Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*

Sriram Kosuri[a,b,1], Daniel B. Goodman[a,b,c,1], Guillaume Cambray[d,e,f], Vivek K. Mutalik[d,e,f], Yuan Gao[g,h,i], Adam P. Arkin[d,e,f], Drew Endy[d,j], and George M. Church[a,b,2]

[a]Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115; [b]Department of Genetics, Harvard Medical School, Boston, MA 02115; [c]Harvard-MIT Health Sciences and Technology, Cambridge, MA 02139; [d]BIOFAB: International Open Facility Advancing Biotechnology, Emeryville, CA 94608; [e]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [f]Department of Bioengineering, University of California, Berkeley, CA 94720; [g]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [h]Neuroregeneration and Stem Cell Biology Program, Institute for Cell Engineering, Johns Hopkins University, Baltimore, MD 21205; [i]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205; and [j]Department of Bioengineering, Stanford University, Stanford, CA 94305

Edited by Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved July 2, 2013 (received for review February 11, 2013)

The inability to predict heterologous gene expression levels precisely hinders our ability to engineer biological systems. Using well-characterized regulatory elements offers a potential solution only if such elements behave predictably when combined. We synthesized 12,563 combinations of common promoters and ribosome binding sites and simultaneously measured DNA, RNA, and protein levels from the entire library. Using a simple model, we found that RNA and protein expression were within twofold of expected levels 80% and 64% of the time, respectively. The large dataset allowed quantitation of global effects, such as translation rate on mRNA stability and mRNA secondary structure on translation rate. However, the worst 5% of constructs deviated from prediction by 13-fold on average, which could hinder large-scale genetic engineering projects. The ease and scale this of approach indicates that rather than relying on prediction or standardization, we can screen synthetic libraries for desired behavior.

next-generation sequencing | synthetic biology | systems biology

Organisms can be engineered to produce chemical, material, fuel, and medical products that are often superior to nonbiological alternatives (1). Biotechnologists have sought to discover, improve, and industrialize such products through the use of recombinant DNA technologies (2, 3). In recent years, these efforts have increased in complexity from expressing a few genes at once to optimizing multicomponent circuits and pathways (4–7). To attain desired systems-level function reliably, careful and time-consuming optimization of individual components is required (8–11).

To mitigate this slow trial-and-error optimization, two dominant approaches have taken hold. The first approach seeks to predict expression levels by elucidating the biophysical relationships between sequence and function. For example, several groups have modified promoters (12, 13) and ribosome binding sites (RBSs) (14–16) to see how small sequence changes affect transcription or translation. Such studies are fundamentally challenging due to the vastness of sequence space. In addition, because these approaches mostly look at either transcription or translation individually, they are rarely able to investigate interactions between these processes.

The second approach uses combinations of individually characterized elements to attain desired expression without directly considering their DNA sequences (17–25). Current efforts have focused on approaches to limit the number of time-consuming steps required to characterize potential interactions and on identifying existing or engineered elements that act predictably when used in combination (26–28). However, these studies still suggest there are enough idiosyncratic interactions and context effects that it will be necessary to construct and measure many variants of a circuit to achieve desired function (29). For larger circuits, such approaches are necessarily limited in scope due to the difficulty in measuring large numbers of combinations (26, 27).

Here, we overcome previous limitations in generating and measuring large numbers of regulatory elements by combining recent advances in DNA synthesis with novel multiplexed methods for measuring DNA, RNA, and protein levels simultaneously using next-generation sequencing. We use the method to characterize all combinations of 114 promoters and 111 RBSs and quantify how often simple measures of promoter and RBS strengths can accurately predict gene expression when used in combination. In addition, because we measure both RNA and protein levels across the library, we can quantify how translation affects mRNA levels and how mRNA secondary structure affects translation efficiency. Finally, the size of the characterized library also provides a resource for researchers seeking to achieve particular expression levels. In lieu of using standardized elements or prediction-based design, library synthesis and screening allows precise tuning of expression in arbitrary contexts.

Results

Library Design, Construction, and Initial Characterization. To explore the effects of regulatory element composition systematically, we designed and synthesized all combinations of 114 promoters with 111 RBSs (12,653 constructs in total; one combination resulted in an incompatible restriction site). We used 90 promoters from an existing library from BIOFAB: International Open Facility Advancing Biotechnology, 17 promoters from the Anderson promoter library on the BioBricks registry, 6 promoters from common cloning vectors, and a spacer sequence chosen as a negative control. From RBSs, we used 55 RBSs from the BIOFAB library, 31 from the Anderson BioBrick library, 13 from the Salis RBS Calculator (14) expected to give a range of expression, 12 commonly used RBSs from cloning vectors and the BioBrick Registry, and one sequence chosen as a negative control (reverse complement of canonical RBS sequence).

We synthesized the construct library using Agilent's oligo library synthesis (OLS) technology (30) and cloned at ~50-fold coverage into a custom medium-copy vector (pGERC), where the constructs drive expression of superfolder GFP (31) (Fig. S1). pGERC also contains an mCherry (32) reporter under constant expression by

Author contributions: S.K., D.B.G., and G.M.C. designed research; S.K., D.B.G., and Y.G. performed research; S.K., D.B.G., G.C., V.K.M., A.P.A., and D.E. contributed new reagents/analytic tools; S.K. and D.B.G. analyzed data; and S.K., D.B.G., and G.M.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequence reported in this paper has been deposited in the AddGene database (accession no. 47441).

[1]S.K. and D.B.G. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: gchurch@genetics.med.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1301301110/-/DCSupplemental.

$P_{LTetO-1}$ (33) to act as a control for extrinsic noise (Fig. 1*A*). We grew the library to early exponential phase and characterized expression levels by flow cytometry. As expected, cells in the library expressed constant levels of mCherry, whereas expression levels of GFP varied over four orders of magnitude (Fig. 1*B*). We sequence-verified 282 colonies and found that 154 (55%) were error-free. We measured fluorescence levels of 144 of the unique error-free colonies individually to act as a defined set of controls (Fig. 1*C*).

Multiplexed Measurements of DNA, RNA, and Protein Levels. We grew the entire pooled library to early exponential phase and performed multiplexed measurements of the steady-state DNA, RNA, and protein levels. We used sequencing, DNASeq and RNASeq, to obtain steady-state DNA and RNA levels, respectively, across the library (12). For obtaining protein levels, we used FlowSeq, which combines fluorescence-activated cell sorting and high-throughput DNA sequencing and is similar in design to recently published work (34, 35). Briefly, we sorted cells into 12 log-spaced bins of varying GFP/mCherry ratios; isolated, amplified, and barcoded DNA from each of the bins; and then used high-throughput sequencing to count the number of constructs that fell into each bin (Fig. 1 *A* and *D*). Using the read counts from each of the bins, we reconstructed the average expression level for each construct. Because our library contains a mixture of perfect and imperfect constructs, we only use reads that match the fully designed sequences perfectly, and thus filter out the effects of synthesis error.

Using DNASeq, we detected 98.5% of constructs and there was high concordance between technical replicates ($R^2 = 0.997$; Figs. S2 and S3). Most of the missing constructs and constructs with few DNA reads (which prevented accurate RNA level measurements) were expected to have very high expression levels, indicating either growth defects or cloning issues (Figs. S4 and S5). RNA level calculations also showed high concordance between technical replicates ($R^2 = 0.995$; Fig. S5). Overall, RNA levels varied by three orders of magnitude, but within a single promoter, the coefficient of variation was only 0.63 (Fig. 2, *Left* and Fig. S6). RNASeq data also allowed us to identify dominant transcriptional start sites for most promoters (Fig. S7). Eighty-seven percent of all promoters had one dominant start position (>60% of all mapped reads). Two promoters (marked with asterisks in Fig. S7) had very few uniquely mapping reads, did not show a strong start site, and showed unrealistic translation efficiency calculations. These observations indicated that we were missing most of the RNA (but not protein) reads from these promoters, possibly because of transcription starting after

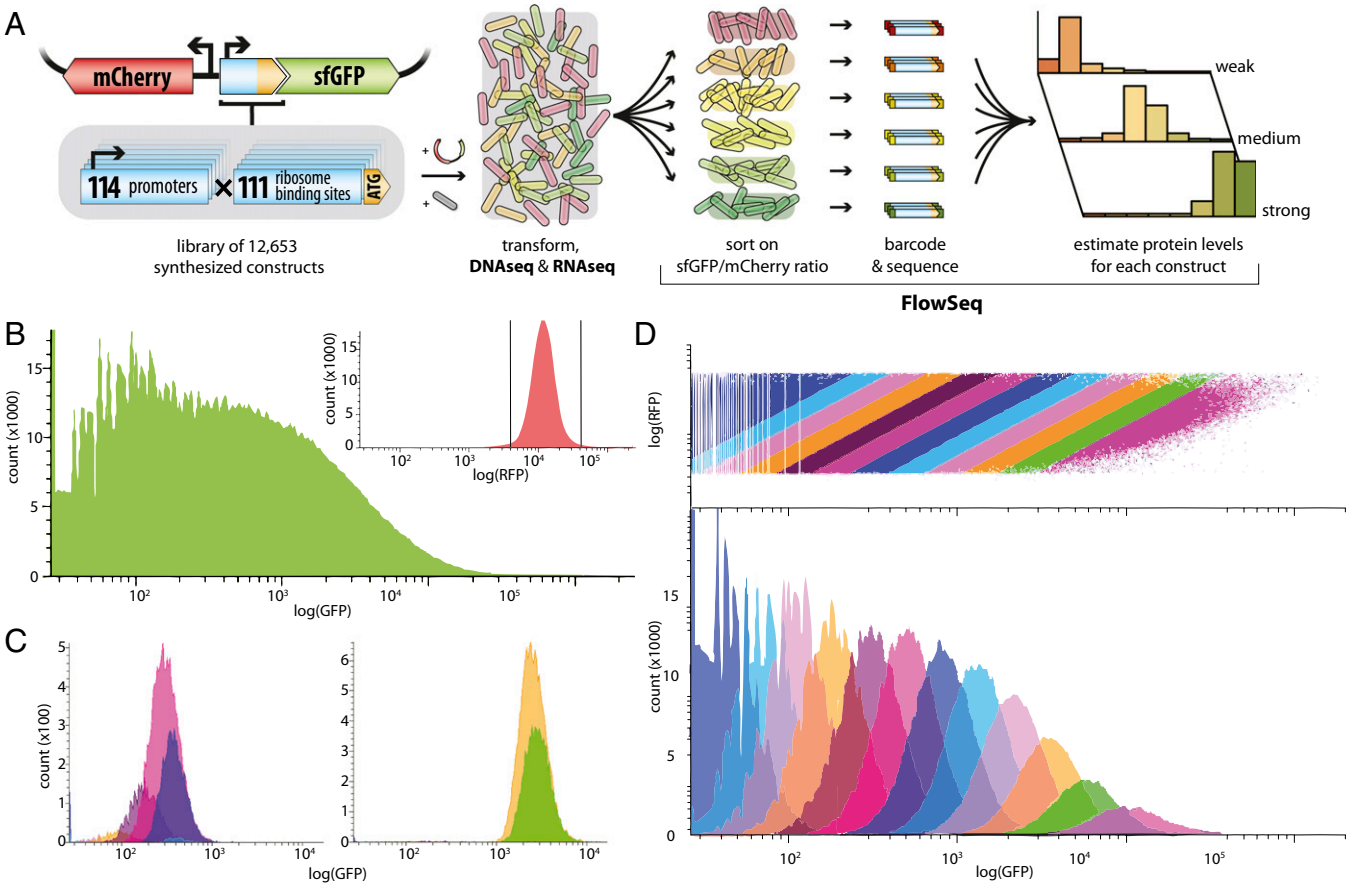

Fig. 1. Library characterization and workflow. (*A*) We synthesized all combinations of 114 promoters and 111 RBS sites to create a library containing 12,653 constructs. The library was then cloned into an expression plasmid to express superfolder GFP, and mCherry was also independently expressed from a constitutive promoter to act as an intracellular control. The cell library was harvested for DNASeq, RNASeq, and FlowSeq to quantify DNA, RNA, and protein levels, respectively, for each construct. In FlowSeq, cells were sorted into bins of varying GFP-to-mCherry ratios, barcoded, and sequenced to reconstruct protein levels for each individual construct. (*B*) GFP expression levels for the library varied over approximately four orders of magnitude compared with relatively constant red fluorescence (*Inset*). (*C*) One hundred forty-four sequence-verified clones were individually subjected to flow cytometry analysis to act as controls. Displayed are GFP levels of two representative clones, P007-R065 (*Left*) and P081-R062 (*Right*), which show that individual constructs generally fall into 2 to 3 bins. (*D*, *Upper*) Library is split into 12 log-spaced bins based on the GFP-to-RFP ratio. (*D*, *Lower*) Individual bins have large differences in the number of cells that fall into each one.

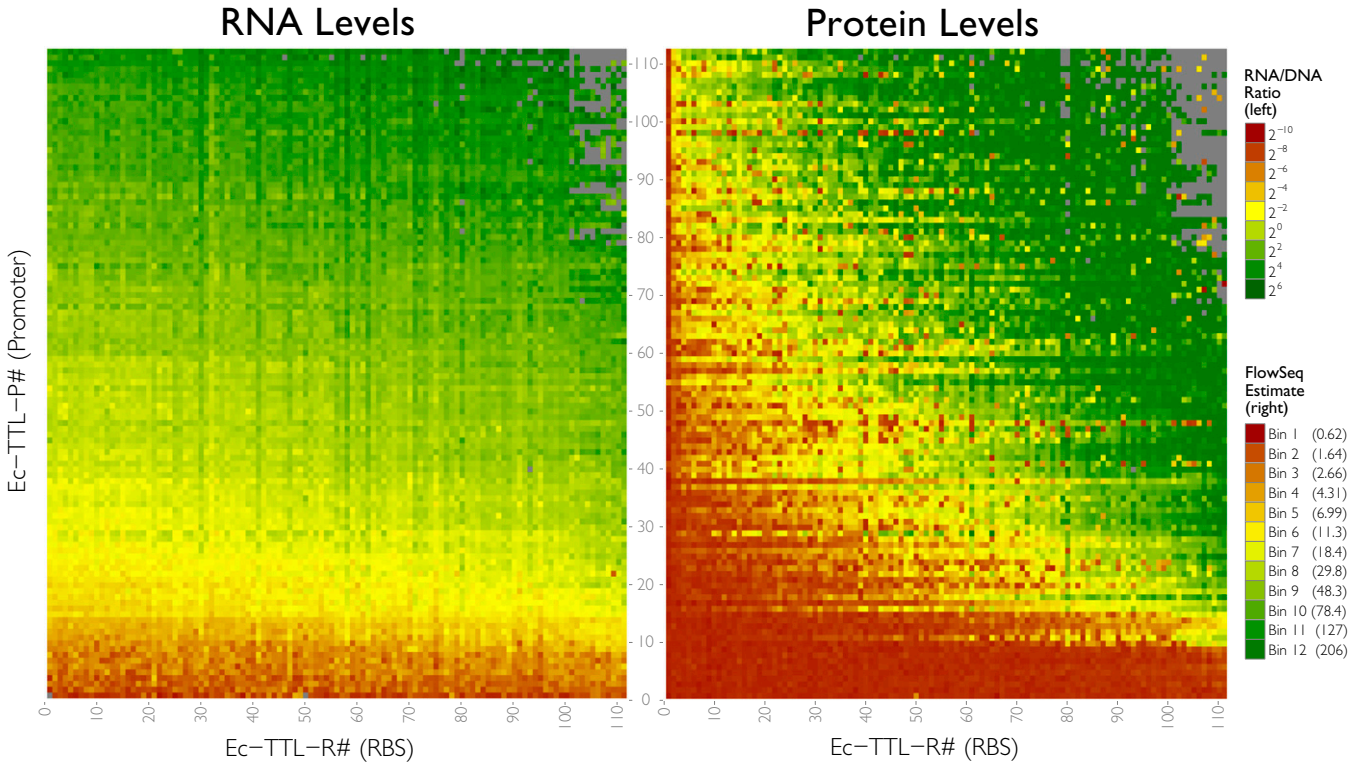

Fig. 2. RNA and protein level grids. The RNA (*Left*) and protein (*Right*) levels for all 12,653 constructs are plotted on a grid according to the identity of construct's promoter (*y* axis) and RBS (*x* axis). Promoters and RBSs are sorted by average RNA and protein abundance, respectively. Gray boxes indicate constructs that were below empirically determined cutoffs. Scale bars for RNA (RNA/DNA ratio) and protein (relative fluorescent units of GFP/RFP ratio) levels are shown to the right.

the end of the barcode sequence preventing unique identification. The 222 constructs (1.7%) containing these promoters were removed from all analyses.

Using FlowSeq, we were able to reconstruct expected protein levels for 94% of the constructs (Fig. 2, *Right*). As expected, individual constructs mostly fell into one to three contiguous

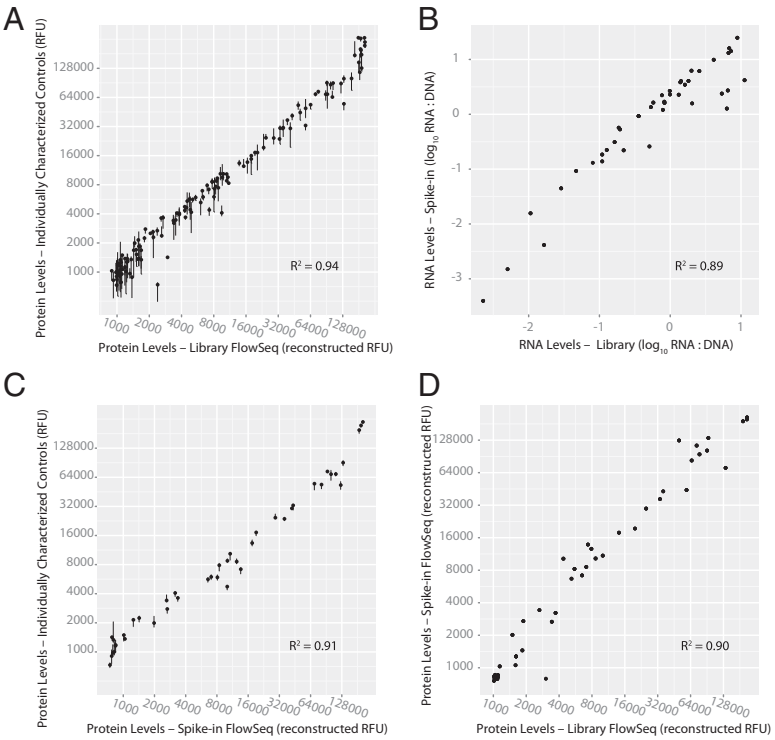

Fig. 3. Library measurements vs. individual colony and spike-in controls. (*A*) Protein levels for 141 sequence-verified constructs characterized by at least two flow cytometry measurements plotted against their FlowSeq-estimated protein levels. One construct of 142 is missing because it had insufficient reads in the FlowSeq analysis. (*B*) RNA levels for 41 constructs as measured in our library plotted against control constructs spiked into a separate library. One construct of 42 is missing because it had no reads in the spike-in data. (*C*) Protein levels for 42 control constructs spiked into a separate library plotted against protein levels for those same constructs measured at least twice by flow cytometry. (*D*) Protein levels for 42 control constructs spiked into a separate library are plotted against protein level measurements as measured in our promoter + RBS library. (All R^2 values for linear regressions pass an *F* test with a *P* value <2.2e-16.). RFU, relative fluorescent units.

Table 1. Lookup table of regulatory elements for given RNA and protein levels

Protein levels	Low RNA, 0.5 ± 0.13	Medium RNA, 2.1 ± 0.53	High RNA, 6.9 ± 1.73
Low protein: 7,393 ± 1,848	**107** P041-R034, P051-R032, P042-R013	**69** P084-R002, P070-R006, P061-R040	**23** P092-R022, P095-R002, P097-R039
Med protein: 39,450 ± 9,863	**95** P055-R032, P017-R107, P022-R096	**178** P070-R031, P035-R107, P060-R089	**157** P086-R028, P109-R015, P094-R006
High protein: 152,484 ± 38,121	**3** P018-R110, P029-R108, P031-R102	**252** P055-R055, P049-R090, P056-R086	**338** P089-R052, P077-R100, P086-R055

We chose three levels of low (17th percentile), medium (50th percentile), and high (83th percentile) RNA and protein levels and determined how many promoter and RBS combinations fall within 25% of those desired levels. The total number of combinations that fall within each range is shown in boldface, along with three examples from each group. RNA levels are given as the measured RNA/DNA ratio, and protein levels are given in relative fluorescence units.

flow-sorted bins (Fig. S8). The average protein expression levels displayed a large range and were highly correlated with the independently characterized constructs (R^2 = 0.94; Fig. 3*A* and Fig. S9). Due to the boundaries of our sorted bins, we determined that accurate quantitation was limited within a maximum and minimum range; 6.5% of the constructs were above and 14% were below this range (Fig. S9). Again, most constructs with missing measurements (insufficient or zero reads) contained combinations of strong promoters and RBSs. We calculated average promoter and RBS strengths by averaging transcription levels and translation efficiency (protein/RNA), respectively (Datasets S1 and S2). Promoters and RBSs were ordered and named based on their relative deviation from the average element (*SI Materials and Methods*).

Finally, we spiked 42 of the individual clones into a separate library (not analyzed here) and performed DNASeq, RNASeq, and FlowSeq to test reproducibility in biological replicates. Once again, protein levels were highly correlated with the individual measurements (R^2 = 0.91; Fig. 3*C*). Reconstructed values for RNA and protein levels also matched well between independent runs (R^2 = 0.89 and R^2 = 0.90, respectively; Fig. 3 *B* and *D*).

Composability of Gene Expression. Our large dataset allows us to measure the extent to which combining regulatory elements led to predictable outcomes. Using a simple model for gene expression, where promoter strengths determine RNA levels and RBS strengths determine translation efficiencies, we reconstructed expected expression across all constructs and compared them with measurements (Fig. 4). We find that 80% of RNA levels and 64% of protein levels fall within twofold of the model predictions, and had R^2 = 0.92 and R^2 = 0.76 for RNA and protein, respectively (Fig. S10 *A* and *B*).

When unexpected levels of expression do occur, they can be quite large; the largest 5% of protein model deviations are off by an average of 13-fold. Such unpredictability makes precise engineering of large systems intractable. The ease and scale of these measurements indicate that rather than using prediction or standardization to construct a single design, we can construct a library to screen for desired expression levels when optimizing large genetic systems. Desired RNA and protein levels for an entire pathway of genes could be chosen from measurements across subsets of promoters and RBSs for each gene. For example, given a desired protein level, we can choose from many sequence-divergent promoter and RBS combinations that achieve desired transcription and translation strengths of GFP (Table 1).

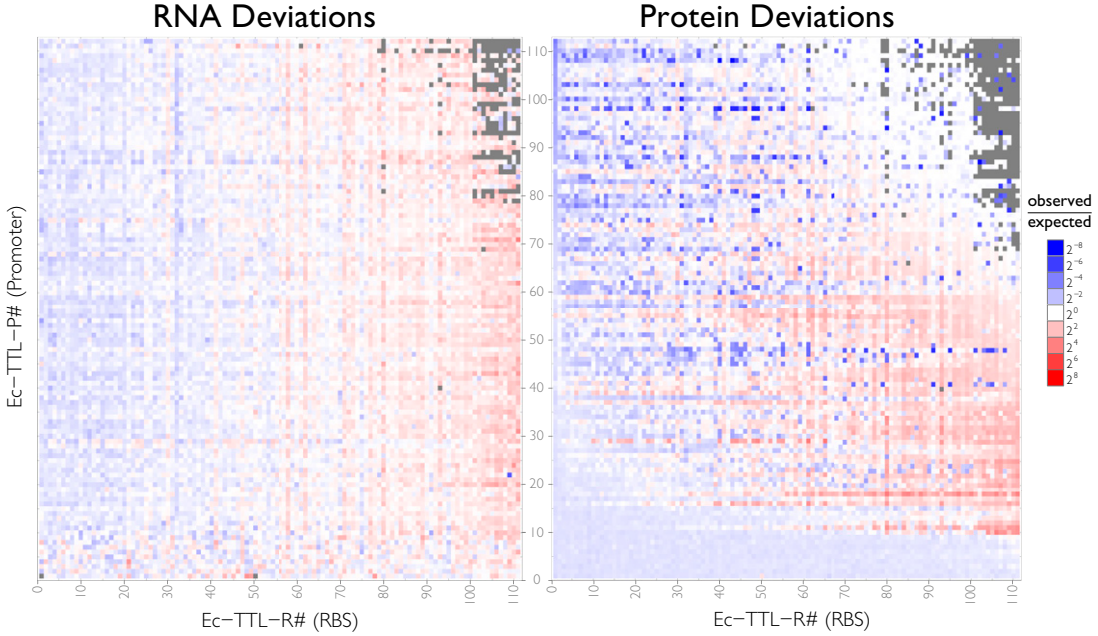

Fig. 4. RNA and protein model deviations. Based the promoter and RBS strengths, we calculated expected RNA (*Left*) and protein (*Right*) levels for each construct. Red and blue denote measured values below and above expectation, and they are plotted on the same scale for both plots. For constructs where expected protein levels are above or below the empirically determined thresholds, we set the prediction to be at the threshold level.

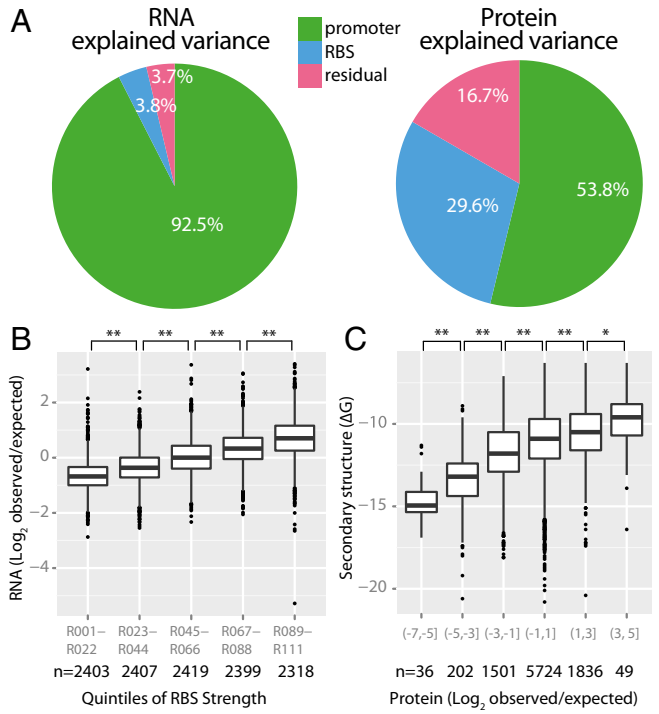

Fig. 5. ANOVA explained variance and composition effects of promoter and RBS pairs. (*A*) Explained variance (as percentages of the sum of squared deviations) for RNA and protein measurements using ANOVA. One pie chart shows partitioned variance for RNA measurements (*Left*), whereas the other chart shows partitioned variance for protein measurements (*Right*). "Residual" indicates the unexplained variance in the model. (*B*) Deviation from expected RNA level is correlated with RBS strength. RBSs are partitioned into five groups based on increasing average translation strength. (*C*) Free energy of a transcript's 5′ secondary structure (transcription start site to +30 of superfolder GFP) is correlated with average deviation from the expected protein level. Average deviations are partitioned into six equal ranges. Brackets at the top indicate two-sample Student *t* tests with *P* values <2e-5 (**) and <0.02 (*). The box plot displays the median, with hinges indicating the first and third quartiles. Whiskers extend to farthest point within 1.5-fold of the interquartile range, with outliers shown as points.

Interactions between RNA and Protein Levels. We conducted a more detailed ANOVA (27), where both RNA and protein levels are independently determined by both promoter and RBS identity. This model is able to take into account effects such as the dependency of RNA levels on the translation rate. We found that the model resulted in a modestly better fit (RNA R^2 = 0.96, protein R^2 = 0.82; Fig. S10 *C* and *D*). Analysis of explained variance showed that 92% of the RNA levels can be explained by the promoter choice, whereas only 4% can be explained by the RBS choice and the remaining 4% are unexplained (Fig. 5*A*). For protein levels, both promoter choice (54% explained variation) and RBS choice (30%) are important, but a larger portion remains unexplained (16.7%).

To understand better how factors such as RBS choice can affect RNA levels, we examined interactions between RNA and protein levels. For example, several previous studies in *Escherichia coli* and *Bacillus subtilis* have shown that for particular model transcripts, increased ribosome binding or occupancy may enhance mRNA stability (36–42). Such studies have been hard to interpret due to the complex interactions between the ribosome, RNA degradation machinery, and transcript. We indeed find a significant and prevalent correlation between mRNA stability and RBS strength across all promoters. Given the size and sequence diversity of our library, it is likely that RBS strength is responsible for increased mRNA levels. Overall, we find that an ∼10-fold increase in translation efficiency correlates to an approximately threefold increase in RNA abundance (Fig. 5*B*). However, the effect is limited at the extremes; the difference between the weakest and strongest RBSs (an 87-fold increase in translation efficiency) corresponds to only an ∼4.3-fold increase in mRNA. As another example, many groups have found that secondary structure across the 5′ UTR and initial coding sequence can hinder effective translation (14, 43–46). In our data, we find that the correlation between secondary structure free energy across the UTR/GFP interface is significant (Fig. 5*C*). However, this metric of secondary structure is neither necessary nor sufficient, because many sequences with high secondary structure do not display reductions in expected expression, and vice versa. Improved models for how secondary structure interacts with ribosome binding could increase this correlation (14).

Discussion

We developed a method to characterize transcription and translation rates of thousands of synthetic regulatory elements simultaneously. We used this method to characterize the extent to which promoters and RBSs can be independently composed. This large RBS-promoter pair library can be used to titrate recombinant protein expression in *E. coli*, and the expression data can be used to refine models of how sequence composition determines levels of gene expression.

We do not examine how expression is altered by a gene's amino acid composition and codon use, which are known to have large effects (26, 43–46). In follow-up work, we explore the influence of these two factors across a matrix of coding sequences, promoters, and RBSs. Another limitation of our current approach is that we do not examine how expression affects cellular growth rate. Highly expressed constructs might impair the growth rate and decrease steady-state dilution of cellular contents, which would lead to an overestimation of transcription and translation strengths. We analyze only promoter and RBS pairings here, but future studies can test large numbers of any composable genetic designs to assess their effectiveness on a broad basis (26, 28).

The methods developed here should be extendable to any organism that is amenable to fluorescence-activated cell sorting and RNASeq, such as other bacteria, yeast, and mammalian cell lines. In addition, our methods can used to optimize more complex phenomena, including inducible expression, gene circuits, and time-dependent responses. Finally, improvements in the quality and length of synthetic oligo pools can also extend such analyses to the characterization of regulatory protein variants or longer range interactions.

Materials and Methods

Strains, Library Construction, and Growth Conditions. We used *E. coli* MG1655 (Yale Coli Genetic Stock Center no. 6300) for all experiments. The oligo library was constructed by Agilent Technologies using the OLS process (30). The design of pGERC is based on the synthetic plasmid pZS-123 (33), which allows independent expression from three promoters, and it was synthesized by DNA2.0, Inc. The amplified OLS pool was subcloned into 5α-electrocompetent *E. coli* (New England Biolabs) (giving an initial library size of ∼600,000 colonies), purified, and retransformed into MG1655, and several aliquots were frozen. Overnight cultures from both pooled experiments and individual clones were first diluted 1,000-fold grown at 30 °C in LB–Miller media shaking at 250 rpm (Infors HT Multitron) for 2–3 h until reaching an OD_{600} of 0.15–0.25. Detailed information can be found in *SI Materials and Methods*.

DNASeq and RNASeq. From a single 300-mL culture of the library, pellets from four 50-mL aliquots of culture were frozen in liquid nitrogen, with the remaining culture saved for FlowSeq. Two technical replicates of DNA and RNA were isolated using Qiagen DNA and RNA Midiprep Kits. Ribosomal RNA was removed by means of a Ribo-Zero rRNA removal kit for metabacteria (Epicentre). The 5′ triphosphates were monophosphorylated by 5′ polyphosphatase (Epicentre) and then ligated to an RNA adaptor using T4

RNA Ligase (Epicentre). First-strand cDNA was made from a specific primer in superfolder GFP. Both DNA and cDNA were amplified and monitored by real-time PCR to prevent overamplification. Illumina adaptors and barcodes were then added, and sequencing was performed on a HiSeq 2000 in two separate PE100 lanes. A separate library that contained spike-ins from the 42 colonies underwent the same procedure. Detailed information can be found in *SI Materials and Methods*.

FlowSeq. We used 50 mL of the library culture as prepared above for analysis by FlowSeq. We flow-sorted the cells into 12 log-spaced bins in three sequential runs sorting 4 bins each. Cells were then grown overnight to saturation and plasmid-prepped using a Qiagen Miniprep kit. A small aliquot was diluted, regrown, and subjected to flow cytometry to verify proper sorting. All data from library measurements are reported in GFP/RFP ratio units, which range from 1 to 255,000. The 12 minipreps were amplified again by real-time PCR, barcoded, and sequenced on a single-lane PE100 using a HiSeq 2000. Detailed information can be found in *SI Materials and Methods*.

Data Analysis. Reads from all experiments were first aligned using SeqPrep (47) to form paired-end contigs for improved accuracy. Custom software was written to identify unique contigs and map them to library members using Bowtie (48) and grep (global search with the regular expression and printing all matching lines). DNASeq and RNASeq contigs were counted, where reads mapped uniquely and contained less than three mismatches. In addition, DNA contamination from RNASeq reads was identified and removed. Statistics, graphs, and tables were all generated using custom software written in Python, R, and the ggplot2 package (49). Detailed information can be found in *SI Materials and Methods*. All values used in intermediate and final calculations are enumerated in Dataset S3.

ACKNOWLEDGMENTS. This work was supported by US Department of Energy Grant DE-FG02-02ER63445 (to G.M.C.), National Science Foundation (NSF) Synthetic Biology Engineering Research Center Grant SA5283-11210 (to G.M.C.), and Office of Naval Research Grant N000141010144 (to G.M.C. and S.K.), as well as by Agilent Technologies and Wyss Institute. D.B.G. is supported by an NSF Graduate Research Fellowship.

1. Organization for Economic Cooperation and Development (2009) *The Bioeconomy to 2030: Designing a Policy Agenda* (OECD Publishing, Paris).
2. Carlson R (2007) Laying the foundations for a bio-economy. *Syst Synth Biol* 1(3): 109–117.
3. Keasling JD (2010) Manufacturing molecules through metabolic engineering. *Science* 330(6009):1355–1358.
4. Wang HH, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460(7257):894–898.
5. Carr PA, Church GM (2009) Genome engineering. *Nat Biotechnol* 27(12):1151–1162.
6. Temme K, Zhao D, Voigt CA (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc Natl Acad Sci USA* 109(18):7085–7090.
7. Tabor JJ, et al. (2009) A synthetic genetic edge detection program. *Cell* 137(7): 1272–1281.
8. Bonnet J, Subsoontorn P, Endy D (2012) Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc Natl Acad Sci USA* 109(23): 8884–8889.
9. Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat Biotechnol* 21(7):796–802.
10. Ro D-K, et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440(7086):940–943.
11. Steen EJ, et al. (2010) Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 463(7280):559–562.
12. Patwardhan RP, et al. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* 27(12):1173–1175.
13. Kinney JB, Murugan A, Callan CG, Jr., Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA* 107(20):9158–9163.
14. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27(10):946–950.
15. Barrick D, et al. (1994) Quantitative analysis of ribosome binding sites in E. coli. *Nucleic Acids Res* 22(7):1287–1295.
16. Na D, Lee S, Lee D (2010) Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst Biol* 4:71.
17. Andrianantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: New engineering rules for an emerging discipline. *Mol Syst Biol* 2:2006.0028.
18. Arkin A (2008) Setting the standard in synthetic biology. *Nat Biotechnol* 26(7): 771–774.
19. Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet* 6(7):533–543.
20. Canton B, Labno A, Endy D (2008) Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol* 26(7):787–793.
21. Endy D (2005) Foundations for engineering biology. *Nature* 438(7067):449–453.
22. Heinemann M, Panke S (2006) Synthetic biology—Putting engineering into biology. *Bioinformatics* 22(22):2790–2799.
23. Serrano L (2007) Synthetic biology: Promises and challenges. *Mol Syst Biol* 3:158.
24. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2007) Accurate prediction of gene feedback circuit behavior from component properties. *Mol Syst Biol* 3:143.
25. Alper H, Fischer C, Nevoigt E, Stephanopoulos G (2005) Tuning genetic control through promoter engineering. *Proc Natl Acad Sci USA* 102(36):12678–12683.
26. Mutalik VK, et al. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods* 10(4):354–360.
27. Mutalik VK, et al. (2013) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat Methods* 10(4):347–353.
28. Qi L, Haurwitz RE, Shao W, Doudna JA, Arkin AP (2012) RNA processing enables predictable programming of gene expression. *Nat Biotechnol* 30(10):1002–1006.
29. Kittleson JT, Wu GC, Anderson JC (2012) Successes and failures in modular genetic engineering. *Curr Opin Chem Biol* 16(3-4):329–336.
30. LeProust EM, et al. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res* 38(8): 2522–2540.
31. Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* 24(1): 79–88.
32. Shu X, Shaner NC, Yarbrough CA, Tsien RY, Remington SJ (2006) Novel chromophores and buried charges control color in mFruits. *Biochemistry* 45(32):9639–9647.
33. Cox RS, 3rd, Dunlop MJ, Elowitz MB (2010) A synthetic three-color scaffold for monitoring genetic regulation and noise. *J Biol Eng* 4:10.
34. Raveh-Sadka T, et al. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* 44(7):743–750.
35. Sharon E, et al. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* 30(6): 521–530.
36. Yarchuk O, Jacques N, Guillerez J, Dreyfus M (1992) Interdependence of translation, transcription and mRNA degradation in the lacZ gene. *J Mol Biol* 226(3):581–596.
37. Jain C, Kleckner N (1993) IS10 mRNA stability and steady state levels in Escherichia coli: Indirect effects of translation and role of rne function. *Mol Microbiol* 9(2): 233–247.
38. Sharp JS, Bechhofer DH (2003) Effect of translational signals on mRNA decay in Bacillus subtilis. *J Bacteriol* 185(18):5372–5379.
39. Hambraeus G, Karhumaa K, Rutberg B (2002) A 5′ stem-loop and ribosome binding but not translation are important for the stability of Bacillus subtilis aprE leader mRNA. *Microbiology* 148(Pt 6):1795–1803.
40. Jürgen B, Schweder T, Hecker M (1998) The stability of mRNA from the gsiB gene of Bacillus subtilis is dependent on the presence of a strong ribosome binding site. *Mol Gen Genet* 258(5):538–545.
41. Wagner LA, Gesteland RF, Dayhuff TJ, Weiss RB (1994) An efficient Shine-Dalgarno sequence but not translation is necessary for lacZ mRNA stability in Escherichia coli. *J Bacteriol* 176(6):1683–1688.
42. Arnold TE, Yu J, Belasco JG (1998) mRNA stabilization by the ompA 5′ untranslated region: Two protective elements hinder distinct pathways for mRNA degradation. *RNA* 4(3):319–330.
43. Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLOS Comput Biol* 6(2): e1000664.
44. Allert M, Cox JC, Hellinga HW (2010) Multifactorial determinants of protein expression in prokaryotic open reading frames. *J Mol Biol* 402(5):905–918.
45. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324(5924):255–258.
46. Welch M, et al. (2009) Design parameters to control synthetic gene expression in Escherichia coli. *PLoS ONE* 4(9):e7002.
47. St. John J (2012) SeqPrep. Available at https://github.com/jstjohn/SeqPrep. Accessed February 1, 2013.
48. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
49. Wickham H (2009) *ggplot2* (Springer, New York).