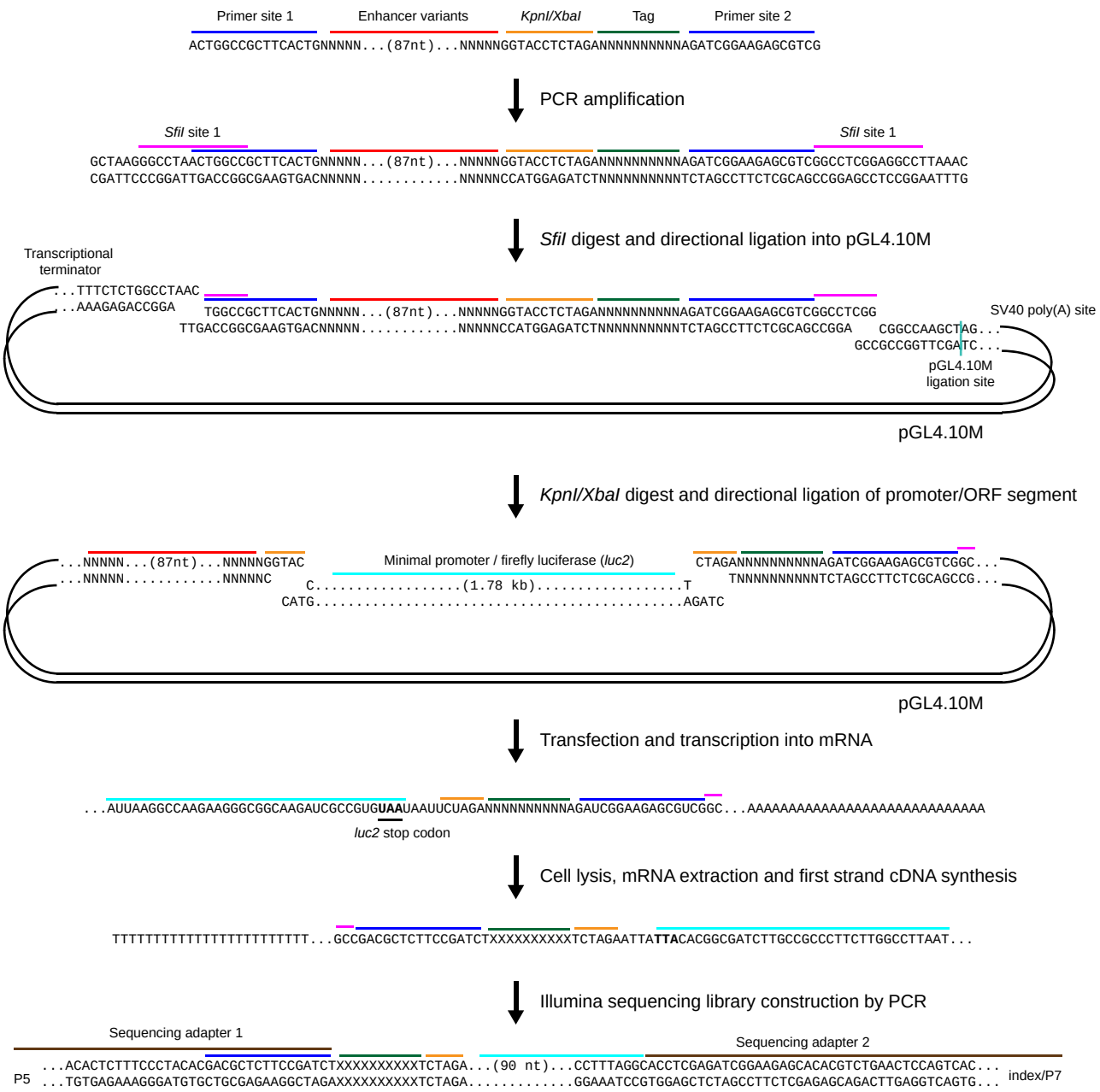
Supplementary Information for

"Rapid dissection and model-based optimization of inducible enhancers in human cells using a massively parallel reporter assay"

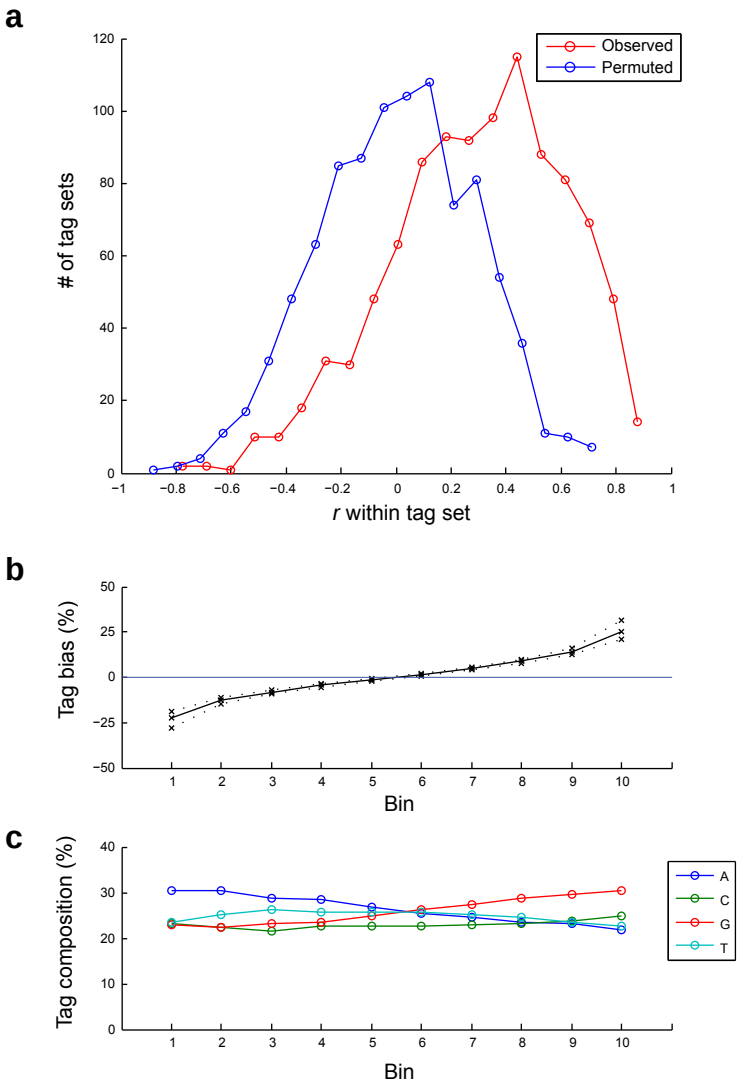
Melnikov, Murugan, Zhang et al. (2012)

Supplementary Figures 1-10
Supplementary Notes
Supplementary Tables 5-6

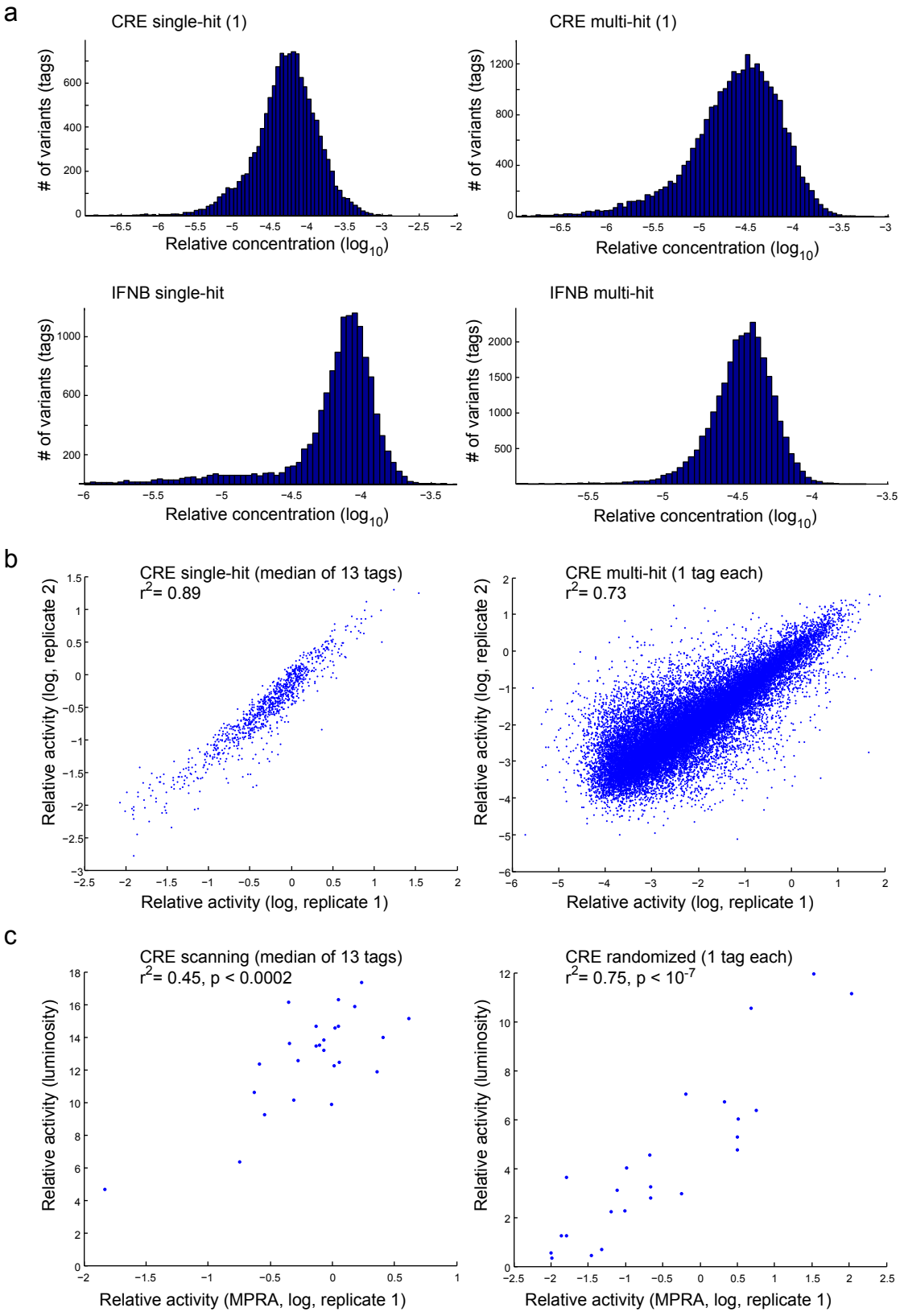
Supplementary Tables 1-4 are distributed in separate Excel spreadsheets
Primary sequence data have been deposited in NCBI GEO under accession GSE31982



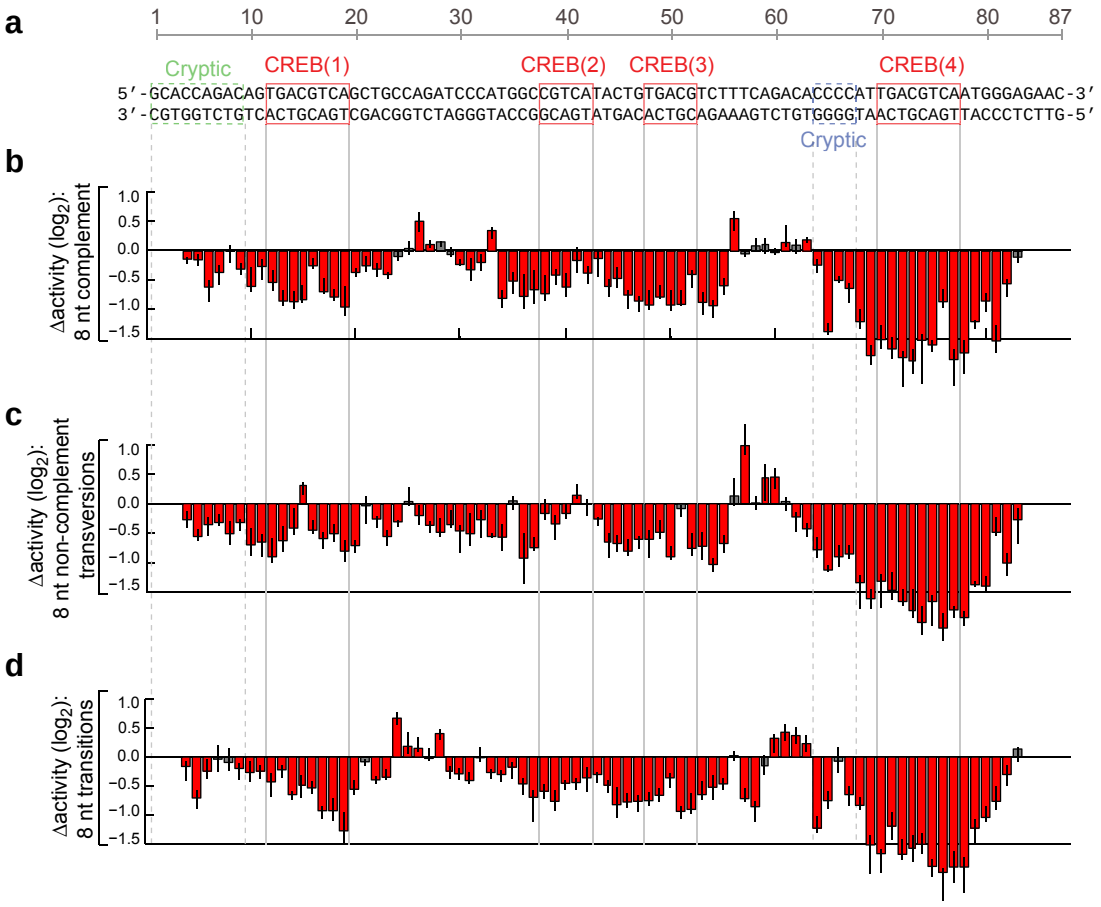
Supplementary Figure 1 – Key steps in the MPRA implementation described in this article.
Critical sequence features are highlighted. See the Methods section for protocol details.



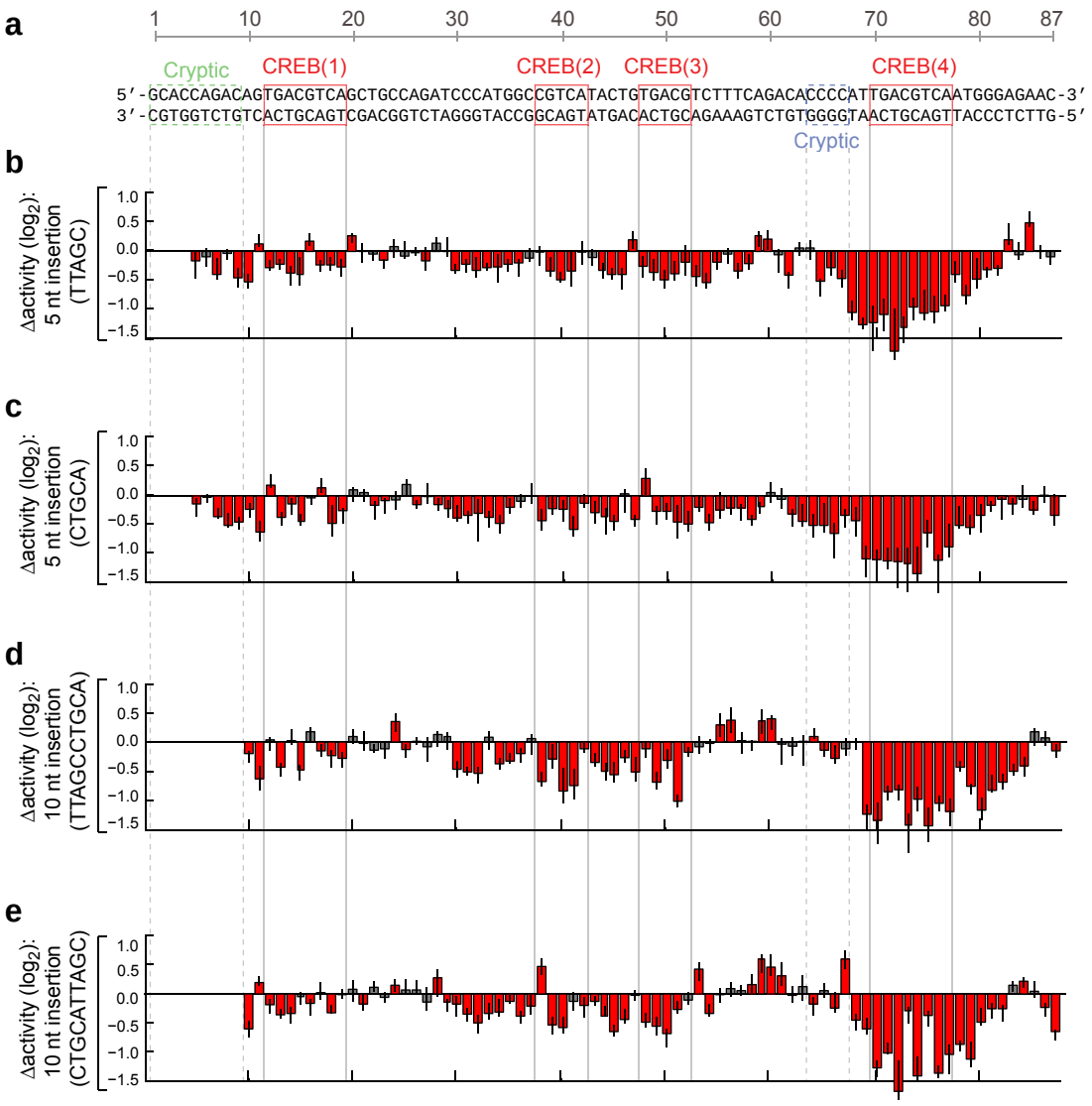
Supplementary Figure 2 – Analysis of tag-related biases in MPRA data. Ideally, the mRNA/plasmid tag ratio (enhancer activity estimate) obtained from each MPRA plasmid should be independent of its specific tag sequence. A variety of sequence-dependent effects, including PCR amplification biases and differential effects on mRNA stability are, however, likely to introduce systematic errors. To estimate the magnitude of tag-related biases, we analyzed the sets of 13 tags that were linked to each of the ~1,000 distinct variants in our single-hit CRE design. In the absence of sequence-dependent biases, the expected correlation between the 13 pairs of mRNA/plasmid tag ratios obtained from our two independent transfection experiments would be zero. (a) The distribution of correlation coefficients (Pearson) between each set of 13 matching mRNA/plasmid tag ratios from the same single-hit CRE variant assayed in two independent MPRA experiments. We observed an excess of r values > 0 (red; median = 0.3) relative to the expected distribution (estimated by permuting the association between tags and ratios within each set; median = 0.0), which indicates a slight tag-related bias. (b) The 'bias' of each of the ~13,000 tags utilized in the single-hit CRE design was estimated as the average of its two observed mRNA/plasmid ratios across the two experiments, divided by the average of the two median ratios from all 13 tags associated with the same variant. The tags were then sorted by their bias and partitioned into ten equally-sized bins. The plot shows the median bias for each bin (solid line; first and third quartiles shown as dotted lines). The majority (~80%) of tags had an estimated bias of less than ±15%. (c) Mean nucleotide composition of tags in each of the ten bins. The tags with the most negative bias (i.e. those that appear to systematically underestimate the activity of their linked variant) tend to be more A-rich than unbiased tags, while the tags with the most positive bias (i.e. those that appear to systematically overestimate the activity of their linked variant) tend to be G-rich. The primary sources of these biases are currently unknown, but we expect that in-depth analysis across additional experiment will be helpful for designing improved MPRA tag sets for future experiments.



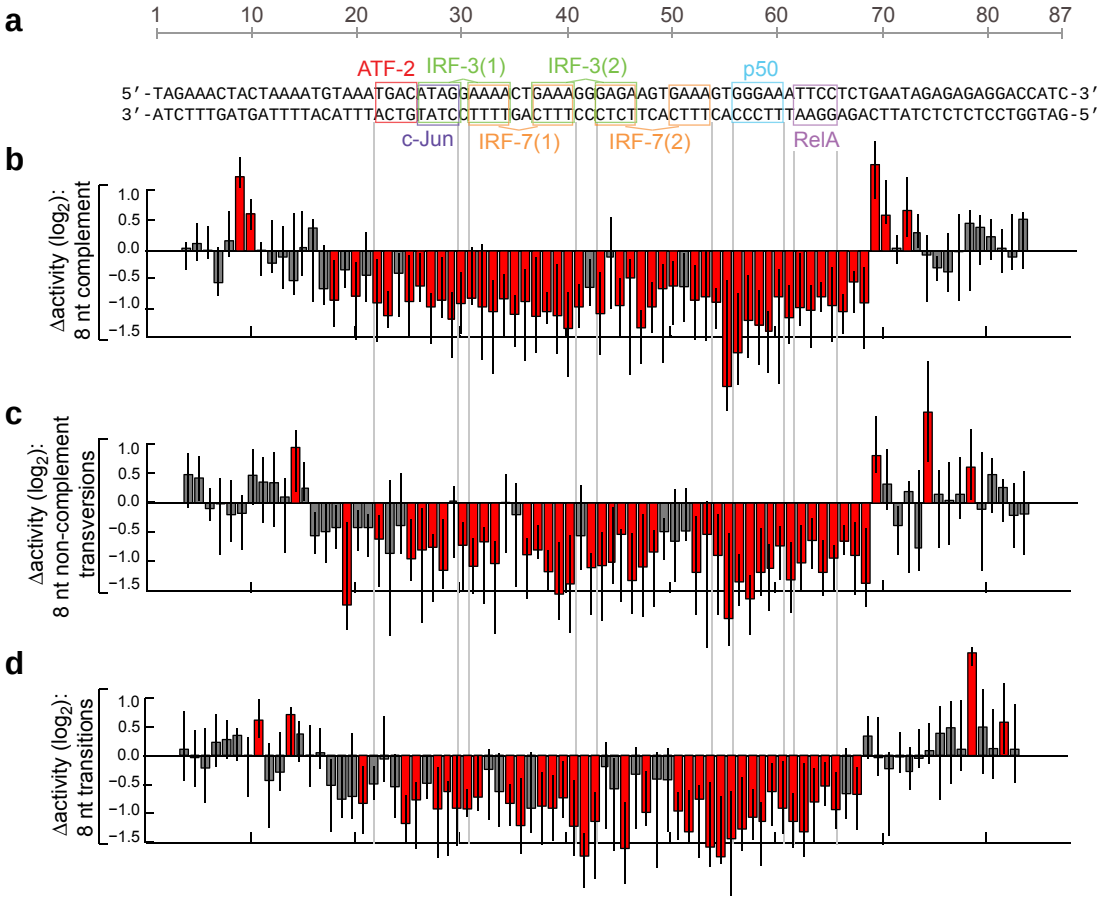
Supplementary Figure 3 – Validation of the MPRA approach. (a) Histograms of the relative concentrations of the designed enhancer variants in each MPRA plasmid pool, as inferred by plasmid Tag-Seq. (b) Concordance between CRE activity estimates from two independent MPRA experiments performed using each of the two mutagenesis designs. (c) Concordance between luciferase-based assays and MPRA for 24 single-hit and multi-hit variants. The lower correlation in the single-hit comparison is likely due to the majority of single-hit subclones containing relatively neutral mutations.



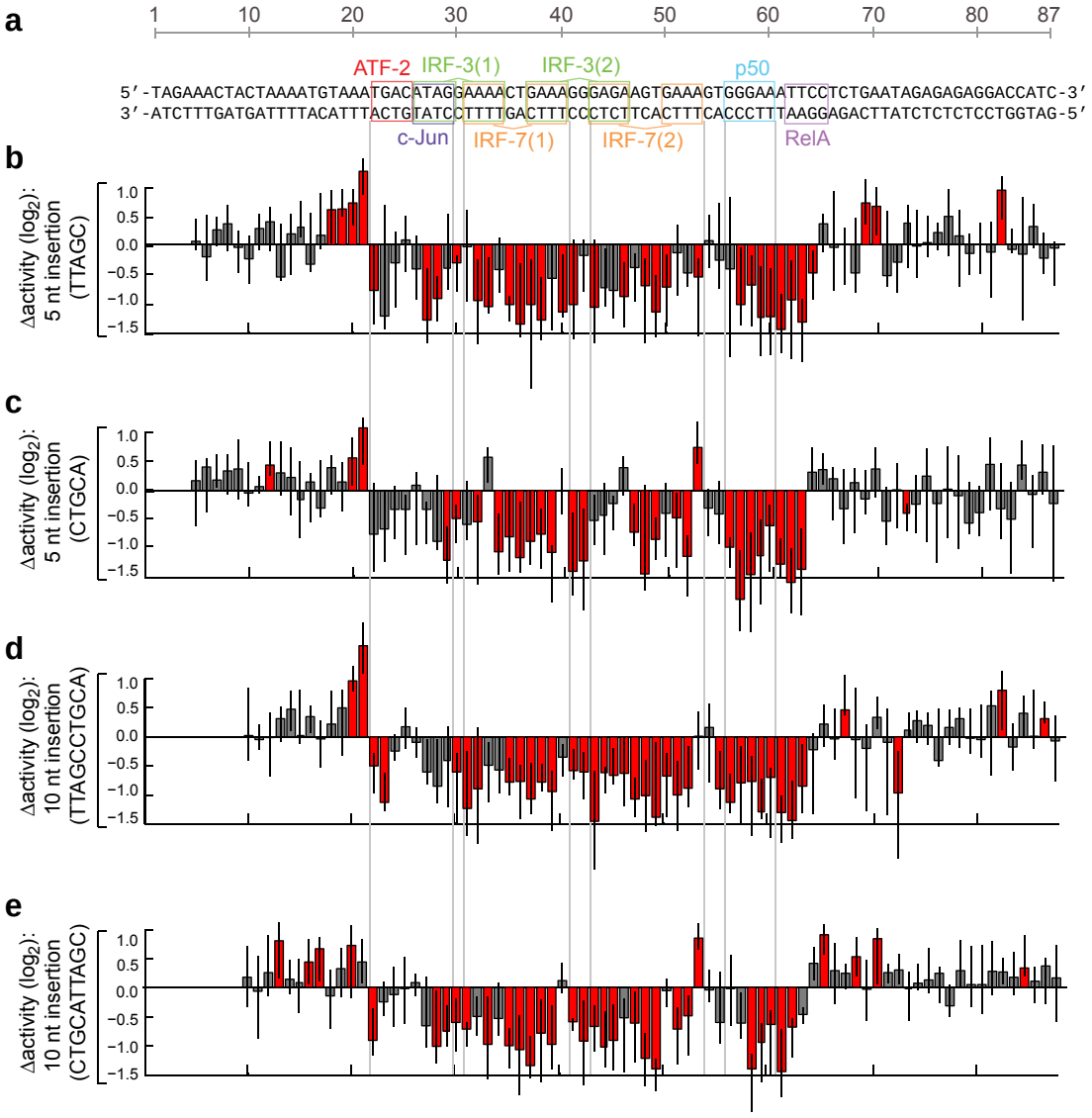
Supplementary Figure 4 – Single-hit scanning mutagenesis of the CRE with 8 consecutive substitutions. (a) The CRE sequence with known and putative transcription factor binding sites indicated. (b) Changes in induced activity due to 8 consecutive complement substitutions (G↔C, A↔T). (c) Changes in induced activity due to 8 consecutive non-complement transversion substitutions (G↔T, A↔C). (d) Changes in induced activity due to 8 consecutive transition substitutions (G↔A, T↔C). Each bar is located at the fourth nucleotide in the corresponding 8 nucleotide substitution. Error bars show the first and third quartile. Red indicates a significant change from wild-type (Mann-Whitney U-test, 5% FDR).



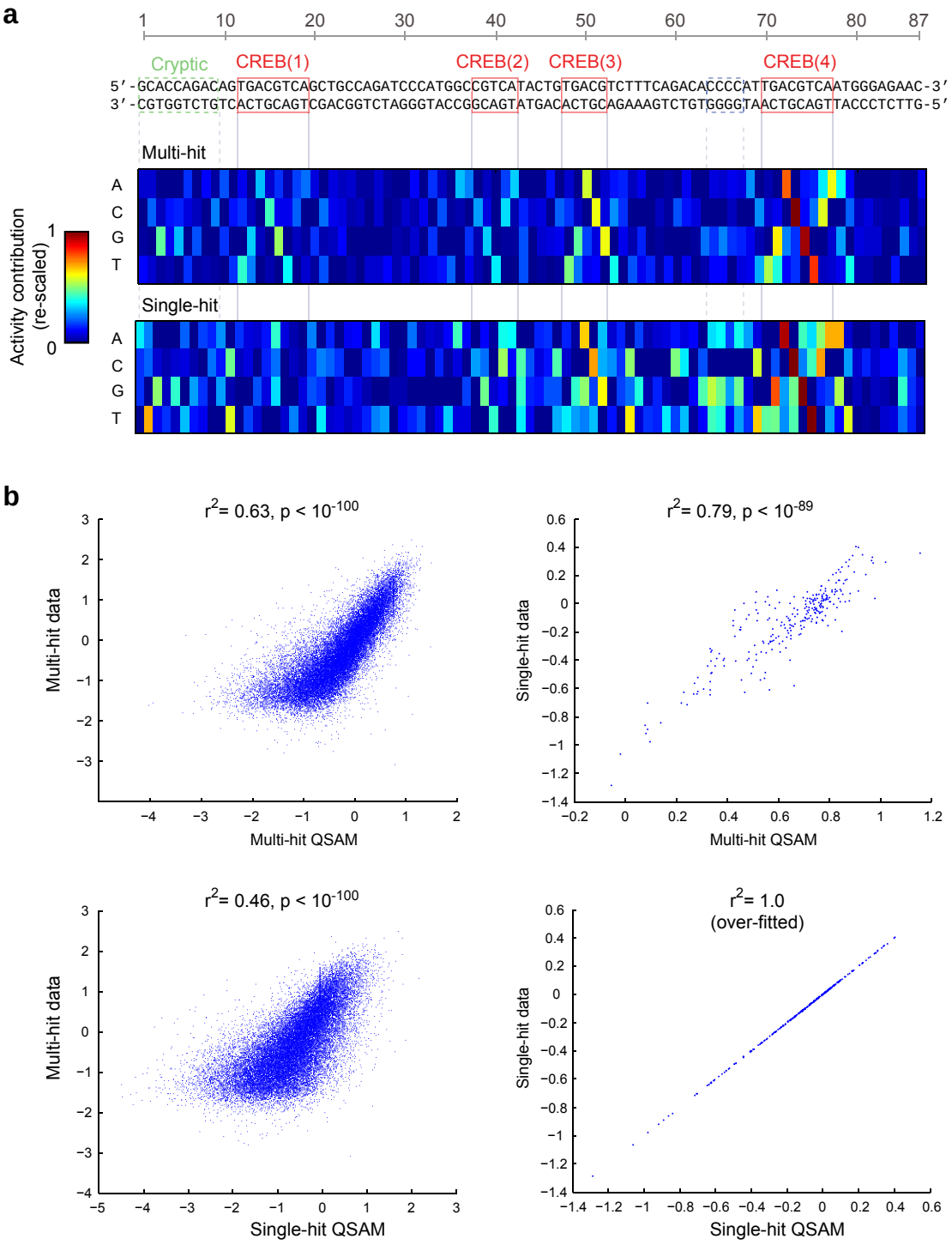
Supplementary Figure 5 – Single-hit scanning mutagenesis of the CRE with small insertions. (a) The CRE sequence with known and putative transcription factor binding sites indicated. (b) Changes in induced activity due to insertion of TTAGC between each pair of consecutive nucleotides. (c) Changes in induced activity due to insertion of CTGCA between each pair of consecutive nucleotides. (d) Changes in induced activity due to insertion of TTAGCCTGCA between each pair of consecutive nucleotides. (e) Changes in induced activity due to insertion of CTGCATTAGC between each pair of consecutive nucleotides. Each bar is located one nucleotide to the right of the insertion. Error bars show the first and third quartile. Red indicates a significant change from wild-type (Mann-Whitney U-test, 5% FDR).



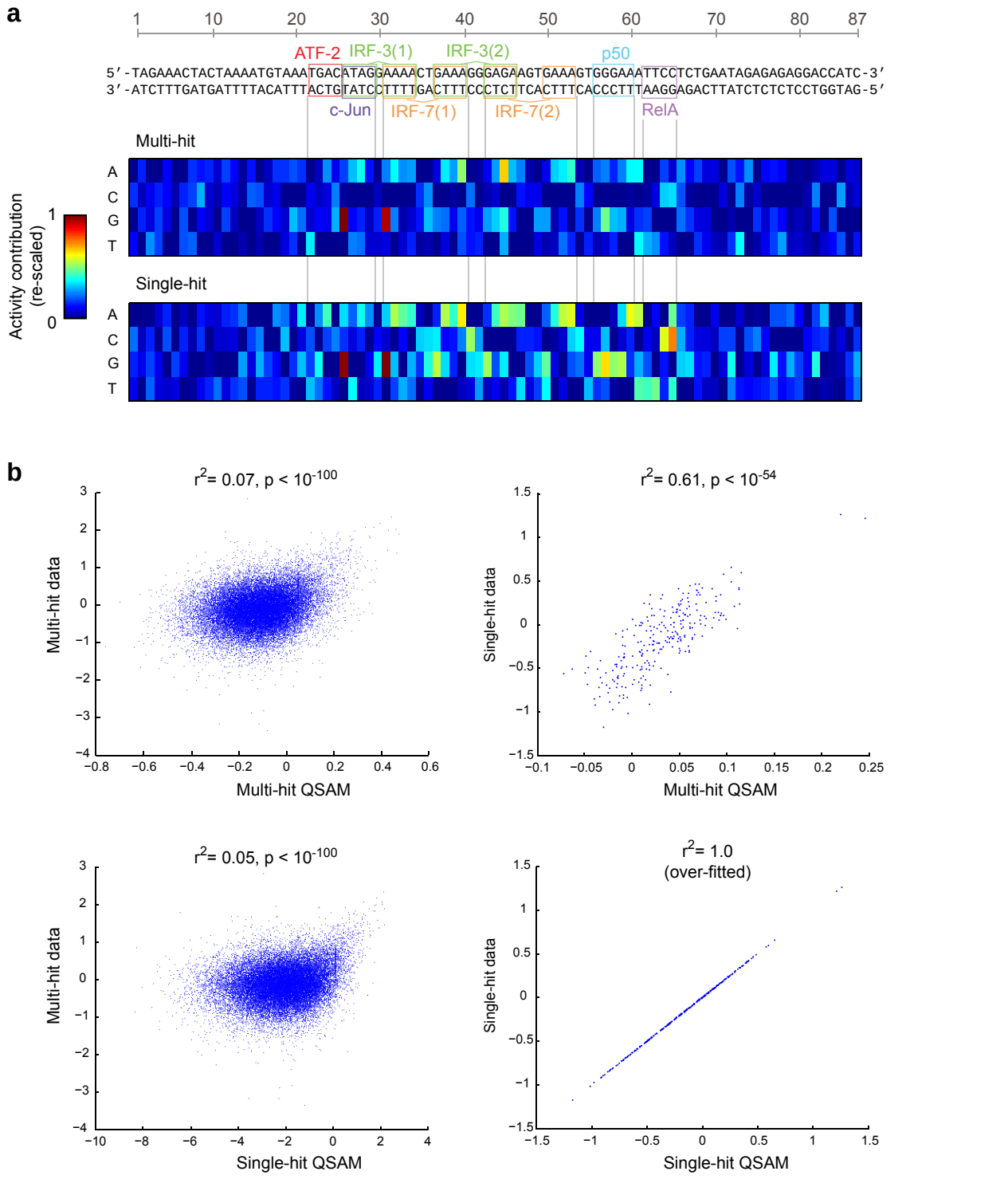
Supplementary Figure 6 – Single-hit scanning mutagenesis of the IFNB enhancer with 8 consecutive substitutions. (a) The IFNB enhancer sequence with known and putative transcription factor binding sites indicated. (b) Changes in induced activity due to 8 consecutive complement substitutions (G↔C, A↔T). (c) Changes in induced activity due to 8 consecutive non-complement transversion substitutions (G↔T, A↔C). (d) Changes in induced activity due to 8 consecutive transition substitutions (G↔A, T↔C). Each bar is located at the fourth nucleotide in the corresponding 8 nucleotide substitution. Error bars show the first and third quartile. Red indicates a significant change from wild-type (Mann-Whitney U-test, 5% FDR).



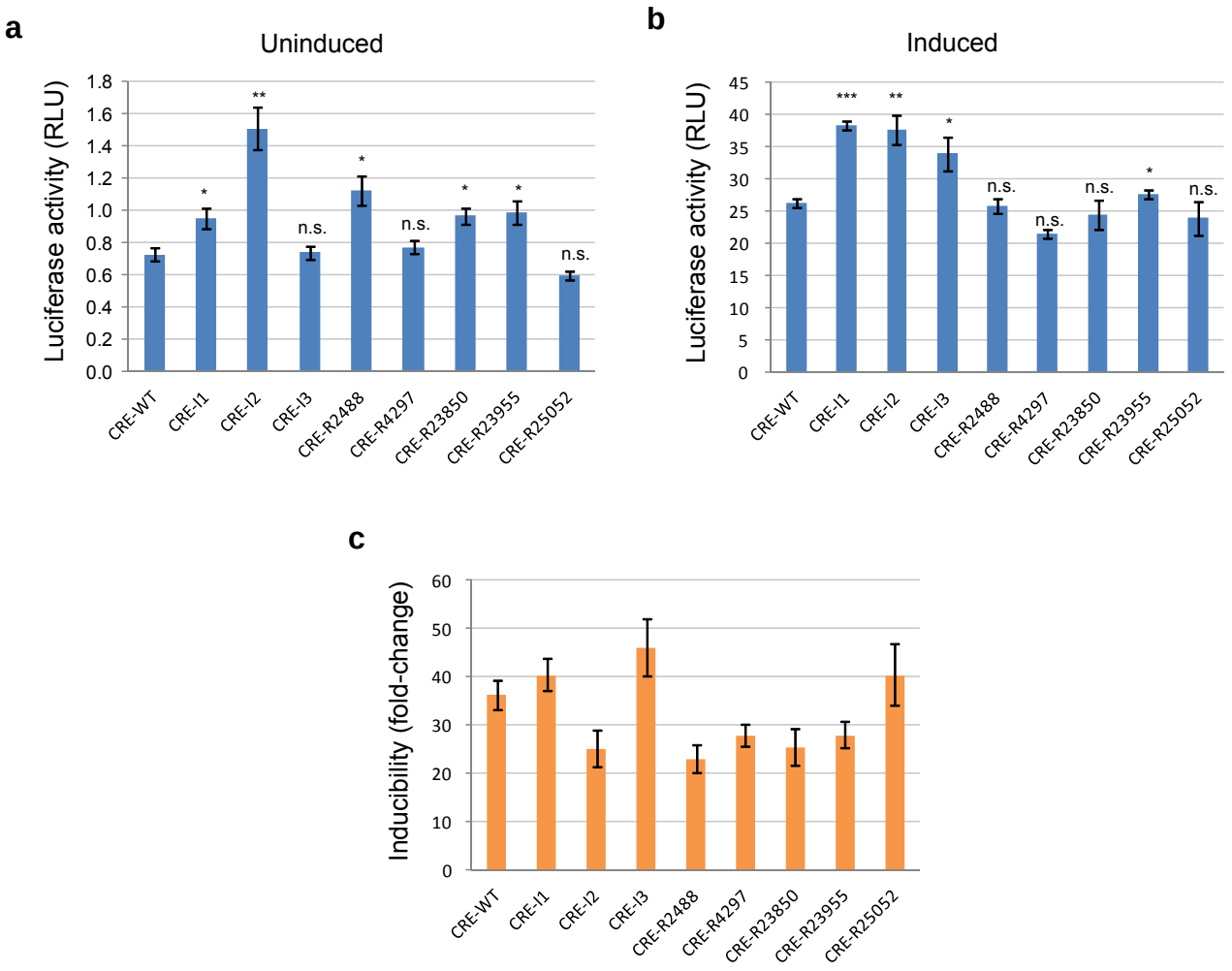
Supplementary Figure 7 – Single-hit scanning mutagenesis of the IFNB enhancer with small insertions. (a) The IFNB enhancer sequence with known and putative transcription factor binding sites indicated. (b) Changes in induced activity due to insertion of TTAGC between each pair of consecutive nucleotides. (c) Changes in induced activity due to insertion of CTGCA between each pair of consecutive nucleotides. (d) Changes in induced activity due to insertion of TTAGCCTGCA between each pair of consecutive nucleotides. (e) Changes in induced activity due to insertion of CTGCATTAGC between each pair of consecutive nucleotides. Each bar is located one nucleotide to the right of the insertion. Error bars show the first and third quartile. Red indicates a significant change from wild-type (Mann-Whitney U-test, 5% FDR).



Supplementary Figure 8 – Comparison of linear QSAMs of the induced CRE trained on multi-hit and single-hit data. (a) Visual representations of QSAMs trained on multi- (top) and single-hit (bottom) substitution data. The color in each entry represents the estimated additive contribution of the corresponding nucleotide to the log-transformed activity of the enhancer. The matrices are re-scaled such that the lowest entry in each column is zero and the highest entry anywhere is one. Both matrices are shown on the same scale. (b) Comparison of log-transformed QSAM-predicted and observed enhancer activities for models trained on multi-hit (top row) and single-hit (bottom row) data and evaluated on multi-hit (right column) or single-hit (left column) sequence variants. Note that the magnitudes of the activity estimates are depended on the specific set of assayed variants and therefore not directly comparable between single-hit and multi-hit data or QSAMs.



Supplementary Figure 9 – Comparison of linear QSAMs of the induced IFNB enhancer trained on multi-hit and single-hit data. (a) Visual representations of QSAMs trained on multi- (top) and single-hit (bottom) substitution data. The color in each entry represents the estimated additive contribution of the corresponding nucleotide to the log-transformed activity of the enhancer. The matrices are re-scaled such that the lowest entry in each column is zero and the highest entry in each matrix is one. The two matrices are not shown on the same scale. (b) Comparison of log-transformed QSAM-predicted and observed enhancer activities for models trained on multi-hit (top row) and single-hit (bottom row) data and evaluated on multi-hit (right column) or single-hit (left column) sequence variants. Note that the magnitudes of the activity estimates are depended on the specific set of assayed variants and therefore not directly comparable between single-hit and multi-hit data or QSAMs.



Supplementary Figure 10 – Inducibility of engineered vs. random variants. An alternative to our model-based optimization approach is to attempt to directly identify CRE variants with increased inducibility from the set of random variants that were tested by MPRA. In principle, this can be done by ranking the variants according to the ratio of their induced and uninduced activities and selecting the top variants – but the relatively high level of noise in individual (single-tag) measurements implies that this is unlikely to be a robust approach unless a significantly larger set of replicate experiments are performed. To directly compare the two approaches, we selected all CRE variants for which, in both replicate data sets, (1) the estimated ratio of their induced and uninduced activities were equal to or higher than the WT ratio, (2) the estimated induced activity was equal to or higher than the WT and (3) the estimated uninduced activity was lower than or equal to the WT. A total of 56 (~0.2%) of the assayed variants met these criteria. We ranked these variants according to their minimum estimated inducibility across the two replicates, synthesized the top five and then tested these against the WT and our three engineered variants using a luciferase assay. (a) Luciferase activity of the wild-type (WT), optimized and random CRE variants in untreated cells. (b) Luciferase activity of the same CRE variants in forskolin-treated cells. None of the top five random variants showed induced activities comparable to the engineered variants. (c) Inducibility of the CRE variants. Only one of the random variants (CRE-R25052) approached the level of inducibility seen for CRE-I1 and -I3, primarily because of its slightly reduced basal activity. Blue bars show mean activity across 3 replicates in the induced or uninduced states. Error bars show standard errors of the means (SE). All statistical comparisons are relative to WT in the same state; n.s., not significant; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; two-tailed t-test. Orange bars show the ratio of the corresponding induced and uninduced mean activities. Error bars show the range from (induced mean - induced SE)/(uninduced mean + uninduced SE) to (induced mean + induced SE)/(uninduced mean - uninduced SE).

Supplementary Notes

Here we describe and compare the various QSAMs that we fit to our data. QSAMs attempt to identify features of enhancer sequence that are predictive of the transcriptional activity of the regulated promoter. We considered several classes of models that instantiate, at varying levels of complexity, familiar ideas about how regulatory proteins can affect gene expression by binding to enhancer DNA. Some of these QSAMs are motivated by heuristic considerations while others, as in ref. 9, instantiate specific thermodynamic models.

QSAMs were fit to both CRE and IFNB data gathered in both inducing and non-inducing conditions. Specific formulas defining these QSAMs are displayed in Supplementary Table 5, and information about model performance is displayed in Supplementary Table 6. The models were in all cases fit to the copious multi-hit data. The quality of fit to this training data, as well as model performance on the sparser but independent single-hit data, was used to evaluate each QSAM's predictive power.

One of two objective functions, least squares or maximal mutual information, was used to optimize the parameters of each QSAM. For least squares, we sought parameters that minimized the sum of square deviations between model predictions and measured log activities. Least-squares-optimal parameters can easily be found using linear regression when a model's predictions depend linearly on these parameters. However, least squares has a maximum likelihood interpretation only when experimental noise is uniformly Gaussian.

In some cases, we also sought parameters that maximized the mutual information between model predictions and measured activities (ref. 9). Mutual information is equivalent, in the large data limit, to maximum likelihood whenever the quantitative form of experimental noise is uncertain (Kinney et al., 2007). Because of this, maximal mutual information is a more meaningful objective function than least squares when fitting QSAMs to MPRA data. However, mutual information cannot be maximized analytically. We therefore used the computationally intensive parallel tempering Monte Carlo (PTMC) algorithm from ref. 9 to infer parameter values when using this objective function. We also used PTMC to perform least squares optimization on models for which simple linear regression could not be applied. MATLAB files containing the fitted model parameters are available from the authors upon request.

In general the CRE models performed much better than the IFNB models on their respective multi-hit training data, while both performed similarly on their respective single-hit test data. We believe this difference is largely due to the IFNB enhancer, with its more compact enhanceosome structure, being more sensitive to multiple mutations than is the billboard-like CRE enhancer. Still, it is surprising that IFNB models that perform poorly on their multi-hit training data fit the single-hit test data so well.

Objective functions and optimization strategies

Linear: A linear QSAM, F_{lin}, is defined by parameters A_{bi} representing additive contributions of the different bases *b* at each enhancer position *i* to log transcriptional activity. This is a generalization of a widely used method of assessing the effect of a single transcription factor acting at a single DNA binding site to the case where multiple transcription factors assemble on an extended enhancer. The model has 4 x 87 = 348 A_{bi} parameters, but because one of the four

bases must be present at every position there are only 1+3 x 87=262 independent degrees of freedom. The primary virtue of linear QSAMs is their simplicity, but it is not a priori obvious that such models can capture the complex response of multi-site enhancers. Nonetheless, for induced CRE and IFNB, linear QSAMs performed nearly as well or better than the more complex models we fit.

We also defined a "sites-only" linear QSAM in which the A_{bi} parameters were fixed at zero for positions *i* outside identified transcription factor binding sites. This simplification was motivated by the assumption that discrete binding sites dominate model predictions. Such a model was fit to the induced CRE data, with nonzero positions restricted to the four CREB binding sites shown in Fig. 4 (but including two extra nucleotides included on each side of CREB site 4). Doing this reduced the number of model parameters from 262 to 90.

Heuristic linear: The heuristic linear QSAM, F_{hlin}, assumes that the effect of a binding site on log transcription is entirely determined by whether or not that site has at least one mutation with respect to wild type. When at least one mutation is present, a contribution A_s is added to log activity. An advantage of this model is the very small number of parameters needed to describe it. Even with only 7 parameters (4 CREB sites, 2 "cryptic" sites and 1 overall constant), this model was able to achieve an r^2 value equal to 85% (65%) of that achieved by the linear QSAM on the induced CRE training (test) data.

Linear-nonlinear: In the linear-nonlinear QSAM, F_{lnl}, a sigmoidal transformation specified by parameters *B* and *C* is applied to the prediction of a linear QSAM having parameters A_{bi} as defined above. This type of model is widely used to describe systems where multiple inputs are combined to generate a response that interpolates monotonically, but not linearly, between minimum and maximum values. For the induced CRE data, this two-parameter nonlinearity increased r^2 by 16% as compared to the linear QSAM. Because monotonic transformations have no effect on mutual information, this quantity was not meaningfully affected. Nevertheless, this linear-nonlinear model has the virtue of being able to predict an upper limit to the expression level that can be achieved by reengineering the enhancer sequence.

Nearest neighbor dinucleotide: In modeling the binding specificity of individual transcription factors, the simple linear model can sometimes be improved upon -- at the price of substantially increasing the number of parameters -- by allowing for dependence on nucleotide pairs. To limit model complexity, it is convenient (and physically reasonable) to limit attention to nearest neighbor dinucleotides. We therefore defined a nearest neighbor dinucleotide QSAM, F_{nn}, in which parameters A_{bci} give the additive contribution to log activity of the dinucleotide consisting of base *b* at position *i* and base *c* at position *i+1.* The simple mononucleotide model is included in this formulation as a special case. When applied to the induced CRE and IFNB data, the nearest neighbor dinucleotide model performed as well as, or better than, the simple linear model on both the training and test sets.

Arbitrary dinucleotide: To explore whether improvements in fit over the nearest neighbor model could be achieved with non-nearest neighbor interactions, we defined a hybrid dinucleotide QSAM, F_{arb}, consisting of a linear QSAM, defined by parameters A_{bi} for all positions *i*, together with dinucleotide contributions B_{bcij} describing interactions between bases *b* and *c* respectively occurring at selected pairs of positions *i* and *j*. To avoid overfitting due to an explosion of parameters, we limited nonzero B_{bcij} values to at most 40 pairs of positions

(*i,j*). Finding the 40 best pairs of positions, and the associated optimal parameter values, presented a combinatorial optimization problem, which we approached using PTMC. As the data in Supplementary Table 6 indicate, these models performed similarly to the nearest neighbor dinucleotide models.

Heuristic interaction: The heuristic interaction QSAM, F_{hint}, consists of a linear QSAM with parameters A_{bi}, a heuristic linear model having parameters B_s with a mutation threshold of 2, and additional interaction terms C_{st} which contribute when both sites *s* and *t* have at least 1 mutation. For the CRE model, the 6 sites annotated in Fig. 4 were used. For the IFNB model, the 8 boxed regions (representing both sites and half-sites) were treated as separate sites. These models have the advantage of implementing interactions between proteins in a way that allows model parameters to be analytically inferred using linear regression. Modest improvements in fit as compared to the linear model were obtained.

Thermodynamic: The thermodynamic QSAM for the induced CRE enhancer, F_{therm}, is based on previously published models (Bintu et al., 2005) in which transcriptional activity is assumed to be proportional to the equilibrium occupancy of the RNA polymerase site. Given a specific picture of how the regulatory proteins assemble on the enhancer, the polymerase site occupancy is determined by a partition function involving the binding free energies of transcription factors to their respective sites in the enhancer and the interaction free energies between both bound proteins and between these bound proteins and the polymerase. This sort of model has a complicated formula and cannot be fit with linear regression, but is important because it relates transcriptional response to a well-defined physical picture of molecular interactions. If a physically accurate model can be identified, it might facilitate the prediction of phenomena that could otherwise only be fit empirically. We attempted to fit one such model to the CRE data. This was not done for the IFNB data because the overlapping binding sites made it less clear what the structure of a reasonable thermodynamic model of that enhancer might be. In the formula for F_{therm}, ε_s represents the binding free energy to site *s*, in natural thermal energy units ($k_B T$), of the cognate CREB protein. This free energy depends on sequence through a linear QSAM with parameters A_{bi}^{s}, and these parameters are nonzero only within the extent of site *s* (defined as for the linear sites-only CRE model). The ω parameters describe the energetic interactions between DNA-bound CREB proteins: ω_{st} is the interaction between proteins bound to sites *s* and *t*, ω_{stu} is the total interaction free energy between three proteins bound to sites *s, t,* and *u* and ω_{1234} is the total interaction free energy when all four CREB proteins are bound. Note that this model allows for irreducible 3-protein and 4-protein interactions, in addition to pairwise interactions between proteins. A constant of proportionality τ relates transcription to an effective RNA polymerase occupancy, which is determined by a protein-DNA interaction free energy ε_P, as well as interaction free energies γ_s, γ_{st}, γ_{stu} and γ_{1234} between RNA polymerase and the various possible CREB-enhancer complexes. Model parameters were fit using PTMC. This model fit the training set reasonably well but performed significantly worse than the simple linear model when predicting the single-hit test data.

References

Bintu L, Buchler N, Garcia H, Gerland U, Hwa T, Kondev J, Phillips R. Transcriptional regulation by the numbers: applications. Curr. Opin. Genet. Dev. (2005) vol. 15 (2) pp. 125-35.

Kinney J, Tkacik G, Callan C. Precise physical models of protein-DNA interaction from high-throughput data. Proc. Natl. Acad. Sci. U S A (2007) vol. 104 (2) pp. 501-6.

Strong S, Koberle R, de Ruyter van Steveninck R, Bialek W. Entropy and Information in Neural Spike Trains. Phys. Rev. Lett. (1998) vol. 80 (1) pp. 197-200.

Formula for log expression from enhancer sequence σ	Parameters
$F_{lin}(\sigma) = \Sigma_{b,i}\, A_{bi}\, x_{bi}$	A_{bi}
$F_{lnl}(\sigma) = log\left\{B + C\left[1 + exp\left(\Sigma_{b,i}\, A_{bi}\, x_{bi}\right)\right]^{-1}\right\}$	$A_{bi}\,, B, C$
$F_{hlin}(\sigma) = \mathrm{B} + \Sigma_s\, A_s\, x_s^{(1)}$	$A_s\,, B$
$F_{nn}(\sigma) = \Sigma_{b,c,i}\, A_{bci}\, x_{b,i}\, x_{c,i+1}$	A_{bci}
$F_{arb}(\sigma) = \Sigma_{b,i}\, A_{bi}\, x_{bi} + \Sigma_{b,c,i,j}\, B_{bcij}\, x_{b,i}\, x_{c,j}$	$A_{bi}\,, B_{bcij}$
$F_{hint}(\sigma) = \Sigma_{b,i}\, A_{bi}\, x_{bi} + \Sigma_s\, B_s\, x_s^{(2)} + \Sigma_{s<t}\, C_{st}\, x_s^{(1)} x_t^{(1)}$	$A_{bi}\,, B_s\,, C_{st}$
$F_{therm}(\sigma) = log\left(\tau\, \frac{Z_{on}}{Z_{on}+Z_{off}}\right)$ where $Z_{on} = e^{-\epsilon_P}\,[1 + \Sigma_s\, e^{-\epsilon_s-\gamma_s} + \Sigma_{s<t}\, e^{-\epsilon_s-\epsilon_t-\gamma_{st}-\omega_{st}}$ $+\Sigma_{s<t<u}\, e^{-\epsilon_s-\epsilon_t-\epsilon_u-\gamma_{stu}-\omega_{stu}} + e^{-\epsilon_1-\epsilon_2-\epsilon_3-\epsilon_4-\omega_{1234}-\gamma_{1234}}]$ $Z_{off} = [1 + \Sigma_s\, e^{-\epsilon_s} + \Sigma_{s<t}\, e^{-\epsilon_s-\epsilon_t-\omega_{st}}$ $+\Sigma_{s<t<u}\, e^{-\epsilon_s-\epsilon_t-\epsilon_u-\omega_{stu}} + e^{-\epsilon_1-\epsilon_2-\epsilon_3-\epsilon_4-\omega_{1234}}]$ $\epsilon_s = \Sigma_{b,i}\, A_{bi}^s\, x_{bi}$	A_{bi}^s $\omega_{st}\,, \omega_{stu}\,, \omega_{1234}$ $\gamma_s\,, \gamma_{st}\,, \gamma_{stu}\,, \gamma_{1234}$ $\tau\,, \epsilon_P$

Supplementary Table 5: Specific formulas for the various QSAMs described in the Supplementary Notes. Parameter indices are defined as follows: *b,c* ∈ { A,C,G,T } index different nucleotides; *i, j* ∈ {1, 2, .. , 87} index positions within the mutagenized enhancers; *s, t, u* index protein binding sites (see Supplementary Notes for details). As described in Methods, x_{bi} = 1 (0 otherwise) if base *b* occurs at position *i* in the sequence σ. In the heuristic models, $x_s^{(n)}$ = 1 (0 otherwise) if site *s* exhibits *n* or more mutations from wild type. ϵ_P is the RNAP binding free energy to its site, and ϵ_s is the binding free energy of a transcription factor (in this case CREB) to one of its specific binding sites indexed by *s*. The logic behind the different expressions is explained in the Supplementary Notes.

Multi-hit training dataset	Model description	Formula	No. of par-ameters	Objective function	Fitting method	r^2 on multi-hit data	r^2 on single-hit data	MI (bits) on multi-hit data
CRE, uninduced	linear	F_{lin}	262	LS	LR	0.359	-	$0.355 \pm .007$
CRE, induced	linear	F_{lin}	262	LS	LR	0.630	0.792	$0.826 \pm .008$
CRE, induced	linear	F_{lin}	262	MMI	PTMC	0.621	0.811	$0.861 \pm .008$
CRE, induced	linear (sites only)	F_{lin}	90	LS	LR	0.559	0.652	$0.677 \pm .006$
CRE, induced	linear/nonlinear	F_{lnl}	264	LS	LR	0.723	0.825	$0.849 \pm .008$
CRE, induced	heuristic linear	F_{hlin}	7	LS	LR	0.526	0.528	$0.513 \pm .007$
CRE, induced	n.n. dinucleotide	F_{nn}	1036	LS	LR	0.681	0.797	$0.901 \pm .007$
CRE, induced	arb. dinucleotide	F_{arb}	622	LS	PTMC	0.696	0.812	$0.886 \pm .006$
CRE, induced	heuristic int'n	F_{hint}	283	LS	LR	0.676	0.816	$0.875 \pm .008$
CRE, induced	thermodynamic	F_{therm}	122	LS	PTMC	0.655	0.688	$0.717 \pm .007$
IFNB, uninduced	linear	F_{lin}	262	LS	LR	0.021	-	$0.017 \pm .001$
IFNB, induced	linear	F_{lin}	262	LS	LR	0.071	0.616	$0.058 \pm .002$
IFNB, induced	linear	F_{lin}	262	MMI	PTMC	0.062	0.596	$0.074 \pm .003$
IFNB, induced	heuristic linear	F_{hlin}	9	LS	LR	0.034	0.425	$0.064 \pm .004$
IFNB, induced	n.n. dinucleotide	F_{nn}	1036	LS	LR	0.102	0.639	$0.074 \pm .002$
IFNB, induced	arb. dinucleotide	F_{arb}	622	LS	PTMC	0.104	0.607	$0.073 \pm .003$
IFNB, induced	heuristic int'n	F_{hint}	298	LS	LR	0.084	0.634	$0.064 \pm .003$

Supplementary Table 6: Summary of the QSAMs fit to multi-hit MPRA data. For each QSAM we report the following: the data set modeled; a description of the model that was fit (linear, heuristic linear, linear covering specific sites only, linear-nonlinear, nearest neighbor dinucleotide, arbitrary dinucleotide, heuristic interaction, and thermodynamic); the specific QSAM formula as described in Supplementary Table 5 and Supplementary Notes; the number of independent parameters fit; the objective function used for model optimization, i.e. least squares (LS) or maximal mutual information (MMI); the computational method used to optimize parameters, i.e. linear regression (LR) or parallel tempering Monte Carl (PTMC); the squared Pearson correlation r^2 achieved by the model on the multi-hit training set and the single-hit test set (all values shown are highly significant, i.e. $p < 10^{-100}$); the mutual information between model predictions and multi-hit measurements, computed using the method of Strong et al., 1998. The induced CRE models were all fit to replicate 2 of the CRE multi-hit dataset (Main Text).