Genome-wide analysis of CRISPR perturbations indicates that enhancers act multiplicatively and without epistatic-like interactions

Jessica Zhou[†1,2], Karthik Guruvayurappan[†1,3,4], Hsiuyi V. Chen[1,5], Aaron R. Chen[1,6], Graham McVicker*[1]

[1]Integrative Biology Laboratory, Salk Institute for Biological Studies, 10010 N Torrey Pines Rd, La Jolla, CA, 92037, USA

[2]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, 92093, USA

[3]School of Biological Sciences, University of California San Diego, La Jolla, CA, 92037, USA

[4]Halicioglu Data Science Institute, University of California San Diego, La Jolla, CA, 92093, USA

[5]A*STAR Infectious Diseases Labs, Agency for Science, Technology and Research (A*STAR), Singapore 138648, Singapore

[6]Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, 11794, USA

*Correspondence: gmcvicker@salk.edu

[†] These authors contributed equally to this work.

Abstract

A single gene may be regulated by multiple enhancers, but how they work in concert to regulate transcription is poorly understood. Prior studies have mostly examined enhancers at single loci and have reached inconsistent conclusions about whether epistatic-like interactions exist between them. To analyze enhancer interactions throughout the genome, we developed a statistical framework for CRISPR regulatory screens that utilizes negative binomial generalized linear models that account for variable guide RNA (gRNA) efficiency. We reanalyzed a single-cell CRISPR interference experiment that delivered random combinations of enhancer-targeting gRNAs to each cell and interrogated interactions between 3,808 enhancer pairs. We found that enhancers act multiplicatively with one another to control gene expression, but our analysis provides no evidence for interaction effects between pairs of enhancers regulating the same gene. Our findings illuminate the regulatory behavior of multiple enhancers and our statistical framework provides utility for future analyses studying interactions between enhancers.

Introduction

Cis-regulatory elements (CREs), which include enhancers, direct transcription and shape cellular identity, growth, and biological function. Most genes are regulated by multiple enhancers[1,2], yet we lack a detailed understanding of how enhancers act together to influence gene expression. When multiple enhancers for a gene are active in the same cell type, it is often assumed that they act additively—that is, their combined effect is equal to the sum of their individual effects[3]. However, enhancers may also act non-additively, and interactions between regulatory elements may modulate their effects on gene expression[3–10].

To date, most studies of regulatory elements have examined their effects independently, and studies of regulatory element interactions have focused on a small number of loci and have reached contradictory conclusions[4–8]. For example, a study of the *α-globin4* gene found that its expression is best explained by simple additivity between constituent elements of its super enhancer[7]. In addition, a study that deleted three constituent enhancers of a super enhancer for *Wap3* found no evidence of synergy between the studied enhancers and differences in the magnitudes of effect that each constituent enhancer had on the target gene, with all three enhancers necessary to induce full induction of the gene during pregnancy[8]. Reexamination of both of these super enhancer datasets found that the effects of the constituent enhancers on the target genes were best described by a logistic generalized linear model (GLM), but that beyond this there was no significant evidence for interactions between enhancers[5]. Contrary to these findings, a recent study of the *MYC* locus described both synergistic and additive enhancer-enhancer interactions, where enhancers separated from one another by larger genomic distances are more likely to have synergistic interactions and enhancers located closer to one another are more likely to have additive interactions[9]. Altogether, these studies have been limited to the examination of a small number of genes and enhancers and their results are difficult to interpret due to their conflicting findings and the lack of explicit definitions and consistent terminology for different models of enhancer activity.

Recent technological advances have made it possible to couple CRISPR-induced genome perturbations with single-cell RNA sequencing[10–16]. Because single-cell CRISPR perturbation experiments can induce multiple genomic perturbations in each cell, they can be used to identify interactions, or epistatic-like effects, between targeted sequences. Specifically, when such experiments are designed to target regulatory elements, they yield cells wherein multiple regulatory elements are simultaneously perturbed. This feature of these datasets can be

harnessed to measure the combined effects of multiple regulatory elements, such as enhancers, on gene expression.

Here, we present GLiMMIRS (Generalized Linear Models for Measuring Interactions between Regulatory Sequences), a statistical analysis framework that can be applied to single cell CRISPR perturbation experiments to quantify the effects of multiple regulatory elements on gene expression and identify interactions between them. GLiMMIRS has both data simulation and modeling components and importantly, accounts for variations in gRNA efficiency, a key variable in the interpretation of CRISPR experiments that has typically been ignored when analyzing data from them[10,11,17–19]. We applied GLiMMIRS to a multiplexed, single-cell CRISPR interference (CRISPRi) experiment that targeted putative enhancers in K562 cells[11]. We conducted a power analysis, which found that this dataset provides sufficient power to detect strong interactions between enhancers, but low power to detect weak interactions. Our analysis strongly supports a model in which most enhancers act multiplicatively to affect the expression of their target genes, but we find no evidence for the presence of additional interactions between them.

Results

Variation in guide efficiency should be considered when estimating enhancer effects from CRISPR perturbations

To analyze the combined effect of multiple enhancers on gene expression, we leveraged data from a multiplexed, single-cell CRISPRi screen performed in K562 cells[11] (Fig. S1). In this screen, gRNAs were designed to target putative enhancers and enhancer-gene pairs were

identified by associating gRNAs with differences in the expression levels of nearby genes (Fig. 1a). Due to the high multiplicity of infection (MOI) used in this experiment, many gRNAs targeting different enhancers are present within each cell (Fig. 1a). While the high MOI was intended to increase power to detect enhancer-gene pairs, we leveraged this feature of the dataset to quantify how pairs of enhancers regulate the expression of common target genes and to detect potential interaction effects between them (Fig. 1b). In particular, we focused on cells which received gRNAs perturbing pairs of enhancers within 1Mb of the same gene, which we designate as the putative target gene[20–22].

Most enhancers in this dataset were targeted by two different gRNAs. The original study did not distinguish between gRNAs that targeted the same enhancer; however, it is important to consider differences in guide efficiency when examining the combined effects of multiple enhancers in a CRISPR screen. This is because the joint effect of both enhancer perturbations can appear smaller or larger than expected if one of the targeting guides has low efficiency. To illustrate this concept, we examined two enhancers of *NMU*, which were among the most significant enhancer-gene pairs discovered by the original study. We performed CRISPRi experiments to perturb the enhancers of *NMU* using guide designs from the paper (Fig. 1c, Supplementary Dataset 1). We quantified gene expression following each perturbation using reverse transcription-quantitative polymerase chain reaction (RT-qPCR) and found that one of the two gRNAs targeting the first enhancer (enhancer A, gRNAs A1 and A2) caused much larger reductions in *NMU* expression (Fig. 1d). Differences in guide efficiency like the ones we observed for gRNAs A1 and A2 can give false signals of epistatic-like interactions if different guides targeting the same enhancers are treated as equivalent. For example, if by chance most of the cells which contained guides targeting both enhancer A and B contained gRNA A1 (rather than the inefficient A2), then the joint effect of targeting both enhancers could be greatly overestimated.

To examine variation in guide efficiency, we estimated the efficiency of the gRNAs included in the experiment using GuideScan 2.0[23]. Predicted guide efficiency varies substantially across the guide library (Fig. 1e), indicating that it is important to consider this variable when analyzing enhancer interactions using this dataset.

GLiMMIRS modeling and simulation framework for enhancer effects and CRISPR screens

We developed GLiMMIRS, a dual modeling and simulation framework for analyzing data from CRISPR screens to evaluate the effects of regulatory elements on target genes. We first sought to evaluate the utility of a model that incorporates guide efficiency by testing a simple model that considers just one enhancer acting on one gene, which we refer to as the GLiMMIRS baseline model (GLiMMIRS-base) (Fig. 2a). For each enhancer and gene of interest, we fit a generalized linear model (GLM) with a negative binomial distribution to the observed scRNA-seq counts. The predictor of interest in this model is the probability that the enhancer is perturbed, $X_{perturb}$. We calculated the value of $X_{perturb}$ using the efficiencies of the targeting sgRNAs which are present in each cell (see Methods). In addition to considering guide efficiency, we also included covariates to account for cell cycle (Fig. S2)[24] and other relevant variables (see Methods).[24]

To evaluate the performance of GLiMMIRS-base, we developed a simulation framework for single-cell CRISPRi screens (Fig. 2b, see Methods) and used it to generate a dataset resembling the Gasperini et al.[11] experimental dataset, with gRNAs targeting the enhancers of predetermined target genes (Fig. S3-10). This is the simulation component of GLiMMIRS (GLiMMIRS-sim), designed for evaluation of our baseline scenario. This provided us with a set of ground truth coefficient values which we could use to benchmark our model. We generated

scRNA-seq counts for each gene by sampling from a negative binomial distribution defined by gene-specific parameters (Fig. 2b, Methods). We then fit our baseline model to the simulated count data and compared the estimated model coefficients to the "ground truth" values used in the simulation. The coefficient of determination (R^2, see Methods) between the estimated enhancer effect coefficients and the ground truth values was higher ($R^2 = 0.657, MSE = 0.52$, Pearson's $r = 0.862$) when we implemented our model with a perturbation probability, $X_{perturb}$ (see Methods), compared to a model that used a simple indicator value representing the presence or absence of targeting gRNAs for the enhancer being modeled ($R^2 = -0.449, MSE = 2.195$, Pearson's $r = 0.811$) (Fig. 2c, Table S1). This is because the model that uses the indicator value systematically underestimates the enhancer effect, by assuming that the presence of a gRNA completely inhibits the target site even when the gRNA has low efficiency. We also generated "noisy" guide efficiency values with GLiMMIRS-sim (Fig. S11) to account for uncertainty in predicted guide efficiencies[25–28]. These noisy guide efficiency values were calculated as a function of true guide efficiency and a noise-controlling constant D (see Methods), where D is inversely related to the amount of noise in the efficiency value. We found that fitting to the simulated data using the values of $X_{perturb}$ computed from the noisy guide efficiencies still performed better than an indicator variable under low noise ($D = 100$; $R^2 = 0.642, MSE = 0.542$, Pearson's $r = 0.854$) and medium noise ($D = 10$; $R^2 = 0.499, MSE = 0.752$, Pearson's $r = 0.789$). Under a simulation with very noisy guide efficiencies, the coefficient estimates correlated very poorly with the ground truth due to the presence of some extreme outliers ($D = 1$; $R^2 = -6107.575, MSE = 8937.909$, Pearson's $r = 0.03$) (Fig. S12, Table S2). In summary, accounting for guide efficiency improves the accuracy in coefficient estimates and is robust to moderate noise in the guide efficiency estimates.

We then applied GLiMMIRS-base to the Gasperini et al.[11] dataset and compared the p-values obtained from our GLM to those from the published analysis. We detected a similar number of significant enhancer-gene pairs (588 validated by GLiMMIRS-base out of the 664 reported by Gasperini et al.[11]), but with lower p-values for most of the highly significant pairs. Our p-values are well-calibrated, and when applied to permuted data (where gRNA identities are assigned to different cells) the p-value distribution matches the null expectation (Fig. 2d, Fig. S13). These results establish that accounting for guide efficiency offers advantages over an indicator variable for gRNA presence, and also suggest that including cell cycle scores as additional covariates in GLiMMIRS may further boost power to detect enhancer-gene pairs. Having established the validity of our approach for the simpler scenario of single enhancers acting on single genes, we proceeded to study the effects of pairs of enhancers on single genes.

Detection of interactions between pairs of enhancers with GLiMMIRs

To model the effects of pairs of enhancers on a target gene, we modified GLiMMIRS-base by replacing the enhancer term $\beta_{enhancer}X_{perturb}$ with three new terms to represent: 1) the first enhancer in the pair ($\beta_A X_A$); 2) the second enhancer in the pair ($\beta_B X_B$); and 3) an epistatic-like interaction between the enhancers ($\beta_{AB} X_{AB}$). As with the baseline model above, the values of the X_A and X_B predictors are the probability that the respective enhancers are perturbed. Likewise, the value of X_{AB} is the probability that both enhancers are simultaneously perturbed, and is also estimated from the predicted guide efficiencies. This new model, which evaluates interaction effects between pairs of enhancers, is the GLiMMIRS interactions model (GLiMMIRS-int).

To identify pairs of enhancers to test in the experimental data, we identified target sites from the Gasperini et al. experiment, or putative enhancers, which were both located within 1MB of a common target gene as testable enhancer pairs. We found a total of 795,616 testable enhancer pairs from the set of enhancers targeted in the Gasperini et al.[11] study. Since cells must contain perturbations of multiple enhancers to determine whether there is an interaction effect between the enhancers, we evaluated the number of cells containing gRNAs targeting both enhancers within testable pairs. While the majority of testable enhancer pairs are simultaneously perturbed in fewer than 10 cells, several hundred enhancer pairs are simultaneously targeted in at least 10 cells (Fig. 3a).

We performed a power analysis to evaluate our power for detecting interactions at different MOI, represented by different values of λ (see Methods) (Fig. S14a), and different magnitudes of (fixed) interaction effect sizes (Fig. 3b-c, Table S3-4). To do this, we used GLiMMIRS-sim to generate ground truth data for evaluating interactions between enhancer pairs (see Methods). In our power analysis, we defined positive cases as enhancer pairs with a true interaction effect on their target gene and negative cases as pairs of enhancers with individual effects on the target gene but no interaction effect. As expected, we observed that power to detect interaction effects scales with the magnitude of the interaction effect size as well as the MOI, which controls the number of testable cells (Fig. S14b-c). Our power analysis indicated that we have low power (<25%) to detect interactions of small effect sizes (<2), particularly at low MOIs ($\lambda = 15,25$). This is likely due to the fact that the number of testable cells, or cells containing gRNAs targeting both enhancers in a testable pair, are very low (Fig. 3a). The scenario $\lambda = 15$ from our power analysis most closely resembles the empirical data (Fig. S14, Fig. 3a), indicating that we have moderate power (>50%) to detect large interaction effects (≥ 5) and low power to detect smaller effects. Thus, with the experimental dataset analyzed in our study, we expect that we will have

sufficient power to detect strong interaction effects between enhancers, but be unable to draw conclusions about the presence or absence of weak interactions.

Enhancers act multiplicative to control gene expression, but analysis of CRISPR perturbations provide no evidence for for interactions

We next applied GLiMMIRS to the Gasperini et al.[11] CRISPRi dataset to study enhancer-enhancer interactions. To survey for interactions between enhancers, we defined two sets of testable enhancer pairs throughout the genome: a smaller, high-confidence set and a larger, unbiased set of testable pairs (see Methods). The high-confidence set consisted of 330 testable pairs and corresponding target genes where each of the individual enhancers had a previously reported regulatory effect on the target gene. The unbiased set consisted of all testable pairs that were perturbed in a minimum of 20 cells, regardless of any previously established relationship between each individual enhancer and the target gene. The unbiased set contained 3,808 enhancer pairs and target genes.

We first examined whether the combined effects of multiple enhancers on gene expression were better described by a multiplicative or additive model. To this end, we fit two versions of GLiMMIRS-int to the 330 enhancer pairs and their target genes in the high-confidence set: an additive model, in which we used an identity link function and a multiplicative model, in which we used a log link function. We then compared the model fits with Akaike Information Criterion (AIC). This approach is similar to that used by Dukler et al.[5] to compare additive, exponential and logistic models for two genes. In all cases, the multiplicative model provided a better fit,

indicating that the combined effect of enhancers is better described by a multiplicative model (Fig. 4a). Thus, we used the multiplicative form of GLiMMIRS-int in all subsequent analyses.

We applied GLiMMIRS-int to the 330 enhancer pairs in the unbiased set and observed no significant interaction terms (Likelihood Ratio Test, FDR<0.1) (Fig. 4b). When applying GLiMMIRS-int to the 3,808 enhancer pairs where each constituent enhancer did not necessarily have a significant effect on gene expression, we identified 4 significant interaction term effects with this model (Likelihood Ratio Test, FDR<0.1) (Fig. 4b). These interactions were observed at the *EXOC8*, *BABAM2*, *H2BC12*, and the *ZBED9* gene loci, and all significant interaction terms were positive (Fig. 4c)

We examined the distribution of single-cell RNA-seq read counts for the four genes with significant interaction terms, focusing on the cells that received guides targeting both of the corresponding enhancers. For all four genes, we noted that there was a single outlier cell with high read counts that received both guides (Fig. 4d). Since GLM coefficients and p-values can be influenced by outliers, we performed a bootstrap analysis of the interaction coefficients (β_{AB}), which is less sensitive to outliers. For each of the enhancer pairs and their corresponding target genes, we resampled cells with replacement 100 times, fit GLiMMIRS-int to the resampled data, and recorded the β_{AB} estimates. The 99% bootstrap confidence intervals for β_{AB} for all four genes spanned zero (Fig. 4e). We additionally performed a permutation test of β_{AB} to obtain p-values that are more robust to outliers. We shuffled the assignments of gRNAs in cells for the gRNAs targeting both enhancers in each pair jointly 10,000 times, and fit GLiMMIRS-int to the permuted data to obtain a null distribution of interaction coefficients. Two of the p-values obtained by this approach were nominally significant (p=0.0077 and p=0.0003 by two-sided permutation test), but would not withstand multiple testing correction given the total number of tests performed (Fig. 4b). In combination, these results indicate that the four significant

interaction terms are largely driven by cells with outlier expression of the target gene, and that there is insufficient evidence to reject the null hypothesis of no interactions between enhancers.

Discussion

CRISPR perturbations provide a new way to measure how combinations of enhancers regulate gene expression. We reanalyzed data from a single-cell CRISPRi experiment designed to map enhancers to the genes that they regulate. Since this dataset transduced guide RNAs with a high MOI, multiple enhancers near to (within 1MB of) the same gene were sometimes perturbed within the same cells, making it possible to analyze the joint effects of multiple enhancers on a common target gene. Our analysis supports a model in which enhancers act multiplicatively to control gene expression. Such a model was previously proposed by Dukler et al.[5], whose analysis of two loci in the genome supported either a logistic or multiplicative model of regulatory activity over an additive model[5]. Our genome-wide analysis confirms that a multiplicative model of enhancer activity fits the data in our analysis very well. The multiplicative model consistently provides a better fit than an additive model (Fig. 4a) and statistics obtained from applying our multiplicative model to 3,808 unbiased testable pairs in the experimental data closely resemble those expected under the null hypothesis of no enhancer interactions (Fig. 4b). The logistic model would be considered a refinement of a multiplicative model in which the expression of a gene has a maximum threshold that can be achieved by the activity of its enhancers. However, we cannot formally distinguish between logistic and multiplicative models with our dataset, because this would require examining interactions between more than two enhancers for a single gene.

A limitation of the dataset that we analyzed is that even with a high MOI and a large number of sequenced cells, only a small subset of enhancer pairs could be interrogated. Specifically, we

only tested 3,808 out of a possible 795,616 testable enhancer pairs because most enhancer pairs satisfying our testing criteria were not simultaneously perturbed in a sufficient number of cells. Furthermore, we only had sufficient power to detect interactions that exerted at least a moderately strong effect on expression (e.g. 29.4% power to detect interactions with an absolute effect size of 3 or greater at a simulated MOI of $\lambda = 15$). Many of these power limitations could be overcome through CRISPRi experiments designed specifically to probe enhancer interactions. For example, a high MOI CRISPRi experiment could be performed in which a much smaller number of candidate enhancers are targeted so that testable pairs are frequently perturbed simultaneously in the same cells. Multiple guides could also be transduced on the same vectors so that nearby enhancers are guaranteed to be targeted in many cells. This latter approach was recently used to estimate enhancer interactions at the MYC locus[9].

Further limitations of our analysis are that we only analyzed data from a single cell line under a single condition, and it is possible that enhancer interactions are more prevalent under dynamic conditions or in different cell types.

Despite the above limitations, our results argue against the presence of strong epistatic interactions between enhancers. If such interactions do exist they must be infrequent, of small effect, or restricted to specific cell types or conditions. How can these observations be reconciled with prior reports of enhancer redundancy or synergy? A possible explanation is that an interaction term is required by additive models because the combined effects of multiple enhancers is greater (synergistic) or less than (redundant) than expected under an additive model. However, these deviations from additivity may be naturally accounted for by a multiplicative model without the need for an interaction term. For example, under a multiplicative model, perturbation of a weak enhancer may have a small or negligible effect on expression, but would have a much more substantial effect when combined with a perturbation to a strong

enhancer. An additive model would require an interaction term to describe these results and the enhancers would appear to be 'redundant'.

A recent study by Lin et al. analyzed enhancer interactions at the *MYC* locus using pairs of CRISPR guides and reported additive interactions between nearby enhancers, and synergistic interactions between distant enhancers[9]. In our dataset, we did not observe any differences in interactions between enhancers that were close together or far apart (Fig. S16); however, it is difficult to compare our results with those from Lin et al. for two reasons. First, the high-throughput screen in Lin et al. was performed using cell proliferation as readout, rather than gene expression, thereby assuming that proliferation was proportional to *MYC* expression. Second, while Lin et al. examined how selected pairs of enhancers affect the expression of *MYC* and other genes, their analysis relied on log relative expression obtained by RT-qPCR, which is not directly comparable to scRNA-seq expression estimates.

Future studies which examine enhancer interactions will benefit from GLiMMIRS, which uses a generalized linear model that accounts for guide efficiency, differences in per-cell sequencing depth and several covariates. We note that it is important to consider a multiplicative model as the baseline expectation when looking for enhancer interactions, and when interactions are identified it is important to consider the possibility that the results are driven by a small number of outlier cells. To increase power to detect weak interactions, CRISPR experiments that are specifically designed to examine enhancer interactions are desirable. Our study motivates the further study of enhancer interactions in more cell types and conditions, to which GLiMMIRS can be applied to yield novel insights into regulatory element interactions and their effects on transcription.

Methods

CRISPRi perturbation of NMU enhancers

We identified two target sites of interest, A and B, for the gene *NMU*, each of which was targeted by two gRNAs in the Gasperini et al.[11] experiment (A1 and A2 targeting enhancer A; B1 and B2 targeting enhancer B). Pairs of gRNAs were designed by FlashFry[29] to target enhancers A and B at the same time, using 2 gRNAs per site. The gRNA pairs included the following: NMU_tss+NMU_tss (positive control), Safe_harbor (SH)+SH (negative control), A_sgRNA1+SH, A_sgRNA2+SH, SH+B_sgRNA1, SH+B_sgRNA2, A_sgRNA1+B_sgRNA1, A_sgRNA1+B_sgRNA2, A_sgRNA2+B_sgRNA1, A_sgRNA2+B_sgRNA2. Pairs of gRNAs were cloned into pLV-dCas9-KRAB-puro (Addgene #71236) following published methods[30,31]. Briefly, DNA oligos carrying pairs of guides were synthesized by IDT and cloned into pLV-dCas9-KRAB-puro plasmids by Gibson assembly reactions. Lentivirus was generated by co-transfecting the plasmid with PsPAX2 (Addgene #12260) and pMD2.G (Addgene #12259) in 293FT cells obtained from the Salk Institute Stem Cell Core. Lentivirus was harvested 48h post transfection. K562 cells (ATCC #CCL-243) were transduced by the lentiviruses using spinoculation. 72h after transduction, K562 cells with viral genome integration were selected by puromycin for 48 h. Total RNA from live K562 cells was extracted and reverse transcribed using SuperScript IV First-Strand Synthesis System (Thermo Fisher Scientific #18091050) with random hexamers. *NMU* expression was quantified by reverse transcription quantitative PCR (RT-qPCR). CRISPR gRNA designs and PCR primers used in experiment can be found in Table S5.

Data from Gasperini et al.

Data from the at-scale screen in the Gasperini et al. study are available at GEO accession number GSE120861. Guide spacer sequences were obtained from supplementary table 2 in the Gasperini et al. study[11]. The single-cell RNA-seq expression matrix from the at-scale screen was downloaded from the GEO file 'GSE120861_at_scale_screen.exprs.mtx'. The cell barcodes were determined from the GEO file 'GSE120861_at_scale_screen.cells.txt'. Gene names were determined from the GEO file 'GSE120861_at_scale_screen.genes.txt'. The expression matrix had 207,324 cell barcodes and 13,135 gene names. Covariate information as well as cell-guide mapping information was determined from the GEO file: 'GSE120861_at_scale_screen.phenoData.txt.gz'.

Computing guide efficiencies

We first collected the 13,189 guide RNA sequences used in the at-scale screen previously published by Gasperini et al.[11], which were published in Supplementary Table 2 of their study. We then appended 'NGG' to each 20 bp spacer sequence for compatibility with GuideScan 2.0[23]. We then used the GuideScan 2.0 gRNA sequence search tool (https://guidescan.com/grna) with the organism 'hg38' and the enzyme 'cas9' parameters to predict efficiencies for the 20bp guide RNA spacer sequences. We used the "Cutting.Efficiency" values outputted from GuideScan as our guide efficiency values.

Out of the 13,189 guide RNA sequences, 762 guide RNAs were designed to target transcription start sites, 101 guide RNAs were designed as non-targeting controls, 14 guide RNAs were designed as positive controls targeting the globin locus, and the remaining 12,312 guide RNAs were designed to target candidate enhancer sequences.

From the 12,312 enhancer-targeting guide RNAs, 1,415 guide RNAs did not find a match, had multiple off-targets, or had multiple perfect matches in the GuideScan 2.0 database. We excluded these 1,415 guide RNA sequences from downstream analysis.

Computing cell cycle scores

Cell cycle scores were computed from the single-cell RNA-sequencing gene expression matrix from the at-scale screen previously published by Gasperini et al.[11] using the Seurat R package (Fig. S2a-b).

Since the Seurat R package uses gene names from the Hugo Gene Nomenclature Committee, gene names were converted from their Ensembl Gene ID to HGNC symbol (https://www.genenames.org/) using the BioMart[32] tool from Ensembl[33] with the "hsapiens_gene_ensembl" dataset. Of the 13,135 genes in the at-scale expression matrix, 349 genes were not recognized by BioMart and 591 genes did not successfully map from Ensembl Gene ID to HGNC symbol. For the total 940 genes that could not be mapped from Ensembl Gene ID to HGNC symbol, the Ensembl Gene ID was imputed as the gene name for downstream analysis with Seurat.

To determine cell cycle scores, we used pre-defined sets of genes associated with S and G2M phases from the Seurat library. We log-normalized the data, identified variable features, and scaled the expression matrix using functions defined in Seurat. We then used the cell cycle scoring function with the predefined S and G2M gene sets in Seurat to compute cell cycle scores for each cell in the at-scale screen. To visualize the separation of cells based on their cell cycle scores, we performed a principal component analysis in Seurat using the S and G2M gene sets as features (Fig. S2c).

Model fitting and implementation

All models were fitted by maximum likelihood using the `glm.nb()` function from the MASS package in R[34]. Every model described in this work is a negative binomial generalized linear model with a log link function.

Defining a baseline model for a single enhancer acting on a single target gene

Our baseline model tests for the simple case where a single enhancer acts on a single gene. The model is a generalized linear model which assumes a log link function and that the single-cell RNA-seq tag counts of each gene are negative binomially-distributed. In other words, $y = NB(\mu, \phi)$ where y represents the scRNA-seq counts of the genes, ϕ represents the dispersion parameter of the negative binomial distribution, and μ is the mean parameter of the negative binomial distribution. The mean parameter is specified by a linear predictor passed through an exponential (inverse log-link) function: $\mu = exp(\beta_0 + \beta_{enhancer}X_{perturb} + \beta_S X_S + \beta_{G2M}X_{G2M} + \beta_{mito}X_{mito} + \beta_{gRNAs}X_{gRNAs} + \beta_{batch}X_{batch} + ln(s))$.

In this expression, we have gene-specific coefficients and cell-specific predictor values. β_0 is the intercept and represents the baseline gene expression before the influence of any other relevant factors on gene expression. $\beta_{enhancer}$ represents the effect of a perturbed target site (putative enhancer) on its target gene. β_S and β_{G2M} are coefficients that represent the effect of the S and G2M cell cycle states, respectively. β_{mito} is a coefficient representing the effect of percentage of mitochondrial DNA. Finally, β_{gRNAs} is a coefficient representing the effect of total counts of gRNAs observed within a given cell. β_{batch} is a coefficient representing the effect of the prep batch, from the Gasperini et al. 2019 experiment. We incorporate measures of guide efficiency in the variable $X_{perturb}$. This variable is calculated for each cell based on the efficiencies of

every gRNA targeting the target site being modeled which are present in the cell. Specifically, $X_{perturb}$ is calculated for any given cell and target site as $1 - \prod_{k=1}^{K} \quad (1 - g_k)$, where K is the total number of gRNAs targeting the target site found in the cell and g_k is the efficiency of the kth gRNA. Because we interpret guide efficiency as the probability that a gRNA successfully perturbs its designated target site, the expression for $X_{perturb}$ can be interpreted as the joint probability of a perturbation in a given cell based on all of the gRNAs targeting the site that are present in that cell. X_S and X_{G2M} are S and G2M cell cycle scores, respectively, for each cell. X_{mito} is the percentage of mitochondrial DNA in a cell. X_{gRNAs} is the total number of gRNAs observed in a cell. X_{batch} is the prep batch (from Gasperini et al. 2019). Finally, s is an offset term for the model that serves as a scaling factor controlling for variable sequencing depth across cells. It is calculated as $s = \frac{T}{1e6}$, where T is the total scRNA-seq counts in a cell summed across all genes in the expression count matrix. Prior to fitting the models, we added a pseudocount of 0.01 to the scRNA-seq counts of the gene being modeled for all cells to prevent inflation of coefficients (see section: Defining a model for an enhancer pair acting on a single target gene).

Simulating data for single enhancers acting on single genes

To begin, we define some simulation parameters, including the total number of cells, C; the total number of genes, G; the total number of target sites, N; and the number of gRNAs targeting each site, d. Note that the total number of target sites, N, is also the total number of target genes, as this simulation assumes that each target site is a unique enhancer for a unique gene. To generate a simulated dataset, we need to simulate sets of coefficient values for each gene ($\beta_0, \beta_{enhancer}, \beta_S, \beta_{G2M}, \beta_{mito}$) as well as corresponding variable values for each cell ($X_{perturb}, X_S, X_{G2M}, X_{mito}$, and scaling factor s). We also need to simulate the gRNA library and

assign gRNAs to cells, as well as assign guide efficiencies to gRNAs (which will be used to calculate $X_{perturb}$). These values are used to calculate a value of μ for defining a negative binomial distribution from which simulated counts for a given gene will be drawn. Specifically, $\mu = exp(\beta_0 + \beta_{enhancer}X_{perturb} + \beta_S X_S + \beta_{G2M}X_{G2M} + \beta_{mito}X_{mito} + ln(s))$. The terms for total gRNA counts per cell and batch are omitted from the simulation for simplicity, and are also omitted when fitting the baseline model to the simulated data. The dispersion parameter for the negative binomial distribution will be constant across all genes, and estimated from the empirical data. For the simulated dataset described in our paper, we used values of $G = 13000, N = 1000, d = 2$.

We first simulated values of $\widehat{\beta_0}$ for each gene. To do this, we randomly selected a subset of 1,000 genes and 10,000 cells from the Gasperini et al. 2019 at scale experiment and fit the counts for these genes to negative binomial distributions using maximum likelihood estimation (MLE). Specifically, we define the mean parameter of the negative binomial here as $\mu = exp(\widehat{\beta_0} + ln(s))$. Note that here s is calculated from the total counts for the gene across the subset of 10,000 cells using the formula defined in the previous section. This simplified model has no covariates, but does account for the scaling factor, as the goal is to simply get a sense of what coefficient values reflect the empirical data. After modeling the counts from the random subset of data, we visualized the distribution of estimated $\widehat{\beta_0}$(from which μ is calculated) and dispersion parameters for each gene tested (Fig. S3). From what we observed, we picked a fixed dispersion value of $\phi = 1.5$ for defining the negative binomial distribution for generating simulated count data. We also observed that the distribution of $\widehat{\beta_0}$ estimated from the subset of the at scale experiment were roughly normally distributed. Therefore, we fit these estimated $\widehat{\beta_0}$ values to a normal distribution with MLE to obtain parameters for defining a normal distribution from which to sample β_0 values for the simulated dataset. We obtained parameters for the

normal distribution of $\mu \approx 2.24$ and $\sigma \approx 1.8$, so we sampled G times from $N(\mu = 2.24, \sigma = 1.8)$ to yield baseline coefficients for all of the genes in the simulated dataset (Fig. S4).

To assign guides to cells, we first determined the number of gRNAs in each cell in our simulated dataset by sampling from a Poisson distribution defined as $Pois(\lambda = 15)$. This value of λ comes from the fact that in the Gasperini et al. 2019 experiment, they observed a median of approximately 15 unique gRNAs per cell. Thus, we sampled C times from the distribution defined by $Pois(\lambda = 15)$ to obtain the number of unique gRNAs in each cell (Fig. S5). To assign gRNAs to each cell, we sampled g times without replacement from the set of all gRNAs in our library, where g is the total number of gRNAs in a given cell (determined in the previous step) and the gRNA library is denoted as a sequence of integers $1,2,\ldots,dN$. Information about which gRNAs are found in which cells are stored in a one hot encoded matrix.

We defined guide efficiency for each gRNA by sampling from a left-skewed Beta distribution, to represent the fact that an experimental design would select for gRNAs with higher efficiencies). For our simulation we used a Beta distribution defined as $Beta(a = 6, b = 3)$ (Fig. S6).

Next, we created a mapping of gRNAs to target genes. For each target site, or putative enhancer, we randomly select an integer from $1,2,\ldots,G$ to represent the target gene of the candidate enhancer (indexers are used as gene identifiers). This is done without replacement to simulate a case where we are attempting to study enhancers of distinct genes, and yields a vector of length N, which we will replicate d times to yield a complete mapping of gRNAs to target genes. In this vector of length Nd, the index of a given value in the vector represents the gRNA identifier.

Enhancer effect sizes are represented by the coefficient $\beta_{enhancer}$ and are assigned on a per-gene basis. These values represent the effect that an enhancer has on the expression of its target gene. To do this, we sampled from a gamma distribution and multiplied the values by -1 to yield a negative value, representative of the expectation that successful repression of an enhancer will most likely decrease target gene expression. We wanted the values to be on a comparable scale wit the expected baseline expression, β_0, while also not being so small that they would be difficult for the model to detect changes in expression. We chose to sample values of $\beta_{enhancer}$ from a gamma distribution defined by $\Gamma(\alpha = 6, \sigma = 0.5)$, and all values drawn from the distribution were multiplied by -1 to represent a negative effect on target gene expression, which is the expectation when an enhancer is repressed (Fig. S7).

$X_{perturb}$ is calculated for each cell as a function of guide efficiencies for the gRNAs targeting the putative enhancer of interest found in that cell. Specifically, it is calculated for each cell as $X_{perturb} = 1 - \prod_{k=1}^{K} (1 - g_k)$ where K is the total number of gRNAs targeting the putative enhancer of the gene being simulated/modeled that are present in the cell and g_k is the guide efficiency of the kth gRNA in this set of targeting gRNAs. $X_{perturb} = 0$ when $K = 0$ (Fig. 2b). We compared the performance of using this variable in our model against the performance of using a binary indicator variable that simply represents the presence of any gRNA targeting the gene being simulated/modeled in a given cell.

We generated cell cycle scores for each cell in our simulated dataset using a similar approach to the one we used for sampling β_0 values. That is, we first fit models to the empirical data to identify a distribution to draw simulated values from such that they would reflect the distribution of the real data. We first calculated S and G2M cell cycle scores for the empirical data using Seurat's CellCycleScoring() function[35–38]. We observed that while the S cycle scores calculated

from the empirical data appeared to be normally distributed, the G2M scores appeared to show a right skewed distribution (Fig. S2a-b). Thus, we fit the empirical S cycle scores to a normal distribution and the empirical G2M scores to a skew normal distribution with MLE. We used the estimated parameters to define distributions for sampling S and G2M scores for the simulated dataset. Specifically, we sampled C times from a normal distribution defined by $N(\mu = -1.296e-3, \sigma = 0.11)$ and a skew normal distribution defined by $N(\zeta = -0.256, \omega = 0.312, \alpha = 6.29, \tau = 0)$ to obtain simulated S and G2M scores, respectively (Fig. S8).

We generated corresponding values of β_S and β_{G2M} by sampling from the same distribution used to generate the enhancer effect sizes, or the gamma distribution defined by $\Gamma(\alpha = 6, \sigma = 0.5)$ (Fig. S9).

Percentage of mitochondrial DNA per cell is simulated using the same approach used to simulated the cell cycle scores and baseline expression values (β_0). We fit to the empirical percentages of mitochondrial DNA per cell. We fit to a beta distribution using MLE, and used the resulting parameter estimates to define a new beta distribution from which we sampled simulated values of percentage of mitochondrial DNA. This beta distribution was defined as $Beta(a = 3.3, b = 81.48)$ (Fig. S10).

Coefficients for the effect size of percentage of mitochondrial DNA, β_{mito}, were simulated per gene by sampling from the same gamma distribution used to sample the other coefficients ($\beta_{enhancer}, \beta_S, \beta_{G2M}$). This is the gamma distribution defined as $\Gamma(\alpha = 6, \sigma = 0.5)$.

Finally, we simulated scaling factor values, s, for each cell in our simulated experiment, which were used to calculate values of μ for simulating counts for each gene. To do this, we simulated

values of T, or total counts per cell, for each cell by sampling from a Poisson distribution defined by $Pois(\lambda = 50000)$, where 50000 is the expected number of reads observed in a given cell in a scRNA-seq experiment.

Simulating noisy guide efficiencies

The noisy guide efficiency estimate, w, for a given gRNA in our simulated dataset was sampled from a new Beta distribution parameterized by a' and b', which are calculated from the "true" simulated guide efficiency for the gRNA, w, and a dispersion-controlling constant D. We wanted the noisy guide efficiency to be sampled from a Beta distribution whose mean is equivalent to the "true" guide efficiency value; thus, $w = \frac{a'}{a'+b'}$. We defined the dispersion-controlling constant D as $D = a' + b'$. From this, it follows that $a' = Dw$ and $b' = D - a'$. Like so, we calculated values of a' and b' from which to draw the noisy guide efficiency estimate for a given gRNA in our simulated guide library. The magnitude of D is inversely proportional to the amount of noise (Fig. S11).

Fitting baseline model to simulated data

To fit the baseline model to simulated data, we used a negative binomial GLM with a mean defined by the same log-link function described for generating simulated counts: $\mu = exp(\beta_0 + \beta_{enhancer}X_{perturb} + \beta_S X_S + \beta_{G2M}X_{G2M} + \beta_{mito}X_{mito} + ln(s))$. Models were fitted by MLE. Each model can be described as $y = NB(\mu, \phi)$, where y is the simulated counts for the gene being modeled, and all variable values ($X_{perturb}, X_S, X_{G2M}, X_{mito}$) come from the per-cell values from the simulated dataset. We omit β_{gRNA} when fitting to the simulated data for simplicity.

Evaluating performance of baseline model on simulated data

Our simulated dataset had N target sites, or genes that were regulated by an enhancer perturbed in the experiment. For each of these genes, we computed the Pearson correlation (Pearson's r and p-value) between the estimated coefficients, derived from fitting the baseline model to the simulated data, and the "true" coefficients, which were the "ground truth" coefficient values that we generated for the simulation and used to parameterize the distribution from which the simulated counts were drawn. We also calculated MSE for these values. Finally, we calculated the correlation of determination (R^2) as a measure of the model performance, as $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where SS_{res} is the sum of squared residuals and SS_{tot} is the total sum of squares. Specifically, we calculated SS_{res} as the sum of squared differences between the true and estimated coefficient values, and SS_{tot} as the sum of squared differences between each estimated coefficient value and the average of all estimate values for the coefficient. These metrics are summarized in Table S1 for the continuous vs. indicator forms of $X_{perturb}$ and in Table S2 for the three different sets of noisy simulated guide efficiencies.

Fitting baseline model to experimental data

For running a single enhancer-gene pair analysis on the experimental data, we obtained the 664 previously published enhancer-gene pairs from the Gasperini et al.[11] paper using information provided in Supplemental Table 1. Using these 664 previously published enhancer-gene pairs, we retrieved all experimental gRNAs targeting these enhancers, and filtered gRNAs where there was no valid guide efficiency from GuideScan 2.0. We then obtained the preparation batch, cell gRNA count, and percent mitochondrial reads covariates from their experimental data published on GEO, and excluded cells without covariate values for our downstream modeling. To account for sequencing depth, we used the at-scale gene expression matrix and counted the number of

transcripts per cell. We then divided these values by 1e-6 to obtain values for each cell which we included in our linear model through the offset() function. Prior to running the models, a pseudocount of 0.01 was added to the scRNA-seq counts for each cell. Models were then fitted using the nb.glm() function in the MASS R package using a log-link function and optimizing via maximum likelihood estimation. In the at-scale model, there were 207,324 cells total. After filtering for cells without covariate values, there were 205,797 cells that were included in the modeling process. The scrambled perturbation negative control was obtained by scrambling the vector of guide efficiencies prior to modeling. The mismatch gene negative control set was obtained by randomly sampling a gene for a given enhancer from the set of 664 previously published enhancer-gene pairs.

Defining a model for an enhancer pair acting on a single target gene

Our model for an enhancer gene is quite similar to our baseline model, except we replace $\beta_{enhancer}$ with three new coefficients: $\beta_A, \beta_B, \beta_{AB}$. Referring to the two enhancers in the pair being modeled as enhancers A and B: β_A represents the effect of enhancer A on the target gene; β_B represents the effect of enhancer B on the target gene; β_{AB} represents the interaction effect between enhancers A and B on the target gene. The new negative binomial GLM has a mean defined as: $\mu = exp(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_{AB} + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + \beta_{gRNAs} X_{gRNAs} + \beta_{batch} X_{batch} + ln(s))$. Here, X_A, X_B, X_{AB} represent the perturbation probabilities of enhancer A, enhancer B, and both enhancers, respectively. They are calculated in the same manner as $X_{perturb}$.

When fitting linear models, we observed inflated β_{AB} coefficients associated with cases where all cells containing gRNAs for both enhancers A and B showed no expression of the target

gene. To prevent this inflation of the coefficients, we added a pseudocount of 0.01 to all of the gene expression counts. When including a pseudocount in our modeling process, we observed a reduction in outliers in our enhancer effect sizes (Fig. S15).

Defining testable pairs of enhancers for interactions

We defined testable enhancer pairs as any pairs of target sites, or putative enhancers, from the Gasperini et al. 2019 experiment which were located within 1MB of a common target gene. We also defined two subsets of testable pairs based on certain filtering criteria: a smaller, high confidence set of 330 enhancer pairs and their corresponding target genes, and a larger unbiased set of 3,808 enhancer pairs and corresponding target genes. To define our high confidence set, we restricted the set of all testable pairs to those where both individual enhancers in the pair had previously established evidence of a regulatory effect on the target gene based on the analysis performed by Gasperini et al.[11] in their original study. To define our unbiased set, we simply looked for testable pairs that were simultaneously perturbed in a minimum of 20 cells; that is, there must be 20 cells receiving at least one of the gRNAs targeting each of the enhancers in the pair. We did not require either enhancer to have prior evidence of a regulatory effect on the target gene, thereby allowing for the possibility of regulatory effects that only arise in the presence of an interaction with another enhancer. In all cases, we also discarded enhancer pairs if all of the gRNAs for either enhancer in the pair had undefined guide efficiency estimates.

Simulating data for enhancer pairs acting on a single target gene

We adapt the simulation framework used for simulating data for a single enhancer acting on a single gene. However, we have additional parameters to determine the number of “ground truth” enhancer pairs with and without an interaction effect between them. We refer to these as

"positive" (N_{pos}) and "negative" (N_{neg}) pairs, respectively. These are selected from the set of all possible pairwise combinations of N target sites defined for our simulation. Note that for the case of an enhancer pair acting on a single gene, N represents the total number of putative enhancers rather than the total number of target genes. After randomly selecting N_{pos} and N_{neg} pairs without replacement from the set of possible pairs, we then randomly select the same number of genes without selection from the set of possible genes $(1, \ldots, G)$ to be the target genes of those pairs. For the simulation described in this paper, we selected values of $N_{pos} = N_{neg} = 500$ and a total of $N = 1000$ target sites.

Simulating data for power analysis

Most aspects of the data simulation are identical to the data simulation for a single enhancer acting on a single gene. The coefficients β_A and β_B are drawn from the same distribution as $\beta_{enhancer}$. However, for the power analysis, we assign a number of different fixed values of β_{AB} for genes that are acted upon by an interaction effect between enhancers (e.g., the target genes of "positive" enhancer pairs). For genes that are not acted upon by any interaction effect, $\beta_{AB} = 0$. The other parameter that we modulate in the simulations is the value of λ for the Poisson distribution used to sample the number of unique gRNAs found in each cell. This is representative of multiplicity of infection, or MOI, so for each value of λ that we want to test with our power analysis, we generate different numbers of gRNAs per cell (Fig. S14), and use these sets of values to generate different mappings of gRNAs in cells. This yields a different one-hot encoded matrix for each value of lambda, which will also lead to different sets of values of X_A, X_B, and X_{AB}, as greater MOI may result in more gRNAs for a target site found in a given cell and greater perturbation probabilities. Simulated counts are generated from a negative binomial distribution parameterized by $NB(\mu, \phi)$, where $\mu = exp(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_{AB} + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + ln(s))$ and $\phi = 1.5$ (determined from modeling empirical data, see

Methods for simulating data for single enhancers acting on a single gene). We generated a set of simulated counts for each value of λ and interaction effect size. For our power analysis, we used values of $\lambda = 15,25,50,75,100$ and $\beta_{AB} = 0.5,1,3,5,7$.

Power analysis

For our power analysis, we fit our model to the simulated data for the "positive" and "negative" pairs to obtain true positive rates (TPR) and true negative rates (TNR), respectively. We calculated the proportion of models that correctly called significant interaction terms, β_{AB}, for the "positive" cases to obtain TPR. We calculated the proportion of models that correctly called no significant interaction terms, β_{AB}, for the "negative" cases to obtain TNR.

Comparing multiplicative to additive model

To compare the fits of multiplicative vs. additive models of enhancer pair activity, we defined each model under the null hypothesis (H_0), where there is no interaction term (for simplicity). For the multiplicative model under H_0, we use the canonical log-link function and define the mean of the negative binomial, μ, as:

$\mu = exp(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + \beta_{gRNAs} X_{gRNAs} + \beta_{batch} X_{batch} + ln(s))$. For the additive model under H_0, we use the identity link function where the mean is simply equivalent to the linear predictor without transformation, defined as:

$\mu = s(\beta_0 + \beta_A X_A + \beta_B X_B + \beta_{AB} X_{AB} + \beta_S X_S + \beta_{G2M} X_{G2M} + \beta_{mito} X_{mito} + \beta_{gRNAs} X_{gRNAs} + \beta_{batch} X_{batch})$. We applied each model to the 330 testable pairs from the experimental data where each enhancer in the pair had evidence of being an enhancer for the target gene based on the analysis by Gasperini et al. We compare model fits by examining the Akaike Information Criterion (AIC), with a lower AIC indicating a better fit. We calculated ΔAIC by subtracting the

AIC of the lesser model from the AIC of the best fitting model. Since we found that the multiplicative model fit better in every case we tested, every ΔAIC reported in our study reflects the AIC of the additive model subtracted from the AIC of the multiplicative model.

Fitting interaction model to empirical data

For analyzing both sets of enhancer pairs tested in our analysis, we followed an identical procedure to the baseline model scenario, with the exception of adding a second enhancer effect vector, and allowing for interactions between the two enhancer vectors using built-in functionality within the glm.nb() function in the MASS R package.

Bootstrapping of significant interaction coefficients

We first performed bootstrapping to generate empirical distributions for the four significant interaction terms identified in our genome-wide analysis of enhancer pairs. We resampled all the cells in our dataset with replacement, and refit our enhancer pair linear models with their associated covariates to obtain the bootstrapped empirical interaction coefficients. We then used the bootstrapped interaction coefficient estimates to derive 99% confidence intervals for the interaction coefficient using quantiles.

Permutation test for significant interaction coefficients

To determine permutation-based p-values associated with the observed significant interaction coefficients, we generated a null distribution of interaction coefficients by shuffling the perturbation probability vectors for enhancer 1 and enhancer 2 jointly, such that the same numbers of cells would have both enhancers perturbed. After performing 1000 permutations, we computed two-tailed p-values by counting the number of interaction coefficients with a

magnitude greater than our observed significant interaction coefficient and dividing by the total number of permutations performed (Fig. S17). Schematic figures

All schematic figures created with BioRender.com.

Data availability

Data from the Gasperini et al. experiment can be found under GEO accession number GSE120861. Our *NMU* RT-qPCR experiment results are provided as a spreadsheet (Table S1).

Code availability

All relevant code and documentation can be found at https://github.com/mcvickerlab/GLiMMIRS.

Acknowledgements

We thank the research groups of Dr. Roland Schwarz and Dr. Christoph Lippert for hosting J.L.Z. during her time as a visiting Fulbright scholar and supporting her work on this project. J.L.Z. was supported by a Fulbright Research Award, an NIH F31 individual predoctoral fellowship (F31DA056226), and the Chapman Charitable Trust Fellowship. G.M. was supported by the NIH/NHGRI (R35HG011315) and the Frederick B. Rentschler Developmental Chair. H.V.C. was supported by the 2020 Salk Women & Science Award and the 2020 Salk Alumni Fellowship Award.

Author Information

Authors and Affiliations

Integrative Biology Laboratory, Salk Institute for Biological Studies, 10010 N Torrey Pines Rd, La Jolla, CA, 92037, USA

Jessica L. Zhou, Karthik Guruvayurappan, Hsiuyi V. Chen, Aaron R. Chen & Graham McVicker

Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, 92093, USA

Jessica L. Zhou & Graham McVicker

School of Biological Sciences, University of California San Diego, La Jolla, CA, 92037, USA

Karthik Guruvayurappan

Halicioglu Data Science Institute, University of California San Diego, La Jolla, CA, 92093, USA

Karthik Guruvayurappan

A*STAR Infectious Diseases Labs, Agency for Science, Technology and Research (A*STAR), Singapore 138648, Singapore

Hsiuyi V. Chen

Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, 11794, USA

Aaron R. Chen

Contributions

G.M. conceived the project. J.L.Z. and K.G. designed and conducted all data analysis, interpretation, modeling and programming with input from G.M.. A.R.C. and H.V.C. performed the *NMU* RT-qPCR CRISPR experiments. J.L.Z. and K.G. drafted the initial version of the manuscript. J.L.Z, K.G. and G.M. revised and edited the manuscript. All authors reviewed the final manuscript prior to submission.

Corresponding authors

Correspondence to Graham McVicker.

Ethics declarations

Competing interests

The authors declare no competing interests.

Figures

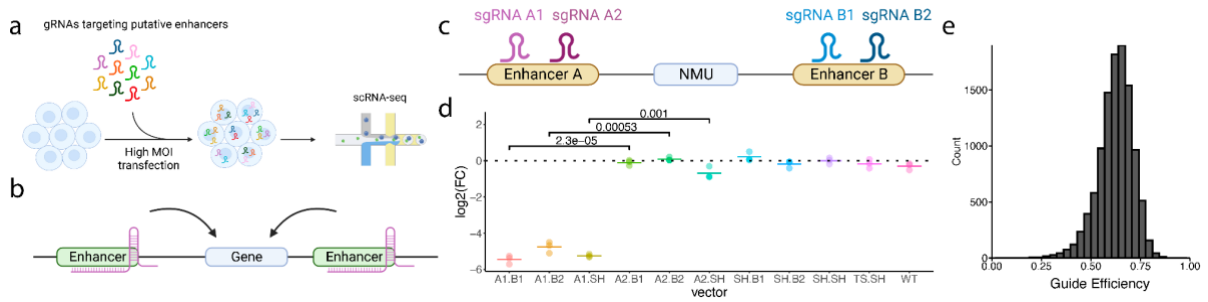

Fig.1: Conceptual schematic and summary of *NMU* RT-qPCR experiment demonstrating effects of variable gRNA efficiency. a) Schematic of the Gasperini et al. experiment. A library of gRNAs targeting putative enhancers was transduced into cells with a high multiplicity of infection (MOI), resulting in multiple perturbations per cell. The identities of the gRNAs and their effects on gene expression were read out with single-cell RNA-seq (scRNA-seq). **b)** Schematic of two enhancers acting on the same gene. We seek to quantify the effect on multiple enhancers acting on a single gene. **c)** Schematic of CRISPR perturbation experiment targeting enhancers of *NMU* with two gRNAs per enhancer. **d)** Results of CRISPRi RT-qPCR experiment perturbing *NMU* enhancers for three technical replicates. For each *NMU* enhancer (enhancers A and B), two gRNAs were used (A1, A2 and B1, B2, respectively) and delivered on the same vector. Vectors containing gRNA A1 resulted in larger fold changes in *NMU* expression than their counterparts containing gRNA A2 instead (denoted p-values come from unpaired Welch's two-sided t-tests against the null hypothesis that there is no difference in mean fold change (FC) between vectors using gRNA A1 vs. gRNA A2. SH = safe harbor). TS = *NMU* transcription start site, WT = wild type K562 cells expressing dCas9-KRAB without any gRNAs, horizontal bar = mean $\log_2$(FC). **e)** Distribution of guide efficiency values predicted by GuideScan 2.0 for the gRNAs used in the Gasperini et al. experiment.

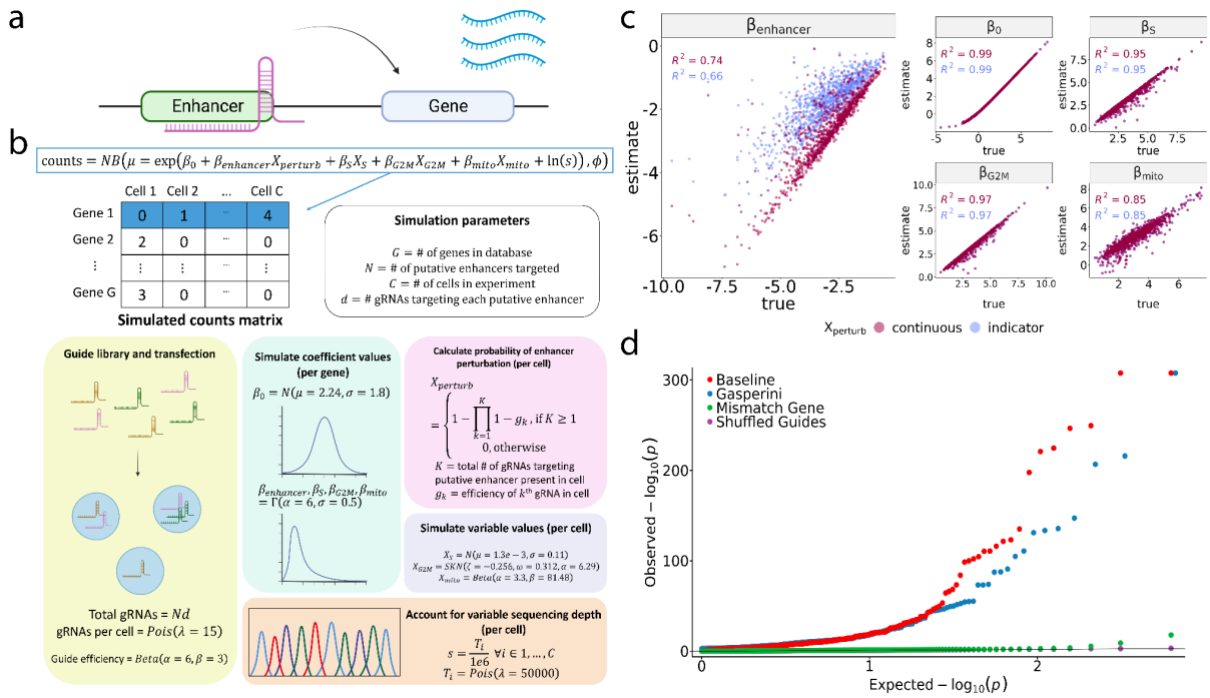

Fig.2: GLiMMIRS-base schematic and performance on simulated and experimental data. a) A schematic of our baseline model, wherein we evaluate the effect of a single putative enhancer on a single target gene. We model count data with a negative binomial generalized linear model (GLM). **b)** Schematic of how data was simulated to represent a single-cell CRISPRi experiment perturbing enhancers. Coefficient values (β) were simulated for each gene and corresponding variable values (X) were simulated for each cell. $X_{perturb}$ was calculated as a function of simulated guide efficiency values. Values were sampled from distributions that resembled the empirical data whenever possible. We also simulated a per cell scaling factor to account for sequencing depth. **c)** Scatterplot comparing true versus estimated coefficient values for each gene evaluated. These plots summarize the results of fitting the baseline model to 1000 genes in the simulated dataset which were designated as "true" target genes; that is, genes whose enhancers were perturbed by gRNAs in the

simulated experiment. Shown here are the results of fitting to simulated data using a value of $X_{perturb}$ calculated from guide efficiency (continuous) versus an indicator variable (indicator), with a pseudocount of 0.01 added to the counts. Coefficients of determination (R^2) are shown. **d)** Quantile-quantile plot of observed versus expected -$\log_{10}$p indicates similarity between GLiMMIRS-base and the results published by Gasperini et al. The baseline values (orange) indicate the results of GLiMMIRS-base. The Gasperini values (green) indicate the previously published results. Mismatch gene and scrambled perturbation are negative control models. Mismatch gene (purple) compares an enhancer with a randomly assigned gene expression vector, while scrambled perturbation (yellow) shuffles the vector of guide perturbation probabilities.

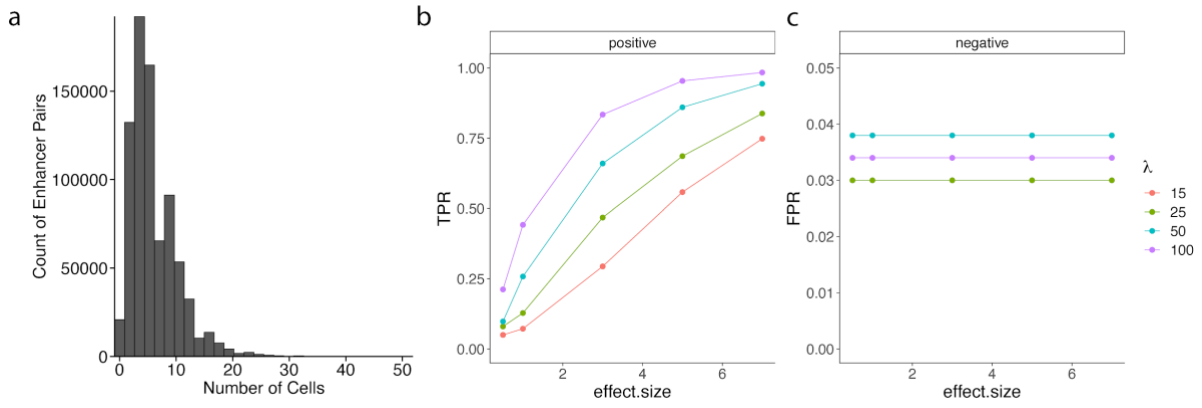

Fig. 3: GLiMMIRS-int power analysis at different simulated MOIs and interaction effect sizes. a) Distribution of the frequency of all testable target site pairs in the Gasperini et al. dataset. Criteria for testable pairs are defined as pairs of target sites, or putative enhancers, located within 1MB of a common target gene that are simultaneously perturbed in the same cells. **b-c)** Results of power analysis for ability to detect interaction effects in simulated datasets with varying multiplicities of infection (MOI) (λ) and effect sizes (x-axis). We calculated **b)** true positive rate (TPR), or power, from the "positive" ground truth enhancer pairs with interaction effects that we simulated, and **c)** false positive rate (FPR) from the "negative" control enhancer pairs without interaction effects that we simulated.

758

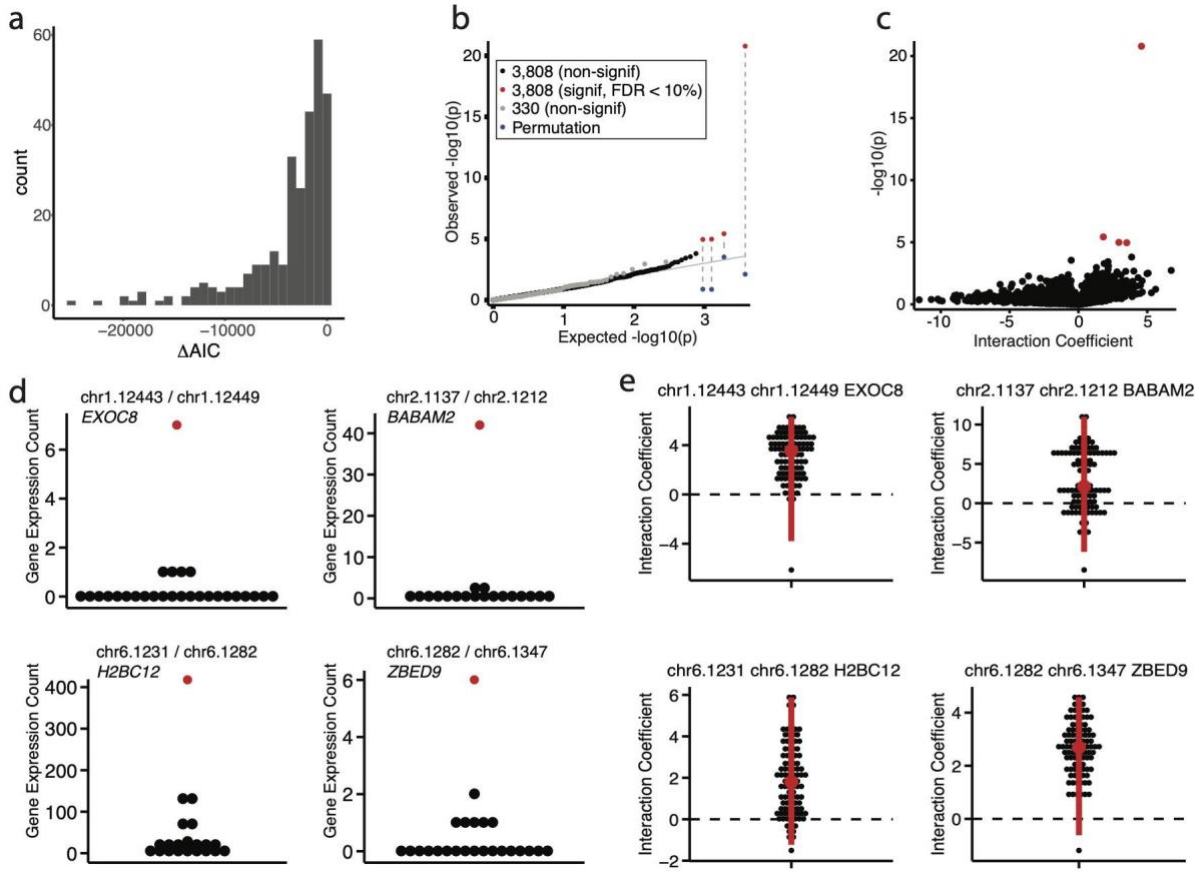

Fig. 4: Analysis of experimental data with GLiMMIRS-int supports a multiplicative model with no interaction effects. a) Distribution of ΔAIC, calculated as the difference in Akaike Information Criterion (AIC) between the best fitting model and the lesser model for 330 testable enhancer pairs and corresponding target genes from the Gasperini et al. 2019 data. In every case we evaluated, the multiplicative model fit better than the additive model. **b)** QQ-plot of interaction coefficient p-values for 330 enhancer pairs where each individual enhancer had significant effects on the target gene expression, and 3,808 enhancer pairs where each constituent enhancer did not necessarily have a significant effect on gene expression. No enhancer pairs had significant interaction coefficients after multiple testing correction for the

330 enhancer pairs (gray). Four significant interactions were observed for the 3,808 enhancer pairs at the *EXOC8*, *BABAM2*, *H2BC12,* and *ZBED9* gene loci (red) (FDR<0.1). Permutation test p-values for the same four loci are shown in blue. Non-significant cases from the 3,808 enhancer pairs are shown in black. **c)** Volcano plot of interaction coefficients for the 3,808 enhancer pairs tested, where significant interaction coefficients (FDR<0.1) are indicated in red. **d)** Gene expression counts from cells containing guides targeting both enhancers in a testable pair for the four genes with significant interaction terms. For all four genes, among the cells containing gRNAs targeting both enhancers in a pair, there contained a single outlier cell with extreme gene expression counts, indicated in red. **e)** Results from bootstrapping analysis of the four significant enhancer interactions. Red dots indicate the median coefficient estimate, and red lines indicate 99% confidence intervals.

759

References

1. Hong, J.-W., Hendrix, D. A. & Levine, M. S. Shadow Enhancers as a Source of Evolutionary Novelty. *Science* **321**, 1314 (2008).
2. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
3. Visel, A. *et al.* Functional Autonomy of Distant-Acting Human Enhancers. *Genomics* **93**, 509–513 (2009).
4. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**, 934–947 (2013).
5. Dukler, N., Gulko, B., Huang, Y.-F. & Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nature Genetics* **49**, 2–3 (2017).
6. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
7. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat Genet* **48**, 895–903 (2016).
8. Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **48**, 904–911 (2016).
9. Lin, X. *et al.* Nested epistasis enhancer networks for robust genome regulation. *Science* **377**, 1077–1085 (2022).
10. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).
11. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **0**, (2019).
12. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Res.* **29**, 464–471 (2019).

13. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods* **14**, 297–301 (2017).
14. Replogle, J. M. *et al.* Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology* 1–8 (2020) doi:10.1038/s41587-020-0470-y.
15. Hill, A. J. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nature Methods* **15**, 271–274 (2018).
16. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell* **66**, 285-299.e5 (2017).
17. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
18. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).
19. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309–315 (2017).
20. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
21. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485**, 376–380 (2012).
22. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature* **485**, 381–385 (2012).
23. Perez, A. R. *et al.* GuideScan software for improved single and paired CRISPR guide RNA design. *Nat Biotechnol* **35**, 347–349 (2017).
24. Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* **25**, 1860–1872 (2015).
25. Kim, H. K. *et al.* SpCas9 activity prediction by DeepSpCas9, a deep learning–based model

with high generalization performance. *Science Advances* **5**, eaax9249 (2019).

26. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184–191 (2016).
27. Xiang, X. *et al.* Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat Commun* **12**, 3238 (2021).
28. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research* **50**, 3616–3637 (2022).
29. McKenna, A. & Shendure, J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biology* **16**, 74 (2018).
30. Diao, Y. *et al.* A tiling-deletion-based genetic screen for *cis*-regulatory element identification in mammalian cells. *Nature Methods* **14**, 629–635 (2017).
31. Chen, H. V. *et al.* Deletion mapping of regulatory elements for GATA3 in T cells reveals a distal enhancer involved in allergic diseases. *Am J Hum Genet* **110**, 703–714 (2023).
32. Smedley, D. *et al.* BioMart – biological queries made easy. *BMC Genomics* **10**, 22 (2009).
33. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50**, D988–D995 (2022).
34. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, 2002).
35. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
36. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018).
37. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495–502 (2015).
38. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).