Conditional resampling improves calibration and sensitivity in single-cell CRISPR screen analysis

Timothy Barry[1], Xuran Wang[1], John A. Morris[2,3], Kathryn Roeder[1,4], and Eugene Katsevich[5,*]

[1]*Department of Statistics and Data Science, Carnegie Mellon University*
[2]*New York Genome Center, New York USA*
[3]*Department of Biology, New York University*
[4]*Computational Biology Department, Carnegie Mellon University*
[5]*Department of Statistics, Wharton School, University of Pennsylvania*
**Corresponding author*

Single-cell CRISPR screens are the most promising biotechnology for mapping regulatory elements to their target genes at genome-wide scale. However, the analysis of these screens presents significant statistical challenges. For example, technical factors like sequencing depth impact not only expression measurement but also perturbation detection, creating a confounding effect. We demonstrate on two recent high multiplicity of infection single-cell CRISPR screens how these challenges cause calibration issues among existing analysis methods. To address these challenges, we propose SCEPTRE: analysis of single-cell perturbation screens via conditional resampling. This methodology, designed to avoid calibration issues due to technical confounders and expression model misspecification, infers associations between perturbations and expression by resampling the former according to a working model for perturbation detection probability in each cell. SCEPTRE demonstrates excellent calibration and sensitivity on the CRISPR screen data and yields hundreds of new regulatory relationships, supported by orthogonal functional evidence.

The noncoding genome plays a crucial role in human development and homeostasis: over 90% of loci implicated by GWAS in diseases lie in regions outside protein-coding exons[1]. Enhancers and silencers, segments of DNA that modulate the expression of a gene or genes in *cis*, harbor many or most of these noncoding trait loci. While millions of *cis*-regulatory elements (CREs) have been nominated through biochemical annotations, the functional role of these CREs, including the genes that they target, remain essentially unknown[2]. A central challenge over the coming decade, therefore, is to unravel the *cis*-regulatory landscape of the genome across various cell types and diseases.

Single-cell CRISPR screens (implemented by Perturb-seq[3,4], CROP-seq[5], ECCITE-seq[6], and other protocols) are the most promising technology for mapping CREs to their target genes at genome-wide scale. Single-cell CRISPR screens pair CRISPR perturbations with single-cell sequencing to survey the effects of perturbations on cellular phenotypes, including the transcriptome. High multiplicity of infection (MOI) screens deliver dozens perturbations to each cell[7–9], enabling the interrogation of hundreds or thousands of CREs in a single experiment. Single-cell screens overcome the limitations of previous technologies for mapping CREs[9]: unlike eQTLs, single-cell screens are high-resolution and can target rare variants, and unlike bulk screens, single-cell screens measure the impact of perturbations on the entire transcriptome.

Despite their promise, high-MOI single cell CRISPR screens pose significant statistical challenges. In particular, researchers have encountered substantial difficulties in calibrating tests of association between a CRISPR perturbation and the expression of a gene. Gasperini et al.[9] found considerable inflation in their negative binomial regression based p-values for negative control perturbations. Similarly, Xie et al.[8] found an excess of false positive hits in their rank-based Virtual FACS analysis. Finally, Yang et al.[10] found that their permutation-based scMAGeCK-RRA method deems almost all gene-enhancer pairs significant in a reanalysis of the Gasperini et al. data. These works propose ad hoc fixes to improve calibration, but we argue that these adjustments are insufficient to address the issue. Miscalibrated p-values can adversely impact the reliability of data analysis conclusions by creating excesses of false positive and false negative discoveries.

In this work we make two contributions. We (i) elucidate core statistical challenges at play in high-MOI single-cell CRISPR screens and (ii) present a novel analysis methodology to address them. We identify a key challenge that sets single-cell CRISPR screens apart from traditional differential expression experiments: the "treatment"—in this case the presence of a CRISPR perturbation in a given cell—is subject to measurement error[3,11,12]. In fact, underlying this measurement error are the same technical factors contributing to errors in the measurement of gene expression, including sequencing depth and batch effects. These technical factors therefore act as confounders, invalidating traditional nonparametric calibration approaches. On the other hand, parametric modeling of single-cell

expression data is also fraught with unresolved difficulties.

To address these challenges, we propose SCEPTRE (analysis of Single-CEll PerTurbation screens via conditional REsampling; pronounced "scepter"). SCEPTRE is based on the conditional randomization test[13], a powerful and intuitive statistical methodology that, like parametric methods, enables simple confounder adjusment, and like nonparametric methods, is robust to expression model misspecification. We used SCEPTRE to analyze two recent, large-scale, high-MOI single-cell CRISPR screen experiments. SCEPTRE demonstrated excellent calibration and sensitivity on the data and revealed hundreds of new regulatory relationships, validated using a variety of orthogonal functional assays. In the Discussion we describe an independent work conducted in parallel to the current study in which we leveraged biobank-scale GWAS data, single-cell CRISPR screens, and SCEPTRE to dissect the *cis* and *trans* effects of noncoding variants associated with blood diseases[14]. This work highlights what we see as a primary application of SCEPTRE: dissecting regulatory mechanisms underlying GWAS associations.

Results

Analysis challenges. We examined two recent single-cell CRISPR screen datasets – one produced by Gasperini et al.[9] and the other by Xie et al.[8] – that exemplify several of the analysis challenges in high-MOI single-cell CRISPR screens. Gasperini et al. and Xie et al. used CRISPRi to perturb putative enhancers at high MOI in K562 cells. They sequenced polyadenylated gRNAs alongside the whole transcriptome and assigned perturbation identities to cells by thresholding the resulting gRNA UMI counts.

Both Gasperini et al. and Xie et al. encountered substantial difficulties in calibrating tests of association between candidate enhancers and genes. Gasperini et al. computed p-values using a DESeq2[15]-inspired negative binomial regression analysis implemented in Monocle2[16], and Xie et al. computed p-values using Virtual FACS, a nonparametric method proposed by these authors. Gasperini et al. assessed calibration by pairing each of 50 non-targeting (or negative) control gRNAs with each protein-coding gene. These "null" p-values exhibited inflation, deviating substantially from the expected uniform distribution (Figure 1a, red). To assess the calibration of Virtual FACS in a similar manner, we constructed a set of *in silico* negative control pairs of genes and gRNAs on the Xie et al. data (see Methods). The resulting p-values were likewise miscalibrated, with some pairs exhibiting strong conservative bias and others strong liberal bias (Figure 1a, gray-green).

A core challenge in the analysis of single-cell CRISPR screens is the presence of confounders, technical factors that impact both gRNA detection probability and gene expression. The total number of gRNAs detected in a cell increases with the total num-

ber of mRNA UMIs detected in a cell ($\rho = 0.35, p < 10^{-15}$ in Gasperini et al. data; $\rho = 0.25, p < 10^{-15}$ in Xie et al. data; Figures 1b-c). Techical covariates, such as sequencing depth and batch, induce a correlation between gRNA detection probability and gene expression, even in the absence of a regulatory relationship (Figure 1d). This confounding effect can lead to severe test miscalibration and is especially problematic for traditional nonparametric approaches, which implicitly (and incorrectly) treat cells symmetrically with respect to confounders.

Parametric regression approaches, like negative binomial regression, are the most straightforward way to adjust for confounders. However, parametric methods rely heavily on correct model specification, a challenge in single-cell analysis given the heterogeneity and complexity of the count data. We hypothesized that inaccurate estimation of the negative binomial dispersion parameter was (in part) responsible for the p-value inflation observed by Gasperini et al. Monocle2 estimates a raw dispersion for each gene, fits a parametric mean-dispersion relationship across genes, and finally collapses raw dispersion estimates onto this fitted line (Figure 1e). We computed the deviation from uniformity of the negative control p-values for each gene using the Kolmogorov-Smirnov (KS) test, represented by the color of each point in Figure 1e. Circled genes had significantly miscalibrated p-values based on a Bonferroni correction at level $\alpha = 0.05$. Genes significantly above the curve showed marked signs of p-value inflation, suggesting model misspecification.

Gasperini et al. and Xie et al. incorporated ad hoc adjustments into their analyses to remedy the observed calibration issues. On closer inspection, however, these efforts were not satisfactory to ensure reliability of the results. Gasperini et al. attempted to calibrate p-values against the distribution negative control p-values instead of the more standard uniform distribution. This adjustment lead to overcorrection for some gene-enhancer pairs (false negatives) and undercorrection for others (false positives) (Figure S1). Along similar lines Xie et al. compared their Virtual FACS p-values to gene-specific simulated null p-values to produce "significance scores" that were used to determine significance. These significance scores were challenging to interpret and could not be subjected to multiple hypothesis testing correction procedures, as they are not p-values.

Improvements to the negative binomial approach. We attempted to alleviate the miscalibration within the negative binomial regression framework by following the recommendations of Hafemeister and Satija, who recently proposed a strategy for parametric modeling of single-cell RNA-seq data[17]. First, we abandoned the DESeq2-style size factors of Monocle2 and instead corrected for sequencing depth by including it as a covariate in the negative binomial regression model. Second, we adopted a more flexible dispersion estimation procedure: we (i) computed raw dispersion estimates for each gene, (ii) regressed the raw dispersion estimates onto the mean gene expressions via kernel regres-

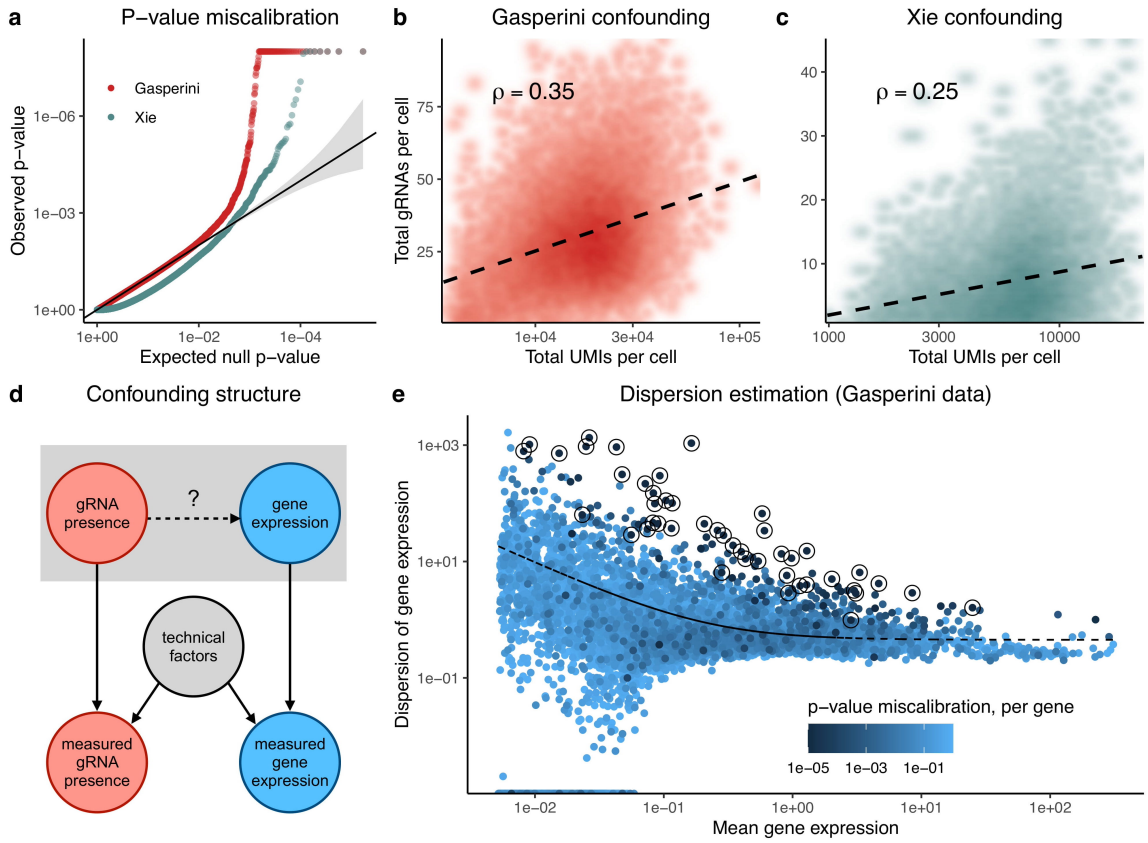


Figure 1: **CRISPR screen analysis challenges can lead to false positives and false negatives**. **a**, QQ-plot of negative control p-values produced by Gasperini et al. (red; downsampled for visualization) and Xie et al. (gray-green). These p-values deviate substantially from the expected uniform distribution, indicating test miscalibration. **b-d**, Technical factors, such as sequencing depth and batch, impact gRNA detection probability and observed gene expression levels in both Gasperini et al. (b) and Xie et al. (c) data. Thus, technical factors act as confounders (d), differentiating CRISPR screens from traditional differential expression applications. **e**, Monocle2 estimates the dispersion of each gene by projecting each gene's raw dispersion estimate onto the fitted raw dispersion-mean expression curve. This estimation procedure leads to miscalibration for high-dispersion genes.

sion, and (iii) projected the raw dispersion estimates onto the fitted nonparametric regression curve.

We reanalyzed the Gasperini et al. negative control data using the improved negative binomial regression approach. In addition to sequencing depth, we included as covariates

in the regression model the total number of expressed genes per cell, as well as the technical factors accounted for in the original analysis (total number of gRNAs detected per cell, percentage of transcripts mapped to mitochondrial genes, and sequencing batch). We likewise applied the improved negative binomial approach to the Xie et al. negative control data, including sequencing depth, number of detected gRNA UMIs per cell, and sequencing batch as regression covariates in the latter analysis. While the improved negative binomial regression exhibited better calibration than Monocle regression on the Gasperini et al. data (Figure 3b), it showed clear signs of p-value inflation on both datasets (Figure 3b-c). We concluded that parametric count models likely are challenging to calibrate to high-MOI single-cell CRISPR screen data.

SCEPTRE: Analysis of single-cell perturbation screens via conditional resampling. To address the challenges identified above, we propose SCEPTRE, a methodology for single-cell CRISPR screens based on the simple but powerful conditional randomization test[13] (Figure 2). To assess the association between a given gRNA and gene, we first fit the improved negative binomial statistic described above. This yields a z-value, which typically would be compared to a standard normal null distribution based on the parametric negative binomial model. Instead, we build a null distribution for this statistic via conditional resampling. First, we estimate the probability that the gRNA will be detected in a given cell based on the cell's technical factors, such as sequencing depth and batch. Next, we resample a large number of "null" datasets, holding gene expression and technical factors constant while redrawing gRNA assignment independently for each cell based on its fitted probability. We compute a negative binomial z-value for each resampled dataset, resulting in an empirical null distribution (gray histogram in Figure 2). Finally, we compute a left-, right-, or two-tailed probability of the original z-value under the emprical null distribution, yielding a well-calibrated p-value. We note that SCEPTRE in principle is compatible with any test statistic that reasonably tracks the expression data, including, for example, statistics based on machine learning algorithms.

We leverage several computational accelerations to enable SCEPTRE to scale to large single-cell CRISPR screen datasets. First, we approximate the null histogram of the resampled test statistics using a skew-t distribution to obtain precise p-values based on a limited number of resamples (500 in the current implementation). Second, we employ statistical shortcuts that reduce the cost of each resample by a factor of about 100 (see Methods). Finally, we implement the algorithm such that it can run in parallel on hundreds or thousands of processors on a computer cluster. (We used this approach in our independent study of noncoding blood trait GWAS loci[14].) Overall, we estimate that SCEPTRE can analyze 2.5 million gene-gRNA pairs on a dataset of 200,000 cells in a single day using 500 processors.

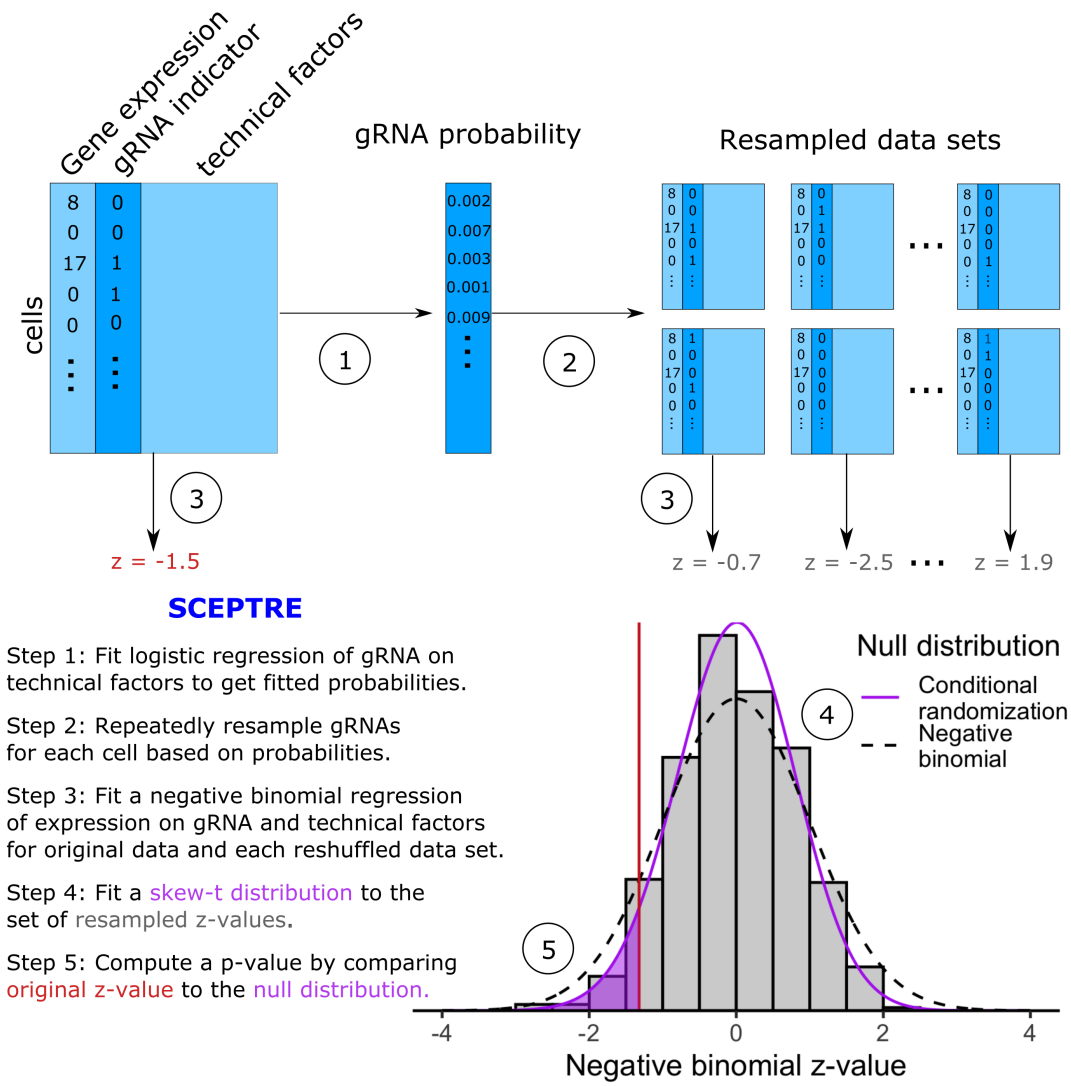


Figure 2: **SCEPTRE: Analysis of single-cell perturbation screens via conditional resampling.** A schematic and outline of the SCEPTRE methodology for one gene and one gRNA is shown. SCEPTRE estimates the probability of gRNA detection in each cell based on its technical factors. It then builds a null distribution for the negative binomial z-value by independently resampling gRNA presence or absence for each cell according to these probabilities to form "negative control" datasets. A skew-t distribution is fit to the resulting histogram to obtain precise p-values based on a limited number of resamples, against which the original NB z-value is compared. The dashed line shows the standard normal distribution, against which the NB z-value typically would be compared.

SCEPTRE demonstrates good calibration and sensitivity on real and simulated data. First, we demonstrated the calibration of SCEPTRE in a small, proof-of-concept simulation study (Figure 3a). We considered a class of negative binomial regression models with fixed dispersion and two technical covariates (sequencing depth and batch). We simulated expression data for a single gene in 1000 cells using four models selected from this class: the first with dispersion $= 1$, the second with dispersion $= 0.2$, the third with dispersion $= 5$, and the last with dispersion $= 1$, but with 25% zero-inflation. We also simulated negative control gRNA data using a logistic regression model with the same covariates as the gene expression model. We assessed the calibration of three methods across the four simulated datasets: SCEPTRE, improved negative binomial regression, and scMAGeCK-LR[10], a recently-proposed, permutation-based nonparametric method. To assess the impact of model misspecification on SCEPTRE and the improved negative binomial method (on which SCEPTRE relies), we fixed the dispersion of the negative binomial method to 1 across all four simulated datasets. The negative binomial method worked as expected when the model was correctly specified but broke down in all three cases of model misspecification. scMAGeCK-LR exhibited poor calibration across all simulated datasets, likely because, as a traditional nonparametric method, it failed to adequately account for confounders. Finally, SCEPTRE was well-calibrated in all settings.

Next, to assess the calibration of SCEPTRE on real data, we used SCEPTRE to test the association between negative control gRNAs and genes in the Gasperini et al. data (Figure 3b) and Xie et al. data (Figure 3c). We compared SCEPTRE to the improved negative binomial method, as well as to the original analysis methods (i.e., Monocle regression on the Gasperini et al. data and Virtual FACS on the Xie et al. data). We did not apply scMAGeCK-LR to these data given its poor calibration on the simulated data. SCEPTRE showed excellent calibration on both datasets; by contrast, Monocle regression and improved negative binomial regression demonstrated signs of severe p-value inflation, while Virtual FACS exhibited a bimodal p-value distribution.

To assess the sensitivity of SCEPTRE, we applied it to test the 381 positive control pairs of genes and TSS-targeting gRNAs assayed by Gasperini et al. (Figure 3d). Allowing for the fact that the empirical correction employed by Gasperini et al. limited the accuracy of p-values to about 10^{-6}, the SCEPTRE p-values for the positive controls were highly significant, and in particular, almost always more significant than the original empirical p-values, indicating greater sensitivity. Finally, we assessed the sensitivity of SCEPTRE on the Xie et al. data. Xie et al. conducted an arrayed CRISPR screen with bulk RNA-seq readout of ARL15-enh, a putative enhancer of gene *ARL15*. Both SCEPTRE and the bulk RNA-seq differential expression analysis rejected *ARL15* (and only *ARL15*) at an FDR of 0.1 after a Benjamini-Hochberg (BH) correction, increasing our confidence in the calibration and sensitivity of SCEPTRE (Figure 3e).

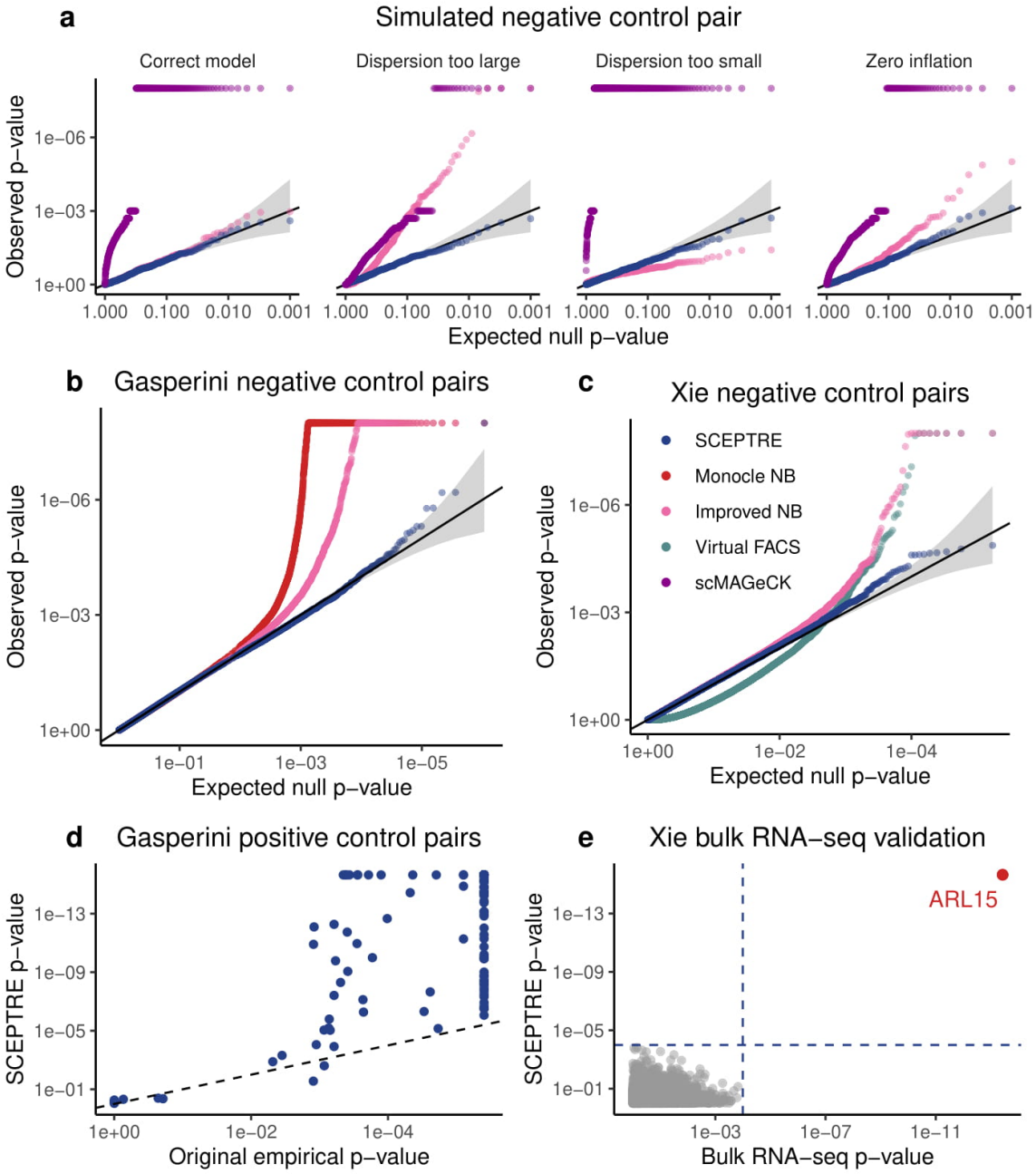
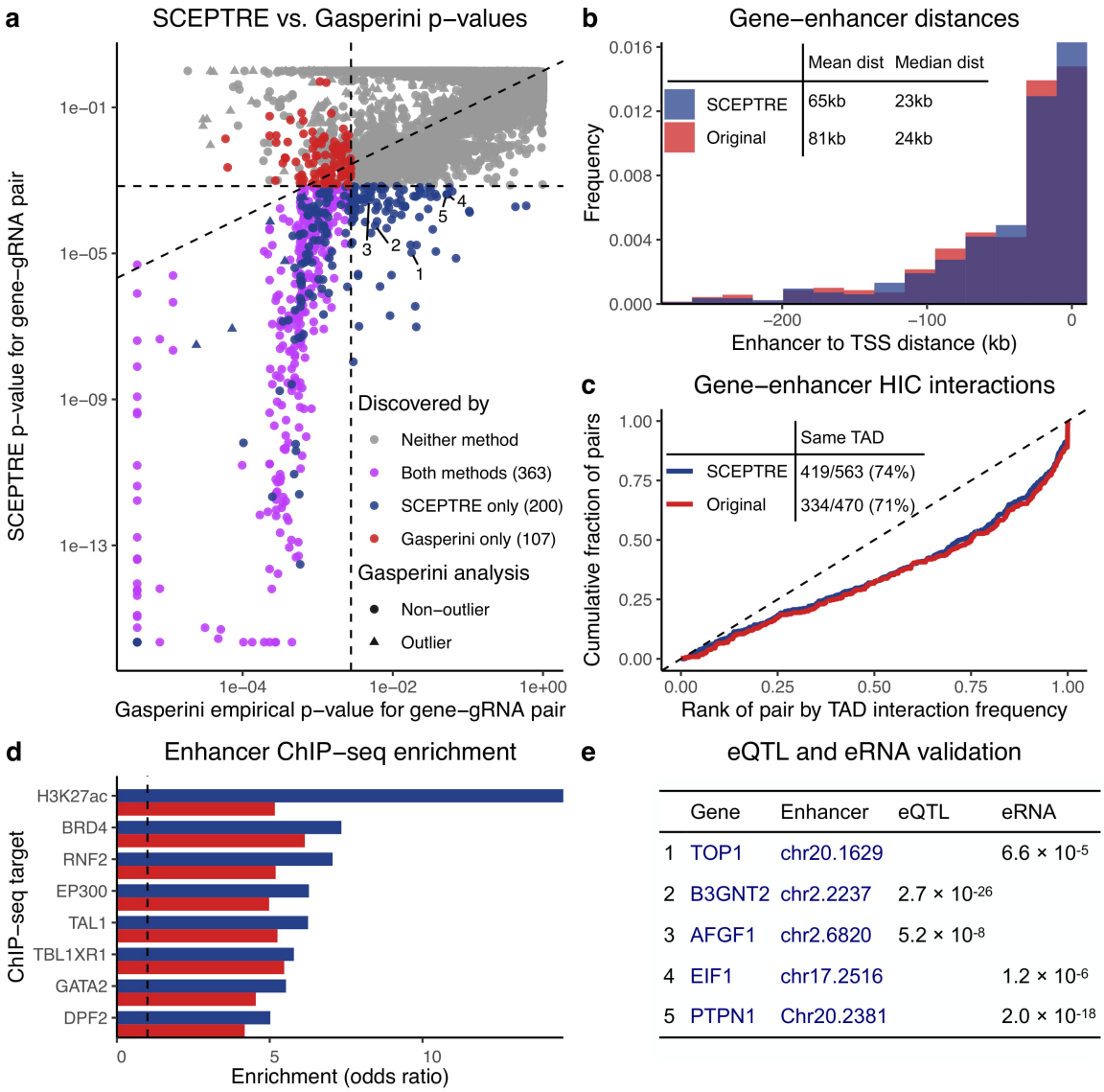

a
Simulated negative control pair
Correct model
Dispersion too large
Dispersion too small
Zero inflation
Observed p−value
1e−06
1e−03
1e+00
1.000
0.100
0.010
0.001
Expected null p−value
b
Gasperini negative control pairs
Observed p−value
1e−06
1e−03
1e+00
1e−01
1e−03
1e−05
Expected null p−value
c
Xie negative control pairs
SCEPTRE
Monocle NB
Improved NB
Virtual FACS
scMAGeCK
Observed p−value
1e−06
1e−03
1e+00
1e+00
1e−02
1e−04
Expected null p−value
d
Gasperini positive control pairs
SCEPTRE p−value
1e−13
1e−09
1e−05
1e−01
1e+00
1e−02
1e−04
Original empirical p−value
e
Xie bulk RNA−seq validation
ARL15
SCEPTRE p−value
1e−13
1e−09
1e−05
1e−01
1e−03
1e−07
1e−11
Bulk RNA−seq p−value

Figure 3: **SCEPTRE demonstrates good calibration and sensitivity under known ground truth. a,** Numerical simulation comparing three methods – SCEPTRE, improved negative binomial regression, and scMAGeCK-LR – on four simulated datasets. Only SCEPTRE maintained calibration despite model misspecification and confounder presence. **b-c**, Application of SCEPTRE, improved negative binomial regression, Monocle regression, and Virtual FACS to pairs of negative control gRNAs and genes in (b) the Gasperini et al. data, and (c) the Xie et al. data. Compared to the other methods, SCEPTRE showed excellent calibration. **d,** SCEPTRE p-values for Gasperini et al. TSS-targeting controls were highly significant, and in general, more significant than the original empirical p-values. **e,** Comparison of p-values produced by SCEPTRE for ARL15-enh to p-values produced by an arrayed, bulk RNA-seq CRISPR screen of ARL15-enh. The results of the two analyses coincided exactly, with both analyses rejecting gene *ARL15* (and only gene *ARL15*) after a BH correction. Dotted blue lines, rejection thresholds.

Analysis of candidate *cis*-regulatory pairs. We applied SCEPTRE to test all candidate *cis*-regulatory pairs in the Gasperini et al. ($n = 84,595$) and Xie et al. ($n = 3,511$) data. A given gene and gRNA were considered a "candidate pair" if the gRNA targeted a site within one Mb the gene's TSS. SCEPTRE discovered 563 and 135 gene-enhancer links at an FDR of 0.1 on the Gasperini et al. and Xie et al. data, respectively. We used several functional assays to quantify the enrichment of SCEPTRE's discovery set for regulatory biological signals, and we compared the SCEPTRE results to those of the original methods.

SCEPTRE's discovery set on the Gasperini et al. data was highly biologically plausible, and in particular, more enriched for biological signals of regulation than the original discovery set. Gasperini et al. discovered 470 gene-gRNA pairs at a reported FDR of 0.1. The SCEPTRE p-values and original empirical p-values diverged substantially: of the 670 gene-enhancer pairs discovered by either method, SCEPTRE and the original method agreed on only 363, or 54% (Figure 4a). Gene-enhancer pairs discovered by SCEPTRE were physically closer (mean distance $= 65$ kb) to each other than those discovered by the original method (mean distance $= 81$ kb; Figure 4b). Furthermore, SCEPTRE's gene-enhancer pairs fell within the same topologically associating domain (TAD) at a higher frequency (74%) than the original pairs (71%). Pairs within the same TAD showed similar levels of HI-C interaction frequency across methods, despite the fact that SCEPTRE discovered 85 more same-TAD pairs (Figure 4c). Finally, enhancers discovered by SCEPTRE showed improved enrichment across all eight cell-type relevant ChIP-seq targets reported by Gasperini et al. (Figure 4d). When we compared discoveries unique to SCEPTRE ($n = 200$) against those unique to the original method ($n = 107$), the disparities became

more extreme (Figure S3). For example, only 57% of pairs unique to the original method fell within the same TAD, compared to 73% unique to SCEPTRE. We concluded that many pairs in the Gasperini et al. discovery set likely were false positives.



	Gene	Enhancer	eQTL	eRNA
1	TOP1	chr20.1629		6.6×10^{-5}
2	B3GNT2	chr2.2237	2.7×10^{-26}	
3	AFGF1	chr2.6820	5.2×10^{-8}	
4	EIF1	chr17.2516		1.2×10^{-6}
5	PTPN1	Chr20.2381		2.0×10^{-18}

Figure 4: **Application of SCEPTRE to Gasperini et al. data yields biologically plausible gene-enhancer links. a,** Comparison of the original empirical p-values to those obtained from SCEPTRE. The two analysis methods differed substantially, with 200 gene-enhancer links discovered only by SCEPTRE and 107 discovered only by the original

analysis. Annotatations correspond to pairs in panel (b). **b,** Distribution of distances from TSS to upstream paired enhancers. Compared to the original analysis, SCEPTRE paired genes with nearer enhancers on average. **c,** For those gene-enhancer pairs falling in the same TAD, the cumulative distribution of the fractional rank of the HI-C interaction frequency compared to other distance-matched loci pairs within the same TAD. SCEPTRE showed similar enrichment despite finding 25% more within-TAD pairs. Inset table shows gene-enhancer pairs falling in the same TAD. SCEPTRE found 93 more total pairs, and a higher percentage of pairs fell within the same TAD. **d,** Enrichment of ChIP-seq signal from seven cell-type relevant transcription factors and one histone mark (H3K27ac) among paired enhancers. SCEPTRE showed stronger enrichment across all ChIP-seq targets. **e**, Five gene-enhancer pairs discovered by SCEPTRE but not the original analysis, each supported by a whole blood GTEx eQTL or FANTOM enhancer RNA correlation p-value.

We highlight several especially interesting gene-enhancer pairs discovered by SCEPTRE. Five discoveries (Figure 4a labels 1-5; Figure 4e) were nominated as probable gene-enhancer links by eQTL[18] and eRNA[19] p-values in relevant tissue types. The SCEPTRE p-values for these pairs were 1-2 orders of magnitude smaller than the original empirical p-values, hinting at SCEPTRE's greater sensitivity. Additionally, six pairs (Figure 4a, blue triangles) were discovered by SCEPTRE but discarded as outliers by the original analysis, underscoring SCEPTRE's ability to handle genes with arbitrary expression distributions.

We repeated the same functional analyses for the SCEPTRE discoveries on the Xie et al. data, comparing SCEPTRE's results to those of Xie et al. Xie et al.'s analysis method, Virtual FACS, outputted "significance scores" rather than p-values (see section "Analysis challenges"). Because significance scores cannot be subjected to multiple hypothesis testing correction procedures (like BH), we compared the top 135 Virtual FACS pairs (ranked by significance score) against the set of 135 (FDR $= 0.1$) SCEPTRE discoveries (Figure 5a; see Methods). Of the 186 pairs in either set, SCEPTRE and Virtual FACS agreed on only 84, or 45%. The SCEPTRE discoveries were more biologically plausible: compared to the Virtual FACS pairs, the SCEPTRE pairs were (i) physically closer (Figure 5b), (ii) more likely to fall within the same TAD (Figure 5c), (iii) more likely to interact when in the same TAD (Figure 5c), and (iv) more enriched for six of eight cell-type relevant ChIP-seq targets (Figure 5d).

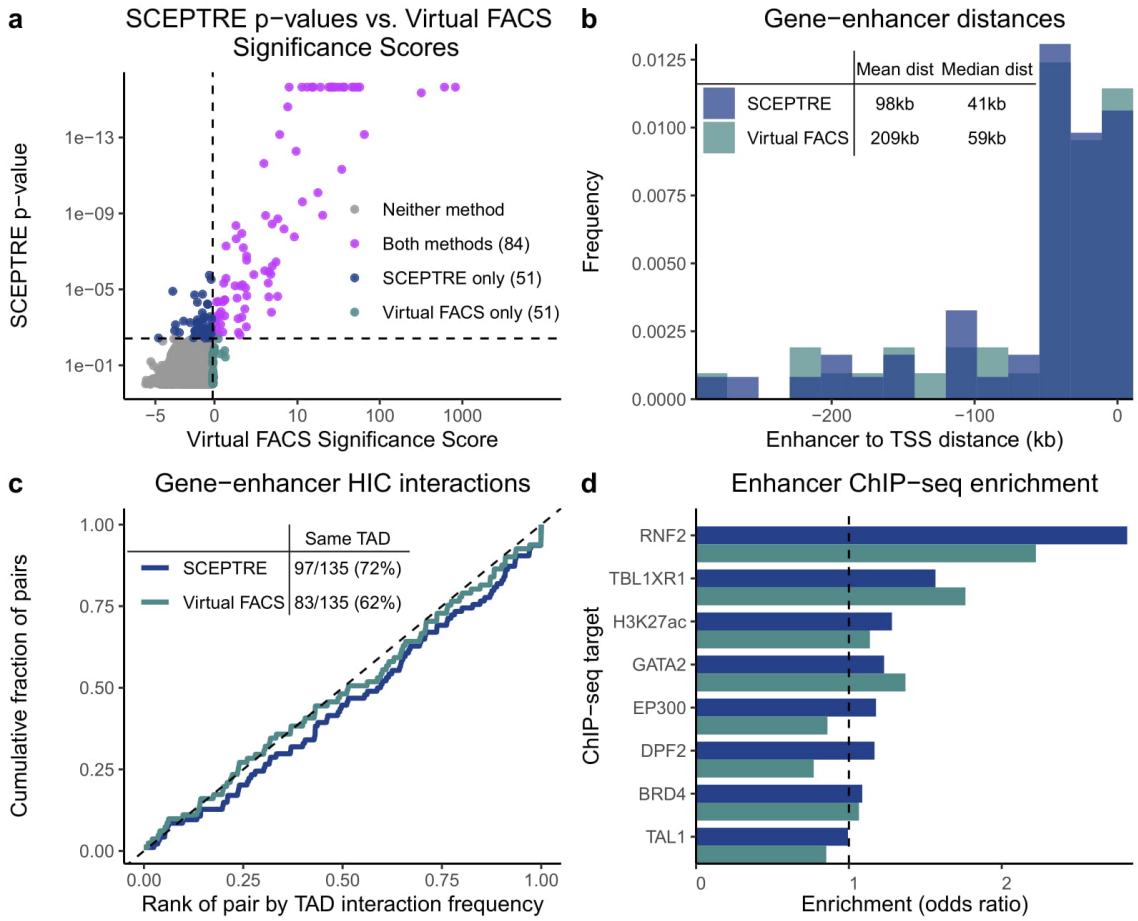


Figure 5: **SCEPTRE discovers biologically plausible gene-enhancer links on Xie et al. data. a,** Comparison of SCEPTRE p-values to Virtual FACS significance scores. Significant SCEPTRE p-values $(n = 135)$ are colored in blue and purple, and the top 135 Virtual FACS pairs, as ranked by significance score, are colored in gray-green and purple. **b-d,** These panels are similar to the corresponding panels in Figure 4. The SCEPTRE pairs were more enriched for biological signals associated with gene-gRNA links.

Discussion

In this work we illustrated a variety of statistical challenges arising in the analysis of high-MOI single-cell CRISPR screens, leaving existing methods (based on parametric expres-

sion models, permutations, or negative control data) vulnerable to miscalibration. To address these challenges, we developed SCEPTRE, a resampling method based on modeling the probability a given gRNA will be detected in a given cell, based on that cell's technical factors. We found that SCEPTRE exhibited very good calibration despite the presence of confounding technical factors and misspecification of single-cell gene expression models. We implemented computational accelerations to bring the cost of the resampling-based methodology down to well within an order of magnitude of the traditional negative binomial parametric approach, making it quite feasible to apply for large-scale data. We used SCEPTRE to reanalyze the Gasperini et al. and Xie et al. data. While our analysis replicated many of their findings, it also clarified other relationships, removing a large set ($> 20\%$ for Gasperini) of pairs that exhibited a weak relationship and adding an even larger set ($> 40\%$ for Gasperini) of new, biologically plausible gene-enhancer relationships. These links were supported by orthogonal evidence from eQTL, enhancer RNA co-expression, ChIP-seq, and HI-C data.

As an example application of SCEPTRE, we highlight *STING-seq*, a platform that we developed in parallel to the current work in an independent study[14]. STING-seq leverages biobank-scale GWAS data and single-cell CRISPR screens to map noncoding, disease-associated variants at scale. First, we used statistical fine-mapping to identify a set of 88 putatively causal variants from 56 loci associated with quantitative blood traits. We perturbed the selected variants at high MOI in K562 cells using an improved CRISPRi platform and sequenced gRNAs and transcriptomes in individual cells using ECCITE-seq[6], a protocol that enables the profiling of multiple modalities and the direct capture of gRNAs.

We then applied SCEPTRE to quantify associations between perturbations and changes in gene expression in *cis* (within 500 kb) and *trans*. SCEPTRE confidently mapped 37 noncoding variants to their *cis* target genes, in some cases identifying a causal variant among a set of candidate variants in strong LD. Nine variants were found to regulate a gene other than the closest gene, and four variants were found to regulate multiple genes, an apparent example of pleiotropy. Several perturbations lead to widespread changes in gene expression, illuminating *trans*-effects networks. For example, two variants that were found to regulate the transcription factor *GFI1B* in *cis* altered the expression of hundreds of genes in *trans* upon perturbation; these differentially expressed genes were strongly enriched for GFI1B binding sites and blood disease GWAS hits. We concluded on the basis of this study SCEPTRE can power the systematic dissection regulatory networks underlying GWAS associations.

Despite these exciting results, key challenges remain in the analysis of single-cell CRISPR screens. Currently, SCEPTRE does not estimate the effect sizes of perturbations, disentangle interactions among perturbed regulatory elements[20,21], or leverage information

from off-targeting gRNAs to improve power. Such extensions could be implemented by harnessing more sophisticated, multivariate models of gRNA detection or applying methods for estimating variable importance in the presence of possibly misspecified models[22]. The statistical challenges that we identified in this study – specifying an accurate expression model, accounting for technical factors – and the solutions that we proposed – conditional resampling, massively parallel association testing – will help guide the development of future versions of SCEPTRE.

Single-cell CRISPR screens will play a key role in unravelling the regulatory architecture of the noncoding genome. Technological improvements and methodological innovations will increase the scope, scale, and variety of theses screens over the coming years. For example, screens of candidate CREs could be extended to different, disease-relevant cell types and tissues (although this remains a challenge); new combinatorial indexing strategies, such as scifi-RNA-seq, could enable the scaling-up of such screens to millions of cells[23]; different CRISPR technologies, such as CRISPRa, could enable the activation, rather than repression, of candidate CREs, yielding new insights; and information-rich, multimodal single-cell readouts could strengthen conclusions drawn about regulatory relationships[24]. SCEPTRE is a flexible, robust, and efficient method: it has now successfully been applied to three single-cell CRISPR screen datasets, across two technologies (CROP-seq and ECCITE-seq), to map regulatory relationships both in *cis* and in *trans*. We expect SCEPTRE to facilitate the analysis of future single-cell screens of the noncoding genome, advancing understading of CREs and enabling the detailed interpretation of GWAS results.

References

1. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics* **102**, 717–730 (2018). URL https://doi.org/10.1016/j.ajhg.2018.04.002.

2. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* (2020).

3. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866 (2016).

4. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867–1882.e21 (2016).

5. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods* **14**, 297–301 (2017).

6. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods* **16**, 409–412 (2019).

7. 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 (2017).

8. Xie, S., Armendariz, D., Zhou, P., Duan, J. & Hon, G. C. Global Analysis of Enhancer Targets Reveals Convergent Enhancer-Driven Regulatory Modules. *Cell Reports* **29**, 2570–2578.e5 (2019).

9. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390.e19 (2019).

10. Yang, L. *et al.* Linking genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biology* **21** (2020).

11. Hill, A. J. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nature Methods* **15**, 271–274 (2018).

12. Replogle, J. M. *et al.* Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology* (2020).

13. Candès, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577 (2018).

14. Morris, J. A. *et al.* Discovery of target genes and pathways of blood trait loci using pooled crispr screens and single cell rna sequencing. *bioRxiv* (2021).

15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

16. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nature Methods* **14**, 309–315 (2017).

17. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 1–15 (2019).

18. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

19. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

20. Zamanighomi, M. *et al.* GEMINI: A variational Bayesian approach to identify genetic interactions from combinatorial CRISPR screens. *Genome Biology* **20**, 1–10 (2019).

21. Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).

22. Zhang, L. & Janson, L. Floodgate : inference for model-free variable importance. *arXiv* 1–67 (2020).

23. Datlinger, P. *et al.* Ultra-high throughput single-cell rna sequencing by combinatorial fluidic indexing. *BioRxiv* (2019).

24. Pierce, S. E., Granja, J. M. & Greenleaf, W. J. High-throughput single-cell chromatin accessibility crispr screens enable unbiased identification of regulatory networks in cancer. *bioRxiv* (2020).

25. Towns, J. *et al.* XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **16**, 62–74 (2014).

26. Nystrom, N. A., Levine, M. J., Roskies, R. Z. & Scott, J. R. Bridges: A Uniquely Flexible HPC Resource for New Communities and Data Analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15, 30:1—-30:8 (ACM, New York, NY, USA, 2015).

27. Liu, M., Katsevich, E., Ramdas, A. & Janson, L. Fast and Powerful Conditional Randomization Testing via Distillation. *arXiv* (2020). URL https://arxiv.org/abs/2006.03980.

28. Finner, H. & Roters, M. On the false discovery rate and expected type I errors. *Biometrical Journal* **43**, 985–1005 (2001).

29. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

30. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

31. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database : the journal of biological databases and curation* **2017**, 1–17 (2017).

Acknowledgements We are indebted to Molly Gasperini, Jacob Tome, and Andrew Hill for clarifying several aspects of their data analysis[9] and to the Shendure lab for providing extensive feedback on an earlier draft of this paper. We thank Shiqi Xie for providing guidance on using the Xie et al. 2019 data, Wei Li for a helpful discussion on scMAGeCK, and John Morris for useful feedback on the SCEPTRE code. Finally, we thank Tom Norman, Atray Dixit, and Wesley Tansey for useful discussions on single-cell CRISPR screens. This work was supported, in part, by National Institute of Mental Health (NIMH) grant R01MH123184 as well as SFARI Grant 575547. Part of the data analysis used the Extreme Science and Engineering Discovery Environment (XSEDE)[25], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system[26], which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to E.K. (email: ekatsevi@wharton.upenn.edu).

Author contributions GK and KR identified the problem, conceived the research, and provided supervision. GK developed the method with input from TB and KR. TB and GK implemented the

method with assistance from JM. TB, XW, and GK performed the analyses. JM provided guidance on interpretation and communication of results. TB and GK wrote the manuscript with input from all authors.

Methods

Gasperini et al. and Xie et al. data. Gasperini et al. used CROP-seq[5,11] to transduce a library of CRISPR guide RNAs (gRNAs) into a population of 207,324 K562 cells expressing the Cas9-KRAB repressive complex at a high multiplicity of infection. Each cell received an average of 28 perturbations. The gRNA library targeted 5,779 candidate enhancers, 50 negative controls, and 381 positive controls. Xie et al. used Mosaic-seq[7,8] to perturb at a high multiplicity of infection 518 putative enhancers in a population of 106,670 Cas9-KRAB-expressing K562 cells. Each putative enhancer was perturbed in an average of 1,276 cells.

***Cis* and *in silico* negative control pairs for Xie et al. data.** We generated the set of candidate *cis* gene-enhancer relationships on the Xie et al. data by pairing each protein-coding gene with each gRNA targeting a site within 1 Mb of the TSS of the gene. This procedure resulted in 3,511 candidate *cis* gene-enhancer links that we tested using SCEPTRE and Virtual FACS.

To generate the set of *in silico* negative control pairs for calibration assessment, we (i) identified gRNAs that targeted sites far (> 1 Mb) from the TSSs of known transcription factor genes and (ii) paired these gRNAs with genes located on other chromosomes. We excluded all pairs containing genes known to be transcription factors, and we downsampled the pairs so that each gRNA was matched to 500 genes. The final *in silico* negative control set consisted of $85,000$ pairs, the elements of which were not expected to exhibit a regulatory relationship.

Conditional randomization test. Consider a particular gene/gRNA pair. For each cell $i = 1, \ldots, n$, let $X_i \in \{0, 1\}$ indicate whether the gRNA was present in the cell, let $Y_i \in \{0, 1, 2, \ldots\}$ be the gene expression in the cell, defined as the number of unique molecular identifiers (UMIs) from this gene, and let $Z_i \in \mathbb{R}^d$ be a list of cell-level technical factors. Letting $(X, Y, Z) = \{(X_i, Y_i, Z_i)\}_{i=1}^n$, consider any test statistic $T(X, Y, Z)$ measuring the effect of the gRNA on the expression of the gene. The conditional randomization test[13] is based on resampling the gRNA indicators independently for each cell. Letting $\pi_i = \mathbb{P}[X_i = 1 | Z_i]$, define random variables

$$\widetilde{X}_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i). \tag{1}$$

Then, the CRT p-value is given by

$$p_{\text{CRT}} = \mathbb{P}[T(\widetilde{X}, Y, Z) \geq T(X, Y, Z) \mid X, Y, Z]. \tag{2}$$

This translates to repeatedly sampling $\widetilde{X}$ from the distribution (1), recomputing the test statistic with X replaced by $\widetilde{X}$, and defining the p-value as the probability the resampled test statistic exceeds the original. Under the null hypothesis that the gRNA perturbation does not impact the cell (adjusting for technical factors), i.e. $Y \perp\!\!\!\perp X \mid Z$, we obtain a valid p-value (2), *regardless of the expression distribution* $Y|X, Z$ *and regardless of the test statistic* T. We choose as a test statistic T the z-score of X_i obtained from a negative binomial regression of Y_i on X_i and Z_i:

$$Y_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \alpha); \quad \log(\mu_i) = \beta_0 + X_i\beta + Z_i^T\gamma, \tag{3}$$

where α is the dispersion. Following Hafemeister and Satjia[17], we estimate α by pooling dispersion information across genes, and we include sequencing depth as an entry in the vector of technical factors Z_i (see section *Improvements to the negative binomial approach*).

Accelerations to the conditional randomization test. We implemented computational accelerations to the conditional randomization test. First, we employed the recently proposed[27] *distillation* technique to accelerate the recomputation of the negative binomial regression for each resample. The idea is to use a slightly modified test statistic, consisting of two steps:

1. Fit $(\widehat{\beta}_0, \widehat{\gamma})$ from the negative binomial regression (3) except without the gRNA term:

$$Y_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \alpha); \quad \log(\mu_i) = \beta_0 + Z_i^T\gamma. \tag{4}$$

2. Fit $\widehat{\beta}$ from a negative binomial regression with the estimated contributions of Z_i from step 1 as offsets:

$$Y_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \alpha); \quad \log(\mu_i) = X_i\beta + \widehat{\beta}_0 + Z_i^T\widehat{\gamma}. \tag{5}$$

Conditional randomization testing with this two step test statistic, which is nearly identical to the full negative binomial regression (3), is much faster. Indeed, since the first step is not a function of X_i, it remains the same for each resampled triple $(\widetilde{X}, Y, Z)$. Therefore, only the second step must be recomputed with each resample, and this step is faster because it involves only a univariate regression.

Next, we accelerated the second step above using the sparsity of the binary vector $(X_1, \ldots, X_n)$ (or a resample of it). To do so, we wrote the log-likelihood of the reduced negative binomial regression (5) as follows, denoting by $\ell(Y_i, \log(\mu_i))$ the negative binomial log-likelihood:

$$\begin{aligned}\sum_{i=1}^{n} \ell(Y_i, X_i\beta + \widehat{\beta}_0 + Z_i^T\widehat{\gamma}) &= \sum_{i:X_i=0} \ell(Y_i, \widehat{\beta}_0 + Z_i^T\widehat{\gamma}) + \sum_{i:X_i=1} \ell(Y_i, \beta + \widehat{\beta}_0 + Z_i^T\widehat{\gamma}) \\ &= C + \sum_{i:X_i=1} \ell(Y_i, \beta + \widehat{\beta}_0 + Z_i^T\widehat{\gamma}).\end{aligned}$$

This simple calculation shows that, up to a constant that does not depend on β, the negative binomial log-likelihood corresponding to the model (5) is the same as that corresponding to the model with only intercept and offset term for those cells with a gRNA:

$$Y_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \alpha); \quad \log(\mu_i) = \beta + \widehat{\beta}_0 + Z_i^T\widehat{\gamma}, \quad \text{for } i \text{ such that } X_i = 1. \tag{6}$$

The above negative binomial regression is therefore equivalent to equation (5), but much faster to compute, because it involves much fewer cells. For example, in the Gasperini et al. data, each gRNA is observed in only about 1000 of the 200,000 total cells.

SCEPTRE methodology. In practice, we must estimate the gRNA probabilities π_i as well as the p-value p_{CRT}. This is because usually we do not know the distribution $X|Z$ and cannot compute the conditional probability in equation (2) exactly. We propose to estimate π_i via logistic regression of X on Z, and to estimate p_{CRT} by resampling $\widetilde{X}$ a large number of times and then fitting a skew-t distribution to the resampling null distribution $T(\widetilde{X}, Y, Z)|X, Y, Z$. We outline SCEPTRE below:

1. Fit technical factor effects $(\widehat{\beta}_0, \widehat{\gamma})$ on gene expression using the negative binomial regression (4).

2. Extract a z-score $z(X, Y, Z)$ from the reduced negative binomial regression (6).

3. Assume that
$$X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \tau_0 + Z_i^T\tau \tag{7}$$
for $\tau_0 \in \mathbb{R}$ and $\tau \in \mathbb{R}^d$, and fit $(\widehat{\tau}_0, \widehat{\tau})$ via logistic regression of X on Z. Then, extract the fitted probabilities $\widehat{\pi}_i = (1 + \exp(-(\widehat{\tau}_0 + Z_i^T\widehat{\tau})))^{-1}$.

4. For $b = 1, \ldots, B$,

- Resample the gRNA assignments based on the probabilities $\widehat{\pi}_i$ to obtain $\widetilde{X}^b$ (1).
- Extract a z-score $z(\widetilde{X}^b, Y, Z)$ from the reduced negative binomial regression (6).

5. Fit a skew-t distribution $\widehat{F}_{\text{null}}$ to the resampled z-scores $\{z(\widetilde{X}^b, Y, Z)\}_{b=1}^B$.

6. Return the p-value $\widehat{p}_{\text{SCEPTRE}} = \mathbb{P}[\widehat{F}_{\text{null}} \leq z(X, Y, Z)]$.

In our data analysis we used $B = 500$ resamples.

Numerical simulation to assess calibration. We simulated one gene Y_i, five gRNAs $X_{i1}, X_{i2}, \ldots, X_{i5}$, and two confounders Z_{i1}, Z_{i2} in $n = 1000$ cells. We generated the confounders Z_{i1} and Z_{i2} by sampling with replacement the batch IDs and log-transformed sequencing depths of the cells in the Gasperini dataset. The batch ID confounder Z_{i1} was a binary variable, as the Gasperni data included two batches. Next, we drew the gRNA indicators $X_{i1}, X_{i2}, \ldots, X_{i5}$ i.i.d. from the logistic regression model (7), with $\tau_0 = -7, \tau_1 = -2$, and $\tau_2 = 0.5$. We selected these parameters to make the probability of gRNA occurrence about 0.04 across cells. Finally, we drew the gene expression Y_i from the following zero-inflated negative binomial model:

$$Y_i \sim \lambda\delta_0 + (1-\lambda)\text{NegBin}(\mu_i, \alpha), \;\; \log(\mu_i) = \beta_0 + Z_i^T\beta.$$

Note that gRNA presence or absence does not impact gene expression in this model. We set $\beta_0 = -2.5, \beta_1 = -2, \beta_2 = 0.5$ to make the average gene expression about 4 across cells. We generated the four datasets shown in Figure 3a by setting the dispersion parameter α and the zero inflation rate parameter λ equal to the following values:

$$(\lambda_1, \alpha_1) = (0, 1); \; (\lambda_2, \alpha_2) = (0, 5); \; (\lambda_3, \alpha_3) = (0, 0.2); \; (\lambda_4, \alpha_4) = (0.25, 1).$$

For the first, the negative binomial model is correctly specified. For the second and third, the dispersion estimate of 1 is too small and too large, respectively. The last setting exhibits zero inflation.

We applied SCEPTRE, negative binomial regression, and scMAGeCK-LR[10] to the four problem settings, each with $n_{sim} = 500$ repetitions. The negative binomial method, and in turn SCEPTRE, was based on the z statistic from the Hafemeister-inspired negative binomial model (3) with $\alpha = 1$. scMAGeCK-LR differs from SCEPTRE and the negative binomial method in that scMAGeCK-LR computes p-values for all enhancers simultaneously. Thus, to facilitate comparisons across methods, we plotted p-values corresponding to enhancer X_{i1} only. We used $B = 500$ resamples for SCEPTRE and $B = 1000$ permutations for scMAGeCK-LR, the default choices for these methods.

Definition of Gasperini et al. discovery set. Gasperini et al. reported a total of 664 gene-enhancer pairs, identifying 470 of these as "high-confidence." We chose to use the latter set, rather than the former, for all our comparisons. Gasperini et al. carried out their ChIP-seq and HI-C enrichment analyses only on the high-confidence discoveries, so for those comparisons we do the same. Furthermore, the 664 total gene-enhancer pairs reported in the original analysis were the result of a BH FDR correction that included not only the candidate enhancers but also hundreds of positive controls. While Bonferroni corrections can only become more conservative when including more hypotheses, BH corrections are known to become anticonservative when extra positive controls are included[28]. To avoid this extra risk of false positives, we chose to use the "high-confidence" set throughout.

Xie et al. significance scores and discovery set. Xie et al. reported a local (or *cis*) discovery set, which consisted of gene-gRNA pairs with a significance score of greater than zero (see original manuscript for definition of "significance score"[8]; cutoff of zero arbitrary). This discovery set was not directly comparable to the SCEPTRE discovery set. First, the candidate set of *cis* gene-gRNA pairs tested by Xie et al. consisted of gRNAs within *two* Mb of a protein-coding gene *or* long-noncoding RNA. Our candidate *cis* set, by contrast, consisted of gRNAs within *one* Mb of a protein-coding gene. We defined our candidate *cis* set differently than Xie et al. to mantain consistency with Gasperini et al. Second, Xie et al. appear to have used a significantly more conservative threshold than Gasperini et al. in defining their discovery set, but this was challenging to ascertain given the impossibility of FDR correction on the significance scores. To enable a meaningful comparison between Virtual FACS and SCEPTRE, we therefore ranked the Virtual FACS pairs by their significance score and selected the top n pairs, where n was the size of the SCEPTRE discovery set at FDR 0.1.

ChIP-seq, HI-C enrichment analyses. ChIP-seq and HI-C enrichment analyses analyses (see Figures 4e-f and S4) were carried out almost exactly following Gasperini et al. The only change we made is in our quantification of the ChIP-seq enrichment (Figure 4f). We use the odds ratio of a candidate enhancer being paired to a gene, comparing the top and bottom ChIP-seq quintiles.

Data availability

Analysis results are available online at https://drive.google.com/drive/folders/1ynZRMvGtFxfBiD0zAcuIYjNeS8Jj4AP9?usp=sharing. All analysis was performed on publicly available data. The Gasperini et al. CRISPR screen data[9] are available at www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120861. The Xie et al. single-cell and bulk CRISPR screen data are available at www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129837. The ChIP-seq data are taken from the ENCODE project[29] and are available at www.encodeproject.org/. The HI-C enrichment analysis is based on the data from Rao et al.[30], available at www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. The eQTL and eRNA co-expression p-values are taken from the GeneHancer, database[31] available as part of GeneCards (www.genecards.org/).

Code availability

The `sceptre` R package is available at https://github.com/timothy-barry/sceptre. The scripts used to run the analyses reported in this paper are available at https://github.com/timothy-barry/sceptre-manuscript.

Supplementary Figures

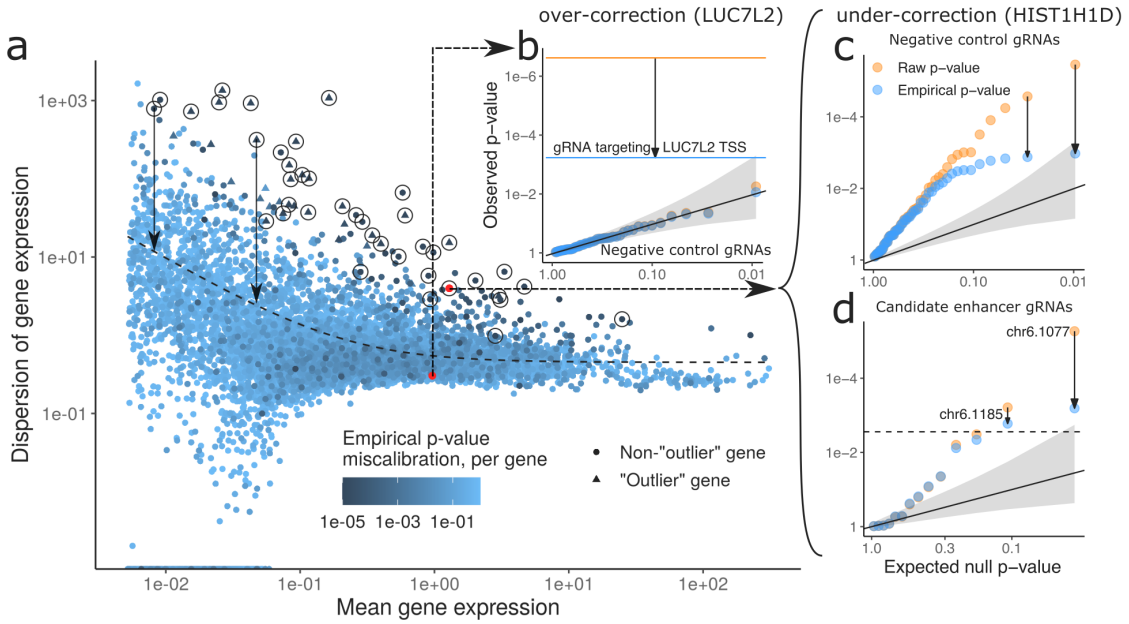


Figure S1: **Gasperini et al.'s empirical correction is insufficient to correct for miscalibration**. **a,** Dispersion estimation procedure employed leads to miscalibration for high-dispersion genes, which the empirical correction does not adequately correct for, as measured by KS test applied to empirical p-values per gene (point colors). **b,** Raw p-values already well-calibrated for *LUC7L2* gene, so empirical correction unnecessarily shrinks the significance of the association with TSS-targeting gRNA, depicted by horizontal lines, by three orders of magnitude. **c,** Empirical correction not strong enough for *HIST1H1D*, which is among circled genes in panel a, which have an NTC-based miscalibration p-value smaller than the Bonferroni threshold. **d**, Under-correction leads to two potential false discoveries for *HIST1H1D*. Dashed horizontal line represents the multiple testing threshold.

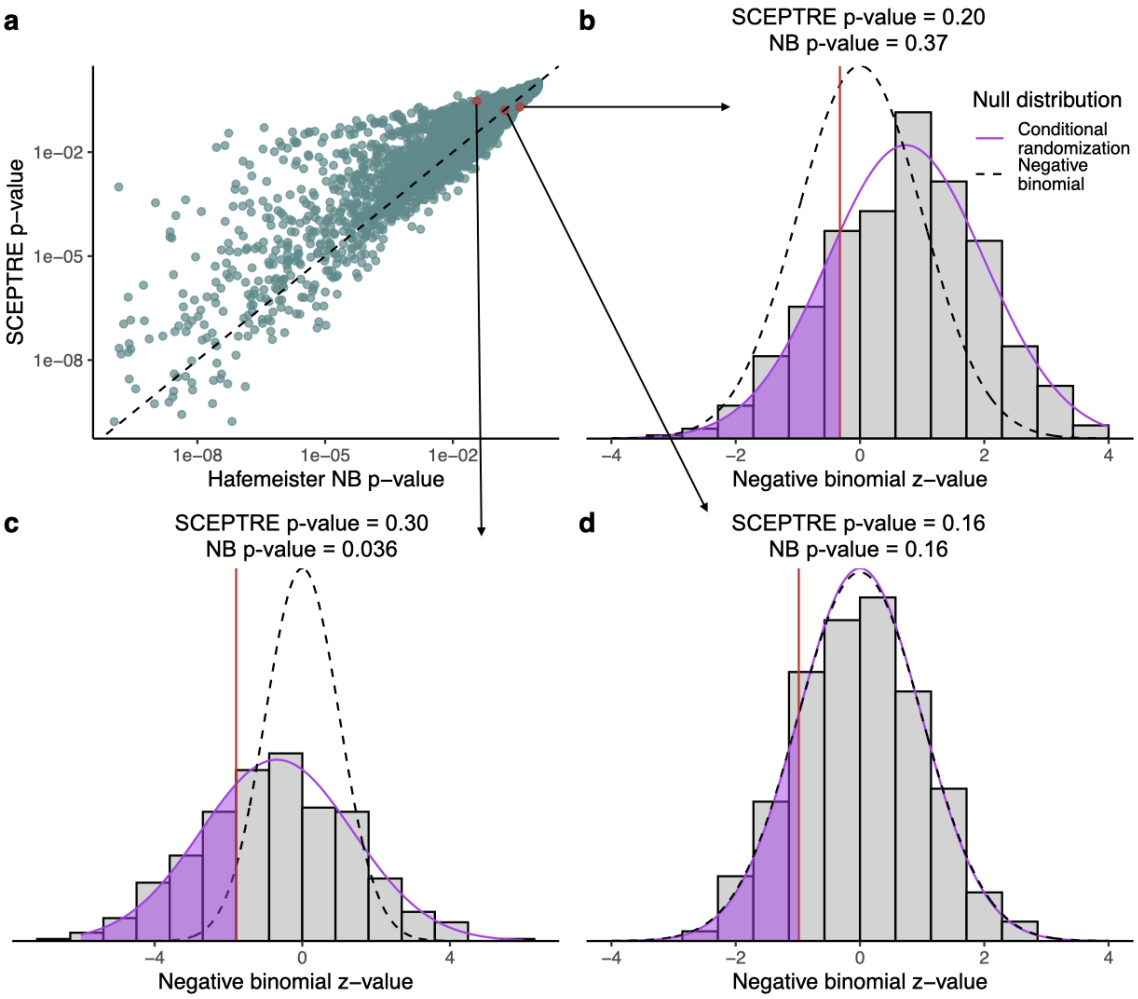


Figure S2: **Comparison of negative binomial and conditional resampling p-values based on the same test statistic. a,** The standard parametric negative binomial p-value versus that obtained from the same test statistic by conditional resampling, for each gene / candidate enhancer pair (both truncated at 10^{-10} for visualization). The two can diverge fairly substantially. **b-d,** Parametric and conditional resampling null distributions for the negative binomial z-value in three cases: the conditional resampling p-value is more significant (b), the parametric p-value is more significant (c), the two p-values are about the same (d).

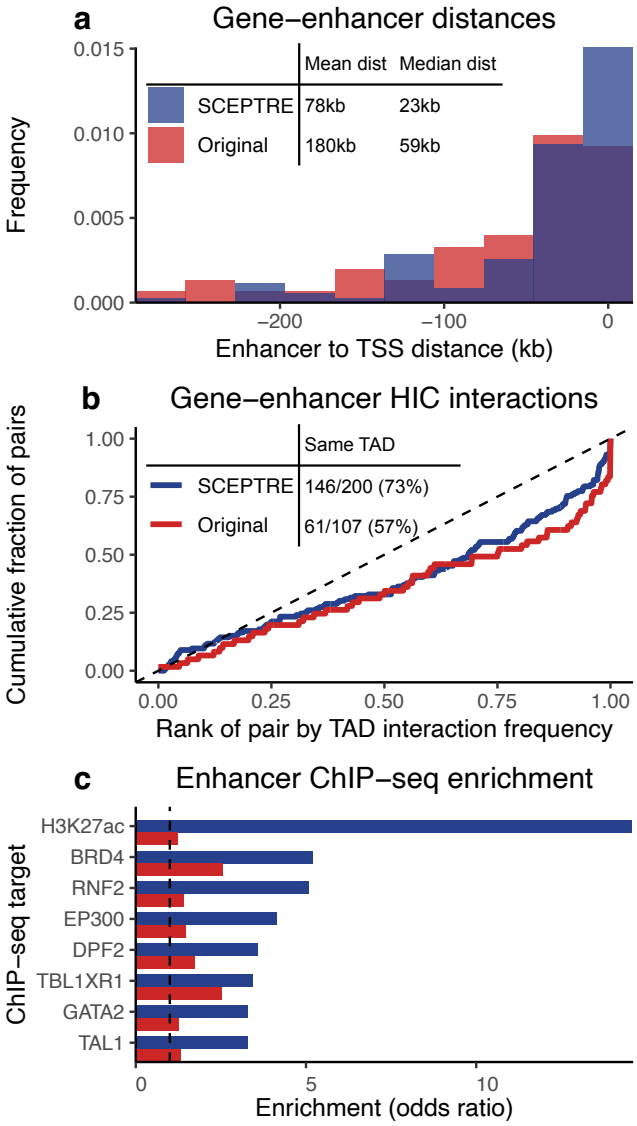


Figure S3: **Discoveries unique to SCEPTRE exhibited greater enrichment for biological signal than those unique to the original method.** This figure shows the 200 discoveries unique to SCEPTRE and the 107 discoveries unique to the original method (in contrast to Figure 4, which shows the entire discovery set of both methods). **a,** On average, enhancers discovered by SCEPTRE were less than half the distance to their target genes than those discovered by Gasperini et al. **b,** 73% of the gene-enhancer pairs discovered by SCEPTRE fell within the same TAD, in contrast to 57% of those discovered by the original method. HI-C interaction frequency was similar across methods (though slightly higher for the original), despite the fact that SCEPTRE found 85 more same-TAD pairs. **c,** Enhancers returned by SCEPTRE showed significantly greater enrichment across all ChIP-seq targets, especially H3K27ac.

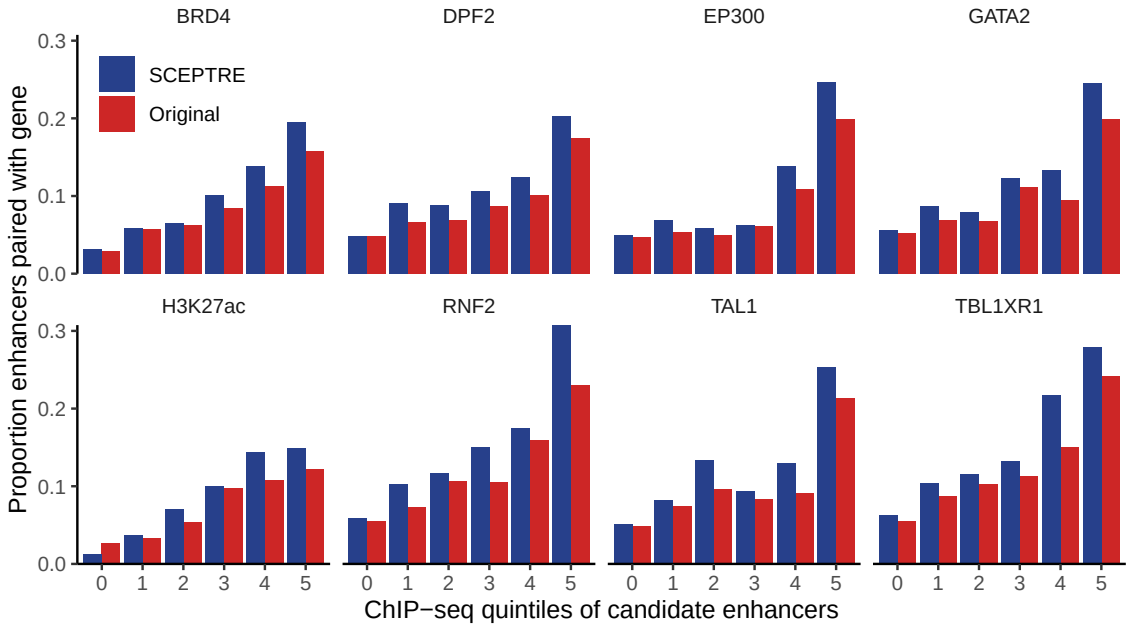


Figure S4: **Details on ChIP-seq enrichment analysis.** Fraction of candidate enhancers paired with a gene, broken down by quintile of ChIP-seq signal (0 means the candidate enhancer did not overlap a ChIP-seq peak), based on which the odds ratios in Figure 4e were computed. Both methods generally pair candidate enhancers in higher ChIP-seq quintiles more frequently, but this enrichment is more pronounced in SCEPTRE across all eight ChIP-seq targets.