Conditional resampling improves sensitivity and specificity of single cell CRISPR regulatory screens

Eugene Katsevich[1] & 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*

Mapping gene-enhancer regulatory relationships is key to unraveling molecular disease mechanisms based on GWAS associations in non-coding regions. Recently developed CRISPR regulatory screens (CRSs) based on single cell RNA-seq (scRNA-seq) are a promising high-throughput experimental approach to this problem. However, the analysis of these screens presents significant statistical challenges, including modeling cell-level gene expression and correcting for sequencing depth. Using a recent large-scale CRS and its original analysis as a case study, we demonstrate weaknesses in existing analysis methodology, which lead to false positives as well as false negatives. To address these challenges, we propose SCEPTRE: analysis of single cell perturbation screens via conditional resampling. This novel method infers gene-enhancer associations by modeling the stochastic assortment of CRISPR gRNAs among cells instead of the gene expression, remaining valid despite arbitrary misspecification of the gene expression model. Applying SCEPTRE to the large-scale CRS, we demonstrate improvements in both sensitivity and specificity. We also discover 217 regulatory relationships not found in the original study, many of which are supported by existing functional data.

Eliciting gene-enhancer regulatory relationships remains a challenging and important problem. According to a recent review[1], "the functional dissection of trait-associated genetic variants from genome-wide association studies (GWAS) is likely to be a major focus of the field for the coming decade," so "it seems key that future efforts should prioritize the linking of enhancers to their target genes." Recently developed pooled CRISPR regulatory screen (CRS) technology is a promising experimental approach to this problem. CRSs initially focused on one gene at a time[2–5], densely tiling perturbations at nearby sites and quantifying their impact based on either cell proliferation or a marker for target gene expression. Even more recently, pairing CRISPR perturbations with single-cell RNA-sequencing (scRNA-seq)[6–9] has facilitated the study of how enhancers impact the entire transcriptome[10–12]. Among these, Gasperini et al[11] assayed thousands of enhancers genome-wide, multiplexing dozens of CRISPR guide RNAs (gRNAs) per cell. Such large-scale regulatory screens hold great promise for disentangling gene-enhancer relationships, providing more direct evidence of regulation than existing methods based on epigenetic data[13,14] or chromatin conformation[15,16].

Despite their promise, the data obtained from such assays pose significant statistical challenges, some inherited from scRNA-seq analysis and some unique to CRSs. Modeling cell-level gene expression is known to be a difficult task for scRNA-seq, and remains an active area of research[17–19]. Issues such as zero-inflation (or lack thereof) and dispersion estimation in the context of negative binomial modeling are as important for scRNA-seq based CRS analysis as they are for traditional scRNA-seq data. Furthermore, unlike traditional (sc)RNA-seq experiments, the "treatment"—in this case gRNA presence—is subject to measurement error[6,12,20]. In particular, we demonstrate that sequencing depth acts as a confounder, since it impacts the measurement of both the treatment (whether a given cell received a given gRNA) and the response (the resulting gene expression in that cell). Improper sequencing depth correction can therefore lead to misleading conclusions.

While there is a fairly substantial literature[21–26] on traditional genome-wide pooled CRISPR screens (testing the impact of gene knockout on cell proliferation), analysis methods for scRNA-seq based CRISPR regulatory screens are still scarce[27]. Gasperini et al. provide a starting point, carrying out a DESeq2[28]-inspired negative binomial regression analysis. However, we demonstrate that this original analysis has several deficiencies, leaving it vulnerable to both false positives and false negatives. A non-parametric approach called virtual FACS[10] has also been proposed to analyze similar data, although it does not easily allow correction for confounders such as cell cycle or sequencing depth.

In this paper, we propose SCEPTRE (analysis of single cell perturbation screens via conditional resampling), a novel methodology addressing the aforementioned challenges. The key idea is to sidestep the issues of modeling single cell gene expression by modeling the stochastic assortment of gRNAs among cells instead. Our approach is based on the conditional randomization test[29], which is a valid test of association despite arbitrary misspecifications of the gene expression distribution. It therefore enjoys the robustness to expression model misspecification of non-parametric approaches as well as the confounder correction abilities of parametric approaches. Applying SCEPTRE to the Gasperini et al. data, we find excellent calibration on negative control gRNAs and discover many novel regulatory relationships supported by a variety of existing functional assays.

Results

Analysis challenges. The original analysis by Gasperini et al[11] was carried out using Monocle2[30], whose differential expression analysis is motivated by DESeq2[28]. The latter, designed for the analysis of bulk RNA-seq data, relies on negative binomial regression. By examining the distribution of negative binomial p-values pairing each gene with each of 50 non-targeting control (NTC) gRNAs, Gasperini et al. find that these p-values are inflated. This trend is illustrated in their Figure 3E and reproduced here in Figure 1a. The inflation evident in the NTC p-values makes it difficult to interpret the p-values for the candidate enhancers. To remedy this issue, Gasperini et al. calibrate the candidate enhancer p-values against the distribution of NTC p-values instead of the uniform distribution. The resulting "empirical" p-values are used in their analysis for determining significance. While appealing in its simplicity, we demonstrate that calibration against the NTC distribution is not sufficient to address the issue. Because the effect of the miscalibration depends on both gene and gRNA, a "one size fits all" approach leads to overcorrection for some gene-enhancer pairs (false negatives) and undercorrection for others (false positives). We illustrate this in the context of two underlying challenging: dispersion estimation and sequencing depth correction.

First, we briefly review the dispersion estimation process employed by Monocle2 (Figure 1b). A raw dispersion is computed for each gene based on its sample variance. Then, a mean-dispersion relationship is fit, depicted as the black dashed line. Monocle2 then collapses each raw dispersion estimate onto this fitted line. While the raw dispersion estimates are noisy and some shrinkage is helpful[18], the collapsing of the dispersion estimates to the fitted line is likely to underestimate (overestimate) the dispersions for points above (below) the line. We compute 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 panel b. Circled genes have significantly miscalibrated empirical p-values based on a Bonferroni correction at level $\alpha = 0.05$; 21 were flagged by Gasperini et al. as outliers but 21 were not. One of these miscalibrated but non-outlier genes is HIST1H1D, whose empirical p-values are still inflated (panel c), leading to two potential false discoveries, one of which is deemed "high confidence" (panel d; see also Figure S7). On the other hand, genes like LUC7L2 have nearly uniformly distributed raw NTC p-values (panel e), making the empirical correction unnecessary. For this gene, the empirical correction decreases the significance of the p-value for a positive control gRNA by three orders of magnitude (the raw and empirical p-values are depicted by horizontal lines in panel e). We conclude that the NTC-based approach is insufficient to address issues with dispersion estimation or other kinds of parametric model misspecification.

Next, we discuss the challenge of normalizing for sequencing depth (measured here as the total number of unique molecular identifiers, or UMIs, per cell). Recently, it has been observed that some of the normalization strategies developed for bulk RNA-seq may not carry over to scRNA-seq[18]. In particular, these authors observed that the estimation of a single "size factor" for each cell (the approach taken by the differential expression module of Monocle2) is not sufficient to correct for sequencing depth for all genes. Instead, these authors advocate for a gene-by-gene correction. Furthermore, we observe that the normalization problem is even more important in the context of CRISPR screens, since sequencing is used for both gRNA detection and gene expression quantification. Despite the use of a targeted amplification protocol for gRNA detection[20], this process is

still imperfect. Indeed, the total number of gRNAs detected in a cell increases with the sequencing depth ($\rho = 0.35, p < 10^{-15}$; Figure 1f). This makes sequencing depth a likely confounder if not properly adjusted for. Examining the distributions of p-values for three NTC gRNAs, computed as in Gasperini et al. but with an improved dispersion estimate (Methods), provides evidence for this confounding (Figure 1g). We see that confounding can cause the p-values to be either too conservative or too liberal. The leftover confounding is tied to the correlation between the gRNA indicator and the sequencing depth, after applying the correction for guide count. We computed a z-score for the direction and strength of this residual confounding (see Methods), annotated in the legend of panel g for each gRNA: scrambled_21 ($z = -3.25$), scrambled_6 ($z = -0.53$), and random_9 ($z = 3.30$). As expected, the residual strength and direction of the confounding impact the direction of the p-value miscalibration. Strong negative residual confounding leads to false positives, as with scrambled_21. On the other hand, strong positive residual confounding leads to false negatives, as with random_9.

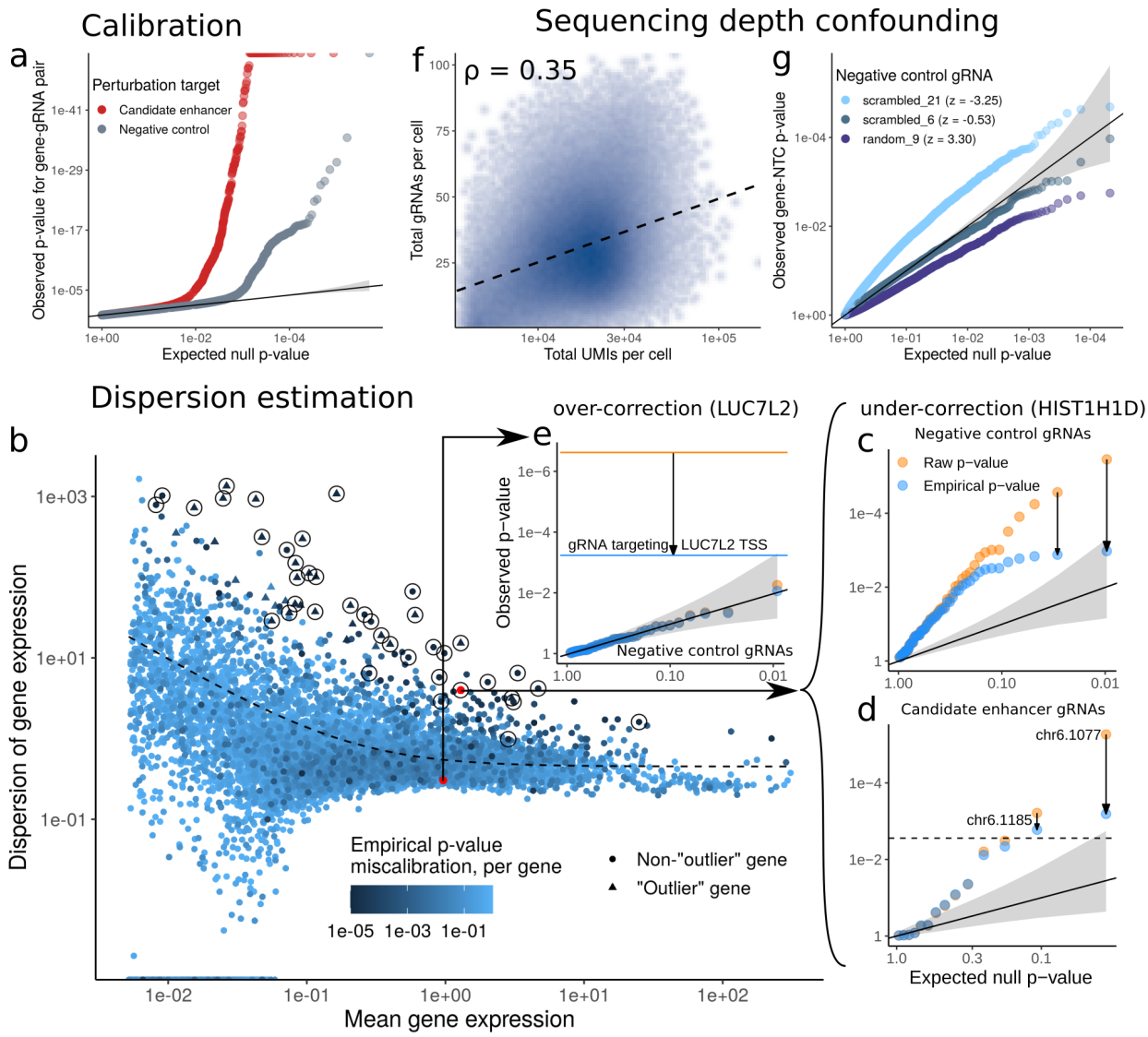


Figure 1: **CRISPR screen analysis challenges can lead to false positives and false negatives. a**, QQ-plot of Gasperini et al. p-values for all gene-gRNA pairs involving either candidate enhancer or negative control gRNAs. Inflation in negative controls makes it harder to interpret associations between genes and candidate enhancers, motivating the empirical p-value correction. **b-d**, Collapsing dispersion estimates to fitted mean-variance relationship (**b**) leads to under-correction for some genes (**c,d**) and over-correction for others (**e**). Circled genes in panel **b** have an NTC-based miscalibration p-value smaller than the Bonferroni threshold. Empirical correction not strong enough for HIST1H1D gene (**c**), leading to two potential false discoveries (**d**). Dashed horizontal line in panel **d** represents multiple testing threshold. 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 (**e**). **f-g**, sequencing depth impacts gRNA detection (**f**) and observed expression levels, and therefore acts as a confounder. Recomputing the negative binomial p-values with improved dispersions to isolate this effect, inadequate sequencing depth correction can cause liberal bias for some gRNAs and conservative bias for others (**g**). The direction and magnitude of this miscalibration for a given gRNA is predicted well by the residual correlation between gRNA presence and sequencing depth after accounting for total gRNAs per cell, quantified by the z-values in parentheses.

Improvements to the negative binomial approach. As a first attempt to alleviate the miscalibration, we introduced a few improvements to the negative binomial model. First, we improved the sequencing depth correction. Instead of relying on DESeq2-style size factors, we corrected for sequencing depth by introducing two additional covariates into the negative binomial regression: the total number of UMIs observed and the number of genes with at least one UMI in a given cell. These are in addition to the confounders Gasperini et al. corrected for (see Methods). This strategy is in line with that advocated by a recent work on scRNA-seq analysis[18]. Second, we replaced the collapsed dispersion estimates by the raw estimates (recall Figure 1b). 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. Third, we replaced the (two-sided) likelihood ratio test employed by Gasperini et al. by a left-tailed z-test, for sensitivity to candidate enhancers that decrease gene expression when perturbed.

To assess the impact of these improvements on calibration, we applied the one-sided negative binomial test with and without the improved confounder correction and dispersion estimation (four total possibilities) to all 50 NTC gRNAs paired with all genes (Figure S5). 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, apparently producing mostly conservative p-values.

There are many possible reasons for the remaining miscalibration. It could be a problem with the dispersion estimates, or misspecification of the negative binomial model itself. While more effort could certainly be invested in further improvements of the parametric model for gene expression, we acknowledge that modeling scRNA-seq expression is a challenging problem that remains open. Furthermore, negative control gRNAs may not always be available to assess calibration. These considerations highlight the appeal of nonparametric approaches, which bypass parametric modeling entirely. This motivates us to propose the following *conditional resampling* approach.

SCEPTRE: Analysis of single cell perturbation screens via conditional resampling. Our main idea is to circumvent the challenge of modeling single cell gene expression by modeling the stochastic assortment of gRNAs to cells instead. We view a CRISPR screen as a kind of randomized experiment where gRNAs are randomly assigned to cells, so a null distribution for any test statistic measuring the effect of an enhancer on a gene can be built by repeatedly reassigning gRNAs to cells. This basic idea dates back to Fisher's exact test[31] and is the basis for permutation tests, which are widely used in genomics and have been proposed for the analysis of CRISPR screens in particular[24]. We must go beyond permutation tests, however, to account for the fact that different cells have different probabilities of receiving a gRNA (we discuss this below in the context of Figure 3). To this end, we propose to use the conditional randomization test (CRT) recently introduced by Candès et al.[29]

SCEPTRE (Figure 2) proceeds one gRNA and one gene at a time. To test the effect of this gRNA on this 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 oc-

currence of the gRNA in a cell, based on its covariates. For each cell, this yields a fitted probability that it contains the gRNA. By analogy with causal inference, this plays the role of the *propensity score*, i.e. the probability of receiving the "treatment." Then, we generate a large number (say 500) of reshuffled datasets, where the expression and the covariates 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 2c). We found that the skew-t distribution, used by CRISPhieRmix[25] for a different purpose, provided a good fit to these null histograms, so we computed a final p-value by comparing the original z-value to this fitted skew-t null distribution. 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 S6).

To mitigate the extra computational cost of resampling, we implement computational accelerations that reduce 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.

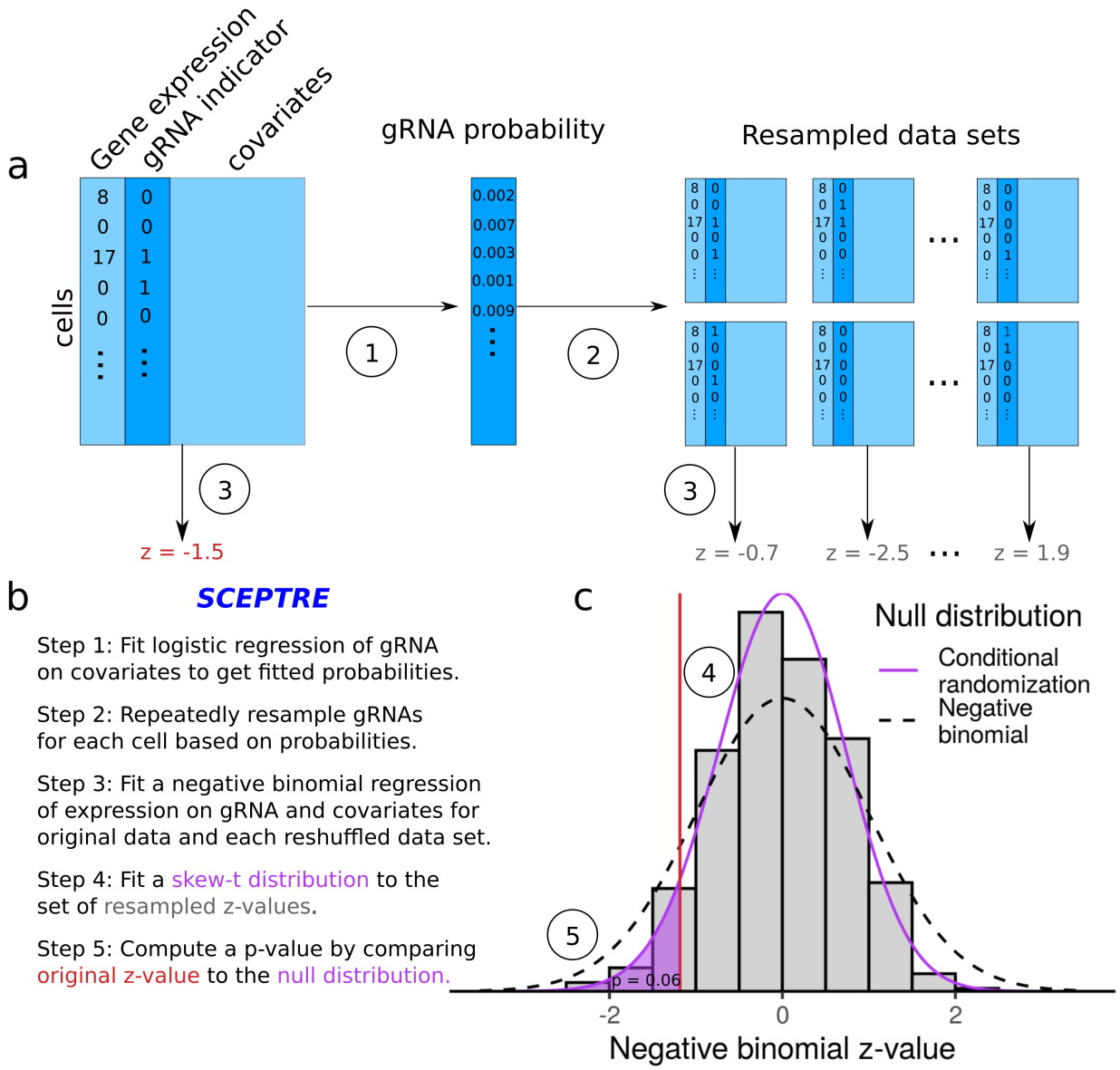


Figure 2: **SCEPTRE: Analysis of single cell perturbation screens via conditional resampling.** Schematic of methodology for one gene and one gRNA. **a,** Applying the conditional randomization test[29] to single cell CRISPR screens. The probability that each cell contains the gRNA is fit using logistic regression based on covariates such as total UMIs and guide count. Then, gRNAs are independently reassigned to each cell according to these probabilities to form "negative control" data sets. **b,** Methodology outline. **c,** Inference using a resampling-based null distribution. The negative binomial z-values from each resampled data set are used to fit a skew-t null distribution, against which the original NB z-value is compared. The dashed line shows the $N(0, 1)$ distribution, against which the NB z-value would normally be compared.

SCEPTRE is well calibrated despite expression model misspecification. By shifting the modeling burden away from the single cell gene expression, SCEPTRE avoids the miscalibration caused by misspecification of parametric models for this quantity. We explored this observation in a small proof-of-concept simulation with 1000 cells, one gene, one NTC gRNA, and one confounder (Figure 3a). We considered the one-sided z test statistic based on a negative binomial expression distribution with mean 5 and dispersion 1. We then generated the data from this model and from three others: one with dispersion 0.2, one with dispersion 5, and one with dispersion 1 but with 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's conditional randomization approach is well-calibrated regardless of model misspecification.

Next, we applied SCEPTRE to test the association between all NTC gRNAs and all genes (Figure 3b-d). 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. We also break the pairs down by NTC, resulting in 50 separate QQ plots. We overlay these QQ plots by taking the median observed p-value for each expected p-value, as well as the corresponding 95% confidence band (Figure 3c). Again, SCEPTRE shows nearly perfect calibration, as evidenced by its colored confidence region coinciding with the gray shaded region representing the pointwise 95% confidence band for the uniform distribution. We find a similar conclusion when breaking the pairs down by gene (Figure 3d), though the other methods perform better with respect to this metric.

To assess the power of SCEPTRE, we applied it to the positive control gRNA / gene pairs assayed in Gasperini et al (Figure 4b). Positive control gRNAs targeted either gene transcription start sites or enhancers previously found to regulate a gene. The proposed method captures strong signal among the positive controls, especially for those targeting TSSs. Compared to the original approach, positive controls of both kinds are generally more significant according to SPECTRE. We note that the empirical correction employed by Gasperini et al. limits the accuracy of p-values to about 10^{-6}, which explains at least part of this observed power difference. We considered a strategy akin to the skew-t fit to alleviate this issue, but the extremely heavy-tailed empirical distribution precluded a simple parametric fit.

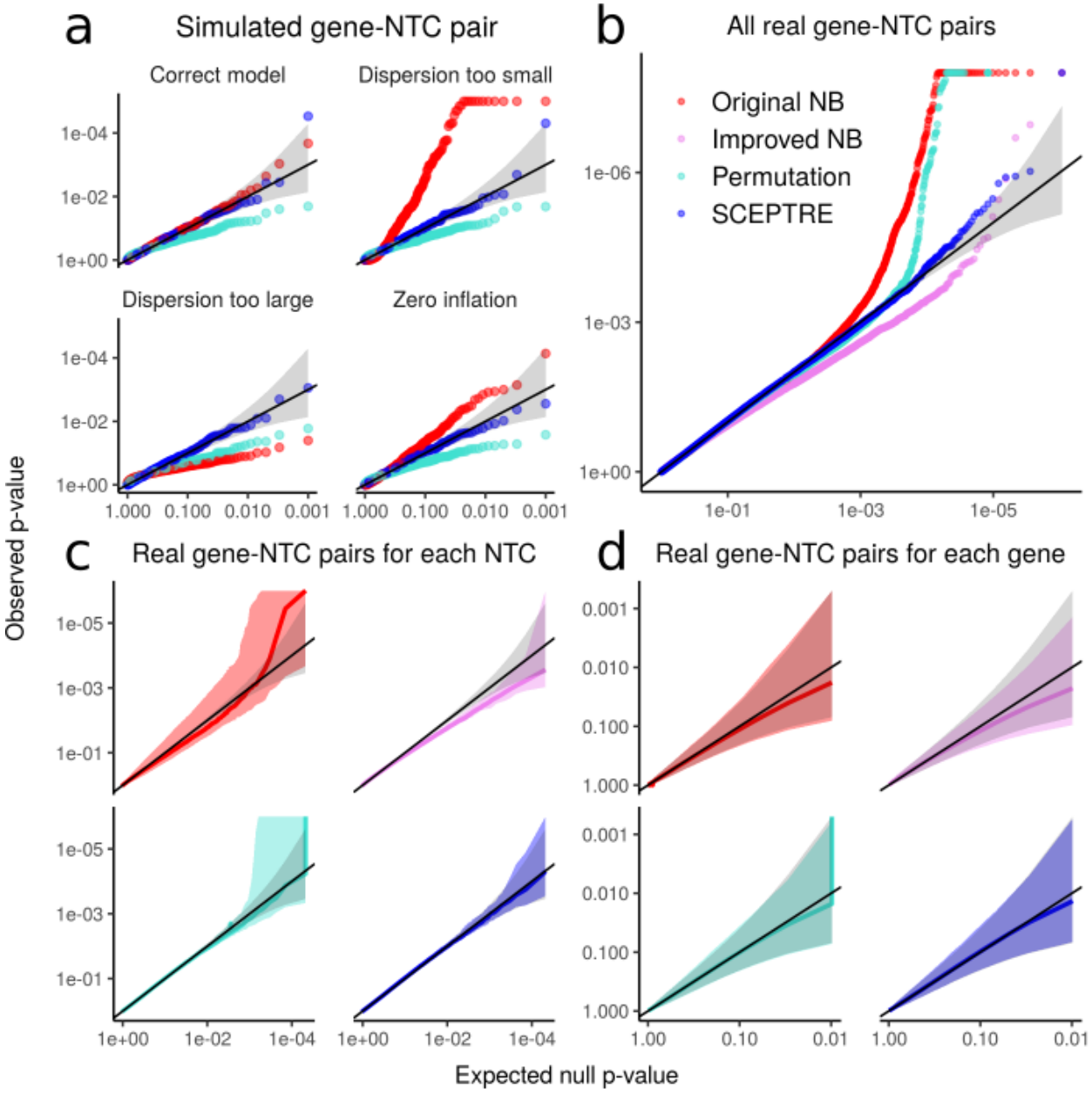
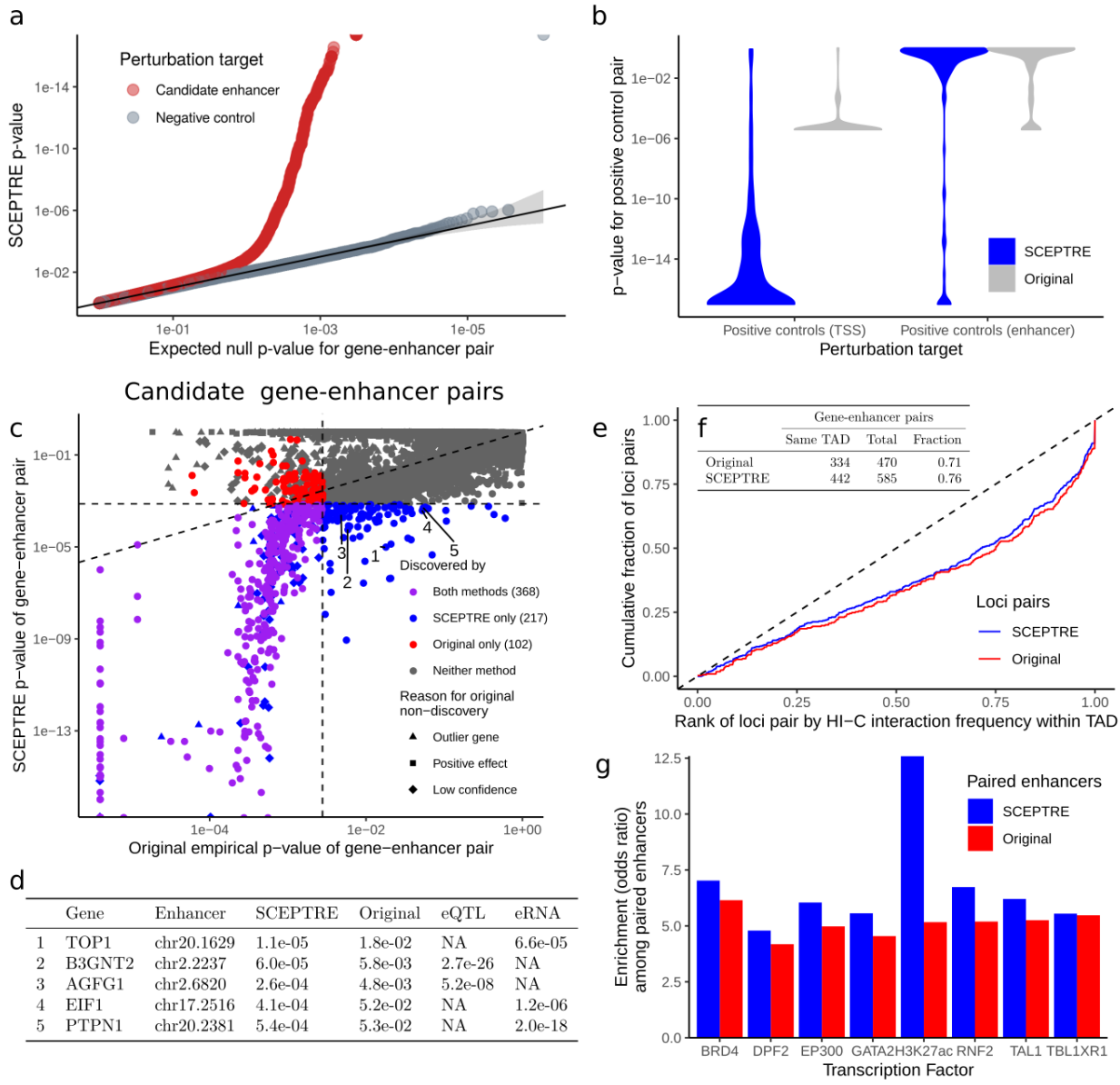


Figure 3: **Calibration of SCEPTRE on simulated and real negative control data. a,** Three ways of calibrating a negative binomial test statistic, when the test statistic is based on the correct model or contains expression model misspecifications. Only the conditional resampling approach maintains calibration despite model misspecification. **b-d,** 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**) or by gene (**d**). The colored lines and shaded regions in panels **c** and **d** are the median and 95% confidence regions of the QQ plots across NTCs or genes, respectively; the gray shaded regions show the corresponding regions under perfect calibration. The matching of SCEPTRE's colored regions with the gray ones in panels **c** and **d**, together with the overall near-uniformity in panel **b**, demonstrates its excellent calibration on real negative control data. Note: p-values in all panels truncated for visualization.

Analysis of candidate gene-enhancer regulatory relationships. We applied SCEPTRE to the 84595 gene-enhancer pairs considered in Gasperini et al., encompassing 10560 genes and 5779 candidate enhancers (Figure 4a). We applied the Benjamini-Hochberg correction at level 0.1 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 4c), 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; see Figure S7.

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 4d) involving genes not paired to any enhancers in the original analysis, which are supported by GTEx[32] eQTL p-values in whole blood or enhancer RNA correlation p-values across tissues from the FANTOM project[33]. These pairs are listed in the GeneHancer database[34], 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.

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. 76% of SCEPTRE's 585 gene-enhancer pairs fell in the same topologically associated domain (TAD), compared to 71% of the 470 pairs discovered in the original analysis. We also repeated Gasperini et al.'s contact frequency enrichment analysis for those pairs falling in the same TAD (see Figure 4e, as well as Figures 6E and S5B,C of Gasperini et al). 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, to see how much more ChIP-seq signal there is in paired enhancers compared to the set of all candidate enhancers. The enrichment was quantified 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 transcription factors considered by Gasperini et al. (Figures 4g and S8).



	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,** The SCEPTRE p-values of the two kinds of positive control gRNAs show significant signal, indicating the power of the proposed approach. Positive control signal is generally stronger than that in the original approach. **c**, 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. **d**, Five gene-enhancer pairs discovered by SCEPTRE but not the original analysis, each supported by a GTeX eQTL or FANTOM enhancer RNA correlation p-value. **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. **f,** 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. **g,** Enrichment of ChIP-seq signal from eight cell-type relevant transcription factors among paired enhancers. SCEPTRE exhibits greater enrichment across all transcription factors.

Discussion

We presented SCEPTRE, a novel method for the analysis of scRNA-seq based CRISPR regulatory screens, which exhibits excellent calibration despite imperfect specification of the single cell gene expression model. We avoid relying on the validity of the expression model by using the simpler process of gRNA assortment among cells for inference instead. Unlike traditional nonparametric analysis methods, our approach seamlessly corrects for important cell-level covariates. Our analysis yielded many new biologically plausible gene-enhancer relationships, supported by evidence from eQTL, enhancer RNA co-expression, ChIP-seq, and HI-C data. We implemented 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.

While SCEPTRE greatly reduces the burden of modeling single cell expression data, there are still reasons to seek good models for gene expression. Even if the model for gRNA occurrence in a cell were perfectly specified, better expression models can improve the sensitivity of the CRT[35]. While gRNA occurrence is much simpler to model than gene expression, in practice we can still only approximate the former. We conjecture that better approximations to the expression model can make SCEPTRE more robust to misspecifications of the gRNA occurrence model; this phenomenon is well-studied in the related contexts[36,37].

There are several directions for improvement in the analysis of scRNA-seq based pooled CRISPR screens, which are not addressed by SCEPTRE. Many of these have in fact been addressed by existing methodologies, mostly for different kinds of CRISPR screens than the one considered here. Such remaining challenges include variable effectiveness of gRNAs[21,25], interactions among enhancers[10,38,39], and the limited resolution of CRISPR interference[40]. Furthermore, many techniques from the increasingly rich literature on scRNA-seq analysis can be brought to bear on this aspect of CRISPR regulatory screen analysis. Therefore, SCEPTRE adds to a growing statistical toolbox for analyzing CRISPR regulatory screens, and in fact other kinds of single cell CRISPR screens as well. Continued methodology development is crucial to fully realize the unprecedented potential of this new technology to reliably elucidate regulatory relationships.

We are optimistic that the rapid advances in CRISPR screen technology, together with the appropriate statistical tools, will facilitate the mapping of regulatory networks and ultimately improve our understanding of disease mechanisms.

References

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

2. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

3. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).

4. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).

5. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods* **14**, 629–635 (2017).

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

7. 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).

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

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

10. 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).

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

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

13. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

14. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

15. 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).

16. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47**, 598–606 (2015).

17. 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).

18. 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).

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

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

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

22. Yu, J., Silva, J. & Califano, A. ScreenBEAM: A novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. *Bioinformatics* **32**, 260–267 (2016).

23. Hart, T. & Moffat, J. BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 1–7 (2016).

24. Jia, G., Wang, X. & Xiao, G. A permutation-based non-parametric analysis of CRISPR screen data. *BMC Genomics* **18**, 1–11 (2017).

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

26. Imkeller, K., Ambrosi, G., Boutros, M. & Huber, W. Modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection. *bioRxiv* (2019). URL https://www.biorxiv.org/content/10.1101/699348v1.

27. 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).

28. 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).

29. 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).

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

31. Fisher, R. *The Design of Experiments* (Oliver and Boyd, Edinburgh, 1935).

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

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

34. 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).

35. 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.

36. 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).

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

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

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

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

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

42. 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).

43. 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.

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

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

Acknowledgements We are indebted to Molly Gasperini and Jacob Tome for clarifying several aspects of their data analysis[11]. This work was supported, in part, by National Institute of Mental Health (NIMH) grants R01MH123184 and R37MH057881. Part of the data analysis used the Extreme Science and Engineering Discovery Environment (XSEDE)[41], which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system[42], 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 K.R. (email: roeder@andrew.cmu.edu).

Methods

Reproducing the negative binomial approach of Gasperini et al. Consider a particular gene/gRNA pair. For each cell $i = 1, \ldots, n$, let $X_i \in \{0, 1\}$ indicate whether the gRNA was present in the cell, let $Y_i \in \{0, 1, 2, \ldots\}$ be the gene expression in the cell, defined as the number of unique molecular identifiers (UMIs) from this gene, and let $Z_i \in \mathbb{R}^d$ be a list of cell-level covariates.

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 covariates: 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. To focus on the issue of sequencing depth confounding, we used raw dispersion estimates (compare the red and blue curves in Figure S5) as well as one-sided z-tests. We then repeated the negative binomial regression analysis underlying Monocle2 with these updated dispersion estimates. These two modifications yielded new negative binomial regression p-values for each gene/gRNA pair (Figure 1g).

Next, we sought to understand the direction and strength of the confounding effect by testing 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 1g.

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[29] 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

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

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 covariates), 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[43] *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, 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 covariate 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[44]. 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 4c-e and S8) were carried out almost exactly following Gasperini et al. The only change we made is in our quantification of the ChIP-seq enrichment (Figure 4e). We use the odds ratio of a candidate enhancer being paired to a gene, comparing the top and bottom ChIP-seq quintiles. Gasperini et al. use a more complicated formula for this enrichment that we were unable to reproduce.

Data availability

Analysis results are available online at https://bit.ly/SCEPTRE. All analysis was performed on publicly available data. The CRISPR screen data[11] 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[45] and are available at https://www.encodeproject.org/. The HI-C enrichment analysis is based on the data from Rao et al[15], 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[34], 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.

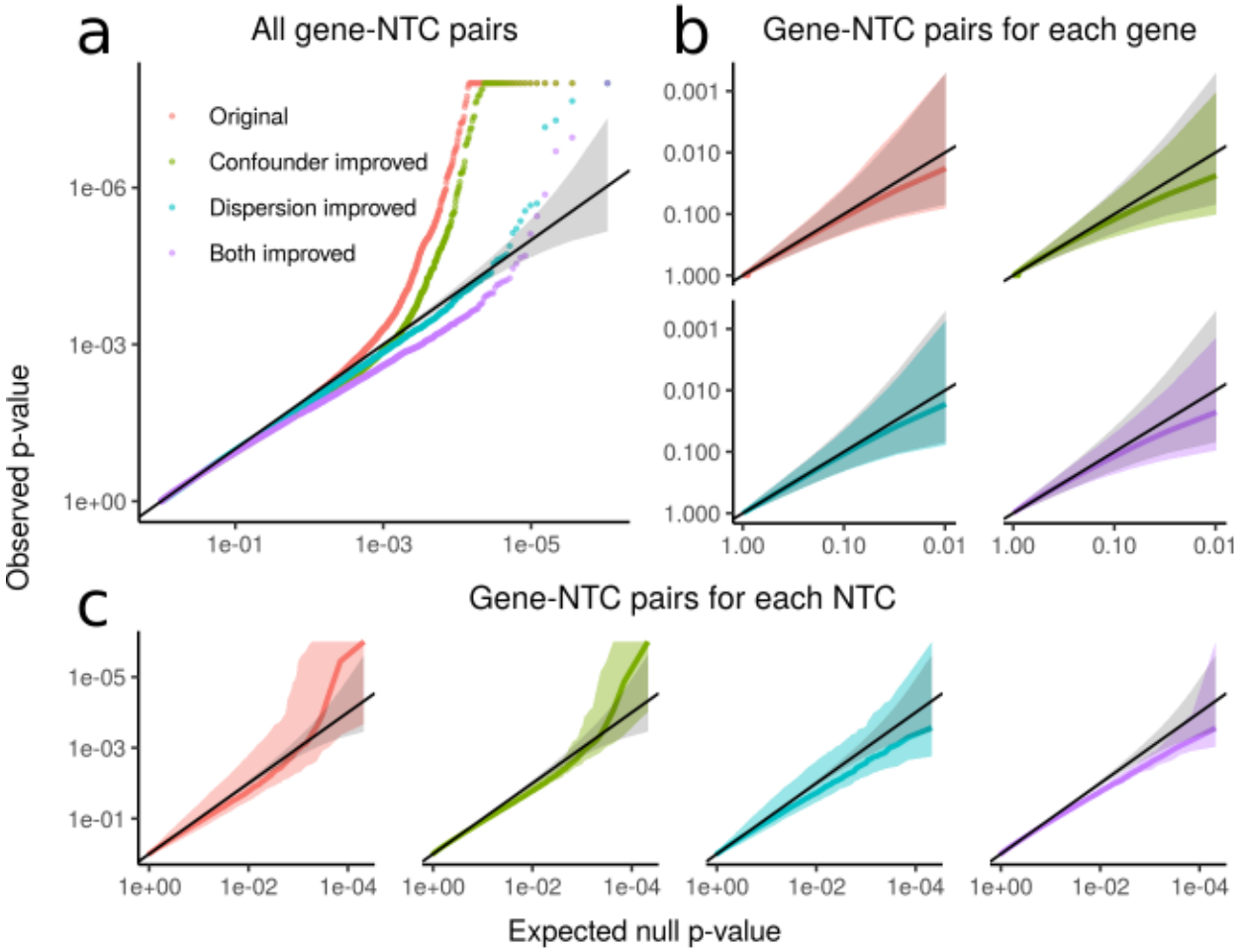


Figure S5: **Improving dispersion estimation and confounder correction.** NTC-based calibration of four negative binomial methods is shown: original, improved confounder correction, improved dispersion estimation, and both improvements. **a,** QQ plot for all gene-NTC pairs. The method with dispersion improvement (shown in blue) appears to be well calibrated, but this is not the case; see panel **c**. **b,** Calibration on a per-gene basis; details as in Figure 3d. All methods perform relatively well on this metric, showing only slight conservative bias on average. **c,** Calibration on a per-NTC basis; details as in Figure 3c. Note that the method in blue shows noticeable miscalibration in both conservative and liberal directions. Also recall Figure 1g, which is based on the same method. In summary, the method with improved confounder correction and dispersion estimation performs best overall, but still is noticeably miscalibrated.

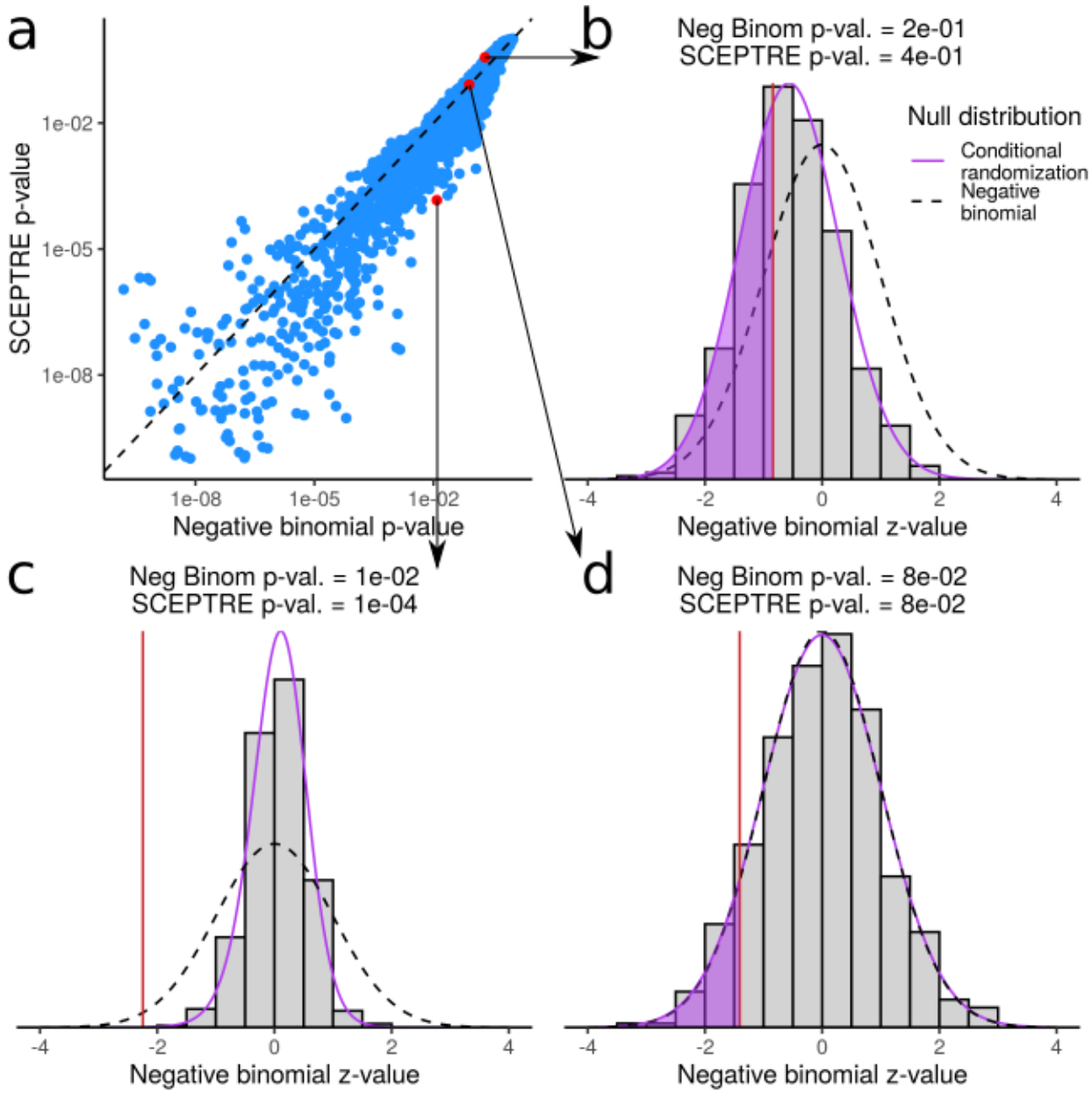


Figure S6: **Comparing negative binomial and conditional randomization p-values based on the same test statistic. a,** The standard parametric negative binomial p-value versus that obtained from the same test statistic by conditional randomization, for each gene / candidate enhancer pair (both truncated at 10^{-10} for visualization). The two can diverge fairly substantially. **b-d,** Parametric and conditional randomization null distributions for the negative binomial z-value in three cases: the two p-values are about the same (**b**), the conditional randomization p-value is more significant (**c**), the parametric p-value is more significant (**d**).

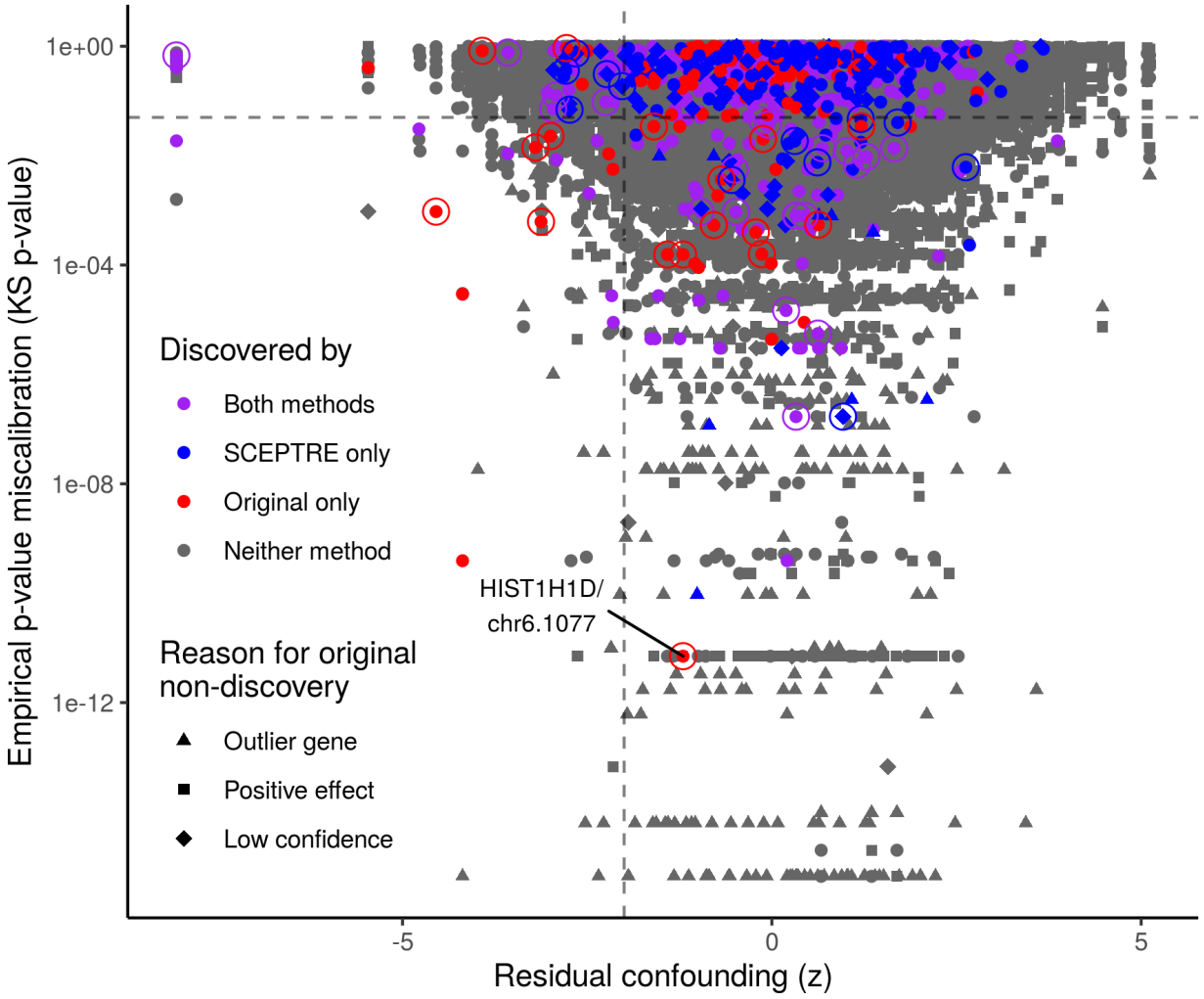


Figure S7: **Examining the sources of potential false positives in the original analysis.** Each point represents a gene / candidate enhancer pair. The x-axis shows the residual confounding z-score (see Methods and Figure 1g), computed for each gRNA, while the y-axis shows the Kolmogorov-Smirnov p-value of the empirical NTC p-values for each gene. Colors and shapes are as in Figure 4c. Circled points correspond to gene / candidate enhancer pairs that do not fall in the same TAD (recall Figure 4f) and either have residual confounding $z < -2$ or KS p-value $p < 0.05$ (these thresholds are denoted by the dashed vertical and horizontal lines). The 18 circled red points therefore represent discoveries in the original analysis that may be false positives; recall the labeled pair HIST1H1D/chr6.1077 from Figure 1d.

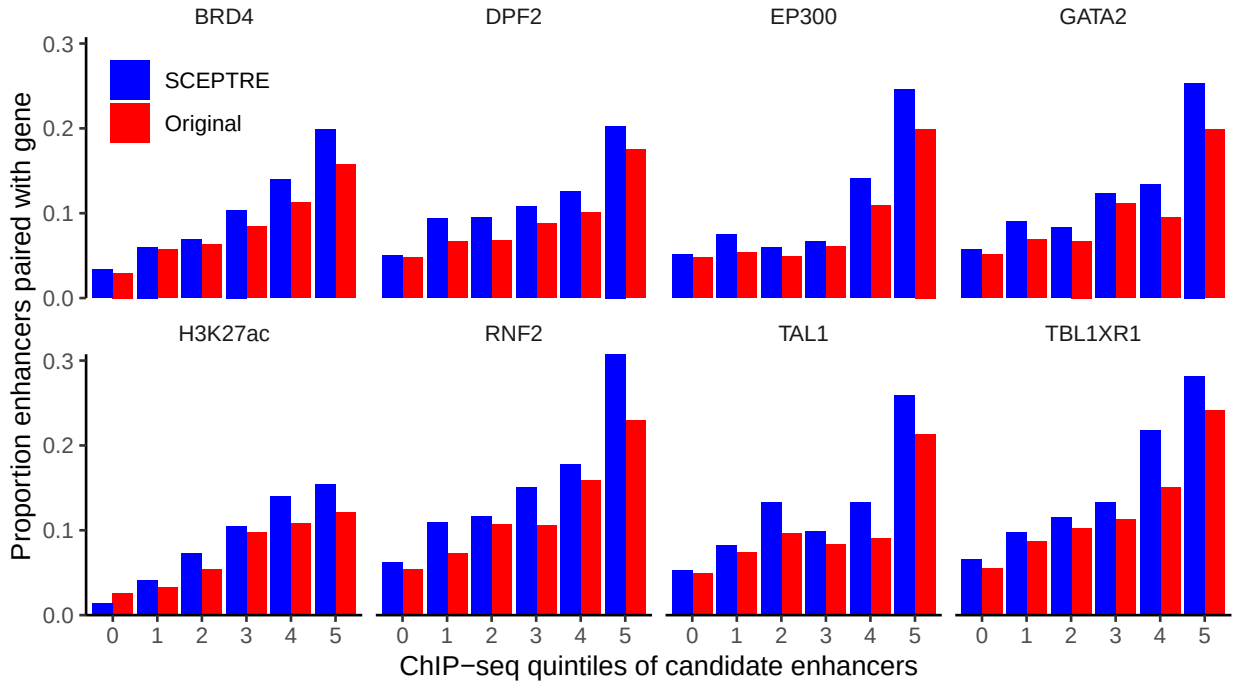


Figure S8: **Details on ChIP-seq enrichment analysis.** Fraction of candidate enhancers paired with a gene, broken down by quintile of ChIP-seq signal (0 means the candidate enhancer did not overlap a ChIP-seq peak), based on which the odds ratios in Figure 4g were computed. Both methods generally pair candidate enhancers in higher ChIP-seq quintiles more frequently, but this enrichment is more pronounced in SCEPTRE across all eight transcription factors.