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

Eugene Katsevich[1], Timothy Barry[2], and Kathryn Roeder[2,3]

[1]*Department of Statistics, Wharton School, University of Pennsylvania*
[2]*Department of Statistics and Data Science, Carnegie Mellon University*
[3]*Computational Biology Department, Carnegie Mellon University*

Single-cell CRISPR screens are an emerging biotechnology promising unprecedented insights into gene regulation. 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 large-scale 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. SCETPRE demonstrates excellent calibration and sensitivity on the CRISPR screen data and yields 200 new regulatory relationships, many of which are supported by existing functional data.

Single cell CRISPR screens (implemented independently as Perturb-seq[1,2], CRISP-seq[3], and CROP-seq[4]) facilitate unprecedented insights into gene regulation by measuring the impact of precisely targeted perturbations on the whole transcriptome. These pooled screens deliver a library of CRISPR perturbations, targeting either genes or non-coding regulatory elements, to a population of cells and then employ single-cell RNA sequencing (scRNA-seq) to identify the perturbations each cell received and measure its transcriptional response. Single-cell CRISPR screens have been scaled up through multiplexing to systematically map gene-enhancer relationships[5–7], applied using combinatorial perturbation libraries to study genetic interactions[8], and extended to single-cell protein expression readout[9,10]. Given their scale, resolution, and versatility, these assays are likely to be a major driver of discovery in functional genomics and its applications to the molecular understanding of human disease in the coming decade[11].

Despite their promise, the data obtained from such assays pose significant statistical challenges. In particular, various difficulties have been encountered in calibrating tests of association between a CRISPR perturbation and the expression of a gene. Gasperini et al.[6] found substantial inflation in their negative binomial regression based p-values for negative control perturbations. Similarly, Xie et al.[7] found an excess of false positive hits in their rank-based Virtual FACS analysis. Finally, Yang et al.[12] found that their permutation-based scMAGeCK-RRA method deems almost all gene-enhancer pairs significant in a reanalysis of the Gasperini et al. data. While these works propose ad hoc fixes to improve calibration, it is clear that new analysis methodology is indispensable to fully realize the potential of single-cell CRISPR screens. However, a recent review[13] noted that few such methods have been proposed so far.

In this work, we explore statistical challenges of single-cell CRISPR screens and present a novel analysis methodology to address them. In addition to known scRNA-seq analysis challenges such as expression normalization and discreteness[14–16], we identify a key challenge that sets 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[1,17,18]. 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 such as those based on ranks or permutations. Indeed, the symmetry among cells assumed by these approaches breaks down in the presence of technical factors varying across cells. On the other hand, parametric modeling of single-cell expression data is also fraught with challenges and remains an active area of research.

To address these challenges, we propose SCEPTRE (analysis of Single-CEll PerTurbation screens via conditional REsampling; pronounced "scepter"). This resampling method-

ology is based on the conditional randomization test[19], a recently proposed statistical methodology that, like traditional nonparametric methods, does not rely on correct specification of the outcome (expression) model. Unlike traditional nonparametric methods, SCEPTRE employs a resampling scheme that explicitly accounts for heterogeneity in technical factors among cells. We analyze data from two recent single-cell pooled CRISPR screen experiments–one produced by Gasperini et al.[6] and one produced by Xie et al.[7]–to illustrate the aforementioned analysis challenges and the performance of the proposed methodology. SCEPTRE demonstrates excellent calibration and sensitivity on the data and reveals many novel regulatory relationships supported by a variety of existing sources of functional evidence.

Results

Analysis challenges. The Gasperini et al.[6] and Xie et al.[7] data exemplify several of the analysis challenges in single-cell CRISPR screens. Gasperini et al. transduced via CROP-seq[4,17] at a high multiplicity of infection a library of CRISPR guide RNAs (gRNAs) into a population of 207,324 K562 cells expressing the Cas9-KRAB repressive complex, each cell receiving 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[5,7] 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. Both Gasperini et al. and Xie et al. sequenced polyadenylated gRNA barcodes alongside the whole transcriptome and assigned perturbation identities to cells by thresholding the resulting gRNA UMI counts.

The authors of these studies computed p-values to test for associations between candidate enhancers and genes. Gasperini et al. obtained p-values from a DESeq2[20]-inspired negative binomial regression analysis implemented in Monocle2[21]. Xie et al., by contrast, obtained p-values from Virtual FACS, a nonparametric rank-based method proposed by the authors. Both approaches produced miscalibrated p-values. Gasperini et al. examined the distribution of negative binomial p-values obtained from pairing each protein-coding gene with each of 50 non-targeting control (or negative control) gRNAs. These "null" p-values exhibited inflation, deviating substantially from the expected uniform distribution (Figure 1a, top panel). Xie et al. examined the distribution of Virtual FACS p-values obtained from pairing gRNAs that targeted an enhancer of gene *ARL15*, ARL15-enh, with all protein-coding genes and long noncoding RNAs (lncRNAs) genome-wide. As ARL15-enh is expected to modulate the expression of ARL15 and neighboring lncRNAs only, these p-values can be used to assess model calibration. Xie et al. found these p-values to be mis-

calibrated, with those corresponding to protein-coding genes exhibiting conservative bias (Figure 1a, bottom panel) and those corresponding to lncRNAs exhibiting liberal bias.

Confounding likely is a cause of the miscalibration present in these analyses. Despite the use of targeted amplification protocols for gRNA detection[17], not all transduced gRNAs are detected in single-cell CRISPR screen experiments. Indeed, the total number of gRNAs detected in a cell increases with the sequencing depth, defined as the total number of mRNA UMIs detected per 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). Therefore, sequencing depth and other technical factors induce correlations between the measured gRNA presence and gene expression, even in the absence of a regulatory relationship (Figure 1d). This confounding effect, if not explicitly accounted for, results in test miscalibration. Confounding is especially problematic for traditional nonparametric methods like Virtual FACS, which implicitly (and incorrectly) assume that there are no technical factors that impact both gRNA detection and gene expression measurement.

Parametric regression approaches, like negative binomial regression, are the easiest way to adjust for confounders. However, parametric methods rely heavily on correct model specification, a challenge in single cell analysis given the heterogeneity, complexity, and occasional multimodality of the 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 fits a raw dispersion for each gene, then fits a parametric mean-dispersion relationship across genes, and finally collapses each raw dispersion estimate 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.

To remedy the observed calibration issues, ad hoc adjustments were incorporated into both analysis approaches. On closer inspection, however, these efforts did not satisfactorily improve calibration. Gasperini et al. attempted to calibrate the candidate enhancer p-values against the distribution of all negative control gRNA p-values instead of the 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 the 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 difficult to interpret and could not be subjected to standard multiple hypothesis testing correction procedures, as they are not p-values. Methods that avoid reliance on ad hoc calibration fixes likely would produce more consistently calibrated p-values.

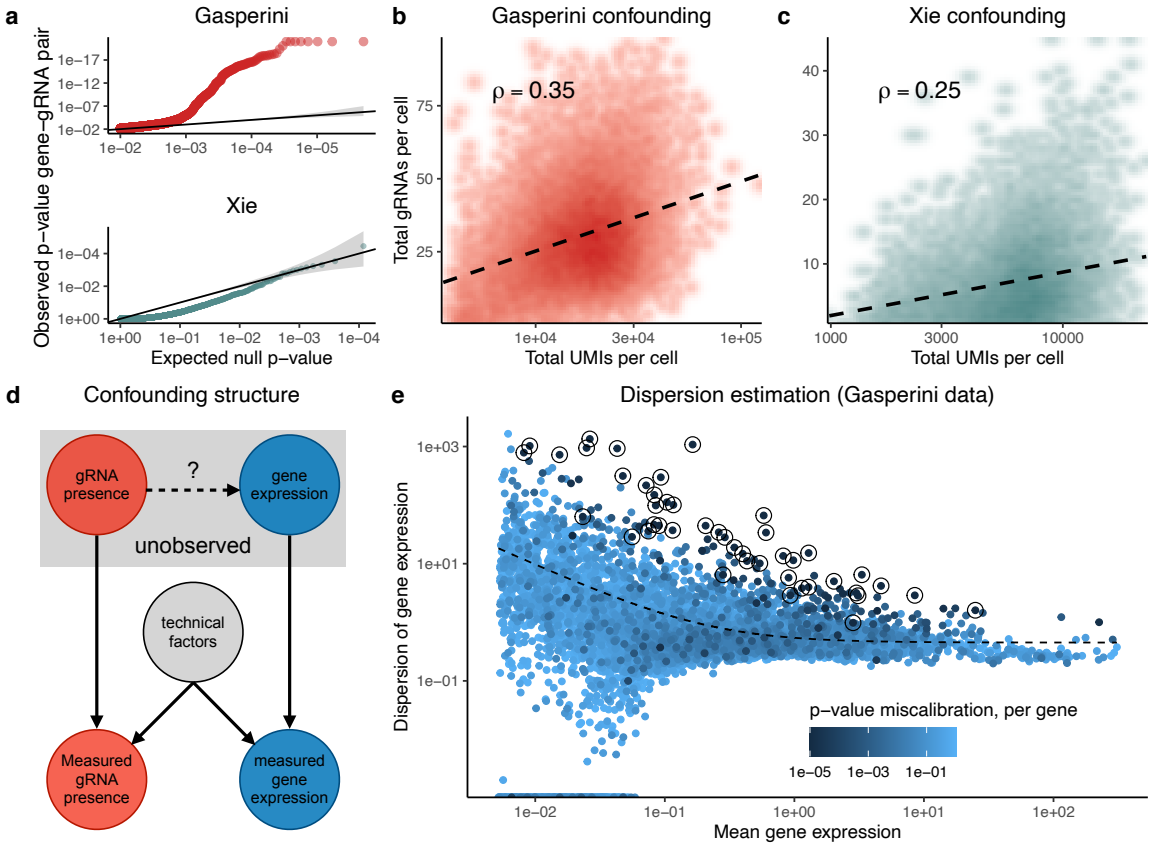


Figure 1: **CRISPR screen analysis challenges can lead to false positives and false negatives**. **a**, QQ-plot of p-values for all pairs of genes and negative control gRNAs (top) and ARL15-enh-targeting gRNAs (bottom; *ARL15* gene omitted). These sets of p-values deviate from the expected uniform distribution, indicating test miscalibration. **b-d**, Sequencing depth (total mRNA UMIs per cell) impacts gRNA detection and observed gene expression levels in both Gasperini et al. (b) and Xie et al. (c) data. Thus, sequencing depth acts as a confounder (d). This analysis challenge distinguishes 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.

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

We reanalyzed the Gasperini et al. negative control gRNA 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 also applied the improved negative binomial approach to the ARL15-enh data from Xie et al. We included sequencing depth, number of detected gRNA UMIs per cell, and sequencing batch as regression covariates in the latter analysis. The improved negative binomial method exhibited better calibration than the Monocle negative binomial method on the negative control data (Figure 3b); however, on both the negative control and ARL15-enh data, the improved negative binomial method showed clear signs of p-value inflation (Figure 3b-c).

There are many possible explanations for the remaining miscalibration, mostly due to misspecification of the negative binomial model. While more effort could be invested into further improvements of a parametric model for gene expression, modeling single-cell data remains a challenging problem outside the scope of the current work. These difficulties highlight the appeal of nonparametric methods for singe cell analysis[22]. However, as discussed in the previous section, traditional nonparametric approaches are not applicable in the presence of technical factors that impact both gRNA detection and gene expression measurement. To resolve this impasse, we propose a resampling-based approach designed to account for such confounders.

SCEPTRE: Analysis of single-cell perturbation screens via conditional resampling. Instead of treating cells symmetrically, we propose to estimate the probability a given gRNA will be observed in a given cell, based on that cell's technical factors. For example, this probability tends to be higher in cells with larger sequencing depths. We propose to generate a null distribution for a given gene and a given gRNA by independently resampling the gRNA assignments for each cell based on their respective perturbation probabil-

ities (Figure 2). This scheme is an instance of the conditional randomization test (CRT), recently proposed by Candès et al[19]. In contrast to standard permutation tests, the CRT explicitly accounts for technical factors that vary from cell to cell.

To quantify the effect of a given gRNA on a given gene, we use the improved negative binomial regression statistic described above. This yields a z-value, which would typically 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. To this end, we first fit a logistic regression model for the observation of the gRNA in a cell based on its technical factors, yielding a fitted probability for each cell. Then, we generate a large number of resampled datasets, where the expression and the technical factors stay the same, while the gRNA assignment is redrawn independently for each cell based on its fitted probability. The negative binomial z-value is then recomputed for each of these datasets, which comprise a null distribution (depicted as a gray histogram in Figure 2). The SCEPTRE p-value is defined as the left tail probability of the original z-value under this null distribution, for sensitivity to perturbations that decrease expression (the right tail could be used for screens targeting repressive elements). The conditional resampling null distribution can differ substantially from that based on the negative binomial model—for the same test statistic—depending on the extent of model misspecification (Figure S2).

To mitigate the extra computational cost of resampling, we implemented several accelerations. We found that the skew-t distribution, used by CRISPhieRmix[23] for a different purpose, provided a good fit to the null histograms. We employed this skew-t approximation to obtain precise p-values based on a limited number of resamples (500 in the current implementation). Furthermore, we implemented computational accelerations that reduced the cost of each resample by a factor of about 100 (see Methods). The original negative binomial regression takes about 3 seconds per gene-gRNA pair, while recomputing the test statistic for 500 resamples takes a total of 16 seconds. Therefore, SCEPTRE takes about 19 seconds per pair, compared to 3 seconds for a single negative binomial regression.

SCEPTRE exhibits 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[12], 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 regardless of expression model misspecification and confounder presence.

Next, to assess the calibration of SCEPTRE on real data, we applied it to test the association between all negative control gRNAs and all genes in the Gasperini et al. data (Figure 3b) and all ARL15-enh-targeting gRNAs and all genes in the 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 negative control gRNA data and Virtual FACS on the ARL15-enh data). We did not apply scMAGeCK-LR to these data given its poor calibration on the simulated data. SCEPTRE showed excellent calibration on the negative control data; Monocle regression and improved negative binomial regression, by contrast, demonstrated signs of severe p-value inflation. Similarly, on the ARL15-enh data, SCEPTRE exhibited nearly perfect calibration, whereas Virtual FACS and the improved negative binomial method showed conservative and liberal bias, respectively.

Last, we assessed the sensitivity of SCEPTRE using positive control gRNA data and confirmatory bulk RNA-seq data. We applied SCEPTRE to the 381 positive control gRNA-gene pairs assayed in Gasperini et al. (Figure 3d), where each gRNA targeted the transcription start site of the corresponding gene. 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 these positive controls were highly significant, and in particular, almost always more significant than the original empirical p-values, indicating greater sensitivity. Finally, we compared the SCEPTRE ARL15-enh results to the results of an arrayed CRISPR screen of ARL15-enh in which populations of perturbed cells were compared to populations of unperturbed cells via bulk RNA-seq[7]. Both SCEPTRE and the bulk RNA-seq differential expression analysis rejected gene *ARL15* (and only gene *ARL15*) at an FDR of 0.1 after a Benjamini-Hochberg correction, increasing our confidence in the calibration and sensitivity of SCEPTRE (Figure 3e).

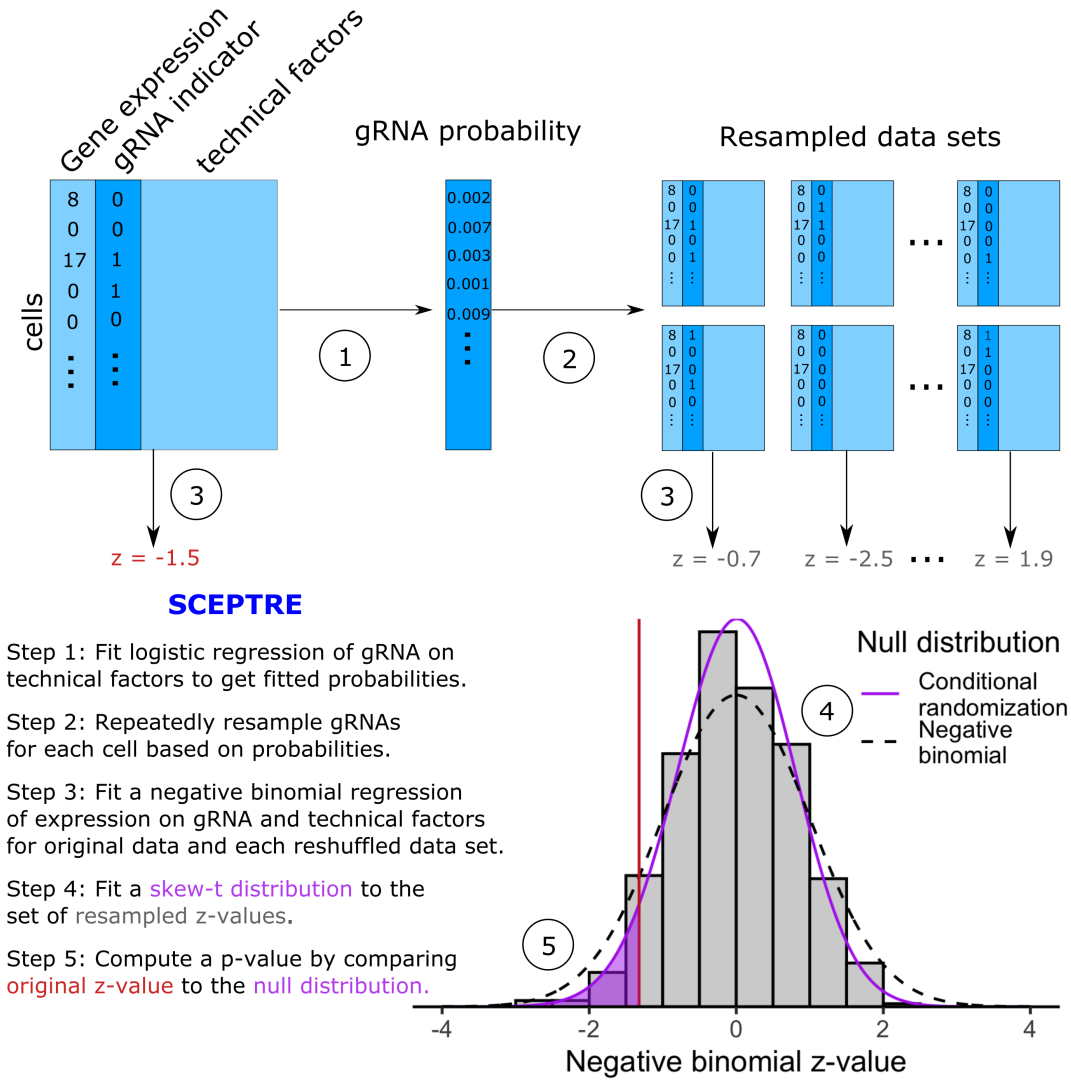


Figure 2: **SCEPTRE: Analysis of single-cell perturbation screens via conditional resampling.** A schematic and outline of the SCEPTRE methodology for one gene and one gRNA is shown, entailing the application of the conditional randomization test[19] to single-cell CRISPR screens. The idea is to assign to every cell the probability of gRNA observation based on its technical factors, and then to build a null distribution for the negative binomial z-value by independently resampling the gRNA indicator for each cell according to these probabilities to form "negative control" data sets. 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 would normally be compared.

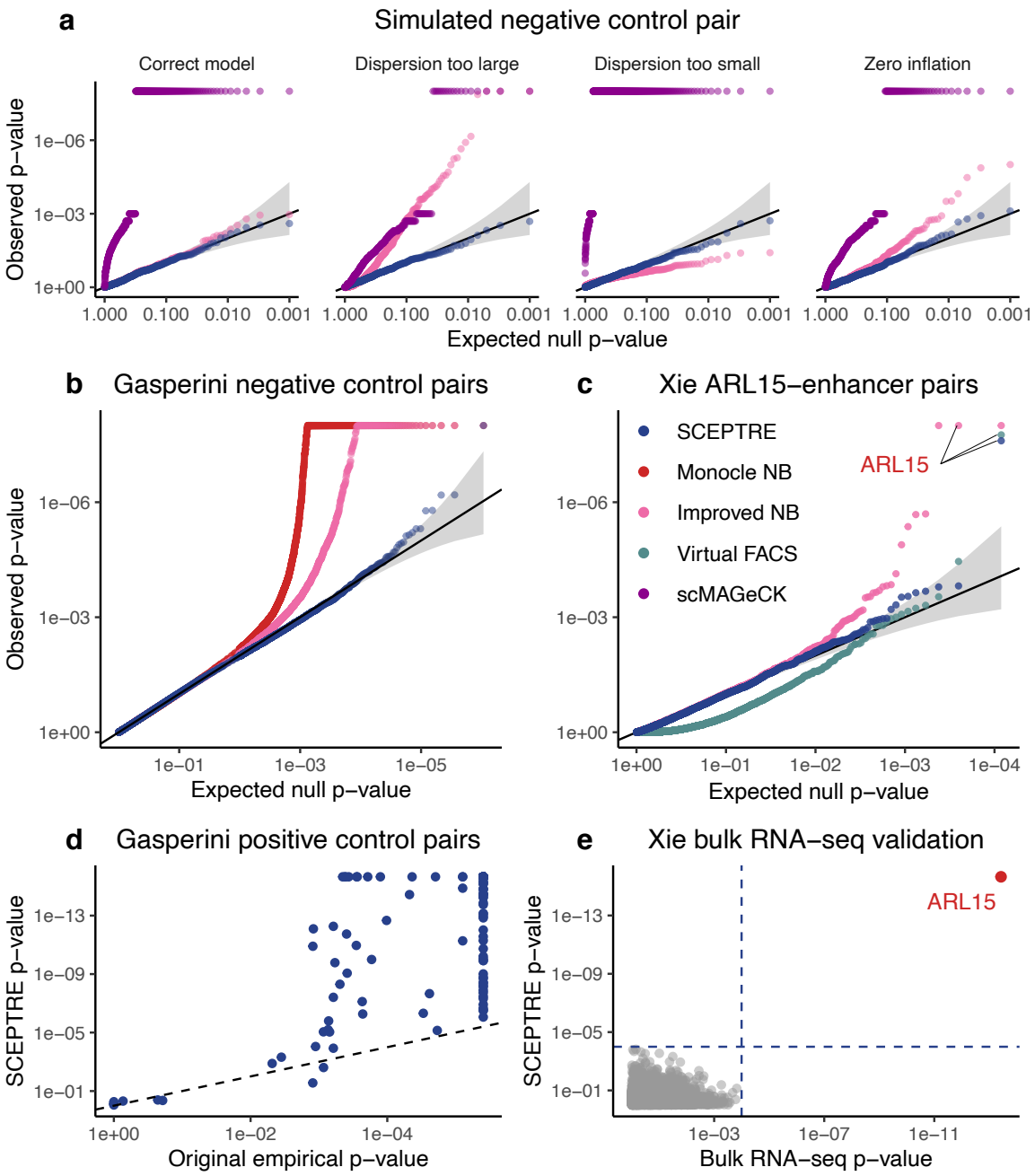
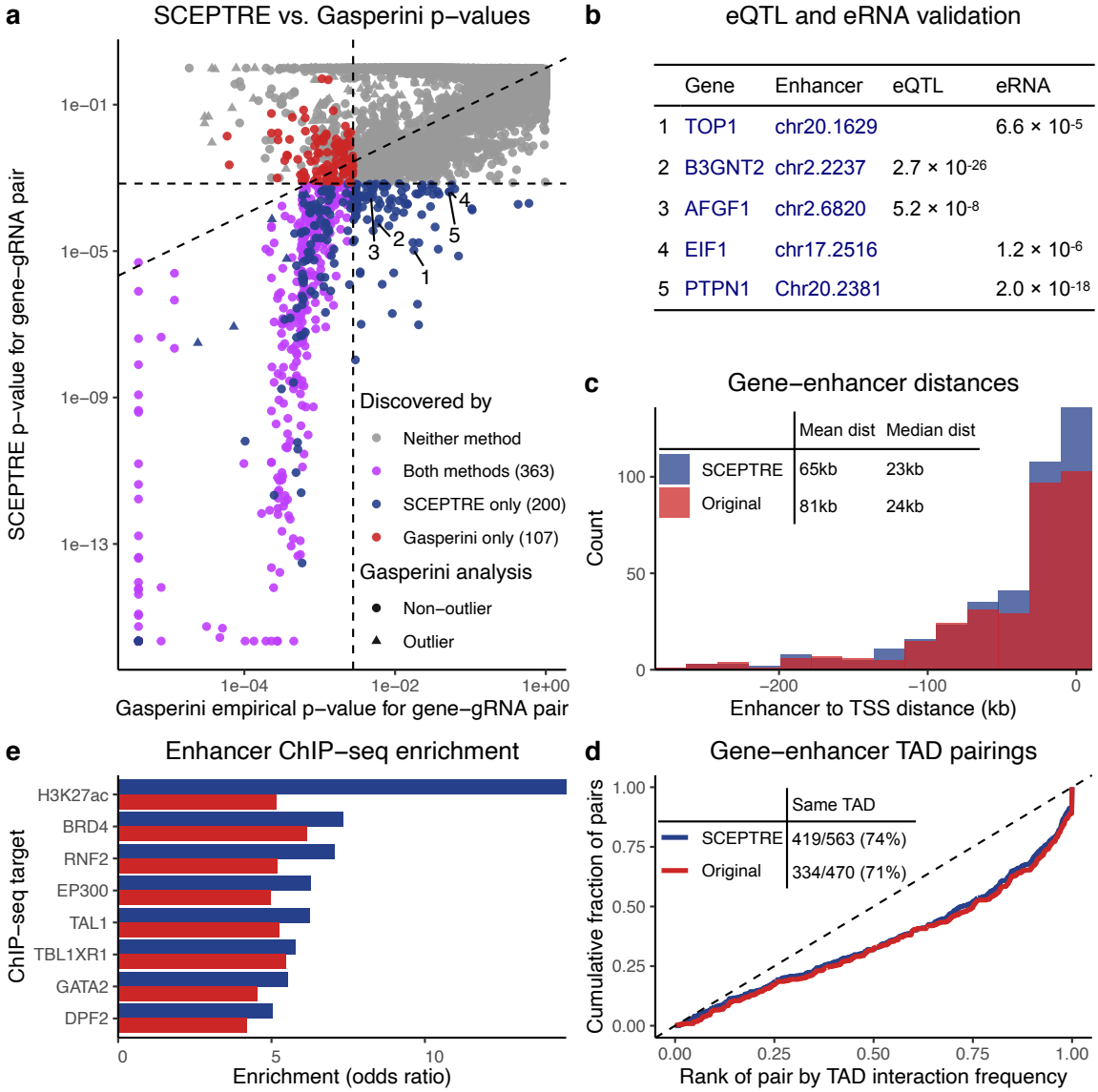


Figure 3: **SCEPTRE demonstrates good calibration and sensitivity under known ground truth. a,** Numerical simulation comparing three methods – SCEPTRE, improved negative binomial regression, and scMAGeCK-LR – on four simulated datasets. Only

SCEPTRE maintained calibration despite model misspecification and confounder presence. **b-c**, Application of SCEPTRE, improved negative binomial regression, Monocle regression, and Virtual FACS to (b) all negative control gRNAs and all genes in the Gasperini et al. data, and (c) all ARL15-enh-targeting gRNAs and all genes in the Xie et al. data. Compared to the other methods, SCEPTRE showed excellent calibration. Gene *ARL15* annotated in (c). **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 Benjamini-Hochberg correction. Dotted blue lines, rejection thresholds.

Analysis of candidate gene-enhancer regulatory relationships. We applied SCEPTRE to the 84,595 gene-enhancer pairs considered in Gasperini et al., encompassing 10,560 genes and 5,779 candidate enhancers. Following Gasperini et al., we based our analysis on the top two gRNAs targeting each enhancer. Some enhancers were also targeted with two additional gRNAs, but we excluded these from the analysis. We applied the Benjamini-Hochberg correction at level 0.1 (the same level used in the original analysis) to the p-values obtained for all of these candidate pairs, obtaining a total of 563 gene-enhancer pairs. By comparison, the original analysis identified 470 high-confidence pairs. Comparing the SCEPTRE p-values against the original empirical p-values (Figure 4a), we observed that the two sets diverged substantially. SCEPTRE found 200 gene-enhancer pairs that the original analysis did not, while 107 were found only by the original analysis. Many of the discoveries found only in the original analysis show signs of the p-value inflation observed in Figure S1.

Among the 200 new gene-enhancer pairs discovered, several were supported by evidence from orthogonal functional assays. In particular, we highlight five of these pairs (Figure 4b) involving genes not paired to any enhancers in the original analysis, which were supported by GTEx[24] eQTL p-values in whole blood or enhancer RNA correlation p-values across tissues from the FANTOM project[25]. These pairs are listed in the GeneHancer database[26], which aggregates eQTL, eRNA, and other sources of evidence of gene-enhancer interactions. The SCEPTRE p-values for these promising pairs were generally 1-2 orders of magnitude more significant than the original empirical p-values. Also among the 200 new gene-enhancer pairs were 6 pairs involving 5 genes that were deemed outliers in the original analysis (blue triangles in Figure 4a), underscoring SCEPTRE's ability to handle genes with arbitrary distributions that may not fit into traditional parametric models.



	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**, 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. **c,** Distribution of distances

from TSS to upstream paired enhancers. Compared to the original analysis, SCEPTRE paired genes with nearer enhancers on average. **d,** 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 fall in the same TAD. **e,** Enrichment of ChIP-seq signal from seven cell-type relevant transcription factors and one histone mark among paired enhancers. SCEPTRE exhibited greater enrichment across all ChIP-seq targets.

We also found that the total set of gene-enhancer pairs discovered was better enriched for regulatory biological signals, including HI-C and ChIP-seq. Gene-enhancer pairs discovered by SCEPTRE tended to be physically closer to each other on average that those discovered in the original analysis (Figure 4c). Furthermore, 74.4% of SCEPTRE's 563 gene-enhancer pairs fell in the same topologically associating domain (TAD), compared to 71.1% of the original 470 pairs. We also repeated Gasperini et al.'s contact frequency enrichment analysis for those pairs falling in the same TAD (Figure 4d). We found similar levels of contact frequency enrichment, despite the fact that SCEPTRE discovered 85 more gene-enhancer pairs in the same TADs. Finally, we repeated the ChIP-seq enrichment analysis of Gasperini et al. for seven cell-type relevant transcription factors and H3K27ac. We quantified enrichment of ChIP-seq signal in paired enhancers as the odds ratio that a candidate enhancer is paired to a gene, comparing those falling in the top quintile of candidate enhancers by ChIP-seq signal and those not overlapping a ChIP-seq peak at all. We found improved enrichment for each of the eight ChIP-seq targets (Figures 4e and S3).

Discussion

In this work we illustrated a variety of statistical challenges arising in the analysis of single-cell CRISPR screens, leaving existing methods (based on parametric expression models, permutations, or negative control data) vulnerable to calibration issues. To address these challenges, we developed SCEPTRE, a novel resampling method based on modeling the probability a given gRNA will be observed in a given cell, based on that cell's technical factors. We found that SCEPTRE exhibits excellent 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 bino-

mial parametric approach, making it quite feasible to apply for large-scale data. We used SCEPTRE to reanalyze the Gasperini et al. data and a subset of the Xie et al. data, yielding many biologically plausible gene-enhancer relationships supported by evidence from eQTL, enhancer RNA co-expression, ChIP-seq, HI-C, and bulk CRISPR screen data.

The main assumption behind SCEPTRE's validity is the accuracy of a model for the resampling of gRNAs conditional on a cell's technical factors. Our calibration results suggest that this is a reasonable assumption on the data that we investigated. However, more work is necessary to assess the quality of gRNA measurement models across a broader range of contexts. Furthermore, we conjecture that misspecifications in the gRNA measurement model can be compensated for by improved gene expression models, a statistical phenomenon referred to as "double robustness"[27,28]. Better expression models can also improve the sensitivity of the conditional randomization test[29].

There are several directions for improvement in the analysis of single-cell CRISPR screens, which are not addressed by SCEPTRE. Many of these have been considered before, especially in the context of bulk CRISPR screens. Such remaining challenges include variable effectiveness of gRNAs[7,23,30], interactions among perturbed elements[5,8,31], and the limited resolution of CRISPR interference[32]. Additionally, SCEPTRE does not address the problem of estimating the effect sizes of CRISPR perturbations on gene expression.

The conditional resampling methodology underlying SCEPTRE is extremely flexible, paving the way for its applicability to future iterations of CRISPR screens and beyond. For example, direct capture Perturb-seq[18] has recently been proposed in order to improve the efficiency of gRNA detection. However, the capture rate can vary substantially across gRNAs. The SCEPTRE framework could be extended to analyze such new CRISPR screen technologies. Finally, confounding technical factors like those pictured in Figure 1d are ubiquitous across genomics; we anticipate the conditional randomization test will complement traditional parametric and nonparametric methods in such contexts as a valuable analysis tool.

References

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

2. 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).

3. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883–1896.e15 (2016).

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

5. 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).

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

7. 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).

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

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

10. Frangieh, C. J. *et al.* Multi-modal pooled Perturb-CITE-Seq screens in patient models define novel mechanisms of cancer immune evasion. *bioRxiv* (2020). URL doi: https://doi.org/10.1101/2020.09.01.267211.

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

12. Yang, L. *et al.* scmageck links genotypes with multiple phenotypes in single-cell crispr screens. *Genome biology* **21**, 1–14 (2020).

13. Lin, X., Chemparathy, A., Russa, M. L., Daley, T. & Qi, L. S. Computational Methods for Analysis of Large-Scale CRISPR Screens. *Annual Review of Biomedical Data Science* **3**, 137–162 (2020).

14. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* **20**, 1–16 (2019).

15. 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).

16. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 142–150 (2020).

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

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

19. 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).

20. 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).

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

22. Zhu, A., Srivastava, A., Ibrahim, J. G., Patro, R. & Love, M. I. Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Research* **47**, e105–e105 (2019).

23. Daley, T. P. *et al.* CRISPhieRmix: A hierarchical mixture model for CRISPR pooled screens. *Genome Biology* **19**, 1–13 (2018).

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

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

26. 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).

27. Robins, J. M. & Rotnitzky, A. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica* **11**, 920–936 (2001).

28. van der Laan, M. J. & Robins, J. M. *Unified methods for censored longitudinal data and causality* (Springer-Verlag, New York, 2003).

29. Katsevich, E. & Ramdas, A. A theoretical treatment of conditional independence testing under Model-X. *arXiv* (2020). URL http://arxiv.org/abs/2005.05506. 2005.05506.

30. Li, W. *et al.* Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biology* **16** (2015).

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

32. Hsu, J. *et al.* CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. *Nature Methods* **15**, 990–992 (2018).

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

34. 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).

35. 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.

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

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

38. 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).

Acknowledgements We are indebted to Molly Gasperini, Jacob Tome, and Andrew Hill for clarifying several aspects of their data analysis[6] 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) grants R01MH123184 and R37MH057881 as well as SFARI Grant 575547. Part of the data analysis used the Extreme Science and Engineering Discovery Environment (XSEDE)[33], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system[34], 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).

Methods

Conditional randomization test. Consider a particular gene/gRNA pair. For each cell $i = 1, \dots, n$, let $X_i \in \{0, 1\}$ indicate whether the gRNA was present in the cell, let $Y_i \in \{0, 1, 2, \dots\}$ 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[19] 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[15], 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[35] *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[12] 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 Benjamini-Hochberg 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[36]. To avoid this extra risk of false positives, we chose to use the "high-confidence" set throughout.

ChIP-seq, HI-C enrichment analyses. ChIP-seq and HI-C enrichment analyses analyses (see Figures 4e-f and S3) 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://bit.ly/SCEPTRE. All analysis was performed on publicly available data. The Gasperini et al. CRISPR screen data[6] 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[37] and are available at www.encodeproject.org/. The HI-C enrichment analysis is based on the data from Rao et al.[38], 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[26] available as part of GeneCards (www.genecards.org/).

Code availability

The `sceptre` R package and data analysis scripts for this paper are available on Github at https://github.com/Timothy-Barry/SCEPTRE.

Supplementary Figures

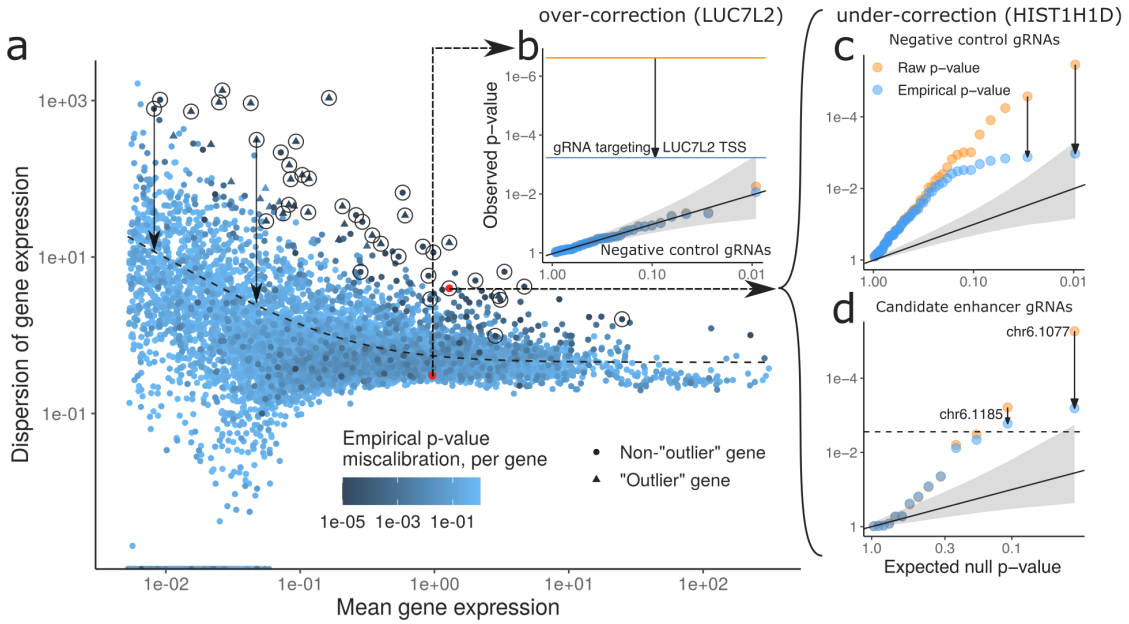


Figure S1: **Gasperini et al.'s empirical correction is insufficient to correct for miscalibration**. **a,** Dispersion estimation procedure employed leads to miscalibration for high-dispersion genes, which the empirical correction does not adequately correct for, as measured by KS test applied to empirical 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.

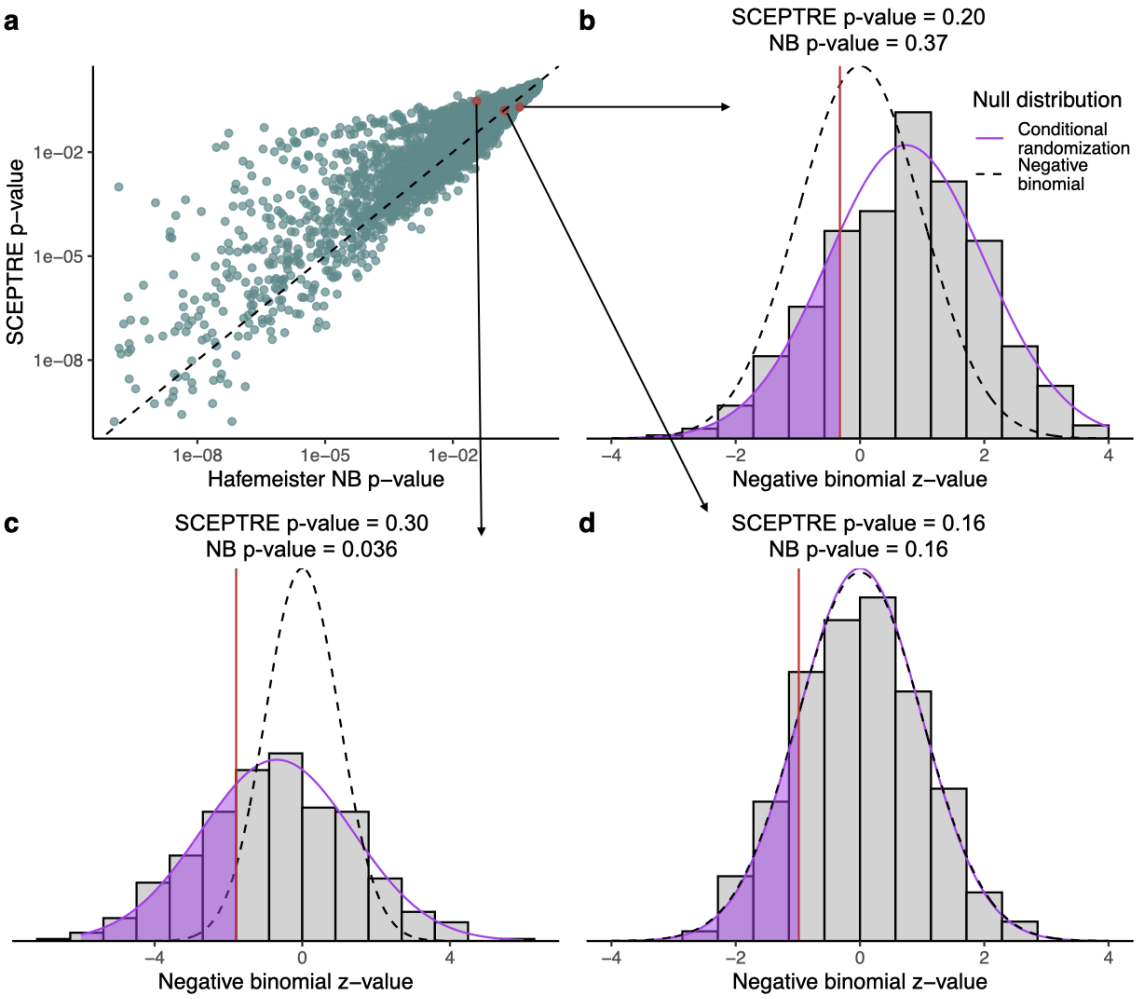


Figure S2: **Comparison of negative binomial and conditional resampling p-values based on the same test statistic. a,** The standard parametric negative binomial p-value versus that obtained from the same test statistic by conditional resampling, for each gene / candidate enhancer pair (both truncated 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).

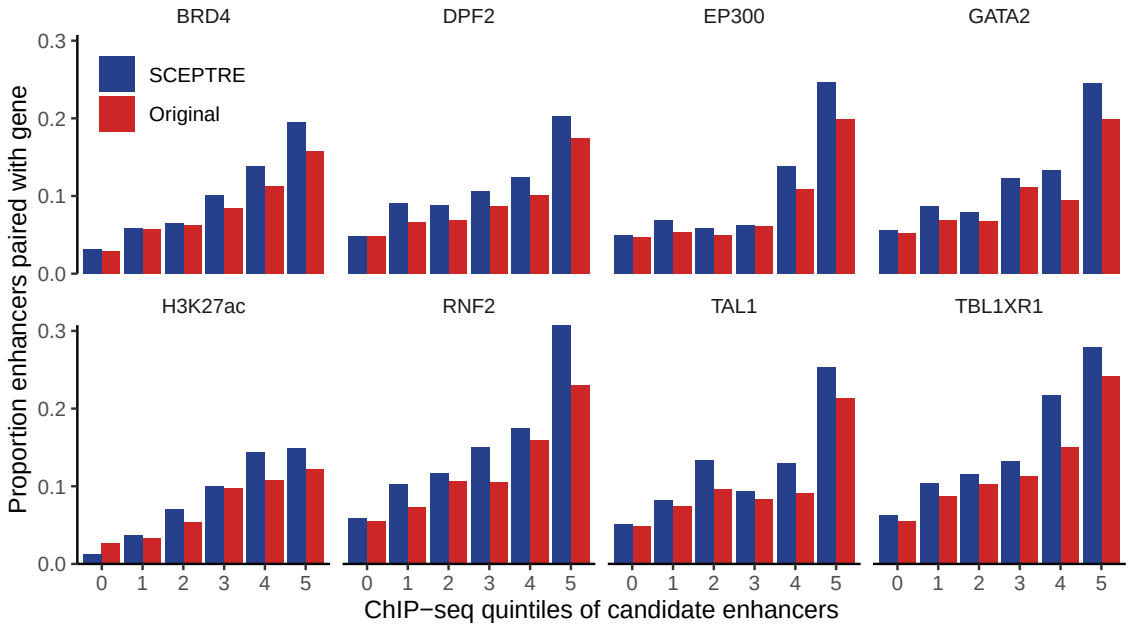


Figure S3: **Details on ChIP-seq enrichment analysis.** Fraction of candidate enhancers paired with a gene, broken down by quintile of ChIP-seq signal (0 means the candidate enhancer did not overlap a ChIP-seq peak), based on which the odds ratios in Figure 4e were computed. Both methods generally pair candidate enhancers in higher ChIP-seq quintiles more frequently, but this enrichment is more pronounced in SCEPTRE across all eight ChIP-seq targets.