For each transformation, 10 ng – 100 ng of linear DNA cassette was added to 50
µL of electrocompetent cells. After electroporation at 1800 V, electroporated cells were
mixed with 1 mL of SOC and then incubated at 30 °C, 250 RPM before plating on LB
agar plate containing tetracycline or appropriate antibiotic. Colonies with correct
insertion of linear DNA were verified by PCR analysis.

**Measurement of intracellular NADPH and NADP⁺ levels**

Intracellular levels of NADPH and NADP⁺ were determined according to a
method described previously [263]. Overnight culture inoculated with isolated colonies
were grown in LB with 50 µg/mL Cm at 30 °C with 250 RPM shaking. 50 mL of M9
minimal media with 0.4% w/v glucose were then inoculated with overnight culture to
OD₆₀₀ of 0.1 and induced with with 0.5 mM IPTG. After 24 h of growth at 30 °C with
250 RPM shaking (OD₆₀₀ of 1.0 to 3.0), cultures were chilled in ice bath for 10 min,
harvested by centrifugation at 4 °C, 10 min, 4750 RPM and resuspended to 0.5 mL of
OD₆₀₀ of 30.

For NADPH (reduced form) analysis, cells were resuspended in 250 µL of 0.3 M
NaOH, incubated at 60 °C for 7 min, then neutralized by 250 µL of of 0.3 M HCl and 50
µL of Tricine-NaOH (pH 8.0). For NADP⁺ (oxidized form) analysis, cells were
resuspended in 250 µL of 0.3 M HCl and 50 µL of Tricine-NaOH (pH 8.0), incubated at
60 °C for 7 min, then neutralized by 250 µL of 0.3 M NaOH. Neutralized samples were then centrifuged at 4 °C, 60 min, 4750 RPM and the resulting supernatant was transferred to new microcentrifuge tube. 80 µL of reduced sample or 40 µL of oxidized sample mixed with 40 µL of 0.1 M NaCl was added to a 96-well microtiter plate for analysis. 2x stock solution of reaction mixture was prepared by mixing equal volume of 1.0 M Tricine-NaOH (pH 8.0), 4.2 mM thiazolyl blue tetrazolium bromide (MTT), 40 mM EDTA, 1.67 mM phenazine ethosulfate (PES), and 25 mM glucose-6-phosphate. Both reduced and oxidized samples were mixed with 80 µL of freshly prepared reaction mixture and then incubated at 37 °C for 15 min. 20 µL of 2.5 U/mL NADP⁺-specific glucose-6-phosphate dehydrogenase (G6PDH) from Saccharomyces cerevisiae was added to initiate reaction and the time-course formation of reduced MTT was measured at 570 nm, 37 °C with Tecan microplate reader. The cofactor concentration was determined from rate of change in absorption when compared to calibration curve constructed by measuring NADP⁺ standards on the same microplate.

**Quantifying NADPH levels with a modified fluorescent reporter**

Plasmid pQE-mBFP expresses a NADPH-dependent metagenomic blue fluorescent protein (mBFP) on a ColE1 vector [283]. The mBFP protein is a short chain dehydrogenase (SDR) that binds specifically to NADPH and emits fluorescence at 451 nm when excited at 395 nm, producing more fluorescence when supplied with more NADPH. We modified pQE-mBFP by replacing its original IPTG-inducible T5 promoter
with a strong constitutive promoter and by increasing the translation initiation rate of mBFP to 356,786 au on the RBS Calculator proportional scale, resulting in pCN-mBFP. The pool of ED-genome strain variants, the strain ED1.0 and the EcNR2 strain were transformed with pCN-mBFP, followed by characterization of isogenic cultures using spectrophotometry (Infinite M1000, TECAN) to record mBFP fluorescence. Overnight cultures in LB with 50 µg/mL of kanamycin (Kan) were used to inoculate 200 µL M9 minimal media with 10 µg/mL Kan in microtiter wells. Cultures were then incubated at 30 °C with high orbital shaking for 6 to 9 hours. Cells were then serially diluted into fresh, pre-warmed M9 minimal media with 10 µg/mL Kan and 1 mM IPTG and grown similarly for another 12 hours. During the IPTG-induced growth period, cell densities (OD\textsubscript{600}) and mBFP fluorescence levels were recorded every 10 minutes. Specific mBFP fluorescence levels were determined by dividing each strain's background-corrected fluorescence levels by their corresponding OD\textsubscript{600} values. Specific mBFP production rates were determined by the slope of mBFP fluorescence levels versus time over a region of linearly increasing fluorescence levels. For normalization, all specific mBFP production rates were divided by the specific mBFP production rate of the EcNR2 control strain. Selected ED-genome variants were sequenced according to their normalized specific mBFP production rates.

**Measurement of NADPH-dependent carotenoid biosynthesis**

The enzymes encoded by \textit{crtEBI} from \textit{Rhodobacter sphaeroides} catalyze the conversion of isopentenyl diphosphate to neurosporene, a brown carotenoid pigment. We
previously constructed plasmids that express an optimally balanced \textit{crtEBI} operon to produce high levels of neurosporene [282]. The operon's expression is controlled by either an IPTG-inducible $P_{\text{lacO1}}$ promoter (pIF-001C or pIF-001K) or an arabinose-inducible $P_{\text{BAD}}$ promoter (pIF-002). Selected ED-genome variants were transformed with pIF-001 or pIF-002, and cultured overnight in 25 mL of LB broth with 50 $\mu$g/mL Cm for EcIF15-derived strains or 50 $\mu$g/mL Kan for EcNR2-derived strains. Cultures were then diluted to a final OD$_{600}$ of 0.1 in 25 mL 2X M9 minimal media supplemented with 0.4% w/v glucose, 10 $\mu$g/mL Cm and 0.2 mM IPTG. For strains transformed by pIF-002, 10 mM arabinose was added. Cultures were incubated at 37°C and 300 RPM shaking for 10 hours and pelleted by centrifugation. Cell pellets were washed, dried, and mixed with 1 mL acetone in pre-weighed microcentrifuge tubes, followed by incubation at 55°C and repeated vortexing for 20 minutes. The extraction solution was centrifuged again, and a 50 $\mu$L supernatant sample was transferred to 1 mL acetone (21-fold dilution) in a fresh microcentrifuge tube. The absorbance of neurosporene in acetone was recorded by a NanoDrop 2000C spectrophotometer at 470 nm, using pure acetone as a blank. Extracted cell pellets were dried open-capped in a 65°C oven for 2 days, equilibrated closed-capped at room temperature, followed by recording of cell pellet masses by comparison of microcentrifuge weights before and after extraction.
Supplementary Information

Content

1. Supplementary Materials and Methods
2. Supplementary Figure S1.
3. Supplementary Figure S2.
4. Supplementary Table S1. Strains and plasmids
5. Supplementary Table S2. Oligonucleotides used in this study
6. Supplementary Table S3. RBS sequences in each RBS library and the corresponding translation rate (au) predicted by RBS Library Calculator
7. Supplementary Table S4. Translation rate for ED variants
8. Supplementary Table S5. Net reaction for neurosporene biosynthesis from glyceraldehyde-3-phosphate and pyruvate
9. Supplementary Table S6. Comparison of neurosporene production from pIF-001C and pIF-002

Supplementary Materials and Methods

1.1. Construction of pgi mutant with co-selection MAGE

*Pgi* mutant was constructed by inserting two consecutive stop codons in the open reading frame of *E. coli* EcNR2 native *pgi* gene using co-selection MAGE (CoS-MAGE) approach. First, *bla* gene of EcNR2 strain was inactivated by performing MAGE with oligo *bla_off* that introduced premature stop codons in *bla* gene. *Pgi_KO* oligo was added with *bla_on* oligo in a 50:1 ratio during electroporation of EcNR2 *bla* strain. Three to four rounds of CoS-MAGE were performed, and the cells were plated on LB plate supplemented with ampicillin. Colonies were screened for genotype of interest by colony PCR with corresponding primers designed according to [225]. Colonies exhibiting the correct band on gel electrophoresis were re-confirmed by sequencing.

1.2. Improving tetA translation rate
The *tetA* gene confers resistance to tetracycline and also sensitivity to nickel salts and fusaric acid. We intended to use *tetA* as selection and counter-selection marker during co-selection MAGE procedure [312]. Higher copy number of *tetA* was more efficient than single/low copy of *tetA* in conferring nickel-sensitivity [311]. Since *tetA* promoter overlaps with *tetR* [313], we increased the expression of *tetA* by modifying its RBS on the EcNR2 strain with chromosomally integrated ED-tetAR cassette as followed. *bla* gene was inactivated in ED strain using *bla_off* oligo. 8-variant RBS library was designed for *tetA* gene using the Genome Editing Mode of RBS Library Calculator. Then, co-selection MAGE (CoS-MAGE) [25] was performed with oligonucleotide mixture of the RBS library of *tetA* gene and *bla_on* oligo that restores the function of the inactivated *bla*. After three rounds of CoS-MAGE, the cells were plated on LB plate supplemented with ampicillin and chloramphenicol. 96 resulting colonies were grow in LB media supplemented with 50 µg/mL of heat-inactivated chlortetracycline (*cTc*), 25 µg/mL of chloramphenicol, and nickel chloride. Expression of *tetA* in single copy at the original translation initiation rate of 228 au did not exhibited significant sensitivity to nickel and fusaric acid (unpublished data). High level of expression of *tetA* expression was found to cause sensitivity to nickel and fusaric acid [311,314,315]. Using growth rate screening in nickel chloride containing LB media, we isolated a colony that exhibited lowest growth rate across four different concentrations of nickel chloride (2.5 mM, 3.0 mM, 3.5 mM and 4.0 mM). Sequencing result verified that it has the highest translation initiation rate for *tetA* among the 8-variant RBS library. This ED strain with predicted *tetA* translation initiation rate of 48,372 au is designated as the ED 1.0 strain.

1.3. Modification of pQE-mBFP plasmid and measurement of mBFP production rate

The pQE-mBFP plasmid harboring metagenomic blue fluorescent protein driven by a IPTG-inducible T5 promoter [283] was modified as followed. *Cat* gene that confers chloramphenicol resistance was removed by inverse PCR of the original plasmid resulting in pQE-mBFP(*Cat*). *E. coli* ER2267 (NEB) was transformed with pQE-mBFP and pQE-mBFP(*Cat*’) resulting in strain CYN014 and CYN019, respectively. These strains were initially constructed for *in vivo* measurement of mBFP production (Supplementary Figure S2).
To propagate plasmid with R6K origin, *E. coli* Pir116 (Transformax) was then used as host strain. Pir116 competent cells were transformed separately with (i.) pQE-mBFP(Cat) and pCN-LA (R6K origin, expressing LacI constitutively and AmtR under pTac promoter) and (ii.) pQE-mBFP(Cat) and pCN-LPab (R6K origin, expressing LacI constitutively and PntAB transhydrogenase under pTac promoter), resulting in (i.) CYN024 and (ii.) CYN027.

We characterized mBFP production of ER2267 or Pir116 derived strains as followed. Isolated colonies were grown overnight in LB broth supplemented with appropriate antibiotic at 37 °C. Overnight culture was diluted in M9 minimal media with 0.4% w/v glucose in culture tube or microtiter wells. Cells were grown at 37 °C with shaking for 4 h and then diluted into 200 µL of fresh M9 minimal media with 0.4% w/v glucose in microtiter wells. IPTG was added between 1 to 3 h after the transition to allow for cell recovery from lag phase. Cell growth (optical density at 600 nm) and blue fluorescence (excitation and emission wavelength of 395 nm and 451 nm, respectively) were continuously monitored using Tecan. Specific mBFP production rate was calculated from the slope of a linear line fitted to specific mBFP fluorescence per unit OD600 versus time data over a period of linear increment after IPTG addition.
Supplementary Figure S1. (a.) Plasmid map of pCN-065 (pCN-LEDT). Labelled arrows indicate genes. RBSs and terminators are represented by yellow boxes and dark red arrow, respectively. The plasmid map is created using ApE - A plasmid Editor v2.0.46 and SnapGene Viewer 2.3.5. (b.) Sequence of the synthetic Entner-Doudoroff operon for ED1.0.
Supplementary Figure S2. Metagenomic blue fluorescent protein (mBFP) as an in vivo NADPH biosensor. (A) Time course specific mBFP fluorescence per cell signal and the corresponding specific mBFP production rate of strain ER2267 (wild-type without mBFP plasmid), CYN014 (ER2267 pQE-mBFP) and CYN019 (ER2267 pQE-mBFP(Cat)) with and without IPTG addition. Increase in blue fluorescence per unit OD$_{600}$ was detectable upon IPTG induction only for the mBFP harboring strains CYN014 and CYN019 but not for the negative control strain ER2267. Values and error bars represent the means and s.d. of two repeats. (B) Measurement of mBFP production rate of strain CYN027 (E. coli Pir116 co-expressing transhydrogenase PntAB and mBFP on an R6K vector) and the control strain CYN024 (E. coli Pir116 co-expressing AmtR and mBFP on an R6K vector) at two different IPTG concentrations. CYN027 expressing the PntAB transhydrogenase exhibited higher specific mBFP production rate than the control strain. Values and error bars represent the means and s.d. of 3-4 repeats. All experiments were conducted at 37°C with M9 minimal glucose media. (Supplementary Materials and Methods S1.3).

Supplementary Table S1. Strains and plasmids. (Chr) indicates chromosomal integration at the yciL-tonB intergenic region.

<table>
<thead>
<tr>
<th>Strains or plasmids</th>
<th>Relevant characteristics</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli strains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG1655</td>
<td>F-, lambda-, rph-1</td>
<td>Thomas Wood's Lab</td>
</tr>
<tr>
<td>MG1655 crtEBI</td>
<td>MG1655 pIF-001C</td>
<td>This study</td>
</tr>
<tr>
<td>DH10B/TOP10</td>
<td>F- mcrA Δ(mrr-hsdRMS-mcrBC) q80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(StrR) endA1 λ-</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>ER2267</td>
<td>K12 F- proA+B+ lacIq Δ(lacZ)M15 zff::mini-Tn10 (KanR)/ Δ(argF-lacZ)U169 glnV44 e14-Δ(mcrA-) rfbD1? recA1 relA? endA1 spoT1? thi-1 Δ(mcrC-mrr)114::IS10</td>
<td>New England Biolabs Inc</td>
</tr>
<tr>
<td>CYN014</td>
<td>Same as ER2267 but with pQE-mBFP</td>
<td>This study</td>
</tr>
<tr>
<td>CYN019</td>
<td>Same as ER2267 but with pQE-mBFP(Cat)</td>
<td>This study</td>
</tr>
<tr>
<td>Pir116 (TransforMax™ EC100D™ pir-116 )</td>
<td>F- mcrA Δ(mrr-hsdRMS-mcrBC) q80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ- rpsL (StrR) nupG pir-116(DHFR)</td>
<td>Epicentre</td>
</tr>
<tr>
<td>CYN024</td>
<td>Same as Pir116 but with pCN-LA pQE-mBFP (Cat-)</td>
<td>This study</td>
</tr>
<tr>
<td>CYN027</td>
<td>Same as Pir116 but with pCN-LPab pQE-mBFP (Cat-)</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Wang et al., 2009)</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>EcNR2</td>
<td>MG1655 bioA/bioB::λ-Red-bla ΔmutS::cat</td>
<td></td>
</tr>
<tr>
<td>EcNR2 mBFP</td>
<td>Same as EcNR2 but with pCN-mBFP</td>
<td>This study</td>
</tr>
<tr>
<td>EcNR2 crtEBI</td>
<td>Same as EcNR2 but with pIF-001K</td>
<td>This study</td>
</tr>
<tr>
<td>Δpgi</td>
<td>Same as EcNR2 but with nonsense inactivation of pgi (two stop codons (taa) replacing codon 27 and 28 of pgi)</td>
<td>This study</td>
</tr>
<tr>
<td>Δpgi crtEBI</td>
<td>Same as Δpgi but with pIF-001K</td>
<td>This study</td>
</tr>
<tr>
<td>EcIF15</td>
<td>E_IF3 (MG1655 bioA/bioB::λ-Red-bla ΔmutS::kanR) with RBS of dxs: AACAATAAGTATAAGGAGGCCCCTG</td>
<td>(Farasat et al., 2014)</td>
</tr>
<tr>
<td>EcIF15 crtEBI</td>
<td>Same as EcIF15 but with pIF-002</td>
<td>(Farasat et al., 2014)</td>
</tr>
<tr>
<td>ED1.0</td>
<td>Same as EcNR2 but with ED-tetAR(Chr)</td>
<td>This study</td>
</tr>
<tr>
<td>ED1.0 mBFP</td>
<td>Same as ED 1.0 but with pCN-mBFP</td>
<td>This study</td>
</tr>
<tr>
<td>ED1.0 crtEBI</td>
<td>Same as ED 1.0 but with pIF-001K</td>
<td>This study</td>
</tr>
<tr>
<td>ED1.0 MEP crtEBI</td>
<td>Same as EcIF15 but with ED-tetAR(Chr) and pIF-002</td>
<td>This study</td>
</tr>
<tr>
<td>EDi, i= 2, … , 23</td>
<td>Same as ED1.0 but with Zm-zwf, Zm-pgi, Zm-pgl, Zm-edd and Zm-edα genomic RBS libraries and pCN-mBFP</td>
<td>This study, refer to Table S4 for details</td>
</tr>
<tr>
<td>EDiR</td>
<td>Same as EcNR2 but with EDi's ED-tetAR(Chr)</td>
<td>This study</td>
</tr>
<tr>
<td>EDiR crtEBI</td>
<td>Same as EDiR but with pIF-001K</td>
<td>This study</td>
</tr>
<tr>
<td>EDiR, MEP crtEBI</td>
<td>Same as EcIF15 but with EDi-tetAR(Chr) and pIF-002</td>
<td>This study</td>
</tr>
</tbody>
</table>

Plasmids

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>(Hwang et al., 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pQE-mBFP</td>
<td>T5 promoter with lac operator; mBFP; ColE1 origin; cmR; ampR</td>
<td></td>
</tr>
<tr>
<td>pQE-mBFP(Cat-)</td>
<td>T5 promoter with lac operator; mBFP; ColE1 origin; ampR</td>
<td></td>
</tr>
<tr>
<td>pCN-mBFP</td>
<td>ColE1 origin; Insulated consitutive sigma 70 promoter (TTCTTTGAGCACAGCTACACACCACGTCGTTCCATATCTGCTGCTTAGGTCTATGAGTG GTTGCTGGATACGACGTCTTCTTTTTTGACAGCTAGCTACAGTCTAGTATAATAATATTCAGGGGACCACAGGATCCACGGGTTC CCACTACAAATAATTTGTATTTAATCTTTTAGACTAG) ; 356,786 au RBS (TAAAGACCAGACGATTAGTTAAGGAGGGA AAAAAAC); mBFP; kanR</td>
<td>This study</td>
</tr>
<tr>
<td>pCN-MCSO</td>
<td>R6K origin; cmR</td>
<td>This study</td>
</tr>
<tr>
<td>pCN-L</td>
<td>R6K origin; constitutive promoter::LacI; cmR</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>R6K origin; constitutive promoter::LacI; pTac::Amr; cmR</td>
<td>This study</td>
</tr>
<tr>
<td>----</td>
<td>--------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>pCN-LPab</td>
<td>R6K origin; constitutive promoter::LacI; pTac::PntA PntB; cmR</td>
<td>This study</td>
</tr>
<tr>
<td>pCN-LED (pCN-053)</td>
<td>R6K origin; pTac::Zm-zwf Zm-pgi(A123S); pTac::Zm-edd Zm-edd Zm-pgl; constitutive promoter::LacI; cmR</td>
<td>This study</td>
</tr>
<tr>
<td>pCN-LEDT (pCN-065)</td>
<td>R6K origin; pTac::Zm-zwf Zm-pgi(A123S); pTac::Zm-edd Zm-edd Zm-pgl; constitutive promoter::LacI; cmR</td>
<td>This study</td>
</tr>
<tr>
<td>pIF-001C</td>
<td>ColE1 origin; IPTG inducible PlacO1 promoter; Rhodobacter sphaeroides crtEBI operons with optimized RBS (N14); cmR</td>
<td>(Farasat et al., 2014)</td>
</tr>
<tr>
<td>pIF-001K</td>
<td>ColE1 origin; IPTG inducible PlacO1 promoter; Rhodobacter sphaeroides crtEBI operons with optimized RBS (N14); kanR</td>
<td>This study</td>
</tr>
<tr>
<td>pIF-002</td>
<td>ColE1 origin; Arabinose inducible PBAD promoter; Rhodobacter sphaeroides crtEBI operons with optimized RBS (N14); cmR</td>
<td>(Farasat et al., 2014)</td>
</tr>
</tbody>
</table>

**Supplementary Table S2.** Oligonucleotides used for MAGE. * indicate phosphorothioated base.

<table>
<thead>
<tr>
<th>Oligonucleotides</th>
<th>Sequence (5’→3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAGE to fix Zm-pgi mutation</strong></td>
<td></td>
</tr>
<tr>
<td>prCY103</td>
<td>T<em>G</em>CCAAAGAATACCATGCACGCATGCGCACCCTGATTGAA GCTATTGATGCTGGTGCATTTGGCGAAGTAAAACACCTGCTG CATATTGG</td>
</tr>
<tr>
<td><strong>MAGE to introduce RBS libraries (degenerate oligos)</strong></td>
<td></td>
</tr>
<tr>
<td>Zm-zwf_dRBS</td>
<td>T<em>G</em>T<em>G</em>AGCGGCTCAACATTGCTAGCTAACTACATATAKWA SGAGKTAACACATGACTAAACCCGATCCACGATATTCTGT TTGGCTCACC</td>
</tr>
<tr>
<td>Zm-kdpga_dRBS</td>
<td>G<em>C</em>T<em>A</em>TCTATGGGCGGGCGTGATTCTAAATAGCAAACCTAA RCRTAASGMGGGTCCCCAAAATGCGATATTCCGATGTG CAGCTTAGCAC</td>
</tr>
<tr>
<td>Zm-pgd_dRBS</td>
<td>T<em>G</em>T<em>G</em>AGCGGCTCAACATTGCGGCGACACGRWAKAGKAGG TACACACGATATTCTTCATCCACCGTGATTCCGATG CAGCTTAGGAT</td>
</tr>
<tr>
<td>Zm-pgi_dRBS</td>
<td>T<em>G</em>A<em>C</em>GGCGTGACTTGGTACGACTAATAGACCCGCCATA ATTAAAGMGKMGWCAATGGCAGCTATGCAATAAAAGCA GCAATTGACGCA</td>
</tr>
<tr>
<td>Zm-pgl_dRBS</td>
<td>T<em>T</em>A*AACGTCGACGGGTGCTTAATAGGATAGAAACGCA WTMAGSASGTAAATATGACGGAAGCTGAATTGATGGGAAATT CAAAACCGTGG</td>
</tr>
</tbody>
</table>
Introduce two consecutive stop codons to knockout pgi gene

**Supplementary Table S3.** RBS sequences in each RBS library and the corresponding translation rate (au) predicted by RBS Library Calculator (Farasat et al., 2014).

<table>
<thead>
<tr>
<th>No.</th>
<th>Sequence</th>
<th>Translation rate (au)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a) Zm-zwf RBS library</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dRBS</td>
<td>TAACTACATATAK\textsubscript{KWASGAGKTAACAC}</td>
</tr>
<tr>
<td>1</td>
<td>TAACTACATATAAGGAGGTAACAC</td>
<td>983028.6</td>
</tr>
<tr>
<td>2</td>
<td>TAACTACATATATTAGGAGGTAACAC</td>
<td>254806.2</td>
</tr>
<tr>
<td>3</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>99028.6</td>
</tr>
<tr>
<td>4</td>
<td>TAACTACATATAACGAGGTAACAC</td>
<td>72267.7</td>
</tr>
<tr>
<td>5</td>
<td>TAACTACATATAAGGAGTTAACAC</td>
<td>63140.6</td>
</tr>
<tr>
<td>6</td>
<td>TAACTACATATAAGGAGGTAACAC</td>
<td>60362</td>
</tr>
<tr>
<td>7</td>
<td>TAACTACATATATTACGAGGTAACAC</td>
<td>20496.4</td>
</tr>
<tr>
<td>8</td>
<td>TAACTACATATATTAGGAGGTAACAC</td>
<td>18732.2</td>
</tr>
<tr>
<td>9</td>
<td>TAACTACATATAGTACGAGGTAACAC</td>
<td>13670.1</td>
</tr>
<tr>
<td>10</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>6360.7</td>
</tr>
<tr>
<td>11</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>5312.8</td>
</tr>
<tr>
<td>12</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>3387.4</td>
</tr>
<tr>
<td>13</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>2585.8</td>
</tr>
<tr>
<td>14</td>
<td>TAACTACATATATTACGAGGTAACAC</td>
<td>2363.2</td>
</tr>
<tr>
<td>15</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>80.5</td>
</tr>
<tr>
<td>16</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>217.6</td>
</tr>
<tr>
<td></td>
<td>(b) Zm-pgi RBS library</td>
<td></td>
</tr>
</tbody>
</table>

---

Zm-tetA\textsubscript{dRBS}

\text{G*T*A*TTACCTCCTCSWTCATGATGAGGAGTGTAAATAACTCTATCAATGATAGA}

Co-selection MAGE

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Translation rate (au)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bla\textsubscript{on}</td>
<td>\text{G<em>C</em>C<em>A</em>CATTAGCAGAATTCTTTAAAGTGCTCTCATCGATGGAAAGCTTTCTCGCTGTGATACCAG}</td>
</tr>
<tr>
<td>bla\textsubscript{off}</td>
<td>\text{G<em>C</em>C<em>A</em>CATTAGCAGAATTCTTTAAAGTGCTCTCATCGATGGAAAGCTTTCTCGCTGTGATACCAG}</td>
</tr>
</tbody>
</table>

---

Zm-pgi\textsubscript{RBS library}

\text{G*G*A*GAACTTAGAAAAACGATCGCCGTCTTTAGCAAAAAATGTTACGAGTGTTCGCTTGAGATCCAG}

**No.**

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16

---

Zm-zwf\textsubscript{RBS library}

\text{TAACTACATATAK\textsubscript{KWASGAGKTAACAC}}

<table>
<thead>
<tr>
<th>No.</th>
<th>Sequence</th>
<th>Translation rate (au)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TAACTACATATAAGGAGGTAACAC</td>
<td>983028.6</td>
</tr>
<tr>
<td>2</td>
<td>TAACTACATATATTAGGAGGTAACAC</td>
<td>254806.2</td>
</tr>
<tr>
<td>3</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>99028.6</td>
</tr>
<tr>
<td>4</td>
<td>TAACTACATATAACGAGGTAACAC</td>
<td>72267.7</td>
</tr>
<tr>
<td>5</td>
<td>TAACTACATATAAGGAGTTAACAC</td>
<td>63140.6</td>
</tr>
<tr>
<td>6</td>
<td>TAACTACATATAAGGAGGTAACAC</td>
<td>60362</td>
</tr>
<tr>
<td>7</td>
<td>TAACTACATATATTACGAGGTAACAC</td>
<td>20496.4</td>
</tr>
<tr>
<td>8</td>
<td>TAACTACATATATTAGGAGGTAACAC</td>
<td>18732.2</td>
</tr>
<tr>
<td>9</td>
<td>TAACTACATATAGTACGAGGTAACAC</td>
<td>13670.1</td>
</tr>
<tr>
<td>10</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>6360.7</td>
</tr>
<tr>
<td>11</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>5312.8</td>
</tr>
<tr>
<td>12</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>3387.4</td>
</tr>
<tr>
<td>13</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>2585.8</td>
</tr>
<tr>
<td>14</td>
<td>TAACTACATATATTACGAGGTAACAC</td>
<td>2363.2</td>
</tr>
<tr>
<td>15</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>80.5</td>
</tr>
<tr>
<td>16</td>
<td>TAACTACATATAGTAGGAGGTAACAC</td>
<td>217.6</td>
</tr>
<tr>
<td>dRBS</td>
<td>ACCCGCCATAATTAAGGAGGTAGACA</td>
<td>209974</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>2</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>44050.1</td>
</tr>
<tr>
<td>3</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>23459.2</td>
</tr>
<tr>
<td>4</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>17907.8</td>
</tr>
<tr>
<td>5</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>17119.8</td>
</tr>
<tr>
<td>6</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>4641.8</td>
</tr>
<tr>
<td>7</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>2363.2</td>
</tr>
<tr>
<td>8</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>1973.9</td>
</tr>
<tr>
<td>9</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>1506.8</td>
</tr>
<tr>
<td>10</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>1150.2</td>
</tr>
<tr>
<td>11</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>260.5</td>
</tr>
<tr>
<td>12</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>181.7</td>
</tr>
<tr>
<td>13</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>173.7</td>
</tr>
<tr>
<td>14</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>121.2</td>
</tr>
<tr>
<td>15</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>ACCCGCCATAATTAAGGAGGTAAGAC</td>
<td>15.3</td>
</tr>
</tbody>
</table>

(c) Zm-pgl RBS library

<table>
<thead>
<tr>
<th>dRBS</th>
<th>GATAGAAAGCAAWTMAGSASGTAGTT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>821079.7</td>
</tr>
<tr>
<td>2</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>599195.9</td>
</tr>
<tr>
<td>3</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>82714.1</td>
</tr>
<tr>
<td>4</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>22426.9</td>
</tr>
<tr>
<td>5</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>21439.9</td>
</tr>
<tr>
<td>6</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>13068.5</td>
</tr>
<tr>
<td>7</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>8332.5</td>
</tr>
<tr>
<td>8</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>4055.6</td>
</tr>
<tr>
<td>9</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>1005</td>
</tr>
<tr>
<td>10</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>489.1</td>
</tr>
<tr>
<td>11</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>427.4</td>
</tr>
<tr>
<td>12</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>190.1</td>
</tr>
<tr>
<td>13</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>132.6</td>
</tr>
<tr>
<td>14</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>92.5</td>
</tr>
<tr>
<td>15</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>17.5</td>
</tr>
<tr>
<td>16</td>
<td>GATAGAAAGCAAWTMAGSASGTAGTT</td>
<td>3.2</td>
</tr>
</tbody>
</table>

(d) Zm-edd RBS library

<table>
<thead>
<tr>
<th>dRBS</th>
<th>GCCGCACACGRWAKAAGKAGGTACAC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GCCGCACACGRWAKAAGKAGGTACAC</td>
<td>971887.9</td>
</tr>
<tr>
<td>2</td>
<td>GCCGCACACGATATAAGGAGGTACAC</td>
<td>783067.5</td>
</tr>
<tr>
<td>3</td>
<td>GCCGCACACGGAATAAGGAGGTACAC</td>
<td>304333.6</td>
</tr>
<tr>
<td>4</td>
<td>GCCGCACACGAAAGAAGGAGGTACAC</td>
<td>118277.1</td>
</tr>
<tr>
<td>5</td>
<td>GCCGCACACGGTATAAGGAGGTACAC</td>
<td>86314.6</td>
</tr>
<tr>
<td>6</td>
<td>GCCGCACACGATAGAAGGAGGTACAC</td>
<td>78884.9</td>
</tr>
<tr>
<td>7</td>
<td>GCCGCACACGGAAGAAGGAGGTACAC</td>
<td>40162</td>
</tr>
<tr>
<td>8</td>
<td>GCCGCACACGAAATAAGTAGGTACAC</td>
<td>32069.3</td>
</tr>
<tr>
<td>9</td>
<td>GCCGCACACGGAATAAGTAGGTACAC</td>
<td>24480.4</td>
</tr>
<tr>
<td>10</td>
<td>GCCGCACACGGTAGAAGGAGGTACAC</td>
<td>7946.7</td>
</tr>
<tr>
<td>11</td>
<td>GCCGCACACGAAAGAAGTAGGTACAC</td>
<td>6066.2</td>
</tr>
<tr>
<td>12</td>
<td>GCCGCACACGGTATAAGTAGGTACAC</td>
<td>2059.8</td>
</tr>
<tr>
<td>13</td>
<td>GCCGCACACGAAAGAAGTAGGTACAC</td>
<td>1255.6</td>
</tr>
<tr>
<td>14</td>
<td>GCCGCACACGAAAGAAGTAGGTACAC</td>
<td>6997.4</td>
</tr>
<tr>
<td>15</td>
<td>GCCGCACACGAAAGAAGTAGGTACAC</td>
<td>3405.8</td>
</tr>
<tr>
<td>16</td>
<td>GCCGCACACGAAAGAAGTAGGTACAC</td>
<td>173.3</td>
</tr>
</tbody>
</table>

(e) Zm-eda RBS library

<table>
<thead>
<tr>
<th>dRBS</th>
<th>CAACCAACACGGAAGAAGTAGGTACAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>2</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>3</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>4</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>5</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>6</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>7</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>8</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>9</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>10</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>11</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>12</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>13</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>14</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>15</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
<tr>
<td>16</td>
<td>CAACCAACACGGAAGAAGTAGGTACAC</td>
</tr>
</tbody>
</table>

Supplementary Table S4. Translation rate for ED variants

<table>
<thead>
<tr>
<th>Name</th>
<th>Predicted translation initiation rate, au</th>
<th>Mean normalized mBFP production rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zm-pgi</td>
<td>Zm-zwf</td>
</tr>
</tbody>
</table>

Zm-pgi | Zm-zwf | Zm-pgl | Zm-edd | Zm-eda |

234
Supplementary Table S5. Net reaction for neurosporene biosynthesis from glyceraldehyde-3-phosphate and pyruvate [285,316,317,318].

<table>
<thead>
<tr>
<th>Gene</th>
<th>Enzyme Name</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dxs</td>
<td>1-deoxy-D-xylulose 5-phosphate synthase</td>
<td>g3p + h + pyr → co2 + dxyl5p</td>
</tr>
<tr>
<td>dxr</td>
<td>1-deoxy-D-xylulose reductoisomerase</td>
<td>dxyl5p + h + nadph → 2me4p + nadp</td>
</tr>
<tr>
<td>ispD</td>
<td>2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase</td>
<td>2me4p + ctp + h → 4c2me + ppi</td>
</tr>
<tr>
<td>ispE</td>
<td>4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase</td>
<td>4c2me + atp → 2p4c2me + adp + h</td>
</tr>
<tr>
<td>ispF</td>
<td>2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase</td>
<td>2p4c2me → 2mecdp + cmp</td>
</tr>
<tr>
<td>ispG</td>
<td>1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase</td>
<td>2mecdp + (2) flxo/nadph + h → h2mb4p + h2o + (2) flxo/nadp</td>
</tr>
<tr>
<td>ispH/idi combined</td>
<td>1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase and isopenentenyl-diphosphate D-isomerase</td>
<td>h2mb4p + nad(p)h + h → ipdp + h2o + nad(p)</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Name</td>
<td>Formula</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>2me4p</td>
<td>2-C-methyl-D-erythritol 4-phosphate</td>
<td>C5H11O7P</td>
</tr>
<tr>
<td>2mecdp</td>
<td>2-C-methyl-D-erythritol 2,4-cyclodiphosphate</td>
<td>C5H10O9P2</td>
</tr>
<tr>
<td>2p4c2me</td>
<td>2-phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol</td>
<td>C14H22N3O17P3</td>
</tr>
<tr>
<td>4c2me</td>
<td>4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol</td>
<td>C14H23N3O14P2</td>
</tr>
<tr>
<td>adp</td>
<td>ADP</td>
<td>C10H15N5O10P2</td>
</tr>
<tr>
<td>atp</td>
<td>ATP</td>
<td>C10H12N5O13P3</td>
</tr>
<tr>
<td>co2</td>
<td>carbon dioxide</td>
<td>CO2</td>
</tr>
<tr>
<td>ctp</td>
<td>cytidine-triphosphate</td>
<td>C9H12N3O14P3</td>
</tr>
<tr>
<td>dmpp</td>
<td>dimethylallyl diphosphate</td>
<td>C5H9O7P</td>
</tr>
<tr>
<td>dxy5p</td>
<td>1-deoxy-D-xylulose 5-phosphate</td>
<td>C5H9O7P</td>
</tr>
<tr>
<td>fad</td>
<td>Flavin adenine dinucleotide oxidized</td>
<td>C27H33N9O15P2</td>
</tr>
<tr>
<td>fadh2</td>
<td>Flavin adenine dinucleotide reduced</td>
<td>C27H35N9O15P2</td>
</tr>
<tr>
<td>flxo</td>
<td>flavodoxin (oxidized)</td>
<td>XH</td>
</tr>
<tr>
<td>flxr</td>
<td>flavodoxin (reduced)</td>
<td>XH2</td>
</tr>
<tr>
<td>frdp</td>
<td>farnesyl diphosphate</td>
<td>C15H25O7P</td>
</tr>
<tr>
<td>g3p</td>
<td>D-glyceraldehyde 3-phosphate</td>
<td>C3H7O6P</td>
</tr>
<tr>
<td>gpp</td>
<td>geranylgeranyl diphosphate</td>
<td>C20H36O7P</td>
</tr>
<tr>
<td>grdp</td>
<td>geranyl diphosphate</td>
<td>C10H20O7P</td>
</tr>
<tr>
<td>h2mb4p</td>
<td>1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate</td>
<td>C5H9O8P2</td>
</tr>
<tr>
<td>h2o</td>
<td>water</td>
<td>H2O</td>
</tr>
<tr>
<td>ipdp</td>
<td>isopentenyl diphosphate</td>
<td>C5H9O7P</td>
</tr>
<tr>
<td>nadp</td>
<td>Nicotinamide adenine dinucleotide phosphate</td>
<td>C21H28N7O17P3</td>
</tr>
<tr>
<td>nadph</td>
<td>Nicotinamide adenine dinucleotide phosphate - reduced</td>
<td>C21H30N7O17P3</td>
</tr>
<tr>
<td>neurosporene</td>
<td></td>
<td>C40H58</td>
</tr>
<tr>
<td>phy</td>
<td>phytoene</td>
<td>C40H64</td>
</tr>
<tr>
<td>ppi</td>
<td>diphosphate</td>
<td>HO7P2</td>
</tr>
<tr>
<td>-----</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>pyr</td>
<td>pyruvate</td>
<td>C3H4O3</td>
</tr>
</tbody>
</table>

**Supplementary Table S6.** Comparison of neurosporene production from pIF-001C (P\text{lacO1}\_crtEBI) and pIF-002 (P\text{BAD}\_crtEBI) in ED17\text{R,MEP}. Values represent average of two replicates.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Strain</th>
<th>Inducer</th>
<th>Average neurosporene content (µg/ gDCW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2x M9, 37°C, 200 RPM, 10 h culture</td>
<td>EcIF15 + ED17’s ED-tetAR + pIF-001C (PlacO1-crtEBI)</td>
<td>0.2 mM IPTG</td>
<td>3154.0</td>
</tr>
<tr>
<td></td>
<td>EcIF15 + ED17’s ED-tetAR + pIF-002 (PBAD-crtEBI)</td>
<td>10mM arabinose + 0.2 mM IPTG</td>
<td>3714.1</td>
</tr>
</tbody>
</table>
VITA
Iman Farasat

EDUCATION

- Ph.D. in Chemical Engineering, 2009-2015
  The Pennsylvania State University; Thesis: Sequence-to-function models for efficient optimization of metabolic pathways and genetic circuits
- Ph.D. minor in Computational Science, 2012-2015
  The Pennsylvania State University
- M.Sc. in Chemical Engineering, 2006-2008
  Sharif University of Technology; Thesis: Synthesis and optimization of heat-integrated distillation columns
- B.Sc. in Chemical Engineering, 2002-2006
  Sharif University of Technology; Thesis: Developing a new model to calculate surface tension of pure components and mixtures using artificial neural networks

AWARDS & RECOGNITIONS

- Safety Improvement of the Month Award, The Pennsylvania State University, 2015
- Best Poster Presentation, SynBERC, 2013
- Departmental Best Ph.D. Candidacy Exam Award, The Pennsylvania State University, 2010
- Leighton Riess Graduate Fellowship, The Pennsylvania State University, 2010

PROFESSIONAL EXPERIENCE

- Process Design Engineer; SEPDCO Consulting Engineers, 2007-2009
- Research Assistant; The Pennsylvania State University, 2009-2015
- Teaching Assistant; Introduction to Biomolecular Engineering, The Pennsylvania State University, 2010
- Internship; Dana System Company, IT & computer programming, summer 2001
- Internship; Pars Oil and Gas Company, summer 2006

OTHER ACTIVITIES

- Treasurer, ISA Student Organization, The Pennsylvania State University, 2010-2011
- Mentored iGEM teams (won gold medals in iGEM Jamboree, 2010 & 2014)
- Trained four graduate and six undergraduate researchers in molecular biology lab techniques