Supporting Information

Kosuri et al. 10.1073/pnas.1301301110

SI Materials and Methods

Reporter Construction. The gene expression reporter construct (pGERC) used in all experiments follows the design of the pZS2-123 plasmid that drives independent expression of three fluorescent proteins from Cox et al. (1). Briefly, we began with the divergent promoter portion of pZS2-123, which has insulated sequences to express CFP with $P_{LtetO-1}$ and YFP with $P_{LlacO-1}$. We replaced the CFP with a codon-optimized version of mCherry (2) and replaced the YFP with a codon-optimized version of superfolder GFP (sfGFP) (3). We replaced $P_{LlacO-1}$ with the EM7 promoter to avoid issues of endogenous regulation by the Lac repressor in MG1655. We also removed an AscI recognition site in the intergenic space and placed an AscI recognition site directly upstream of the EM7 promoter and an NdeI recognition site at the start of the sfGFP sequence. These sites are used for cloning library components upstream of the sfGFP sequence. The whole construct is flanked by XhoI and NotI on the left and by PacI and XbaI on the right, and it was constructed by DNA2.0, Inc. in a pJ251 backbone, which has a low copy number p15A origin of replication and a kanamycin resistance marker.

Library Design, Construction, and Cloning. The library was constructed by combining 114 promoter sequences with 111 RBS sequences. Promoter sequences were chosen from existing libraries, such as the BIOFAB: International Open Facility Advancing Biotechnology (4), a few control promoters (including an inactive spacer), and a set of promoters from Chris Anderson's promoter library from the BioBricks registry (5). We added a five-base barcode and then checked for restriction site compatibility (AscI and NdeI) to generate the final promoter library. The ribosome binding site (RBS) library contains RBSs from BIOFAB (4), control RBSs, Chris Anderson's RBS library from the BioBricks registry (6), and sequences generated by the Salis RBS Calculator (7). The promoters and RBSs were filtered for restriction sites and to ensure that all pairwise Levenshtein distances are greater than 1. In addition, all RBSs have bases "CAT" replacing the terminal three bases before the coding sequence to allow for cloning using the NdeI site for a total of 111 RBSs. Finally, each promoter is crossed by all RBSs to form a final library of 12,653 promoter + RBS combinations. One combination was removed because the junction resulted in a disallowed restriction site. All constructs were flanked by restriction enzyme sites (AscI and NdeI) and the following PCR primer binding sites: skpp-202-F AATCCTTGCGTCAATGGTTC and skpp-202-R GGGTTCTCGGATTTTACACG.

The oligo library was constructed by Agilent Technologies using their oligo library synthesis process (8), and it was delivered as an ~1-pmol lyophilized oligo pool. The library was amplified from the oligo pool using biotinylated primers, digested with AscI and NdeI (New England Biolabs), and the resulting ends were removed by Invitrogen M-270 streptavidin beads. The plasmid backbone was also amplified by PCR using biotinylated primers, digested with the same restriction enzymes, and cleaned again by streptavidin beads. We then ligated the library and plasmid backbone using T4 DNA Ligase (New England Biolabs) and cloned into 5-alpha electrocompetent cells (New England Biolabs), resulting in ~600,000 clones. The library was grown under kanamycin selection, and plasmids were isolated using a Qiagen Miniprep kit. The plasmid library was retransformed into *Escherichia coli* MG1655 (Yale Coli Genetic Stock Center no. 6300) (>3 million clones). We froze several aliquots of this library and used these aliquots for all subsequent experiments.

Control Colonies and Flow Cytometry. We plated the transformed MG1655 library and Sanger-sequenced 282 clones. One hundred fifty-four (55%) of 282 of these clones matched the designed sequences exactly. One hundred forty-four sequence-perfect clones (2 clones were duplicates) were inoculated from glycerol stocks into 200 μL of LB with kanamycin and grown overnight at 30 °C with shaking in 96-well culture plates. The cells were then back-diluted at a ratio of 1:1,000 into 200 μL of LB with kanamycin and grown for 3.5 h, until the cells reached an OD_{600} of ~0.15–0.25. The cells were then immediately put on ice, pelleted by centrifugation, and diluted 1,000-fold in ice-cold PBS. We measured RFP and GFP fluorescence levels using a Becton Dickinson FACS LSRFortessa flow cytometer with a high-throughput sampling attachment (30,000 events per observation). Events were gated on forward and side scatter to exclude debris, dead cells, and doublets. The overnight growth, back-dilution, and flow cytometry procedure were performed four times from different back-dilutions on 2 separate days.

Library Growth and FlowSeq. A 300-mL culture was inoculated with 1 mL of library culture grown overnight at 30 °C from a frozen aliquot. The culture was grown for 3.5 h to an OD_{600} of 0.2 at 30 °C and shaking at 250 rpm (Infors HT Multitron). The culture was quickly brought to 4 °C in an ice slurry. Five 50-mL aliquots were pelleted. Four were snap-frozen in liquid nitrogen, and one was resuspended in 50 mL of ice-cold PBS. The library in PBS was directly subjected to FlowSeq. We conducted three consecutive flow sorts on a Becton Dickinson FACSAria IIu while keeping cells at 4 °C. Each run sorted four nonadjacent log-spaced bins based on the GFP/RFP ratio. We sorted 1 million cells for the first bin (lowest ratio) because it had the most cells in it. For all other bins, we sorted 250,000 cells, except for the last two bins, where we sorted 100,000 cells each. Cells were grown overnight with shaking at room temperature to minimize growth rate differences, and plasmids were isolated using a Qiagen Miniprep kit. Each bin was separately amplified for five cycles by RT-PCR to prevent overamplification using Kapa SybrFast RT-PCR master mix. The reverse primer was an equimolar mixture of five separate sequences to allow frame shifting so as to give better sequence distributions during read 2 of sequencing:

FlowSeq-F: AATGATACGGCGACCACCGAGATCTACACTGAAGCACAGCAGCTCTTCGCCTTTACGCATATG

FlowSeq-R0: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACAATGAAAAGCTTAGTCATGGCG

FlowSeq-R1: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGACAATGAAAAGCTTAGTCATGGCG

FlowSeq-R2: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGACAATGAAAAGCTTAGTCATGGCG

FlowSeq-R3: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCATGACAATGAAAAGCTTAGTCATGGCG

FlowSeq-R4: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCATGACAATGAAAAGCTTAGTCATGGCG

FlowSeq-R5: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCATTGACAATGAAAAGCTTAGTCATGGCG

A final RT-PCR step added barcodes to each binned construct using the following primers:

FlowSeq-F: AATGATACGGCGACCACCGAGATCTACACTGAAGCACAGCAGCTCTTCGCCTTTACGCATATG

Bin 1 FlowSeq-R-index_6nt_1: CAAGCAGAAGACGGCATACGAGATtcaggtGTGACTGGAGTTCAGACGTGT

Bin 2 FlowSeq-R-index_6nt_2: CAAGCAGAAGACGGCATACGAGATaagcgtGTGACTGGAGTTCAGACGTGT

Bin 3 FlowSeq-R-index_6nt_3: CAAGCAGAAGACGGCATACGAGATgtcgatGTGACTGGAGTTCAGACGTGT

Bin 4 FlowSeq-R-index_6nt_4: CAAGCAGAAGACGGCATACGAGATgccttgGTGACTGGAGTTCAGACGTGT

Bin 5 FlowSeq-R-index_6nt_7: CAAGCAGAAGACGGCATACGAGATggtaagGTGACTGGAGTTCAGACGTGT

Bin 6 FlowSeq-R-index_6nt_9: CAAGCAGAAGACGGCATACGAGATgattgcGTGACTGGAGTTCAGACGTGT

Bin 7 FlowSeq-R-index_6nt_11: CAAGCAGAAGACGGCATACGAGATcggtccGTGACTGGAGTTCAGACGTGT

Bin 8 FlowSeq-R-index_6nt_13: CAAGCAGAAGACGGCATACGAGATgcaaccGTGACTGGAGTTCAGACGTGT

Bin 9 FlowSeq-R-index_6nt_15: CAAGCAGAAGACGGCATACGAGATatgaacGTGACTGGAGTTCAGACGTGT

Bin 10 FlowSeq-R-index_6nt_16: CAAGCAGAAGACGGCATACGAGATcttataGTGACTGGAGTTCAGACGTGT

Bin 11 FlowSeq-R-index_6nt_17: CAAGCAGAAGACGGCATACGAGATagcagaGTGACTGGAGTTCAGACGTGT

Bin 12 FlowSeq-R-index_6nt_20: CAAGCAGAAGACGGCATACGAGATcaataaGTGACTGGAGTTCAGACGTGT

The amplified bins were quantitated using the Kapa Library Quantification Kit and mixed in equimolar ratios before sequencing all 12 on a single HiSeq 2000 paired-end 100-bp lane with the following sequencing primers:

Custom Read 1: 5′ GAAGCACAGCAGCTCTTCGCCTTTACGCATATG

Illumina Multiplexing Read 2: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Illumina Multiplexing Index Read: GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Spike-In Controls. A separate library underwent the same procedure. Before back-dilution, we spiked in a subset of 42 of the perfect sequences and performed all procedures, including DNASeq, RNASeq, and FlowSeq, identically.

DNASeq and RNASeq. For DNASeq, we isolated plasmids using a Qiagen Midiprep kit from two frozen cell pellets from the 50-mL library growth culture. We amplified the library as we did in the FlowSeq experiment, using only primers FlowSeq-R-index_6nt_1 and FlowSeq-R-index_6nt_4. We also processed the spike-in libraries similarly, but we used FlowSeq-R-index_6nt_15 and FlowSeq-R-index_6nt_16. All four DNASeq libraries were run on a single lane in the same HiSeq run as the FlowSeq data.

For RNASeq, we used the remaining two cell pellets, first isolating total RNA using a Qiagen RNEasy Midi Kit and then removing ribosomal RNA using an Epicentre Ribo-Zero rRNA Magnetic Removal Kit for Meta-Bacteria according to the manufacturer's instructions. We then used 250 ng of mRNA and removed the 5′ triphosphate group with RNA 5′ Polyphosphatase (Epicentre) as follows:

50 μL of RNA (250 ng)

6 μL of RNA polyphosphatase 10× reaction buffer

1.5 μL of RiboGuard RNase Inhibitor (Epicentre)

3 μL of RNA 5′ Polyphosphatase (60 units)

37 °C for 30 min

The resulting reaction was cleaned up using a Qiagen RNAEasy MinElute Kit. We then ligated the following RNA adaptor to the processed mRNA/RNA ligation primer: GACAAUGAAAAGCUUAGUCAUGGCGNN.

The two trailing N's indicate degenerate bases that are used to reduce biases found in RNA ligation efficiency across different templates (9). We used the following procedure for ligation using T4 RNA Ligase (Epicentre):

10 μL of RNA from the previous step

2 μL of 250 μM RNA oligo

2 μL of 10× ligase buffer

2 μL of 10 units of T4 RNA Ligase

2 μL of 10 mM ATP

1 μL of RiboGuard RNase Inhibitor (Epicentre)

1 μL of DMSO

25 °C for 3 h

The resulting reaction was cleaned up, again using a Qiagen RNAEasy MinElute Kit. To make cDNA, we used Invitrogen's SuperScript III with the following procedure:

i) We added the following components to a nuclease-free microcentrifuge tube:

0.2 μL of 10 μM (2 pmol) reverse transcriptase primer: ACCGTTGACATCACCATCCAGTTCC

12 μL of RNA from RNA ligation reaction

1 μL of 10 mM dNTP mix

ii) We heated the mixture to 65 °C for 5 min and incubated it on ice for >1 min.

iii) We collected the contents of the tube by brief centrifugation and added:

4 μL of 5× first-strand buffer

1 μL of 0.1 M DTT

1 μL of RNaseOUT Recombinant RNase Inhibitor (40 U/μL; Invitrogen)

1 μL of SuperScript III reverse transcriptase (200 U/μL; Invitrogen)

iv) We mixed the contents of the tube by gentle up and down pipetting.

v) We incubated the contents of the tube at 55 °C for 60 min.

vi) We inactivated the reaction by heating at 70 °C for 15 min.

vii) We added 1 μL (2 units) of *E. coli* RNase H and incubated the contents of the tube at 37 °C for 20 min.

The resulting cDNA was amplified using the same procedures as DNASeq and FlowSeq, and using the same barcodes for technical and spike-in replicates on a separate lane of the same HiSeq 2000 run.

Data Analysis: Contig Formation and Trimming. We used a modified version of SeqPrep (10) and custom Python scripts to pair and trim reads into contigs with increased sequencing fidelity for regions of paired-end coverage. Each set of two paired-end 100-bp reads was aligned and merged into a contig based on its overlapping sequence. The adapter and constant primer sequences were trimmed from both ends of the contig. If only a portion of the adapter sequence was identifiable, the cloning restriction sites were used to identify the region for trimming. Reads that did not pair were discarded, because all sequences are under 200 bp; thus, contigs, should be created where the two paired reads overlap. Additionally, the first two bases of RNA contigs were trimmed, corresponding to the two degenerate ligated bases used in the experimental protocols.

Deduplication and Sorting of Unique Contigs to Library. After trimming, occurrences of each unique contig were counted per bin and merged to generate a vector of 12 numbers corresponding to the occurrences per bin per contig. These unique contigs were then aligned to the promoter + RBS sequence library. In the case of the protein data, grep (global search with the regular expression and printing all matching lines) and USEARCH 5.2.32 were used. We aligned all unique contigs but used the intersection of three criteria to filter for downstream analysis. Contigs were required to: (*i*) be perfect end-to-end matches to the library; (*ii*) consist of at least 100 occurrences; and (*iii*) occur in multiple bins, excepting the final bin. In the case of DNASeq and RNASeq data, Bowtie (11) was used. We filtered matching contigs on three criteria: (*i*) contigs were allowed no more than three mismatches; (*ii*) contigs were required to match best to only one library combination; and (*iii*) to remove DNA contamination, contigs were required to begin at least two bases into a library combination and match up until the very end of the RBS (corresponding to the start codon).

Protein Level Calculation. To calculate protein expression levels for each construct, we first normalized the counts from each bin to one another using the total fraction of cells in the library that fell into each particular bin. We defined the fraction of cells sorted in each bin as f_j, so that $\sum_j f_j = 1$, and the number of occurrences of sequence i in each bin j as c_{ij}. Then, normalized fractional contribution of each bin j per sequence i, a_{ij}, is calculated as:

$$a_{ij} = \frac{f_j \cdot c_{ij}}{\sum_i c_{ij}} \Bigg/ \sum_j \frac{f_j \cdot c_{ij}}{\sum_i c_{ij}},$$

so that $\sum_j a_{ij} = 1$.

Once the compensated bin distributions were calculated, we used the median fluorescence level in each bin as the value for all observations in that bin. We defined the center of the measurement range for each sorted bin j as m_j. The protein level, p_i, was then calculated as:

$$p_i = \exp\left[\sum_j a_{ij} \cdot \log(m_j)\right].$$

FlowSeq Minimum and Maximum Cutoffs. Due to the placement of the bin cutoffs during sorting, there were upper and lower boundaries on the linear measurement range for protein level. These thresholds were empirically determined to be twofold the minimum protein level and 99% of the maximum protein level (noted with a dotted line in Fig. S10). In total, 14.3% of constructs were below this range and 6.5% were above. These out-of-range data were not used to calculate ordering or average strength of promoters and RBSs, although we do display them as measured in Figs. 2 and 3.

Calculation of Transcription Start Sites. Using the RNA contigs aligned to the library, we determined the transcription start site (TSS) for each promoter. After filtering RNA contigs as described above, the TSS for each unique sequence was determined, relative to the RBS + promoter junction. In most cases, RNA contigs could be assigned uniquely to an RBS + promoter pair because of the unique barcode appended to the end of every promoter sequence. To calculate a single TSS per promoter, the alignment offset of each RNA contig against its DNA sequence was recorded. Eighty-seven percent of all promoters had one dominant start position (>60% of all mapped contigs). The most prevalent start site was used to calculate the RNA secondary structure as described below. Two promoters (marked with an asterisk in Fig. S6) had very few uniquely mapping contigs, did not show a strong start site, and showed unrealistic translation efficiency calculations. These observations indicated that we were missing most of the RNA data (but not protein data) from these promoters because of transcription starting after the end of the barcode sequence. The 222 constructs (1.7%) containing these promoters were removed from all analyses.

RNA Level Calculation. RNA levels were calculated separately for each technical replicate, using a ratio of normalized RNA to normalized DNA:

$$RNA_i = \frac{c_{i,RNA}}{\sum c_{RNA}} \Bigg/ \frac{c_{i,DNA}}{\sum c_{DNA}},$$

where i is each individual construct; c_i is the number of DNA or RNA contigs for construct i; and $\sum c_{RNA}$ and $\sum c_{DNA}$ are the total number of sequenced and merged RNA and DNA contigs, respectively, before filtering. The RNA levels across the replicates showed a high level of correlation ($R^2 = 0.992$) and were averaged.

Filtering of RNASeq and DNASeq Data. RNA and DNA data were adjusted or discarded from some constructs based on low contig counts. One hundred eighty-four constructs (1.4%) did not have at least 10 DNA contig counts in both replicates and were discarded. Seven additional constructs (0.7%) had fewer than 20 RNA contig counts and also had fewer than 50 DNA contig counts, and they were also discarded. Two hundred seventy-five constructs (2.2%) had sufficient DNA but insufficient RNA contig counts; thus, their RNA contig counts were set to 10 (separately for each technical replicate) for purposes of RNA level calculation as described above.

Calculation of Average Transcription and Translation Levels. Average transcription and translation levels were calculated for all promoters and RBSs, respectively. To calculate the average promoter transcription level, the geometric mean of the RNA level was calculated across each promoter, excluding constructs with insufficient RNASeq/DNASeq contig counts as described above. To calculate the average RBS translation level, the translation efficiency was first calculated per construct as the ratio of protein level to RNA level. The average translation level for each RBS was then calculated as the geometric mean of this translation efficiency. Constructs with protein levels above and below the aforementioned minimum and maximum thresholds were excluded from this calculation, as were constructs with insufficient RNASeq/DNASeq contig counts.

Element Ordering. Because we did not want missing constructs with strongly expressing promoter and RBS elements to influence the element ordering, we used the average deviation from mean values across all elements for ranking purposes.

The naming and ordering of each promoter were determined as:

$$o_p = \frac{1}{n_r}\sum_r\left[\ln\left(RNA_{p,r}\right) - \frac{1}{n_p}\sum_p \ln\left(RNA_{p,r}\right)\right],$$

where n_r and n_p are the number of RBS and promoter elements, respectively, and $RNA_{p,r}$ is the RNA level for a promoter/RBS combination. This ranks the promoters by how much each promoter/RBS construct deviates from the average RNA level across all RBSs. Promoters were sorted and named Ec-TTL-P# (*E. coli* transcription/translation library promoter no.), from 001 to n_p based on their rank-ordered o_p value.

RBSs were ordered similarly, with the equation:

$$o_r = \frac{1}{n_p}\sum_p\left[\ln\left(PROT_{p,r}\right) - \frac{1}{n_r}\sum_r \ln\left(PROT_{p,r}\right)\right],$$

where $PROT_{p,r}$ is the protein level for a promoter/RBS combination. This ranks the RBSs by how much each promoter/RBS construct deviates from the average protein level across all promoters. Individual RBS were ordered from 001 to n_r based on o_r, as Ec-TTL-R# (*E. coli* transcription/translation library RBS no.).

Calculation of Secondary Structure. The 5′ UTRs used for secondary structure free energy determination were taken from the start of the dominant TSS site to 30 bases into the coding sequence of sfGFP. Free energy of 5′ UTR regions was calculated using UNAFOLD's (12) "hybrid-ss-min – NA = RNA" command line program with default parameterizations.

Simple Model of Transcription and Translation Based on Mean Element Strengths. To create a simple prediction for protein level, we took the product of the mean transcription per promoter and the mean normalized translation (i.e., translation efficiency) per RBS:

$$\ln\left(\widehat{TRANSCRIPTION_p}\right) = \frac{1}{n_r}\sum_r \ln\left(RNA_{p,r}\right),$$

$$\ln\left(\widehat{TRANSLATION_r}\right) = \frac{1}{n_p}\sum_p \ln\left(\frac{PROT_{p,r}}{RNA_{p,r}}\right),$$

$$\widehat{PROT_{p,r}} = \exp\left[\ln\left(\widehat{TRANSCRIPTION_p}\right) + \ln\left(\widehat{TRANSLATION}\right)_r\right].$$

In the transcription calculations, we removed constructs that had insufficient RNA or DNA contig counts, as described above. In the translation calculations, we removed all constructs that did not match the previously mentioned RNA, DNA, and protein level filters, as well as constructs that were above and below the protein level linear range described above.

Linear Modeling (ANOVA). We also constructed a linear model to determine the contribution of promoter and RBS to both protein level and RNA expression level:

$$\log\left(PROT_{p,r}\right) = \alpha + P_p + R_r,$$

$$\log\left(RNA_{p,r}\right) = \alpha + P_p + R_r,$$

where α is the average signal, P_p is the pth promoter, and R_r is the rth RBS.

Using this linear model framework, we use a random effects model (type II) ANOVA to calculate the relative contribution of each component to the explained sum of squares for each output (Fig. 4*B*).

Statistical Analysis Software. All statistics and tables described above were generated using custom software written in Python and R. Graphs were generated using the ggplot2 package (13) in R.

1. Cox RS, 3rd, Dunlop MJ, Elowitz MB (2010) A synthetic three-color scaffold for monitoring genetic regulation and noise. *J Biol Eng* 4:10.
2. 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.
3. 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.
4. Endy D, Arkin AP, Keasling JD (2013) BIOFAB. Available at www.biofab.org/. Accessed February 1, 2013.
5. Anderson JC (2006) Anderson Promoter Library Registry of Standard Biological Parts. Available at http://partsregistry.org/Promoters/Catalog/Anderson. Accessed February 1, 2013.
6. Anderson JC (2007) Anderson RBS Library, Registry of Standard Biological Parts. Available at: http://partsregistry.org/Ribosome_Binding_Sites/Prokaryotic/Constitutive/Anderson. Accessed February 1, 2013.
7. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 27(10):946–950.
8. 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.
9. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* 39(21):e141.
10. St. John J (2012) SeqPrep. Available at https://github.com/jstjohn/SeqPrep. Accessed February 1, 2013.
11. 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.
12. Markham NR, Zuker M (2008) UNAFold: Software for nucleic acid folding and hybridization. *Methods Mol Biol* 453:3–31.
13. Wickham H (2009) *ggplot2* (Springer, New York).

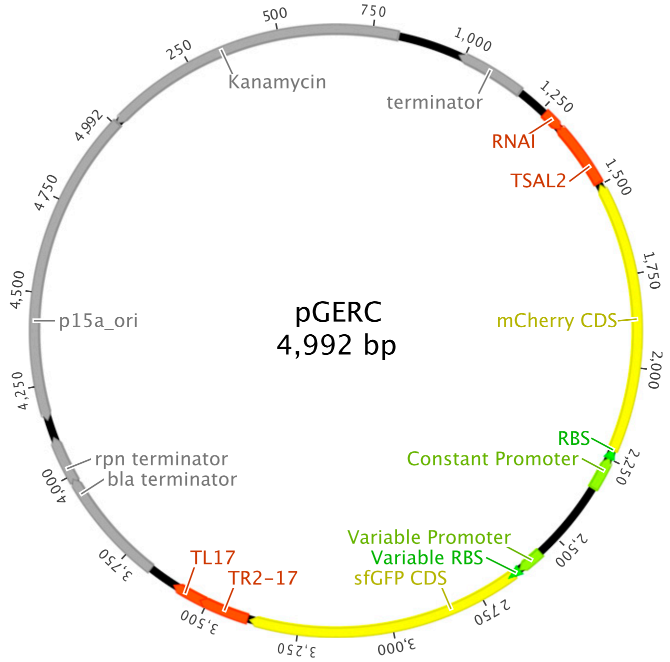

Fig. S1. Plasmid map of pGERC. A plasmid map shows the sequence of pGERC (based on pZS-123), including the plasmid backbone (gray) with the kanamycin resistance cassette, origin of replication, and terminators. The two fluorescent protein coding DNA sequence (CDS) regions are shown in yellow, whereas promoter and RBS regions are shown in green. Terminators for the fluorescent protein coding regions are shown in red.

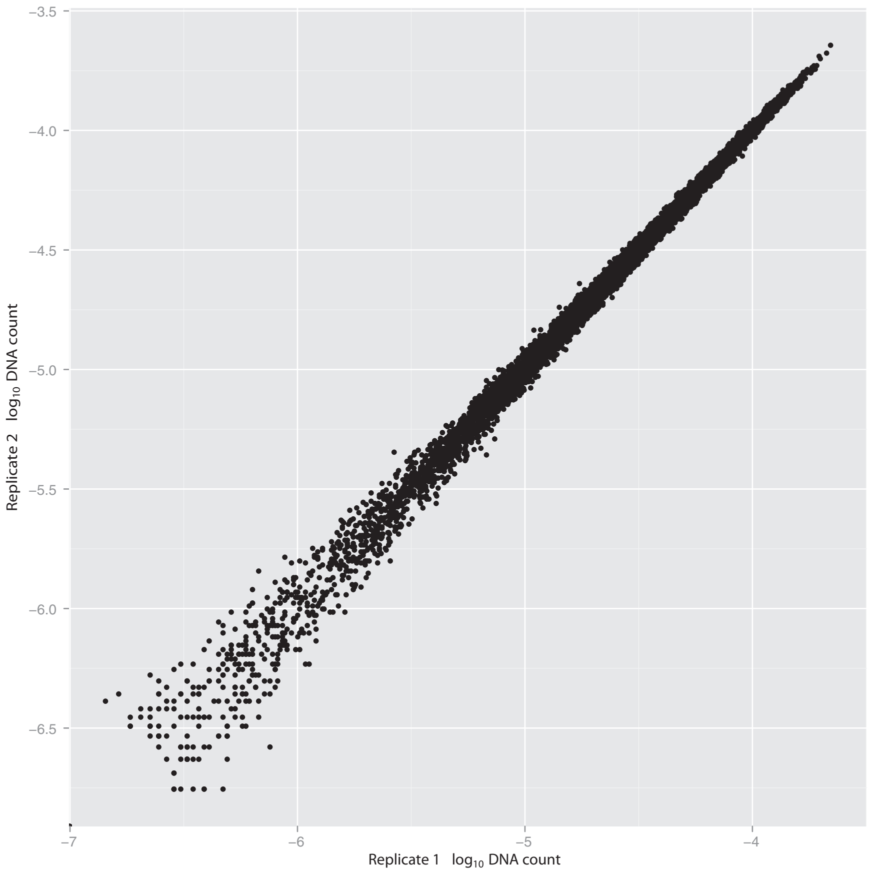

Fig. S2. DNA technical replicates 1 and 2. Observation frequencies of library members across two technical replicates of DNA isolation, amplification, and sequencing are plotted against one another. The R^2 value of the linear model is 0.997 (*F* test, *P* value <2.2e-16).

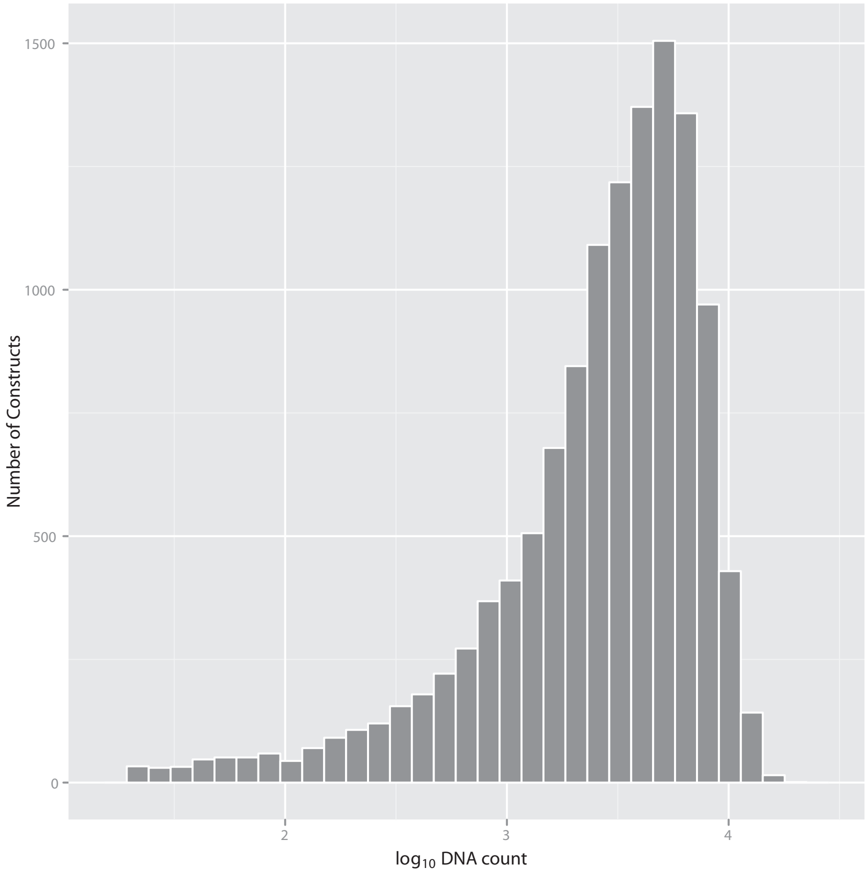

Fig. S3. Distribution of contig counts for observed members of the library. Library members with five or more counts across both replicates are binned and plotted on the histogram. One hundred eighty-three constructs were below the threshold and not plotted.

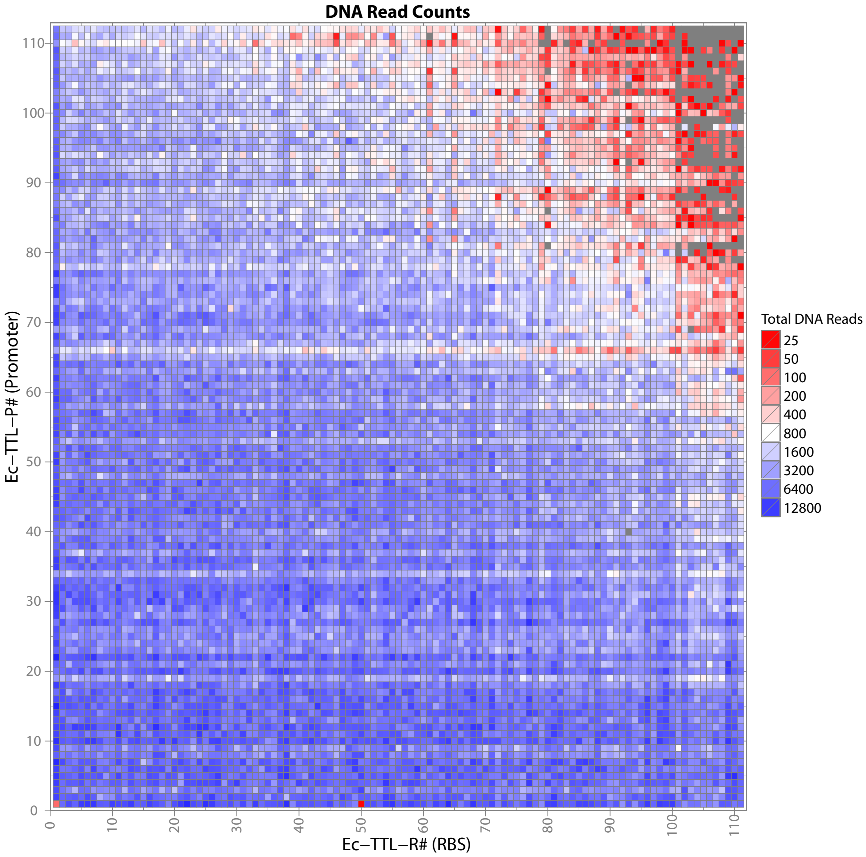

Fig. S4. Distribution of DNA contigs by construct. Contig counts are displayed by color for each construct. Constructs are labeled by promoter (*y* axis) and RBS (*x* axis) and ordered as in Figs. 2 and 3. Dark gray boxes are unobserved contigs, as well as one combination (040P-093R) that was not synthesized due to restriction site incompatibility. Most constructs with few contigs contain combinations of strong promoters and RBSs, potentially indicating that the high level of gene expression from these constructs affects growth and viability.

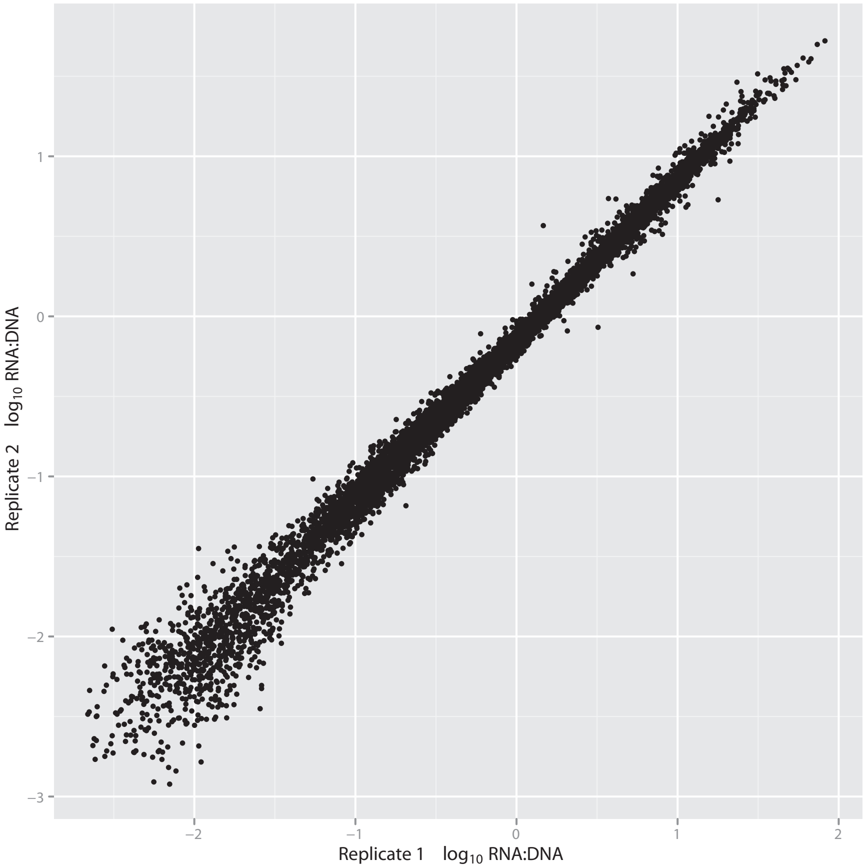

Fig. S5. RNASeq/DNASeq ratio calculated separately for each technical replicate. The RNA levels, as measured by the RNASeq/DNASeq ratios, are plotted for two technical replicates and showed a high degree of concordance (R^2 = 0.99; *F* test, *P* value <2.2e-16).

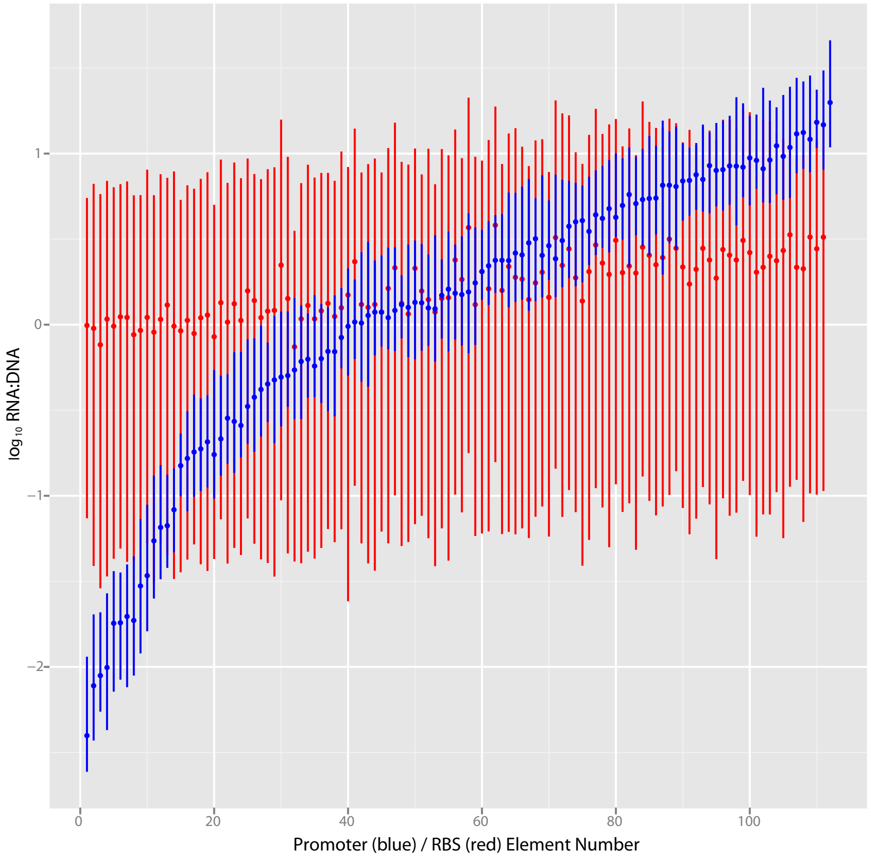

Fig. S6. RNA levels across each promoter and RBS. Mean RNA levels across all promoters (blue circles) and RBSs (red circles) are plotted with lines corresponding to 10th and 90th percentile values. Promoter identity is tightly correlated to RNA level, whereas RBS identity has a slight positive effect, albeit with large variation.

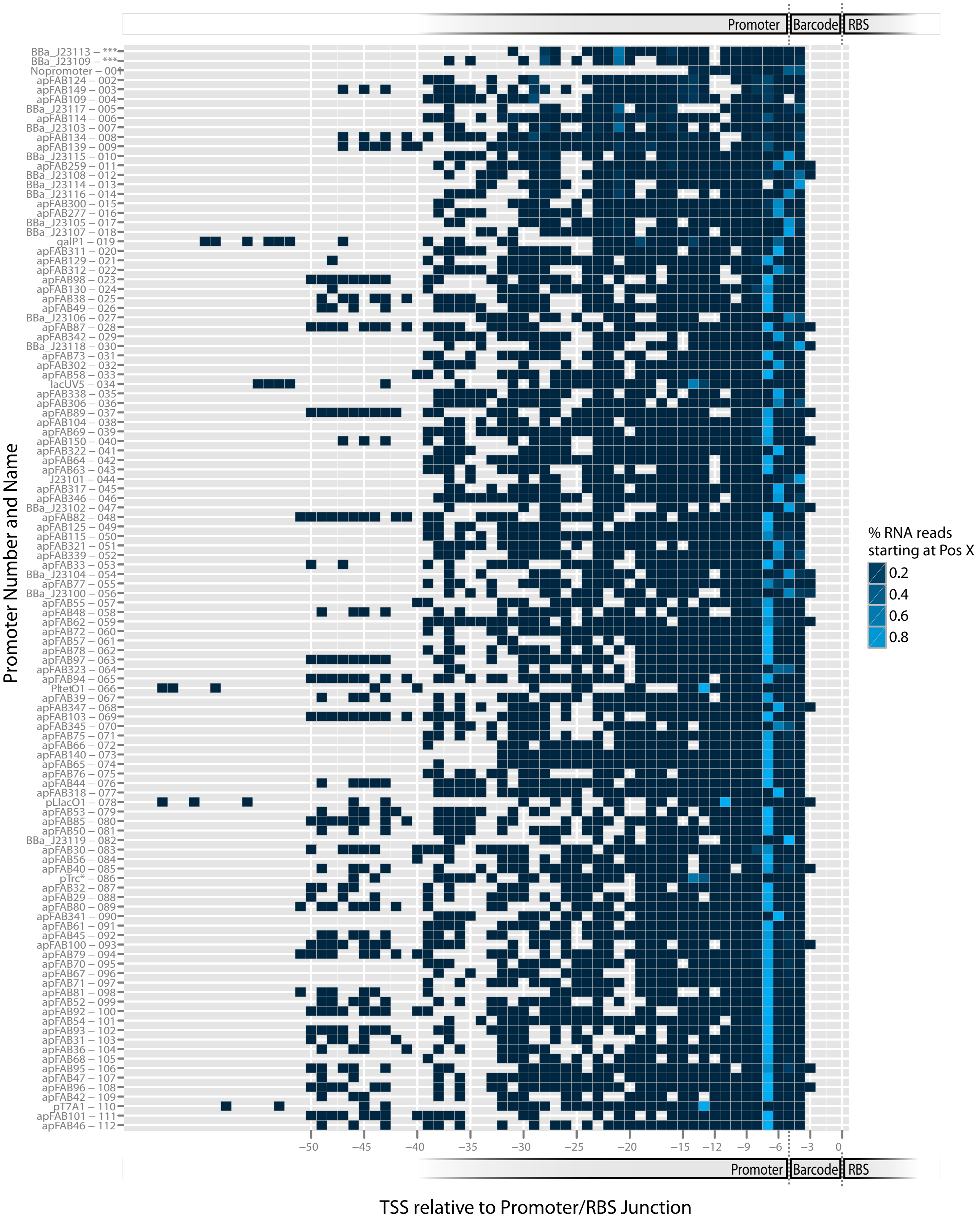

Fig. S7. TSS analysis. The measured start positions from RNA contigs for each promoter are plotted, with more brightly colored squares indicating more common start sites. All positions are relative to the junction between the promoter-specific barcode and the RBS (as shown in the schematic at the bottom of the figure). The five-base promoter-specific barcode sequence allows promoter identification for RNA contigs that begin after the end of the functional promoter region. If an RNA contig begins more than two bases into the barcode, it cannot be mapped uniquely; those contigs are discarded. The first two constructs were removed from further analysis due to start sites that presumably started after the barcode (*SI Materials and Methods*); these are marked by three asterisks. Pos, Position.

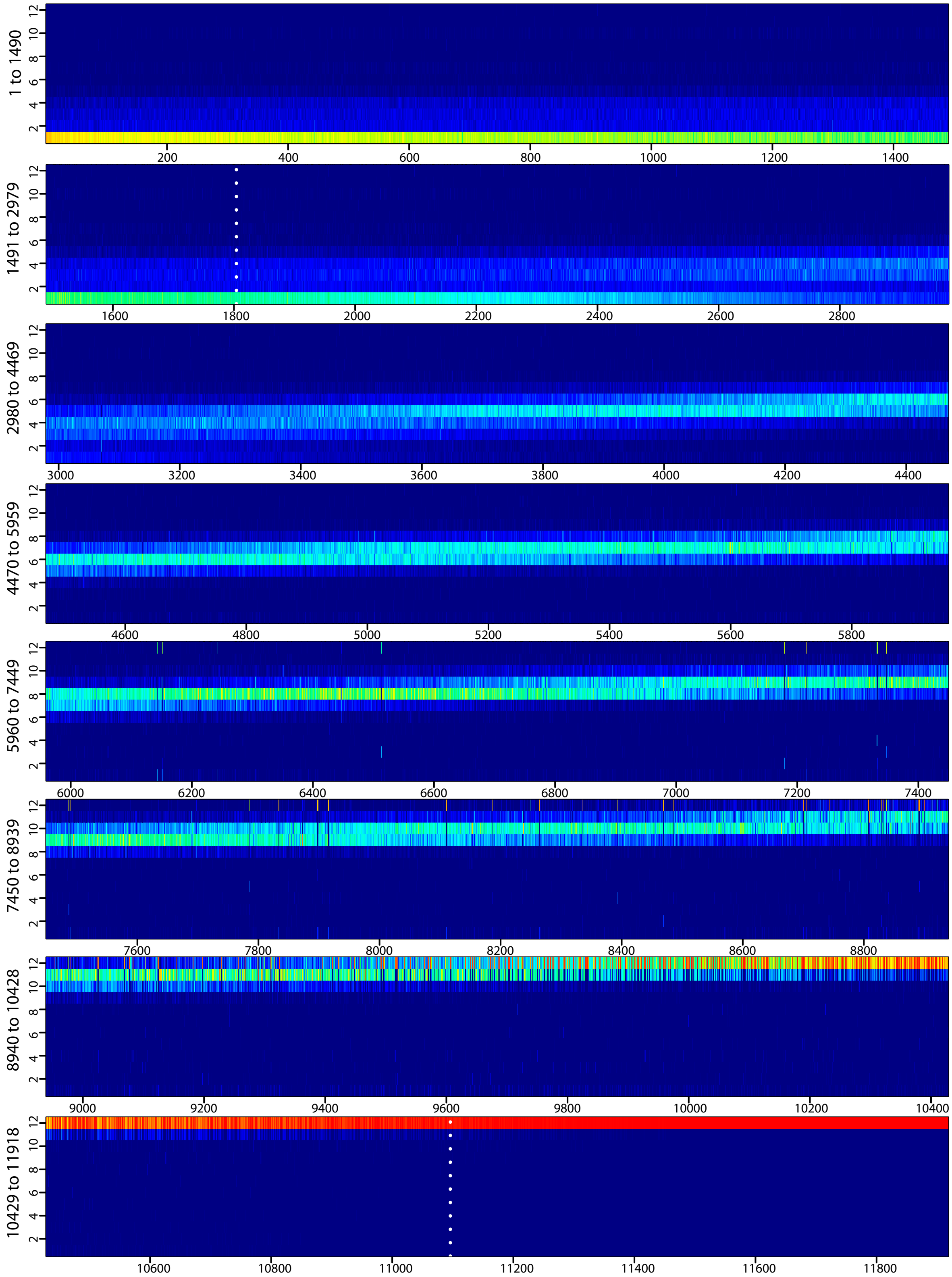

Fig. S8. Percentages of contigs falling into each of the 12 bins across all constructs. A total of 11,981 constructs are shown on the *x* axis, ordered by increasing protein level as estimated by FlowSeq. White dotted lines show the high- and low-protein level cutoffs, beyond which constructs cannot be accurately measured. Contigs for most constructs fall into a few contiguous bins, suggesting a continuous distribution of gene expression level among cells harboring the same construct. Seven hundred thirty-five constructs with fewer than 100 counts or constructs whose contigs fell entirely into one bin (except the final bin) were discarded from analysis and are not shown here.

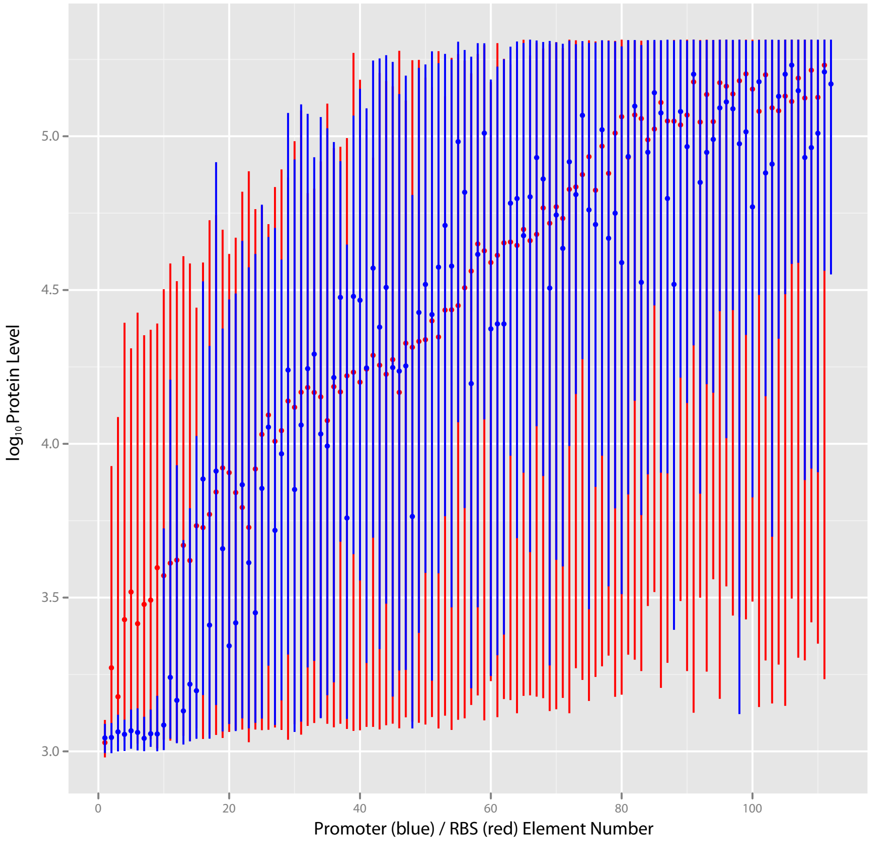

Fig. S9. Protein levels across each promoter and RBS. Mean protein levels across all promoters (blue circles) and RBSs (red circles) are plotted with lines corresponding to 10th and 90th percentile values.

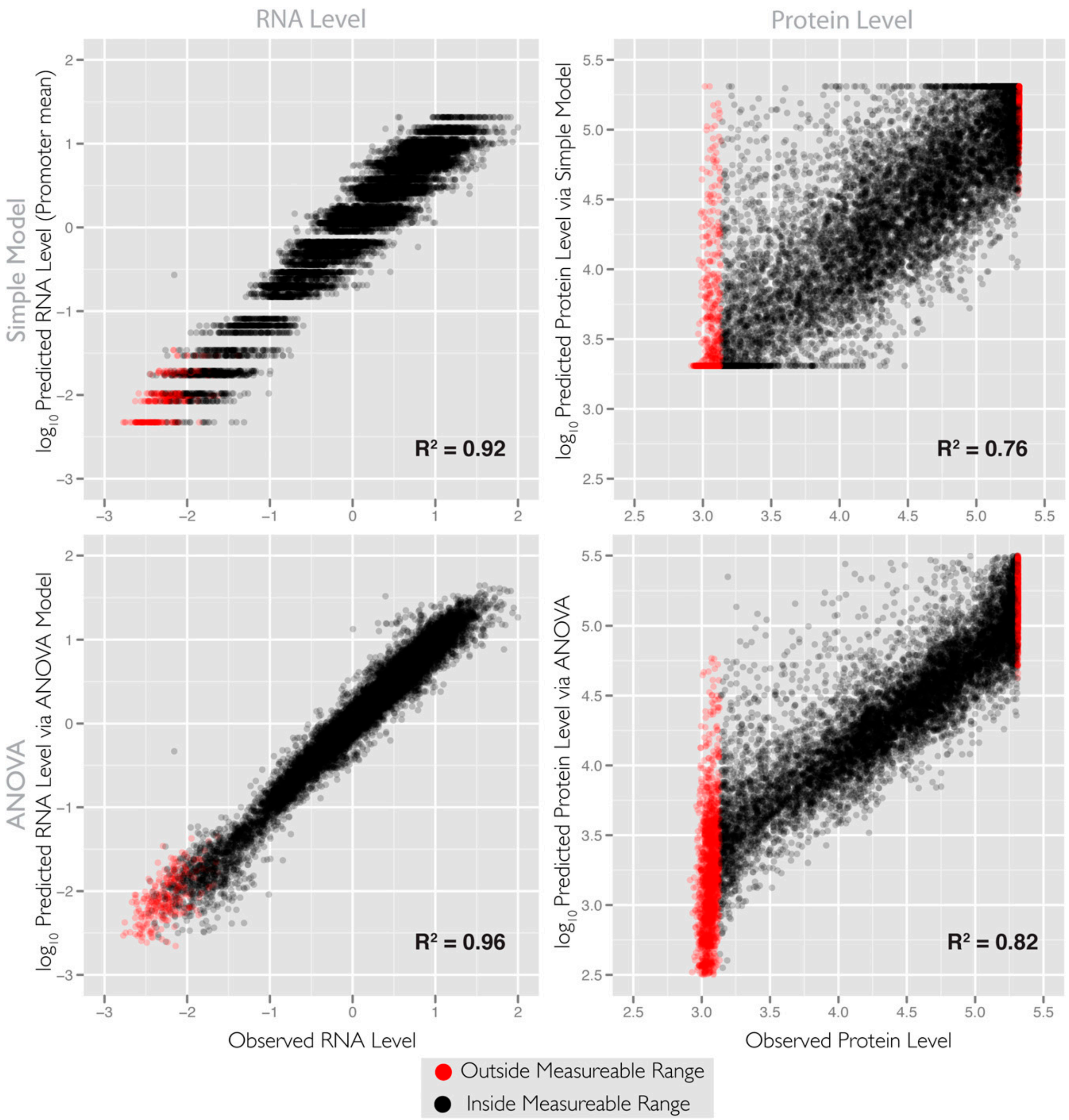

Fig. S10. Comparison of simple and ANOVA models. For each construct, we plotted predicted vs. observed protein and RNA levels for the simple promoter + RBS model (*Upper*) and the ANOVA model (*Lower*). Red points are those outside the linear range of our FlowSeq measurement.

Dataset S1. Promoters

Dataset S1

Name of the promoter, indicating originating sequence or library, is provided. *full.name*, full name of the promoter in our library naming scheme; *num*, number of the promoter in our library naming scheme; *mean.RNA*, geometric mean of the RNA level across all constructs not filtered; *sd.RNA*, geometric SD of RNA level across all constructs not filtered; *mean.prot*, geometric mean of protein level across all constructs not filtered; *sd.prot*, geometric SD deviation of protein level across all constructs not filtered; *TSS.best*, most prevalent TSS relative to promoter/RBS junction (Fig. S6); *TSS.pct_best*, fraction of contigs that begin at the most prevalent TSS; *Sequence*, sequence of the promoter, including the cut site at the beginning, separated by a space, in which the last five bases are the unique promoter barcode for TSS identification.

Dataset S2. RBSs

Dataset S2

Name of the RBS, indicating originating sequence or library, is provided. *full.name*, full name of the RBS in our library naming scheme; *num*, number of the RBS in our library naming scheme; *mean.RNA*, geometric mean of the RNA level across all constructs not filtered; *sd.RNA*, geometric SD of RNA level across all constructs not filtered; *mean.prot*, geometric mean of protein level across all constructs not filtered; *sd.prot*, geometric SD of protein level across all constructs not filtered; *mean.xlat*, geometric mean of translation efficiency (protein/RNA) across all constructs not filtered; *sd.xlat*, geometric SD of translation efficiency (protein/RNA) across all constructs not filtered; *Sequence*, sequence of the RBS, including the cut site at the end, separated by a space.

Dataset S3. Constructs

Dataset S3

Name of the promoter, indicating originating sequence or library, is provided. *RBS*, name of the RBS, indicating originating sequence or library; *target*, full construct name; *count.prot*, total number of FlowSeq contig counts; *bin.1..12*, contig counts in bin 1 to bin 12; *prot*, calculated protein level (as in *SI Materials and Methods*); *count.prot.203*, total number of FlowSeq spike-in counts; *bin.1.203*, counts in bin 1..12 (spike-in); *prot.203*, calculated protein level (as in *SI Materials and Methods*); *count.A.RNA*, RNA counts, replicate A; *count.B.RNA*, RNA counts, replicate B; *count.RNA*, total RNA contig counts; *count.A.DNA*, DNA counts, replicate A; *count.B.DNA*, DNA counts, replicate B; *count.DNA*, total DNA contig counts; *RNA.A*, RNA level, replicate A; *RNA.B*, RNA level, replicate B; *RNA*, mean RNA level; *count.A.RNA.203*, RNA counts, replicate A (spike-in); *count.B.RNA.203*, RNA counts, replicate B (spike-in); *count.RNA.203*, total RNA counts (spike-in); *count.A.DNA.203*, DNA counts, replicate A (spike-in); *count.B.DNA.203*, DNA counts, replicate B (spike-in); *count.DNA.203*, total DNA counts (spike-in); *RNA.A.203*, RNA level, replicate A (spike-in); *RNA.B.203*, RNA level, replicate B (spike-in); *RNA.203*, mean RNA level (spike-in); Δ*G*, calculated secondary structure ΔG (as in *SI Materials and Methods*); *bad.prot*, Boolean: insufficient protein data?; *bad.DNA*, Boolean: insufficient DNA data?; *min.A.RNA*, Boolean: too few RNA counts in replicate A?; *min.B.RNA*, Boolean: too few RNA counts in replicate B?; *min.RNA*, Boolean: too few RNA counts in either replicate?; *bad.A.RNA*, Boolean: both RNA and DNA too low in replicate A?; *bad.B.RNA*, Boolean: both RNA and DNA too low in replicate B?; *bad.RNA*, Boolean: both RNA and DNA too low in either replicate?; *count.A.RNA.raw*, unadjusted RNA count, replicate A; *count.B.RNA.raw*, unadjusted RNA count, replicate B; *count.RNA.raw*, total unadjusted RNA count; *RNA.A.raw*, unadjusted RNA level, replicate A; *RNA.B.raw*, unadjusted RNA level, replicate B; *RNA.raw*, unadjusted RNA level, mean; *bad.promo*, Boolean: Was this promoter removed due to a late TSS? (as in *SI Materials and Methods*)?; *min.prot*, Boolean: below linear protein measurement threshold?; *max.prot* , Boolean: above linear protein measurement threshold?; *mean.rbs.xlat*, mean RBS translation efficiency; *mean.promo.RNA*, mean promoter RNA level; *mean.rbs.RNA*, mean RBS RNA level; *dev.rbs.RNA*, deviation from RBS mean RNA level; *mean.promo.prot*, mean promoter protein level; *dev.promo.prot*, deviation from promoter mean protein level; *RBS.TTL*, RBS full name in library naming scheme; *RBS.num*, RBS number in naming scheme; *Promoter.TTL*, promoter full name in library naming scheme; *Promoter.num*, promoter number in library naming scheme; *model.RNA.simple*, RNA level prediction based on mean RNA level per promoter; *model.prot.simple*, protein level prediction based on mean RNA level per promoter * mean translation efficiency per RBS; *model.prot.avg*, not used; *model.prot.add*, not used; *model.prot.full*, protein level prediction based on ANOVA framework; *model.RNA.full*, RNA level prediction based on ANOVA framework; *model.trans.full*, not used.