Conditional resampling improves calibration 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 a recent large-scale single cell CRISPR screen 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. Applying SCEPTRE to the CRISPR screen data yields excellent calibration on negative control perturbations and 217 regulatory relationships not found in the original study, 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). This resampling methodology is based on

the conditional randomization test[19], a recently proposed statistical methodology which, 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 both synthetic data and that of Gasperini et al.[6] to illustrate the aforementioned analysis challenges and the performance of the proposed methodology. SCEPTRE demonstrates excellent calibration and reveals many novel regulatory relationships supported by a variety of existing sources of functional evidence.

Results

Analysis challenges. The Gasperini et al.[6] data exemplify several of the analysis challenges in single cell CRISPR screens. A library of CRISPR guide RNAs (gRNAs) targeting 5,779 candidate enhancers, 50 negative controls, and 381 positive controls was transduced via CROP-seq[4,17] at high multiplicity of infection into a population of 207,324 K562 cells expressing the Cas9-KRAB repressive complex, each cell receiving an average of 28 perturbations. Polyadenylated gRNA barcodes were sequenced alongside the whole transcriptome, and perturbation identities assigned to cells by thresholding the resulting gRNA UMI counts. Initially, p-values for each gRNA-gene pair were obtained from a DESeq2[20]-inspired negative binomial regression analysis implemented in Monocle2[21]. By examining the distribution of negative binomial p-values pairing each gene with each of 50 non-targeting control (NTC) gRNAs, Gasperini et al. found that these p-values are inflated (Figure 1a). To remedy this issue, Gasperini et al. calibrated the candidate enhancer p-values against the distribution of all gene-NTC p-values instead of the uniform distribution. The resulting "empirical" p-values were used for determining significance.

First, we illustrate the confounding effect discussed in the introduction. Despite the use of a targeted amplification protocol for gRNA detection[17], not all gRNAs are detected. 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 ($\rho = 0.35, p < 10^{-15}$; Figure 1b). We observed a similar phenomenon in the Mosaic-seq data[5] ($\rho = 0.23, p < 10^{-15}$; Figure S.1). 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 1c). Parametric regression approaches like those employed by Gasperini et al. are the easiest way to adjust for confounders, but their heavy reliance on correct model specification creates a different set of calibration issues.

We hypothesized that imperfect dispersion estimation may be responsible for the inflation observed by Gasperini et al. Monocle2 fits a raw dispersion for each gene, then

fits a mean-dispersion relationship across genes, and finally collapses each raw dispersion estimate onto this fitted line (Figure 1d). While the raw dispersion estimates are noisy and some shrinkage is helpful[15], the collapsing of the dispersion estimates to the fitted curve is likely to underestimate the dispersions for points significantly above the curve. On the other hand, points near the curve may not have this issue. Indeed, plotting the 50 raw NTC p-values for genes near the curve, we found good calibration (e.g., LUC7L2; Figure 1e), while genes significantly above the curve showed marked signs of raw p-value inflation (e.g., HIST1H1D; Figure 1f).

This heterogeneity in inflation among gene-gRNA pairs reveals a weakness of the "one-size-fits-all" NTC-based calibration approach employed by Gasperini et al.: building a null distribution using all gene-NTC pairs leads to overcorrection for some gene-enhancer pairs (false negatives) and undercorrection for others (false positives). For example, the empirical correction is unnecessary for genes like LUC7L2. However, we pay a price in sensitivity for this empirical correction; it decreases the significance of the p-value for a positive control gRNA by three orders of magnitude (Figure 1e). On the other hand, the empirical correction is insufficient for high-dispersion genes. We computed the deviation from uniformity of the empirical NTC p-values for each gene using the Kolmogorov-Smirnov (KS) test, represented by the color of each point in Figure 1b. Circled genes have significantly miscalibrated empirical p-values based on a Bonferroni correction at level $\alpha = 0.05$; 21 of these were flagged by Gasperini et al. as outliers but 21 were not. Among the latter group is HIST1H1D, whose empirical p-values are still inflated (Figure 1f), leading to two potential false discoveries, one of which is deemed "high confidence" (Figure 1g, Figure **??**).

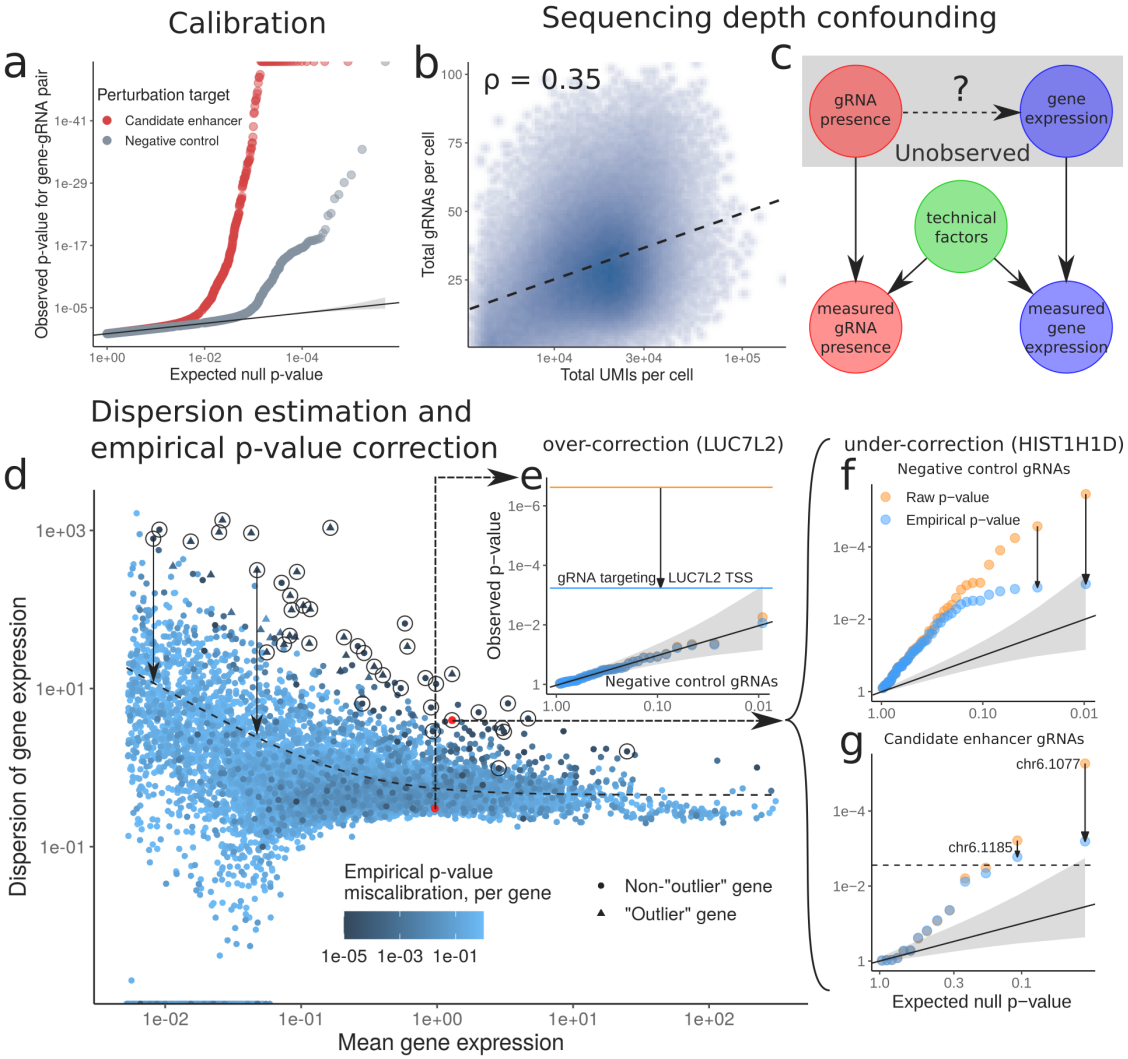


Figure 1: **CRISPR screen analysis challenges can lead to false positives and false negatives.** **a**, QQ-plot of Gasperini et al. p-values for all gene-gRNA pairs involving either candidate enhancer or negative control gRNAs. Inflation in negative controls makes it harder to interpret associations between genes and candidate enhancers, motivating the empirical p-value correction. **b-c**, Sequencing depth (total mRNA UMIs per cell) impacts gRNA detection (b) and observed expression levels, and therefore acts as a confounder (c). This analysis challenge distinguishes CRISPR screens from traditional differential expression applications.

Improvements to the negative binomial approach. We attempted to alleviate the miscalibration within the negative binomial regression framework (Figure **??**). First, we replaced the collapsed dispersion estimates by the raw estimates (recall Figure 1d). This simple modification is based on the intuition that not as much dispersion shrinkage is necessary as in bulk RNA-seq, since in scRNA-seq we have many more samples (one for each cell) to estimate this parameter. This dispersion estimate modification yielded greatly improved calibration over all gene-NTC pairs (blue curve in Figure **??**a), but separating by NTC revealed substantial remaining gRNA-dependent miscalibration (Figure **??**b). This dependency on gRNA further underscores the problem with "one-size-fits-all" NTC-based calibration approaches, also implemented by Xie et al.[7]

We hypothesized that this remaining miscalibration stems from inadequate sequencing depth correction. Indeed, we found we could predict the direction and strength of the miscalibration for each gRNA from an adjusted correlation between the sequencing depth and the gRNA presence (Methods). This yielded z-values as shown in the legend of Figure **??**b; large positive (negative) z-values signal spurious positive (negative) correlations between gRNA presence and sequencing depth, leading to conservative (liberal) bias. Gasperini et al. corrected for sequencing depth using DESeq2-style size factors, whereas recent work on scRNA-seq analysis[15] advocates for including sequencing depth as a covariate in the negative binomial regression. We therefore included as covariates both the total number of UMIs observed and the number of genes with at least one UMI in a given cell, in addition to the technical factors Gasperini et al. corrected for (see Methods).

To assess the impact of these two improvements, we applied the negative binomial test with and without these improvements (four total possibilities) to all 50 NTC gRNAs paired with all genes (Figure **??**a,c). The changes to confounder correction and dispersion estimation both markedly improve calibration, especially when used in conjunction. However, the resulting improved negative binomial method still shows clear signs of miscalibration, producing mostly conservative p-values. There are many possible reasons for the remaining miscalibration, most likely due to remaining misspecification of the negative binomial model. While more effort could certainly be invested in further improvements of this parametric model for gene expression, we acknowledge that modeling scRNA-seq expression is a challenging problem that remains open. This highlights the appeal of nonparametric approaches; however, as discussed before, traditional nonparametric approaches are not applicable in the presence of technical factors impacting 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 probabilities (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 S.2).

To mitigate the extra computational cost of resampling, we implemented several accelerations. We found that the skew-t distribution, used by CRISPhieRmix[22] 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 the original method.

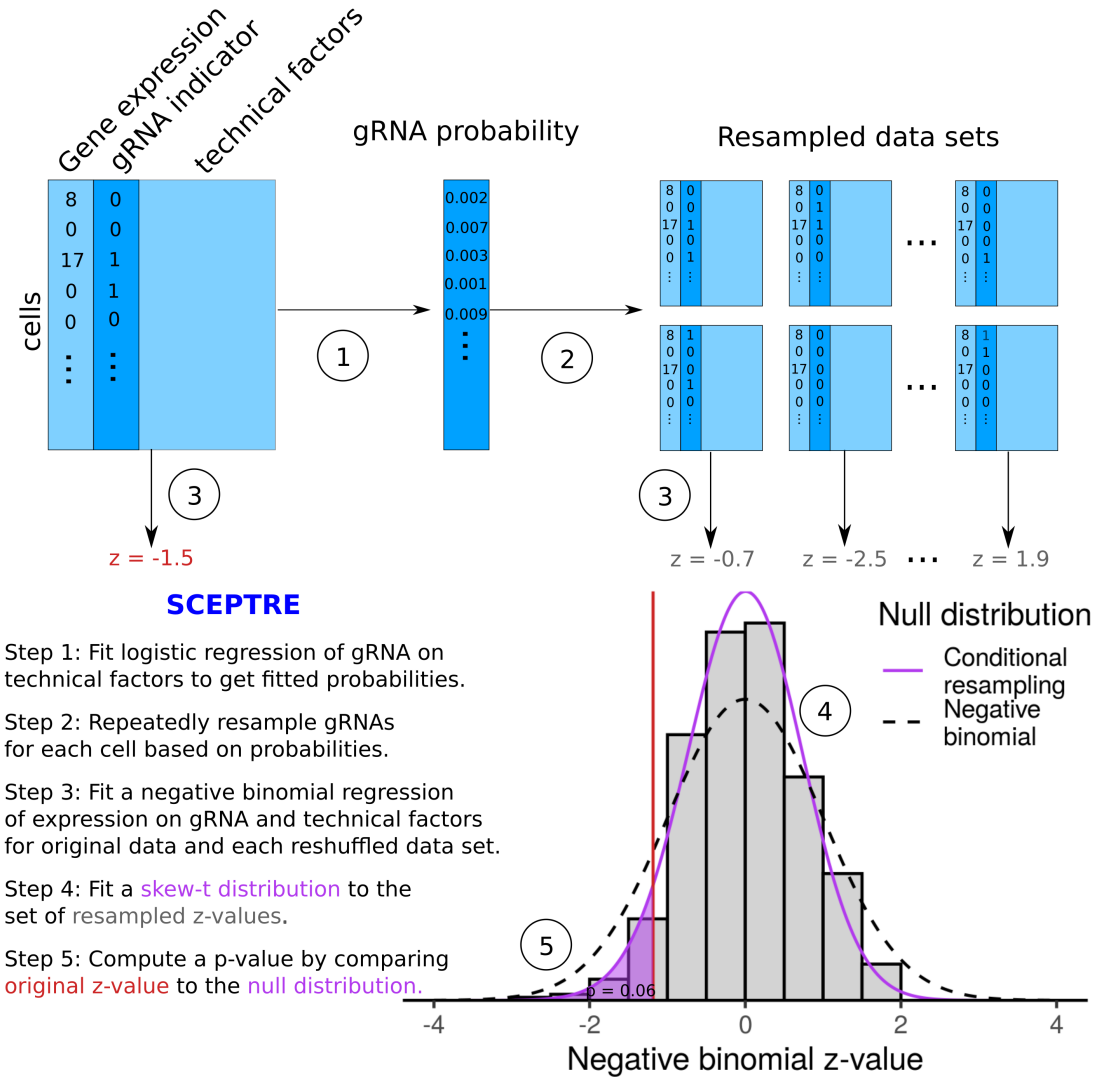


Figure 2: **SCEPTRE: Analysis of single cell perturbation screens via conditional resampling.** A schematic and outline of the SCEPTRE methodology for one gene and one gRNA is shown, entailing the application of the conditional randomization test[19] to single cell CRISPR screens. The idea is to assign 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.

SCEPTRE is well calibrated despite sequencing depth confounding and expression model misspecification. First, we demonstrated SCEPTRE's calibration in a small proof-of-concept simulation with 1,000 cells, one gene, one NTC gRNA, and one confounder (Figure 3a). We considered the one-sided z test statistic based on a negative binomial regression with dispersion 1. We then generated the data from this model as well as from three others: one with dispersion 0.2, one with dispersion 5, and one with dispersion 1 but with 25% zero-inflation. We compared three ways of building a null distribution for the test statistic based on the possibly misspecified negative binomial model: the standard parametric approach, the permutation approach (based on permuting gRNA assignments), and conditional resampling. The parametric approach works as expected when the negative binomial model is correctly specified, but breaks down in all three cases of model misspecification. The permutation approach is systematically conservative because of inadequate confounder correction. Finally, SCEPTRE is well-calibrated regardless of expression model misspecification.

Next, we applied SCEPTRE to test the association between all NTC gRNAs and all genes in the Gasperini et al. data (Figure 3b-c). For comparison, we also applied the original negative binomial method, the improved negative binomial method, and a permutation-based calibration of the latter. The simplest calibration assessment is to compare the resulting 538,560 p-values to the uniform distribution (Figure 3b). SCEPTRE shows excellent calibration, substantially improving on the parametric approach based on the same test statistic. Breaking the pairs down by NTC (Figure 3c), SCEPTRE again shows nearly perfect calibration, as evidenced by its colored confidence region coinciding with the gray shaded region representing the 95% confidence band for the uniform distribution.

To assess the sensitivity of SCEPTRE, we applied it 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. Comparing to the original empirical p-values, SCEPTRE p-values for these positive controls are generally more significant, suggesting greater sensitivity. We note that the empirical correction employed by Gasperini et al. limits the accuracy of p-values to about 10^{-6}, explaining at least part of the difference.

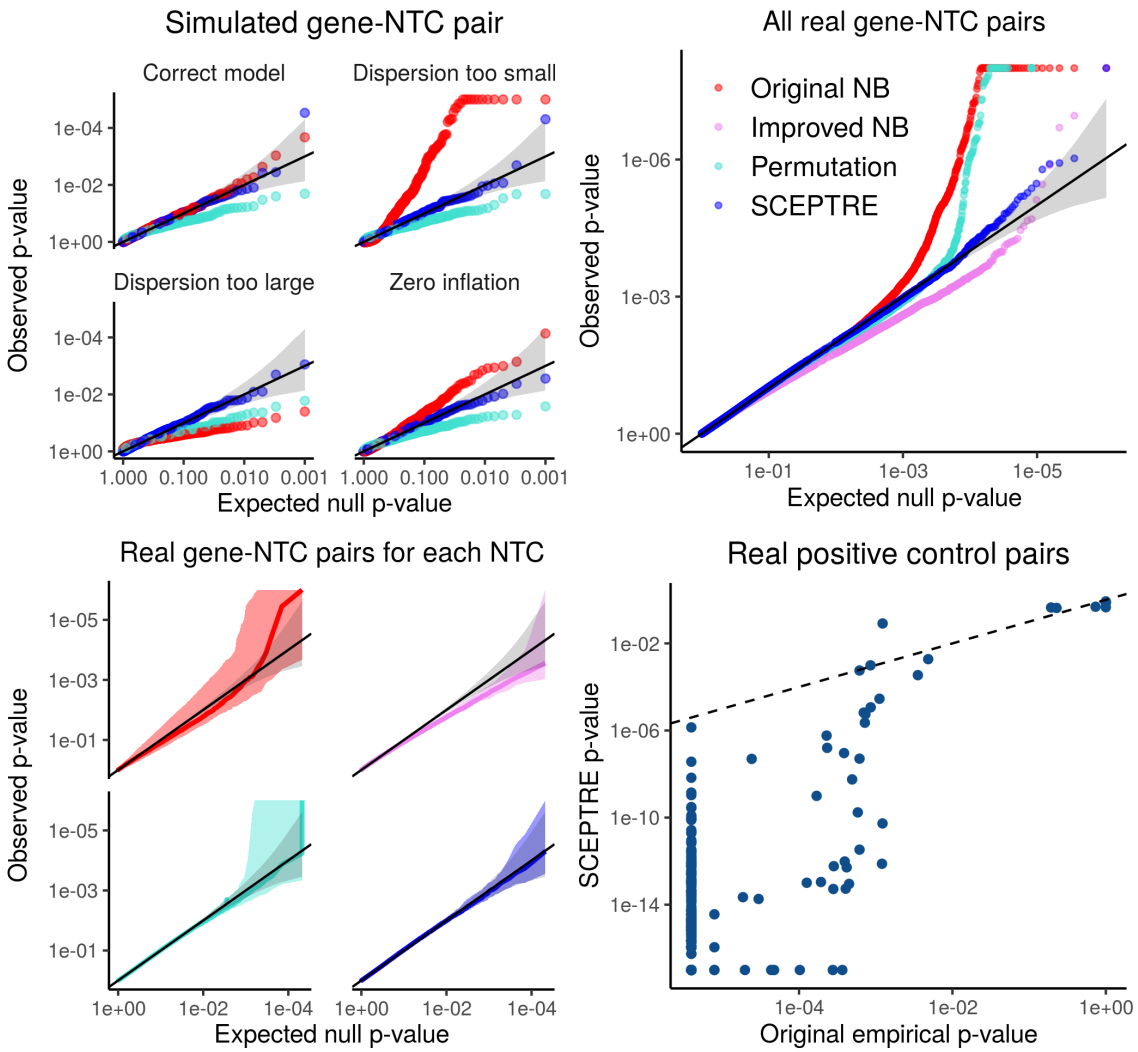
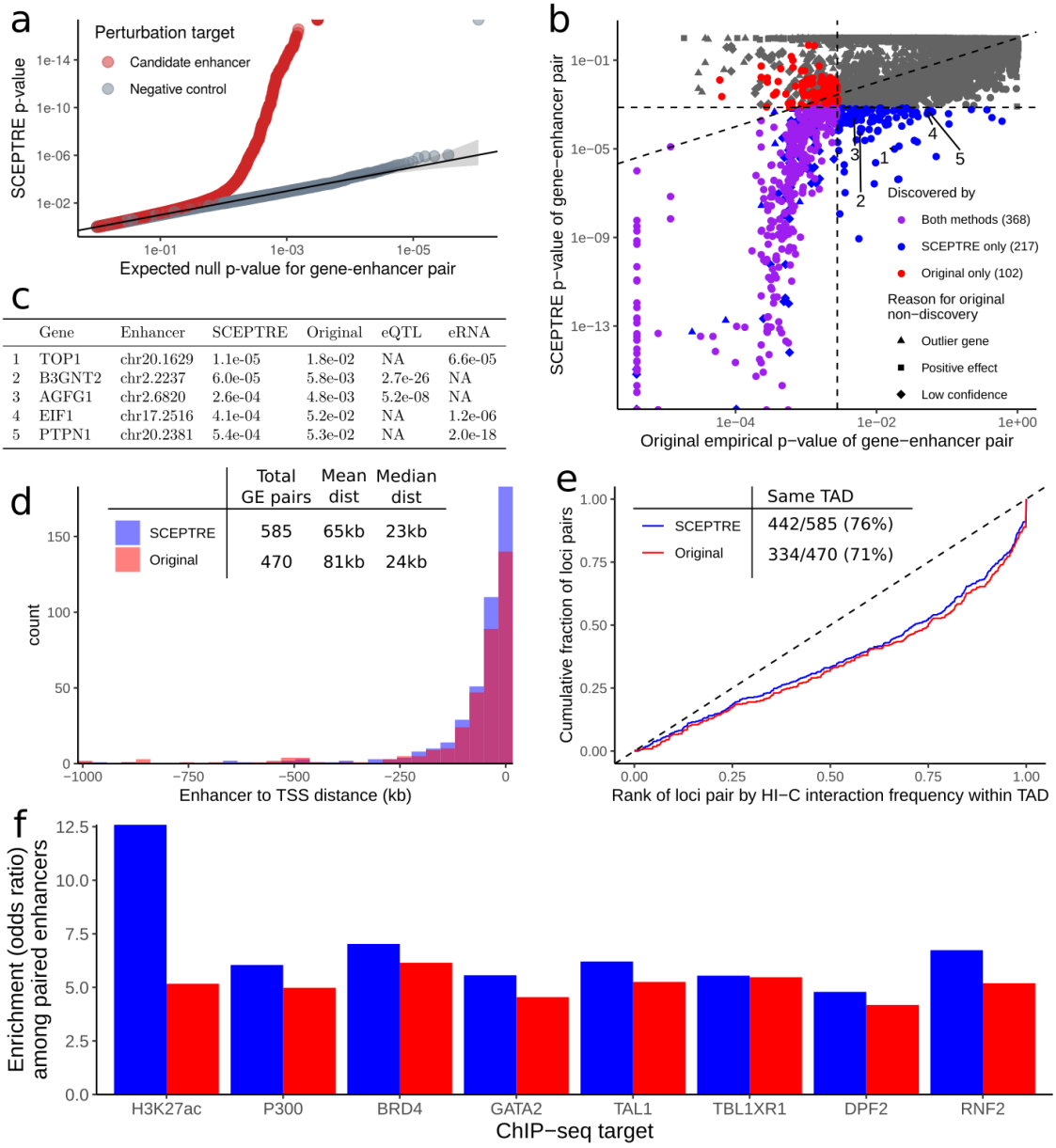


Figure 3: **SCEPTRE improves on the sensitivity and specificity of the parametric approach by correcting for expression model misspecification. a,** Numerical simulation comparing three ways of calibrating a negative binomial test statistic (parametric, permutation, conditional resampling), when the test statistic is based on the correct model or contains expression model misspecifications. Only the conditional resampling approach maintains calibration despite model misspecification. **b,c,** Application of SCEPTRE to Gasperini et al negative control gRNAs, comparing all gene-NTC pairs to the uniform distribution (b) or breaking down by NTC (c); details as in Figure **??**a,c. The matching of SCEPTRE's colored region with the gray one in panel c, together with the overall near-uniformity in panel b, demonstrates its excellent calibration on real negative control data. **d,** SCEPTRE p-values for TSS-targeting controls generally more significant than the original empirical p-values; dashed line is the 45° line. Note: p-values truncated for visualization, and empirical p-values in panel d at least 1/538,560 $\approx 2\times10^{-6}$ by construction.

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 (Figure 4a). 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 585 gene-enhancer pairs. By comparison, Gasperini et al. found 470 high-confidence pairs. Comparing the SCEPTRE p-values against the original empirical p-values (Figure 4b), we see that the two often diverge substantially. Our analysis found 217 gene-enhancer pairs that the original analysis did not, while 102 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 1 (Figure **??**).

Among the 217 new gene-enhancer pairs discovered, several are supported by evidence from orthogonal functional assays. In particular, we highlight five of these pairs (Figure 4c) involving genes not paired to any enhancers in the original analysis, which are supported by GTEx[23] eQTL p-values in whole blood or enhancer RNA correlation p-values across tissues from the FANTOM project[24]. These pairs are listed in the GeneHancer database[25], which aggregates eQTL, eRNA, and other sources of evidence of gene-enhancer interactions. The SCEPTRE p-values for these promising pairs are generally 1-2 orders of magnitude more significant than the original empirical p-values. Also among the 217 new gene-enhancer pairs are 9 pairs involving 7 genes that were deemed outliers in the original analysis (blue triangles in Figure 4b), underscoring SCEPTRE's ability to handle genes with arbitrary distributions that may not fit into traditional parametric models.

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 closer to each other on average that those discovered in the original analysis (Figure 4d). Furthermore, 76% of SCEPTRE's 585 gene-enhancer pairs fell in the same topologically associating domain (TAD), compared to 71% 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 4e). We found similar levels of contact frequency enrichment, despite the fact that we had 108 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 find improved enrichment for each of the eight ChIP-seq targets (Figures 4f and S.3).



	Gene	Enhancer	SCEPTRE	Original	eQTL	eRNA
1	TOP1	chr20.1629	1.1e-05	1.8e-02	NA	6.6e-05
2	B3GNT2	chr2.2237	6.0e-05	5.8e-03	2.7e-26	NA
3	AGFG1	chr2.6820	2.6e-04	4.8e-03	5.2e-08	NA
4	EIF1	chr17.2516	4.1e-04	5.2e-02	NA	1.2e-06
5	PTPN1	chr20.2381	5.4e-04	5.3e-02	NA	2.0e-18

Figure 4: **Application to Gasperini et al. data yields biologically plausible gene-enhancer links. a,** Well-calibrated NTC p-values give confidence in SCEPTRE p-values for candidate enhancers (compare to Figure 1a). **b**, Comparing the original empirical p-values to those obtained from SCEPTRE. The two analysis methods differ substantially, with 217 gene-enhancer links discovered only by SCEPTRE and 102 discovered only by the original. **c**, 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. **d,** Distribution of distances from TSS to upstream paired enhancers. Compared to the original analysis, SCEPTRE pairs genes with nearer enhancers on average. **e,** 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 shows similar enrichment despite finding 32% more within-TAD pairs. Inset table shows gene-enhancer pairs falling in the same TAD. SCEPTRE finds 115 more total pairs, and a higher percentage of pairs fall in the same TAD. **f,** Enrichment of ChIP-seq signal from seven cell-type relevant transcription factors and one histone mark among paired enhancers. SCEPTRE exhibits greater enrichment across all ChIP-seq targets.

Discussion

Here we illustrate 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 develop 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. SCEPTRE exhibits excellent calibration despite the presence of confounding technical factors and despite misspecification of single cell gene expression models. We implement computational accelerations to bring the cost of our resampling-based methodology down to well within an order of magnitude of the traditional approach, making it quite feasible to apply for large-scale data. We use SCEPTRE to reanalyze the Gasperini et al. data, yielding many new biologically plausible gene-enhancer relationships supported by evidence from eQTL, enhancer RNA co-expression, ChIP-seq, and HI-C data.

The main assumption behind SCEPTRE's validity is the accuracy of the model for gRNA observation. Our calibration results suggest that this is a reasonable assumption for the simulated and negative control data considered here. 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 by improved gene expression models, a statistical phenomenon referred to as "double robustness"[26,27]. Better expression models can also improve the sensitivity of the conditional randomization test[28].

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 non-single-cell CRISPR screens. Such remaining challenges include variable effectiveness of gRNAs[7,22,29], interactions among perturbed elements[5,8,30], and the limited resolution of CRISPR interference[31]. Additionally, single cell CRISPR screens would benefit from the adoption of techniques from the increasingly rich literature on scRNA-seq analysis.

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. We look forward to adapting the SCEPTRE framework to the analysis of such new CRISPR screen technologies. Finally, confounding technical factors like those pictured in Figure 1c 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.* Linking genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biology* **21** (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. Daley, T. P. *et al.* CRISPhieRmix: A hierarchical mixture model for CRISPR pooled screens. *Genome Biology* **19**, 1–13 (2018).

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

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

25. 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).

26. 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).

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

28. 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.

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

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

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

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

33. 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).

34. 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.

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

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

37. 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 and Jacob Tome for clarifying several aspects of their data analysis[6], and to the Shendure lab and Timothy Barry for providing extensive feedback on an earlier draft of this paper. We also 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)[32], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system[33], 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

Reproducing the negative binomial approach of Gasperini et al. 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.

Gasperini et al. used a negative binomial regression of Y_i on X_i and Z_i:

$$Y_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \alpha); \quad \log(\mu_i) = \beta_0 + X_i\beta + Z_i^T\gamma, \tag{1}$$

with α being the dispersion and Z_i consisting of three cell-level technical factors: the total number of gRNAs in the cell, the percentage of observed transcripts that are mitochondrial, and the sequencing batch. To correct for sequencing depth, Gasperini et al. replaced Y_i in the above regression by rounding Y_i/s_i to the nearest integer, where s_i are DESeq2-style size factors. Raw dispersions α_{raw} were estimated for each gene based on its sample mean and variance. Gasperini et al fit a mean-dispersion relationship to these raw dispersions, and finally obtained a shrunk estimate of α by projecting α_{raw} onto this curve (Figure 1b). The significance of X_i in the above regression was determined with a likelihood ratio test for $H_0 : \beta = 0$. Since enhancer perturbation decreases gene expression, gene-enhancer pairs with $\widehat{\beta} > 0$ were removed from the analysis.

Exploration of sequencing depth confounding. After improving the dispersion estimates, we sought to understand the direction and strength of the remaining miscalibration

for each gRNA, reasoning it may relate to sequencing depth confounding. To this end, we tested the residual association between each gRNA and the sequencing depth, after accounting for the total number of gRNAs per cell. In particular, for each gRNA, we ran a logistic regression of the gRNA indicator against the total number of gRNAs and the total number of UMIs detected per cell. We then extracted the z-score of the coefficient of the total number of UMIs. We found these agreed well with the direction and strength of the confounding effect, as shown in Figure **??**b.

Conditional randomization test and accelerations. 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

$$\widetilde{X}_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i). \tag{2}$$

Then, the CRT p-value is given by

$$p_{\text{CRT}} = \mathbb{P}[T(\widetilde{X}, Y, Z) \geq T(X, Y, Z) \mid X, Y, Z]. \tag{3}$$

This translates to repeatedly sampling $\widetilde{X}$ from the distribution (2), 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 (3), *regardless of the expression distribution* $Y|X, Z$ *and regardless of the test statistic* T. We choose a test statistic T based on the improved negative binomial regression discussed in the main text, with two computational accelerations.

First, we employed the recently proposed[34] *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 (1) 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 (1), 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 only the thousand or so cells containing the gRNA instead 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 (3) 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 $\widetilde{X}^b$ (2).
   - 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. 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.

Numerical simulation to assess calibration. We simulated one gene Y_i, one gRNA X_i, and one confounder Z_i in $n = 1000$ cells. We drew $Z_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$. Then, we drew X_i based on the logistic model (7), with $\tau = 4$ and $\tau_0 = \log \frac{0.05}{0.95}$, the latter to make the probability of gRNA occurrence about 0.05 on average across cells. Finally, we drew the gene expression Y_i from the following zero-inflated negative binomial model:

$$Y_i \overset{\text{ind}}{\sim} \lambda \delta_0 + (1 - \lambda)\text{NegBin}(\mu_i, \alpha), \quad \log(\mu_i) = \beta_0 + 4 Z_i.$$

Note that for any β_0, λ, α, the gRNA does not impact the gene expression in this setup. We chose $\beta_0 = \log(5)$ to make the average gene expression about 5. The four settings shown in Figure 3a correspond to

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

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 three methods to these four problem settings, each with 500 repetitions. The negative binomial method was based on the z statistic from the standard negative binomial regression (1) with $\alpha = 1$. The permutation method was implemented the same as SPECTRE, except skipping step 3 and defining $\widetilde{X}^b$ to be a random permutation of X. Both the permutation method and SPECTRE used $B = 250$ resamples.

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[35]. To avoid this extra risk of false positives, we chose to use the "high-confidence" set throughout.

ChIP-seq, HI-C enrichment analyses. These analyses (see Figures 4e-f and S.3) 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 CRISPR screen data[6] is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120861. The ChIP-seq data are drawn from the ENCODE project[36] and are available at https://www.encodeproject.org/. The HI-C enrichment analysis is based on the data from Rao et al[37], available at https://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[25], available as part of GeneCards (https://www.genecards.org/).

Code availability

Code to reproduce all data analysis is available on Github at https://github.com/ekatsevi/SCEPTRE.

A Supplementary Figures

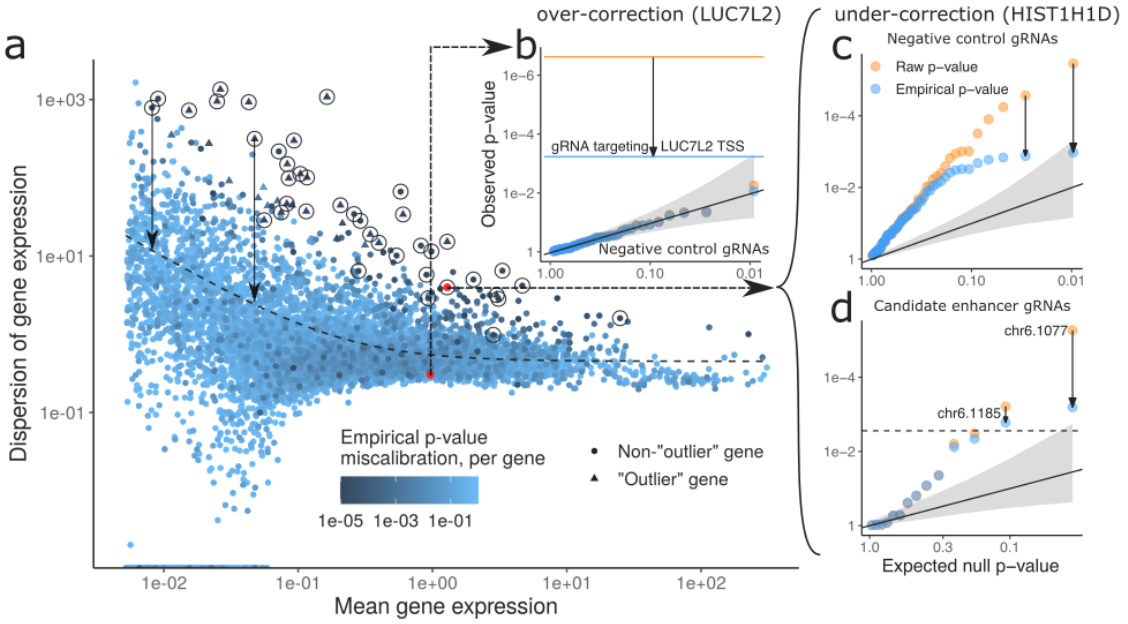


Figure S.1: Gasperini et al.'s empirical correction is insufficient to correct for miscalibration. **a,** Dispersion estimation procedure employed leads to miscalibration for high-dispersion genes, which the empirical correction does not adequately correct for, as measured by KS test applied to empirical 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.

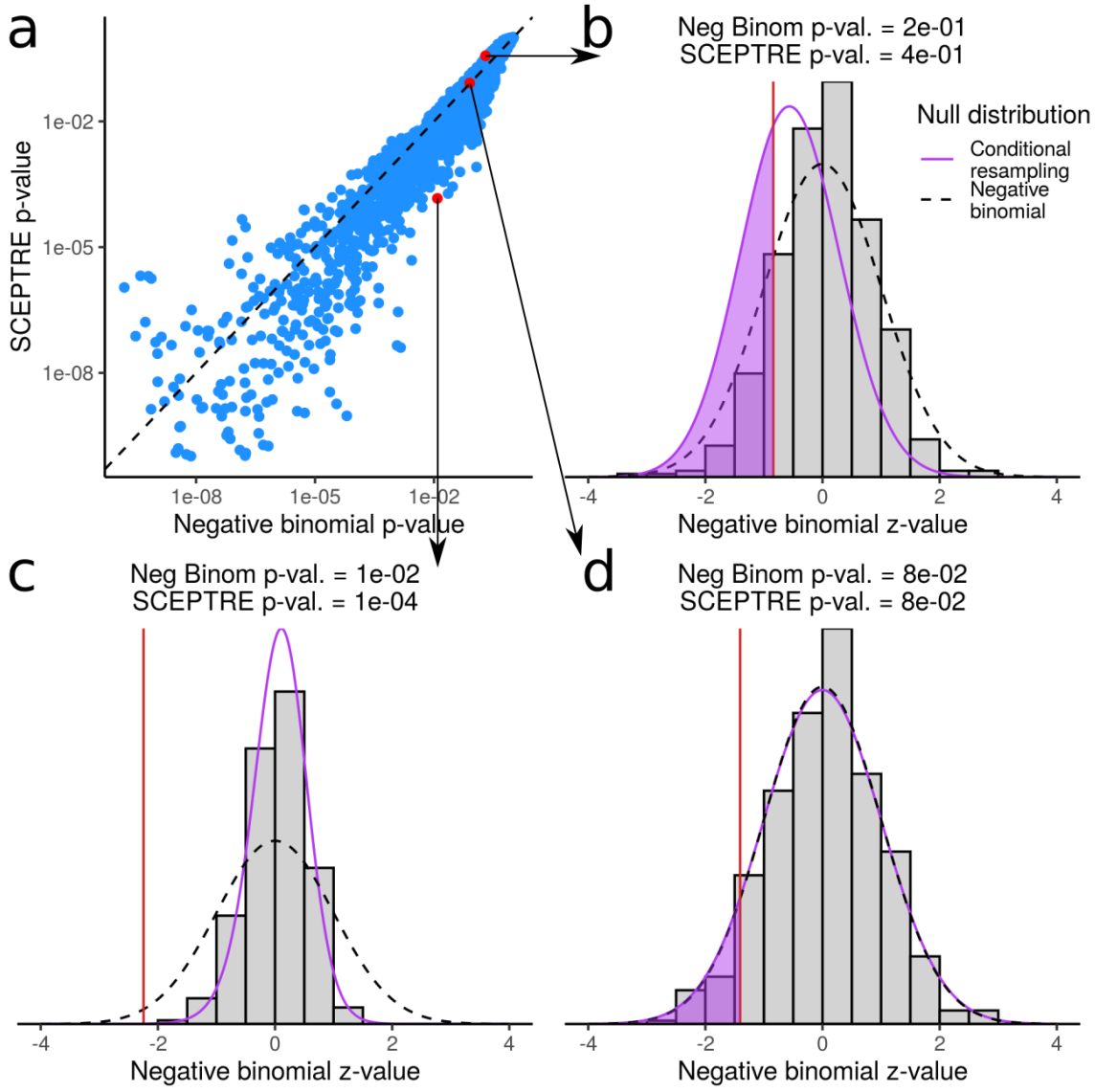


Figure S.2: **Comparing negative binomial and conditional resampling p-values based on the same test statistic. a,** The standard parametric negative binomial p-value versus that obtained from the same test statistic by conditional resampling, for each gene / candidate enhancer pair (both truncated at 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 parametric p-value is more significant (b), the conditional resampling p-value is more significant (c), the two p-values are about the same (d).

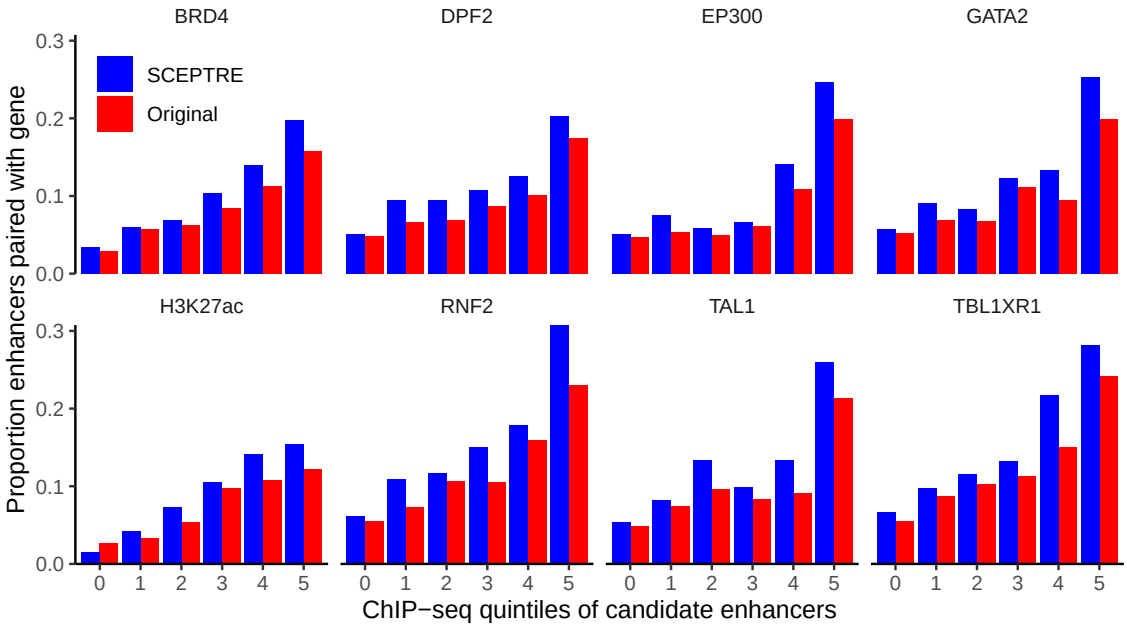


Figure S.3: **Details on ChIP-seq enrichment analysis.** Fraction of candidate enhancers paired with a gene, broken down by quintile of ChIP-seq signal (0 means the candidate enhancer did not overlap a ChIP-seq peak), based on which the odds ratios in Figure 4f were computed. Both methods generally pair candidate enhancers in higher ChIP-seq quintiles more frequently, but this enrichment is more pronounced in SCEPTRE across all eight ChIP-seq targets.