Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically-Encoded Multiplexed Reporter Assay in *E. coli*

Authors and Affiliations:

Guillaume Urtecho[1], Arielle D. Tripp[2], Kimberly D. Insigne[3], Hwangbeom Kim[4], and Sriram Kosuri[4,5*]

[1] Molecular Biology Interdepartmental Doctoral Program, University of California, Los Angeles, CA, 90095, USAvvedens
[2] Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA, 90095, USA
[3] Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, CA 90095, USA
[4] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA
[5] UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed. Tel: +1 310 825 8931; Email: sri@ucla.edu

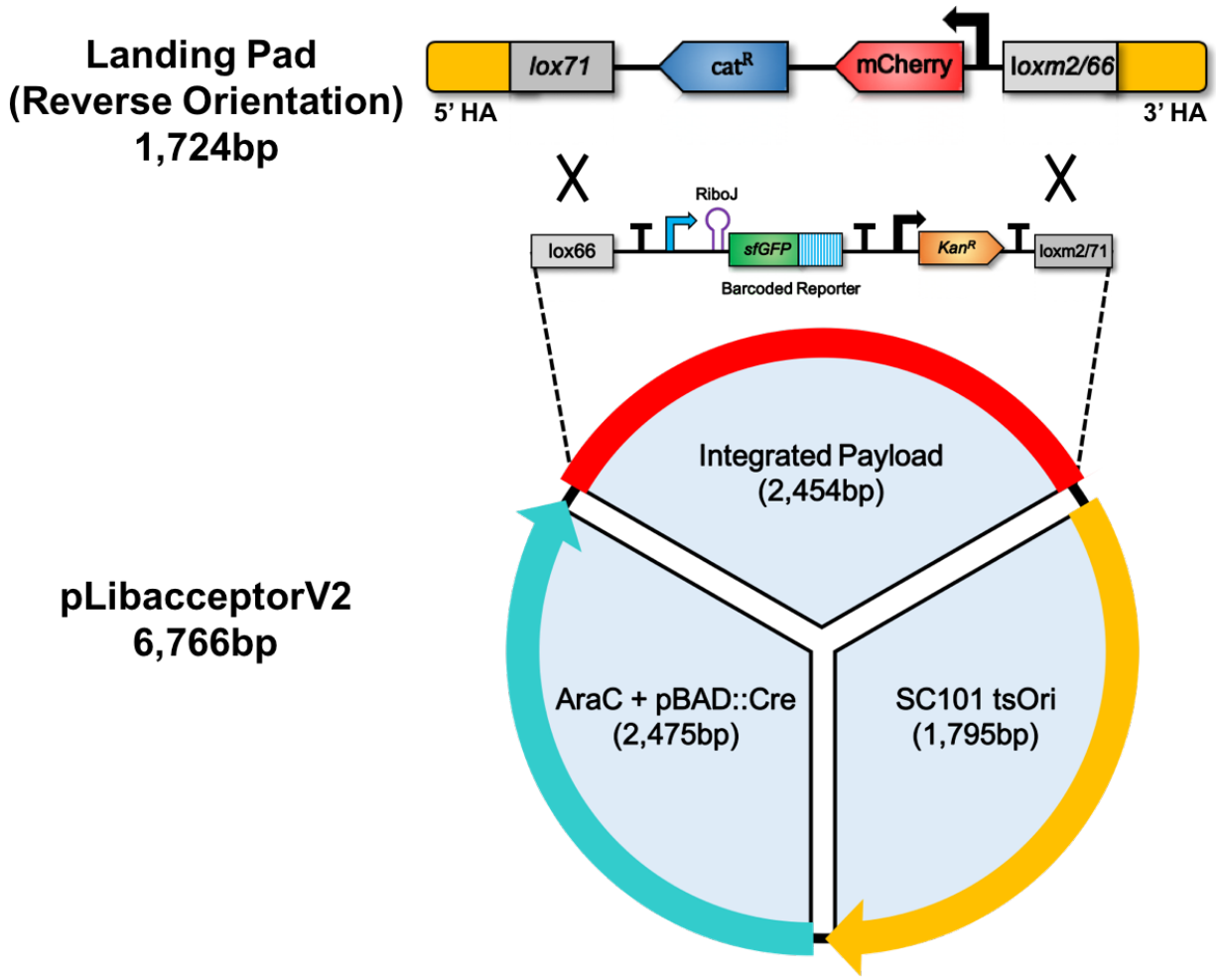

Figure S1. RMCE components. An example of a reversed-orientation landing pad (top) showing the mCherry and *cat*^{R} (chloramphenicol resistance) bicistronic operon. Below that we show the final donor plasmid containing the promoter reporter construct and the boundaries of the cassette exchange. Components of the original integration vector, pLibacceptorV2 are described in the methods. Abbreviations: homology arm (HA).

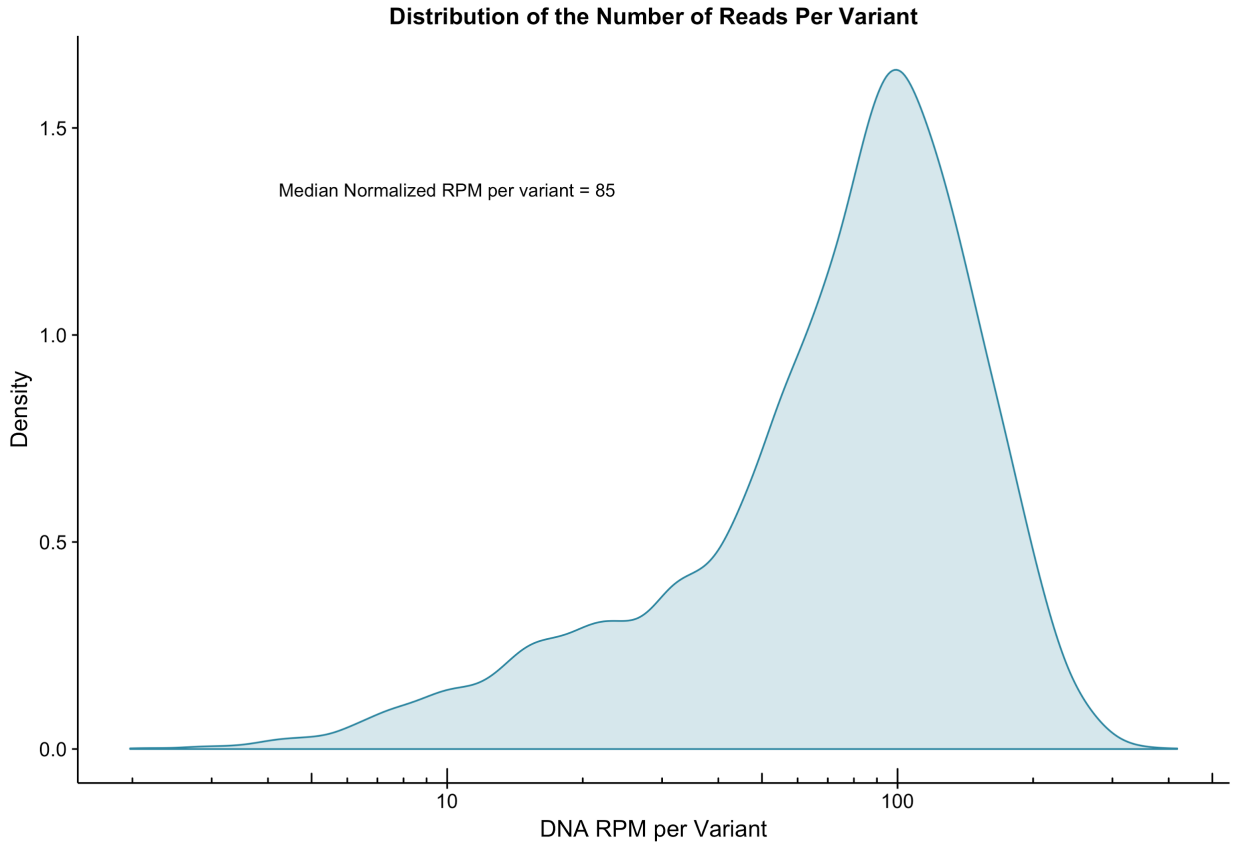

Figure S2. DNA sequencing reads are well distributed amongst integrated variants. Here we plot the distribution of normalized DNA reads per variant. DNA read counts for each barcode were converted to reads per million (RPM) and aggregated for each mapped variant.

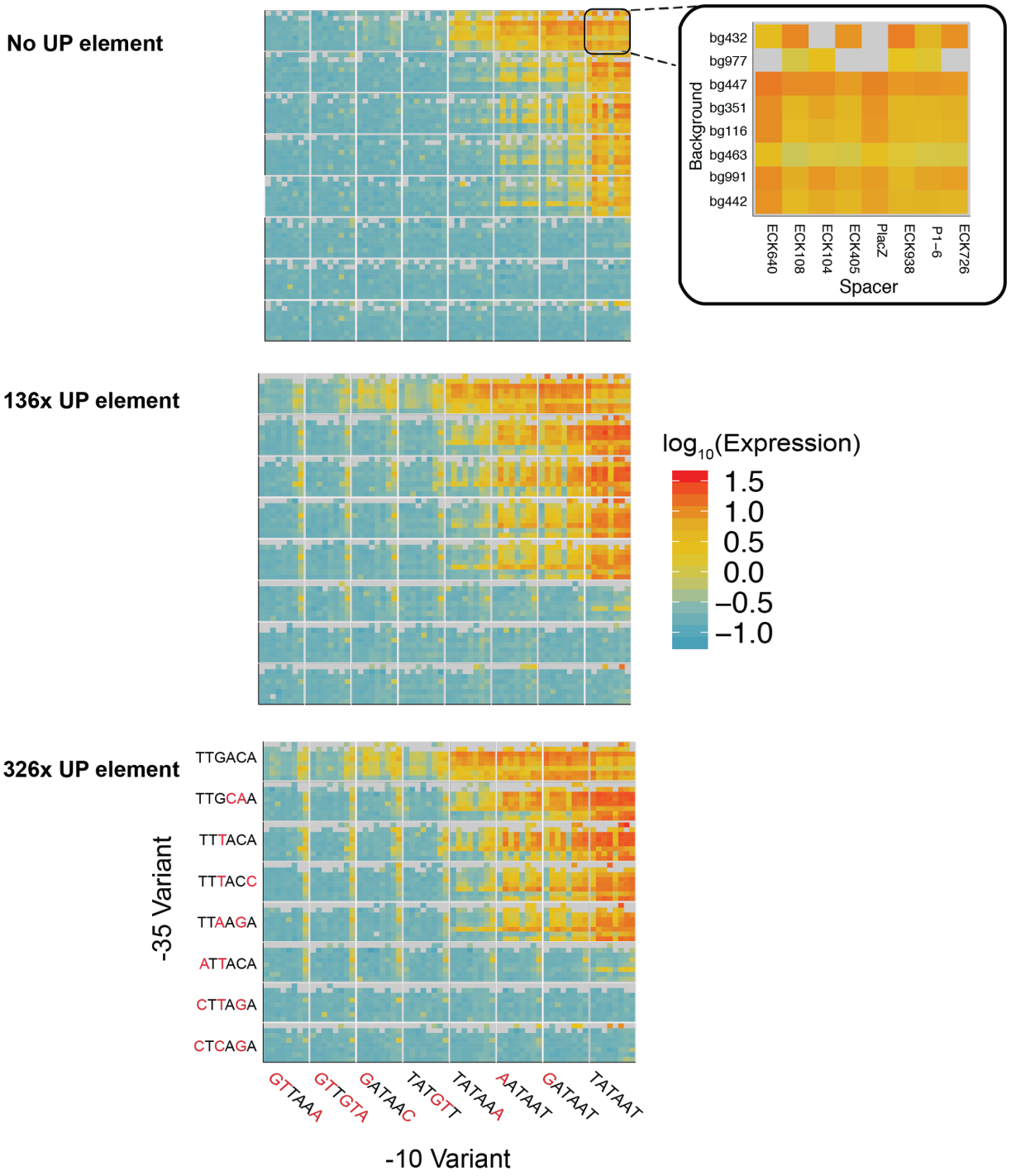

Figure S3. Global trends of promoter expression are consistent between UP element variants. As in Figure 3B, we show that promoter expression increases as -10 and -35 sequences approach the consensus. We include the two backgrounds omitted from Figure 3B (bg432 and bg977) that show reduced mappings possibly due to inefficient oligo synthesis or amplification with these backgrounds. Promoter variants for which we could not detect more than four barcodes were omitted from our analysis and are displayed as grey squares.

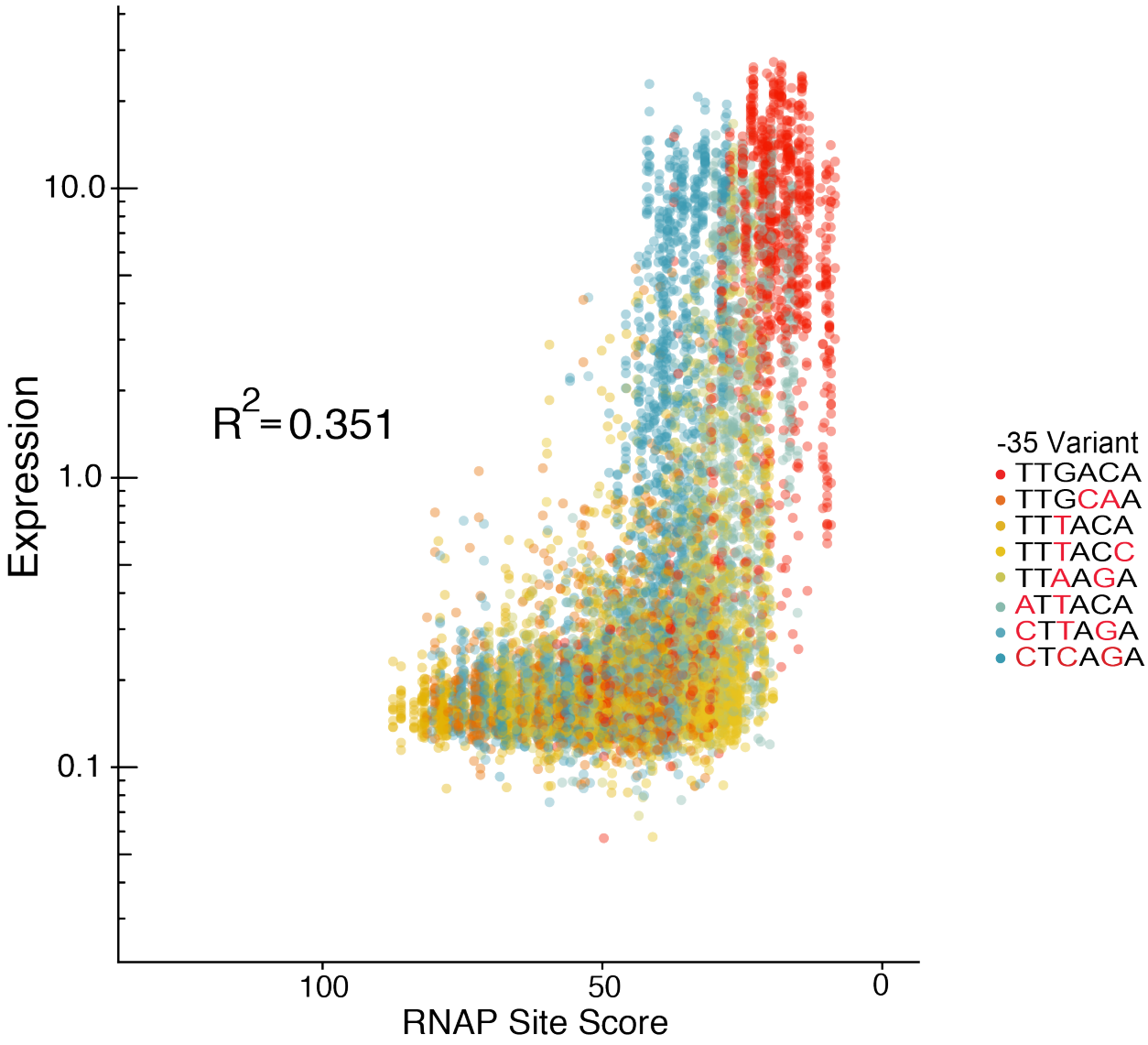

Figure S4. Assessing external binding energy scores against expression. We used a previously described binding energy matrix model of RNAP to the lac promoter to score our minimal library. This model experimentally determined an energy matrix model representing the interaction of RNAP and a promoter sequence of interest as a proxy for expression.

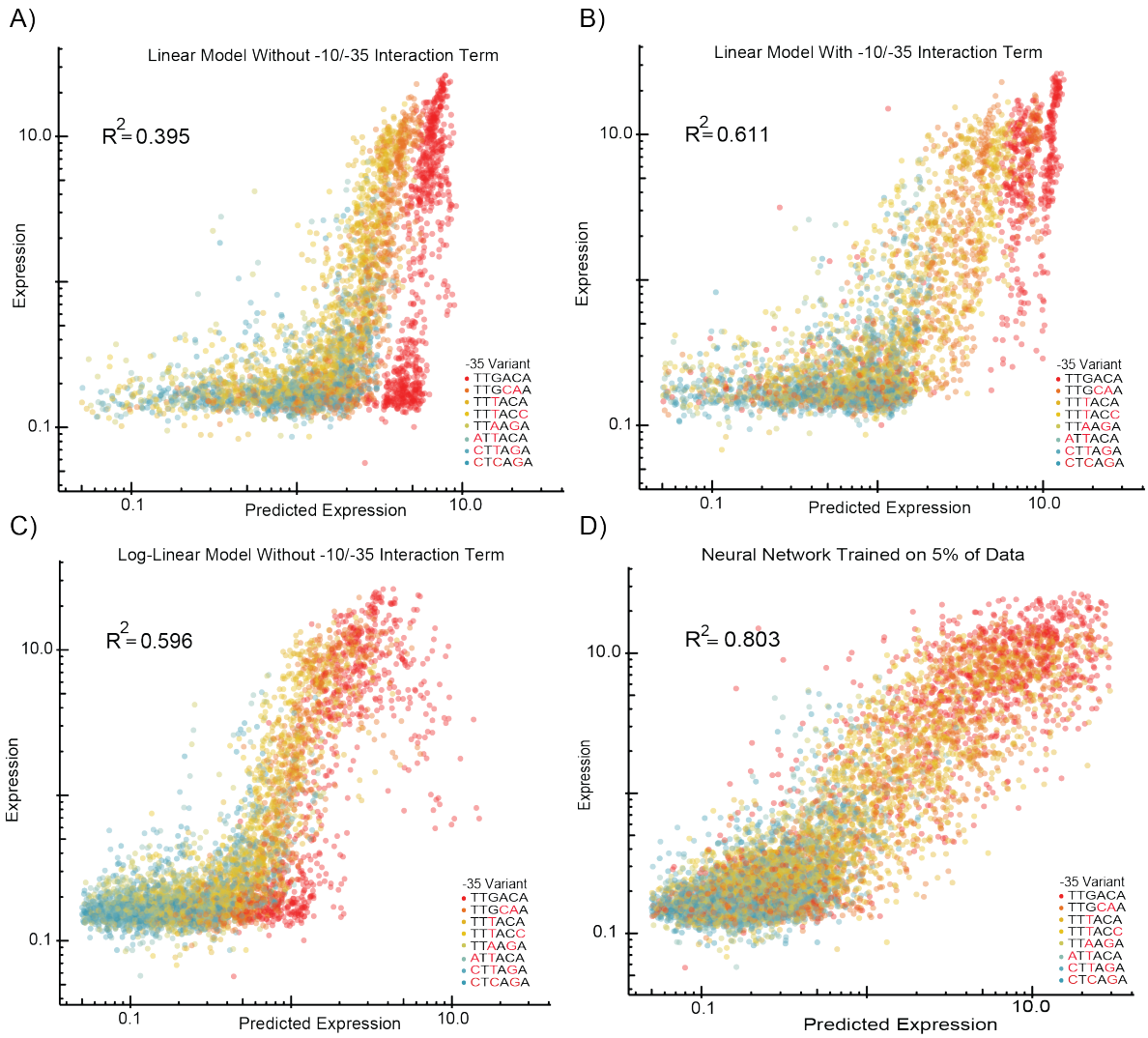

Figure S5. Alternative models to fit experimental data. A) We implemented a basic linear model trained without a -10/-35 interaction term to predict promoter expression. B) Inclusion of the interaction term increases the predictive capacity of linear models. C) Training models on log-transformed data captures the multiplicative effects of sequence elements thereby increasing performance. D) The neural network performs better than any linear model even when trained on as little as 5% of the data. All linear models in this figure were trained on 50% of the data and tested on the remaining data.

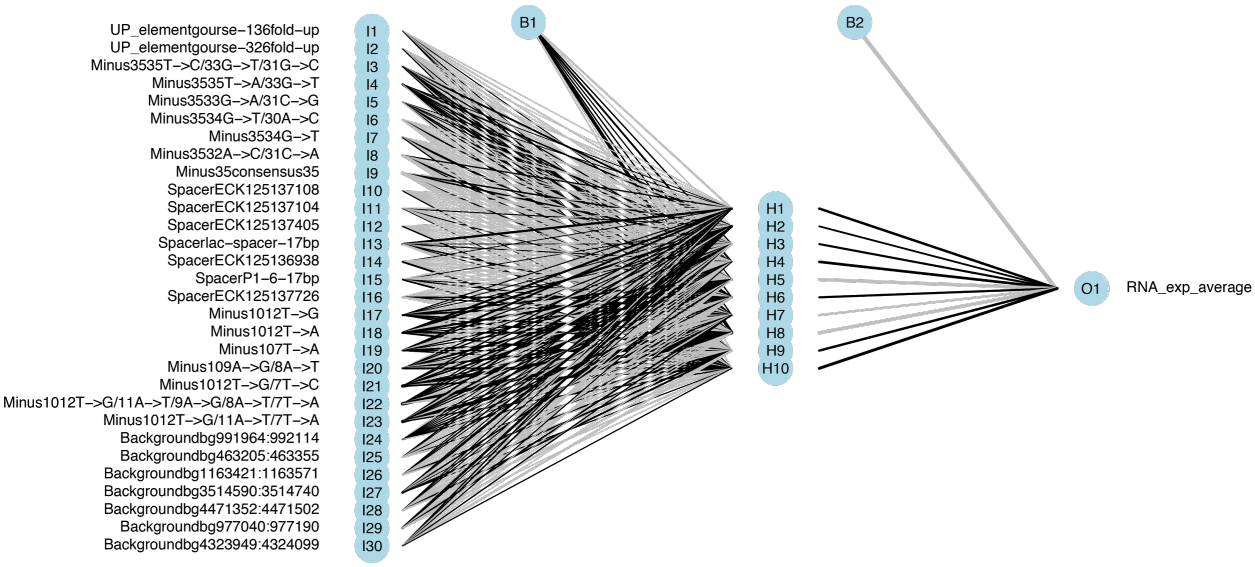

Figure S6. Neural network schematic. The neural network is comprised of an input layer containing 30 nodes, each representing one of the possible Backgrounds, UP elements, -35 elements, spacers, or -10 elements, and a single hidden layer containing 10 nodes. This is a feed-forward network that implements a sigmoid activation function in the hidden layer and a linear output function. Respectively, nodes I1-30, H1-10, B1-2, and O1 represent the input, hidden, bias, and output nodes. The value of the weights connecting each node are represented by their thickness.

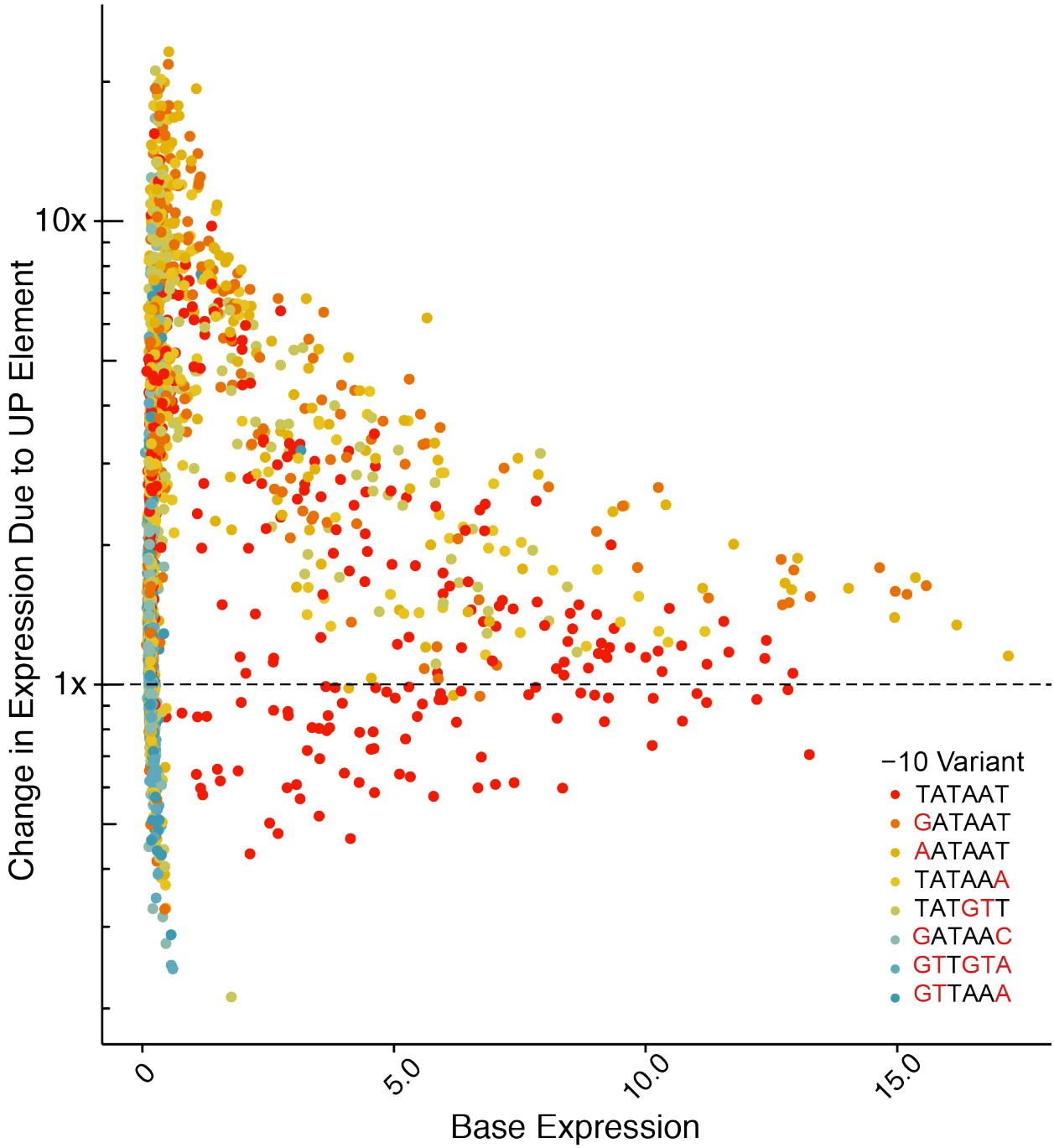

Figure S7. Addition of the consensus UP element can negatively impact expression of promoters containing a consensus -10 element. We determined the change in expression due to the consensus UP element by dividing the expression of promoters with the consensus UP by their expression without an UP element.

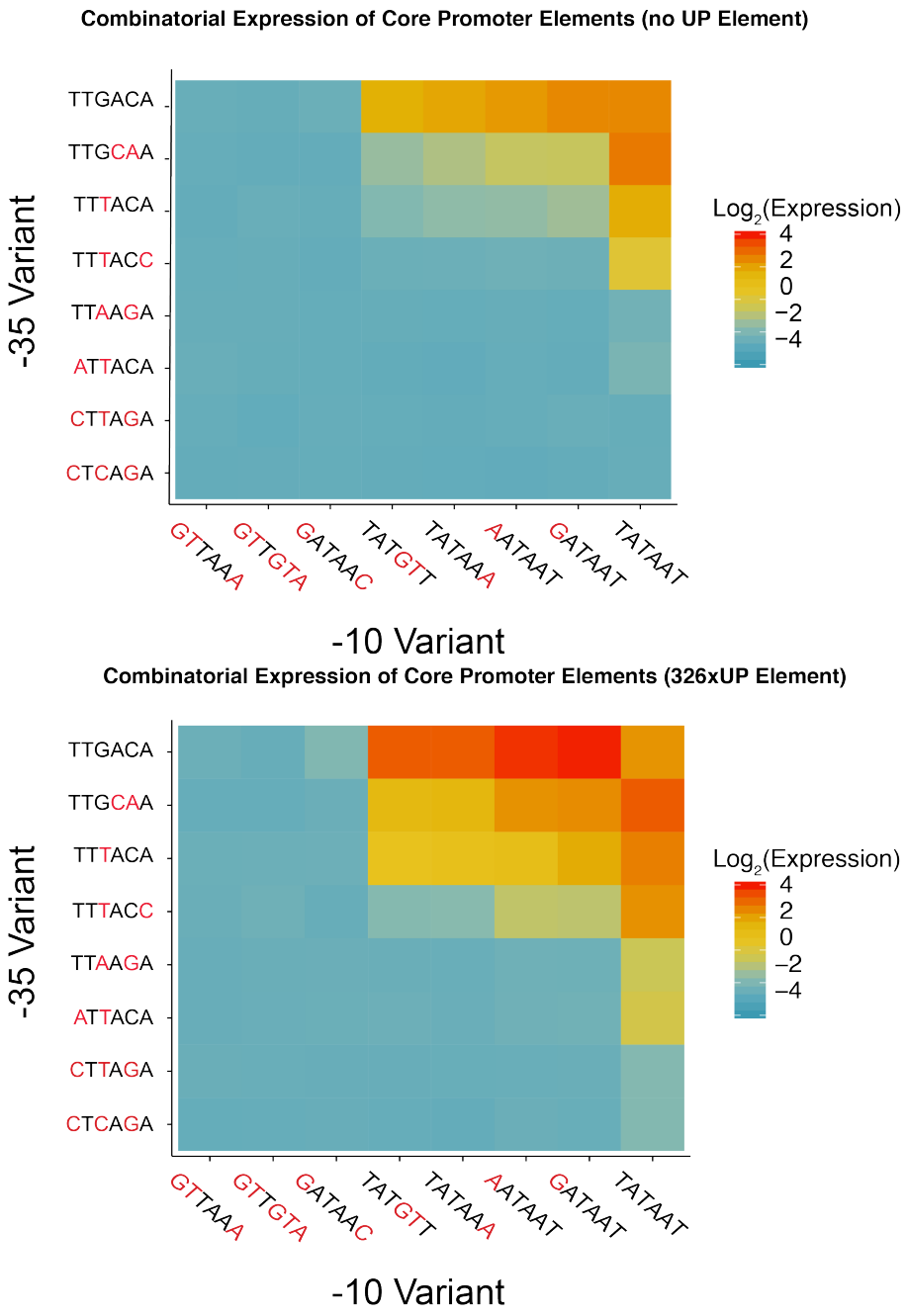

Figure S8. Promoters with consensus -10 but inactive -35 elements are rescued by consensus UP element. The -10 (horizontal) and -35 (vertical) elements are ordered by increasing median expression amongst promoters with the indicated variant. The two lowest expressing -35 variants are inactive unless the consensus -10 and UP elements are present.

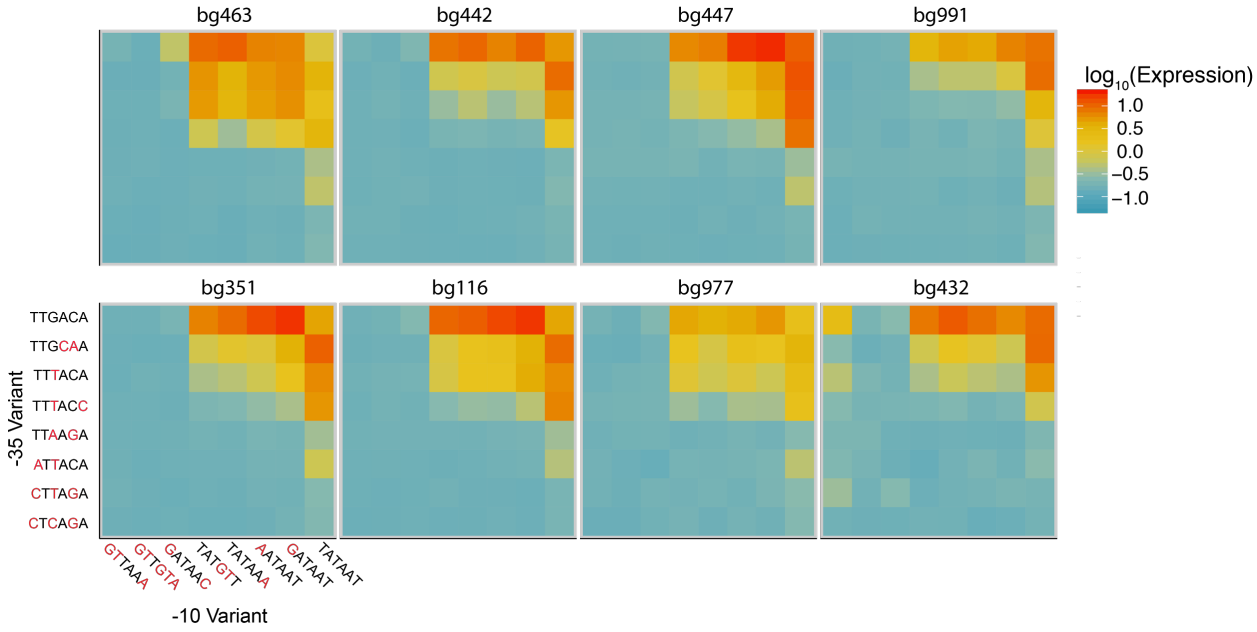

Figure S9. The activity of the core promoter is dependent on local sequence context. Combinations of -10 and -35 elements have variable expression between backgrounds. We separated the library by background and determined the median expression of all promoters with each -10 and -35 combination.

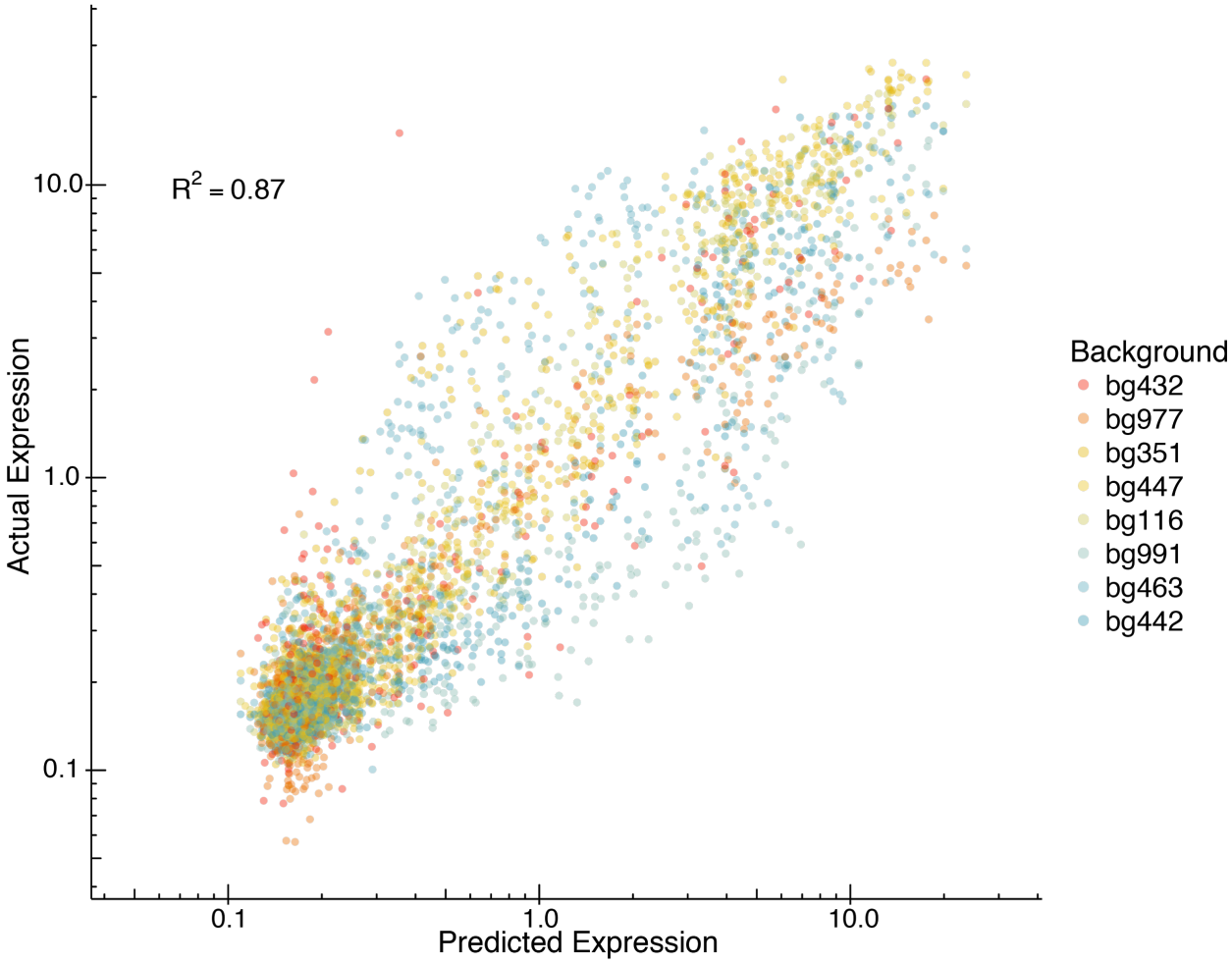

Figure S10. Background sequences are an essential consideration when predicting promoter expression. Here we show the performance of the neural network (as in Figure 4C) trained on all sequence elements aside from the background sequence. The coefficient of determination between predicted and actual expression drops from 0.955 to 0.87 when the background is not taken into account.

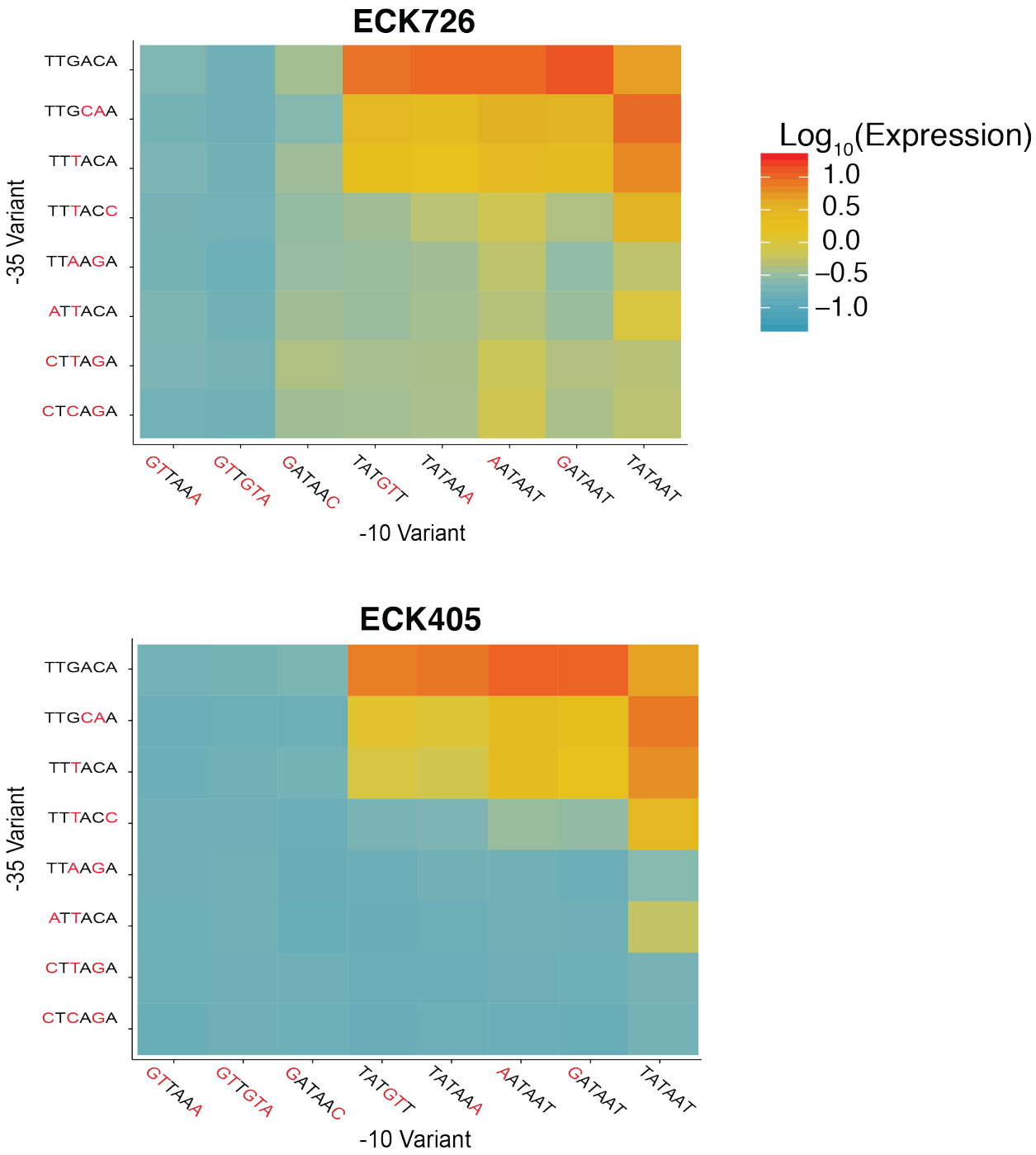

Figure S11. Spacer variant ECK726 permits transcription from promoters with weak core elements. Here we show the median expression of each core promoter combination amongst promoters with spacer variants ECK726 and ECK405, the spacer variant with the strongest median expression. Element variants are arranged in order of median expression amongst promoters containing the indicated variant.

Figure S12. Construction of σ70 variant library. Here we show an example of a constructed variant with its components highlighted. Each variant is composed of a base background sequence in which one of each UP element, Minus 35, Spacer, and Minus 10 were superimposed at their respective positions. Variants without UP Elements maintained their relevant background sequence. See methods for further details.

UP Element	Sequence			Notes
	GAAAATATATTTTTCAAAAGTA			136-fold increase
	GGAAAATTTTTTTTCAAAAGTA			326-fold increase
-35	**Sequence**	**Mismatched Bases**	**Sequence**	**Mismatched Bases**
	TTGACA	0	TTAAGA	2
	TTTACA	1	TTGCAA	2
	ATTACA	2	CTCAGA	3
	TTTACC	2	CTTAGA	3
-10	**Sequence**	**Mismatched Bases**	**Sequence**	**Mismatched Bases**
	TATAAT	0	GATAAC	2
	AATAAT	1	TATGTT	2
	GATAAT	1	GTTAAA	3
	TATAAA	1	GTTGTA	5
Spacer	**Sequence**	**Label**	**Sequence**	**Label**
	AAAACTCATTTTATTTT	ECK125137405 spacer	TCGCGCATGATCGAAAG	ECK125137104 spacer
	AGCACGAAAATGGAAGT	ECK125137108 spacer	TGGCTGAATGGTCTGTC	ECK125137640 spacer
	ATAACTTAGAAAGTAAT	ECK125136938 spacer	CTTTATGCTTCGGCTCG	*lac* spacer
	TTTCCATTAGCGAGTAT	ECK125137726 spacer	CTTTATGCTTTTATGTT	P1-6 spacer
Background	**Sequence**	**Source**	**Sequence**	**Source**
	TTGCGGTTTTTTCGGTTCAATCACCGCCTGCTGACGAGCTGGGCG CGTAGTGGACGGACGTTCAGCGGTTTGCGTTTTCGCCGTGTCACG CGATGCATTGCCATTGCGAACAGGCGCAACCGTTGCGGGTTCAGT CGGCAACGTGGAATT	Genomic region 3514590:3514740	CTGGAAGAAAACGCCAAAAAAGAAGGTGTGAATAGCACCGAATCTGG CCTGCAATTCCGCGTGATCAACCAGGGTGAAGGCGCAATTCCGGCAC GTACCGACCGCGTTCGTGTTCATTACACCGGTAAACTGATCGACGGC ACCGTGTTT	Genomic region 4427287:4427437
	ACTGGACGCGGAAGAGCGTGAAATACTGGCGCATCCGCTGGTGG GAGGGCTGATTCTCTTTACGCGTAACTATCATGATCCTGCCCAGTT ACGTGAACTGGTGCGCCAGATCCGCGCAGCTTCGCGCAATCGTCT GGTGGTGGCGGTTGA	Genomic region 1163421:1163571	AGCCGCTTTTAGCGGACGACGTGAGTAAACAAAACCCAGACATCATG GATAATGGCTGGGCTTAATTGAGCGTAGTCGGTTATGCGCCAAACGC GCCATCAATGGTATGCATCGCGCCGGTAACAAAACTGGCTTCTGGCC CTGCTAACC	Genomic region 4471352:4471502
	GCGGCGTAAACGCCTTATCCGGCCTACGGAGGGTGCGGGAATTTG TAGGCCTGATAAGACGCGCAAGCGTCGCATCAGGCAGTCGGCACC ATTGCCGGATGCGGCGTAAACGCCTTATCCGGCCTACGGAGGGTG CGGGAATTTGTAGGC	Genomic region 4323949:4324099	TTGAGCAGGCTTATCAGCTGGTGGTGGCAATCAACGGCCCACTGGCG CGTAACGAGGCGTGGGATGTCGCTCGCGAACTATTGCGCGAAGGGG TCGATCAGCGTCACCTGGCAGAGCAGGTTCAGCCGCTGCGGATGCGA TTAAGCGAAC	Genomic region 977040:977190
	GTTAACAAACAAGGAGAAAACAGTATGAAACACGGAATTAAAGCAC TGCTCATTACCCTGTCCCTGGCCTGTGCCGGAATGTCTCATAGCGC GCTGGCGGCAGCTTCTGTGGCGAAACCGACGGCGGTAGAAACCA AAGCGGAAGCTCCT	Genomic region 463205:463355	CCTGCGTTTCCTCGCTTTTGGTGAAACGCATCTGGCTGATGTGCTGGT GAGCAAGCAGTTCCACGAAGCAAACAATATGACTGATGCGCTGGCGG CGCTTTCTGCGGCGGTTGCCGCACAGCTGCCTTGCCGTGACGCGCTG ATGCAGGA	Genomic region 991964:992114

Table S1. Minimal library design organized by sequence elements. The library consisted on 12,288 unique σ70 promoters that consisted of a set of all possible combinations for the following sequence elements: three UP elements, eight -35 regions, eight spacer sequences, eight -10 regions, and eight background sequences.

Sample	Number of Reads
Biological Replicate #1-1 (DNA)	31,966,245
Biological Replicate #1-2 (DNA)	25,945,822
Biological Replicate #2 (DNA)	17,009,806
Biological Replicate #3 (DNA)	14,196,597
Biological Replicate #1-1 (RNA)	36,559,121
Biological Replicate #1-2 (RNA)	33,731,384
Biological Replicate #2 (RNA)	15,462,030
Biological Replicate #3 (RNA)	16,449,240

Table S2. Number of reads acquired per barcode sequencing run. Technical replicates are labeled as #(Biological Replicate)-(Technical Replicate).

Primer	Sequence (5’ to 3’)
GU59	CATGTTGTCCACTCCAATCGGTGATGGTCCTG
GU60	GTAATAGCTAAATCCCACCCGATGCCTGCAGG
GU61	CAAGCAGAAGACGGCATACGAGAT ACTGTG CATGTTGTCCACTCCAATCG
GU62	CAAGCAGAAGACGGCATACGAGAT AGCCAT CATGTTGTCCACTCCAATCG
GU63	CAAGCAGAAGACGGCATACGAGAT ATCTCG CATGTTGTCCACTCCAATCG
GU64	CAAGCAGAAGACGGCATACGAGAT CAGTGT CATGTTGTCCACTCCAATCG
GU70	AATGATACGGCGACCACCGAGATCTACACGTAATAGCTAAATCCCACCCG ATGC
GU72	ACCTGTAATTCCAAGCGTCTCGAG
GU73	TCGTATCCCTGCAGGNNNNNNNNNNNNNNNNNNNNGCATGTGAGACCGG ATGCTAACTAAACACCGCTAGC
GU79	CGTGCATAGTGCCATGTTATCCCTGAAGTCGAG
GU82	CAAGCAGAAGACGGCATACGAGATATCTCGCGTGCATAGTGCCATGTTAT C
GU83	CAAGCAGAAGACGGCATACGAGATAGCCATCGTGCATAGTGCCATGTTAT C
GU101	AATGATACGGCGACCACCGAGATCTACACGTAATAGCTAAATCCCACCCG ATGCCTGCAGG
GU102	AATGATACGGCGACCACCGAGATCTACAC
GU116	GGATGCTAACTAAACACCGCTAGC

Table S3. Primers used in this study.