Robust differential expression testing for single-cell CRISPR screens

Timothy Barry[1], Kathryn Roeder[1,2], and Eugene Katsevich[3]
[1]*Department of Statistics and Data Science, Carnegie Mellon University*
[2]*Computational Biology Department, Carnegie Mellon University*
[3]*Department of Statistics and Data Science, Wharton School, University of Pennsylvania*

Single-cell CRISPR screens have emerged as a standard tool for mapping genetic perturbations to phenotypic changes in single cells. A fundamental task in single-cell CRISPR screen data analysis is to test for association between a CRISPR perturbation and a univariate outcome, such as the expression of a gene or protein. We conduct the first-ever comprehensive benchmarking study of single-cell CRISPR screen differential expression (DE) methods, applying five leading methods to analyze six diverse datasets. We find that existing methods exhibit pervasive miscalibration, rendering results unreliable. To search for possible sources of miscalibration, we conduct an extensive investigation of the data, identifying three core analysis challenges: data sparsity, confounding, and model misspecification. We introduce SCEPTRE, a single-cell CRISPR screen DE method based on the statistically principled technique of resampling score statistics. SCEPTRE addresses the core analysis challenges both in theory and in practice, demonstrating improved calibration and power across all datasets.

Single-cell CRISPR screens (e.g., perturb-seq[1]) link genetic perturbations to changes in the transcriptome and other molecular phenotypes in single cells. Among the most fundamental tasks in single-cell CRISPR screen data analysis is to test for association between a CRISPR perturbation and the expression of a gene, protein, or other univariate outcome. A common use case of single-cell CRISPR screens, for example, is to map non-coding regulatory elements of unknown function to target genes by testing for association between CRISPR perturbations that inhibit these regulatory elements and candidate genes[2–4]. Another standard application is to construct causal (as opposed to merely correlative) models of gene-gene and gene-protein regulation by knocking out a gene of interest and then testing for a change in the expression of genes or proteins hypothesized to be regulated — directly or indirectly — by the targeted gene[5,6]. A reliable test of association between a CRISPR perturbation and a univariate outcome, thus, would greatly facilitate the application of single-cell CRISPR screen technology to unravel the regulatory wiring of the genome across cell types and disease states.

Researchers have developed several methods[1,2,5–8] to test for association between a CRISPR perturbation and a univariate outcome on low-MOI (multiplicity of infection)

single-cell CRISPR screen data. However, existing single-cell CRISPR screen differential expression (DE) methods have not undergone rigorous statistical validation. In particular, it is unknown whether existing methods are adequately calibrated (i.e., able to control the rate of false discoveries) or adequately powered (i.e., able to consistently make true discoveries). To shed light on the performance of current methodologies, we introduce a simple framework — the "undercover gRNA procedure" — to rigorously assess the calibration of competing single-cell CRISPR screen DE methods. We leverage this framework to conduct the first-ever comprehensive benchmarking study of low MOI single-cell CRISPR screen DE methods, applying five leading methods to analyze six diverse, recent single-cell CRISPR screen datasets. We find that all existing methods exhibit pervasive miscalibration, indicating that results obtained using these methods may be contaminated by excess false positives.

To uncover possible sources of miscalibration, we conduct an in-depth empirical investigation of the data, uncovering three core analysis challenges: confounding, model misspecification, and data sparsity. No existing method addresses all of these analysis challenges, explaining their collective lack of calibration. Guided by our recent work on high-MOI single-cell CRISPR screen analysis[9], we introduce a new method, SCEPTRE, for association testing and estimation in the low-MOI setting. SCEPTRE is based on the simple and statistically principled technique of resampling negative binomial (NB) score statistics. In effect a robust version of NB regression, SCEPTRE addresses all three core analysis challenges both in theory and in practice. An application of SCEPTRE to negative and positive control data reveals improved calibration and power relative to other methods; in most cases the difference in performance is considerable. In short SCEPTRE is a theoretically sound and empirically promising approach to single-cell CRISPR screen analysis.

Results

Methods and datasets under analysis

Before introducing SCEPTRE we study the performance of five leading single-cell CRISPR screen differential expression (DE) methods: Seurat DE[5,10], MIMOSCA[1,6], "Liscovitch Method[7]," "Schraivogel Method[2]," and "Weissman Method[8]." (The latter three methods were not given names by the original authors; thus, we assign these methods names according to the papers in which they initially were proposed.) The methods vary along several dimensions (Table 1), including how they normalize the data, whether they make parametric assumptions, and whether they rely upon large-sample asymptotics. We systematically benchmark the performance of these methods on six single-cell CRISPR screen

Method	Short description	Parametric assumption	Asymptotic approx.	Control group
Seurat DE	A Mann-Whitney test on the library size-normalized expressions	No	Yes	NT cells
MIMOSCA	Permutation test with elastic net test statistic	No	No	Compliment set
Liscovitch Method	Two-sample t-test on the library size-normalized expressions	Yes	No	NT cells
Schraivogel Method	Implementation of MAST[11] for single-cell CRISPR screens	Yes	Yes	NT cells
Weissman method	Kolmogorov-Smirnov two-sample test	No	Yes	NT cells

Table 1: **A summary of single-cell CRISPR screen DE methods currently in use**. The methods vary along several key axes, including the use (or lack thereof) of parametric assumptions and asymptotic approximations. The "control group" column, which refers to the set of cells used in the DE test, is explicated in the Existing methods details. Asymptotic approx., asymptotic approximation. NT, non-targeting.

datasets, five real and one simulated (Tables S1-S2). The five real datasets come from three recent papers: Frangieh 2021[6] (three datasets), Papalexi 2021[5] (one dataset), and Schraivogel 2020[2] (one dataset). The data are diverse, varying along the axes of CRISPR modality (CRISPRko or CRISPRi), technology platform (perturb-CITE seq, ECCITE-seq, or targeted perturb-seq), cell type (TIL, K562, or THP1), and genomic element targeted (enhancers or gene transcription start sites). Notably, the Papalexi data are multimodal, containing both gene and protein expression measurements. For simplicity we analyze the gene and protein modalities separately throughout.

The undercover gRNA procedure and benchmarking results

Verifying that a single-cell CRISPR screen DE method is correctly calibrated (i.e., that it yields uniformly distributed p-values under the null hypothesis of no association between the gRNA and the gene) is crucial to ensuring reliability of the results. Unfortunately, there does not exist a standard protocol for assessing the calibration of single-cell CRISPR

screen DE methods or for comparing the calibration of competing methods. To help fill this methodological gap, we describe the "undercover gRNA procedure," a simple strategy for generating negative control gRNA-gene pairs on which a method can be applied to assess its calibration (Figure 1a). A single-cell CRISPR screen DE method requires three inputs: (i) the gene and gRNA expression data (not pictured), (ii) a table containing the ID (e.g.,"gRNA 1," "gRNA 2," etc.) and type (either "targeting" or "non-targeting") of each gRNA, and (iii) a table containing the set of gRNA-gene pairs to analyze.

We proceed as follows. First, for a given non-targeting (NT) gRNA (e.g., "gRNA 1" in the figure), we change the ID of this gRNA to "undercover" and its type to "non-targeting." We call this NT gRNA the "undercover gRNA" as it masquerades as a targeting gRNA. Second, we form the table of pairs to analyze by coupling the undercover gRNA to the entire set of genes. Third, we pass the (unaltered) expression data, the swapped table of gRNA types, and the table of pairs to analyze to the single-cell CRISPR screen DE method. The method performs a test of association between the undercover gRNA and each gene. We repeat this procedure for each NT gRNA, yielding a set of $N_{\text{gene}} \cdot N_{\text{NT}}$ null p-values, where N_{gene} is the number of genes and N_{NT} is the number of NT gRNAs. If the method under assessment is correctly calibrated, the null p-values that it yields within this framework are uniformly distributed. Deviations from uniformity — and thus miscalibration of the method — can be detected by plotting the null p-values on a QQ-plot. The undercover gRNA procedure is a highly flexible black box that can be applied to virtually any single-cell CRISPR screen method and dataset.

We leverage the undercover gRNA procedure to assess the calibration of each of the five methods (Tables 1) on the six datasets (Table S1-S2), uncovering widespread (and in many cases extreme) miscalibration. We display the results for the Papalexi (gene modality) and Frangieh IFN-γ negative control data (Figures 1b and 1c, respectively) here; full results are available in the supplementary materials (Figures S1-S3). The left (resp., middle) panel of Figures 1b and 1c displays a QQ-plot of the null p-values plotted on an untransformed (resp., negative log transformed) scale; the null p-values should lie along the diagonal. The right panel, meanwhile, shows the number of rejections that each method makes on the negative control data after a Bonferroni correction at level 0.1; this number should be zero.

An inspection of the results reveals serious problems. On the Frangieh IFN-γ data, for example, the Weissman method is grossly inflated, making over 9,000 false Bonferroni discoveries. The Schraivogel method is similarly inflated on the Papalexi negative control data, rejecting nearly 2,000 pairs. MIMOSCA, meanwhile, exhibits pathological behavior in the bulk of the distribution on both datasets, outputting p-values strictly less than 0.26 across all hypotheses. We observe similar trends on the other datasets (Figures S1-S3). Overall, the best method appears to be Seurat DE. However, given the clear inflation of

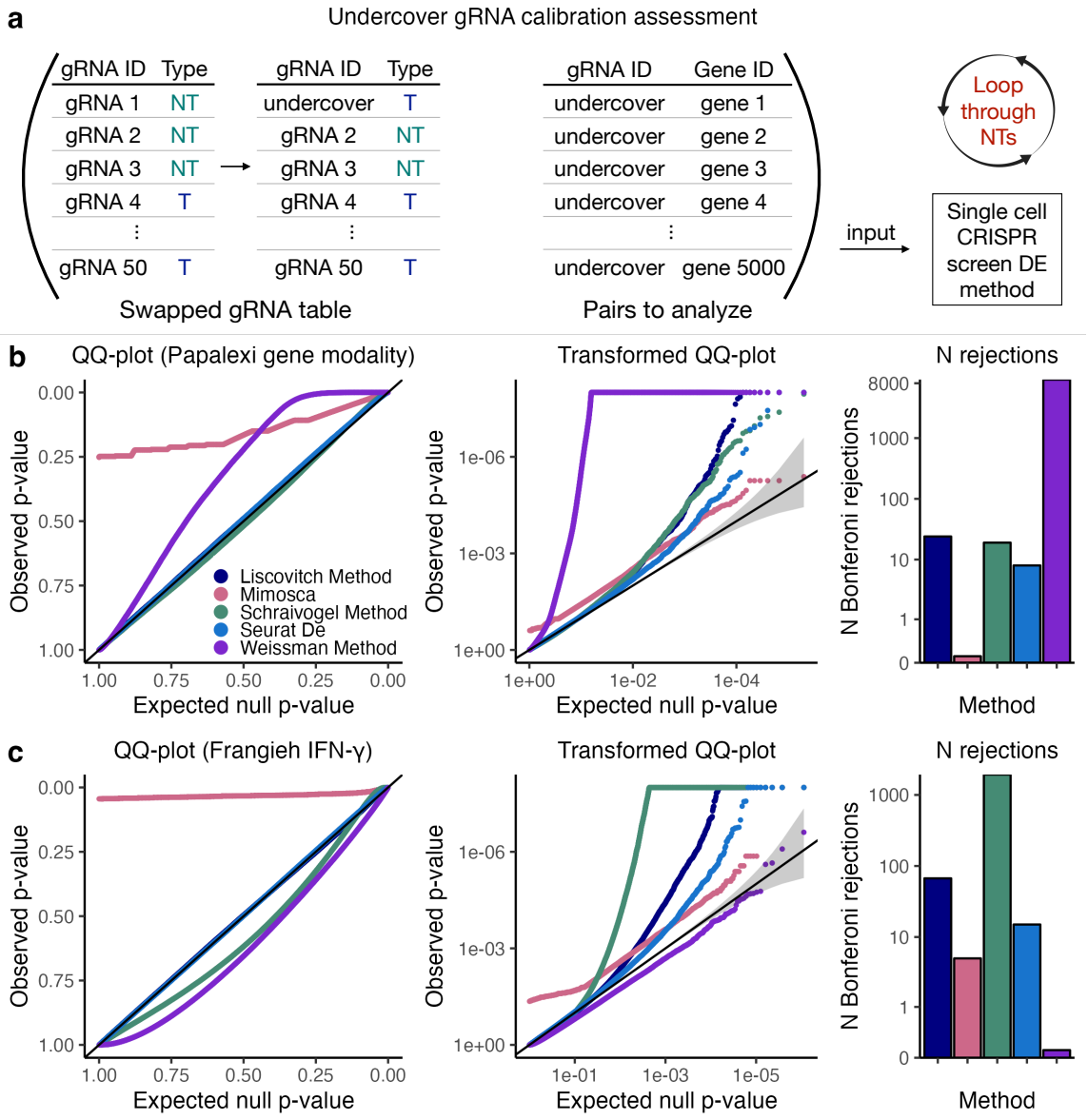

Figure 1: **Leading single-cell CRISPR screen differential expression methods exhibit pervasive miscalibration, rendering results unreliable**. **a**, A schematic of the "undercover gRNA procedure," a strategy for rigorously assessing the calibration of competing single-cell CRISPR screen DE methods. **b** (resp. **c**), Calibration of existing methods on the Papalexi (resp., Frangieh IFN-γ) negative control data. Left, a QQ-plot of the null p-values (colored by method) plotted on an untransformed scale; middle, the same QQ-plot

displayed on a negative log transformed scale; right, the number of Bonferroni rejections that each method makes on the negative control data. Gray region, 95% confidence band; NT, non-targeting; T, targeting.

Seurat DE in the tail, substantial room for improvement is available.

Core analysis challenges

We conduct an extensive empirical investigation of the data to search for possible sources of miscalibration. We uncover three core analysis challenges: sparsity, confounding, and model misspecification. No method that we examine addresses all (or even a majority) of these analysis challenges (Table S3), explaining their collective lack of calibration. First, single-cell CRISPR screen data typically are sparse: most genes are lowly expressed, rendering the "effective sample size" (or amount of information available for statistical inference) small. Methods that rely upon large-sample asymptotic approximations, such as Seurat DE (an implementation of the Mann-Whitney test), break down in sparse settings. We compare the exact null distribution of the Mann-Whitney test statistic (obtained via permutations) to the standard Gaussian distribution; the latter is used as a computationally tractable approximation to the former in large samples. The Gaussian distribution provides a reasonable approximation to the exact null distribution for some pairs (Figure 2a, left) but not others (2a, right). As the effective sample size decreases and the Gaussian approximation degrades, the accuracy of the p-value that one obtains using the Gaussian approximation likewise degrades, leading to miscalibration (Figures 2b-c).

Second, technical factors, such as biological replicate, batch, and sequencing depth, impact not only a cell's expression level, but also its probability of receiving a gRNA, thereby creating a confounding effect that, if not accounted for, leads to spurious associations[9] (Figure 2d, Figure S4). To assess the utility of adjusting for confounders, we apply negative binomial (NB) regression — both with and without confounders, such as biological replicate, included as covariates — to the Papalexi (gene modality) negative control data (Figure 2e). The variant of NB regression with covariates, though not perfectly calibrated, outperforms its counterpart without covariates, highlighting the importance of accounting for confounding. Finally, methods that rely upon parametric models for the gene expression distribution, such as NB regression and the Schraivogel method, yield miscalibrated p-values when those models are misspecified[12]. We test for goodness of fit of the NB regression model to the gene expression data across four datasets; the resulting goodness of fit p-values exhibit inflation (Figure 2f), suggesting poor fit of the NB regression model to some subset of the genes (plausibly due to inadequate estimation of the NB size parameter).

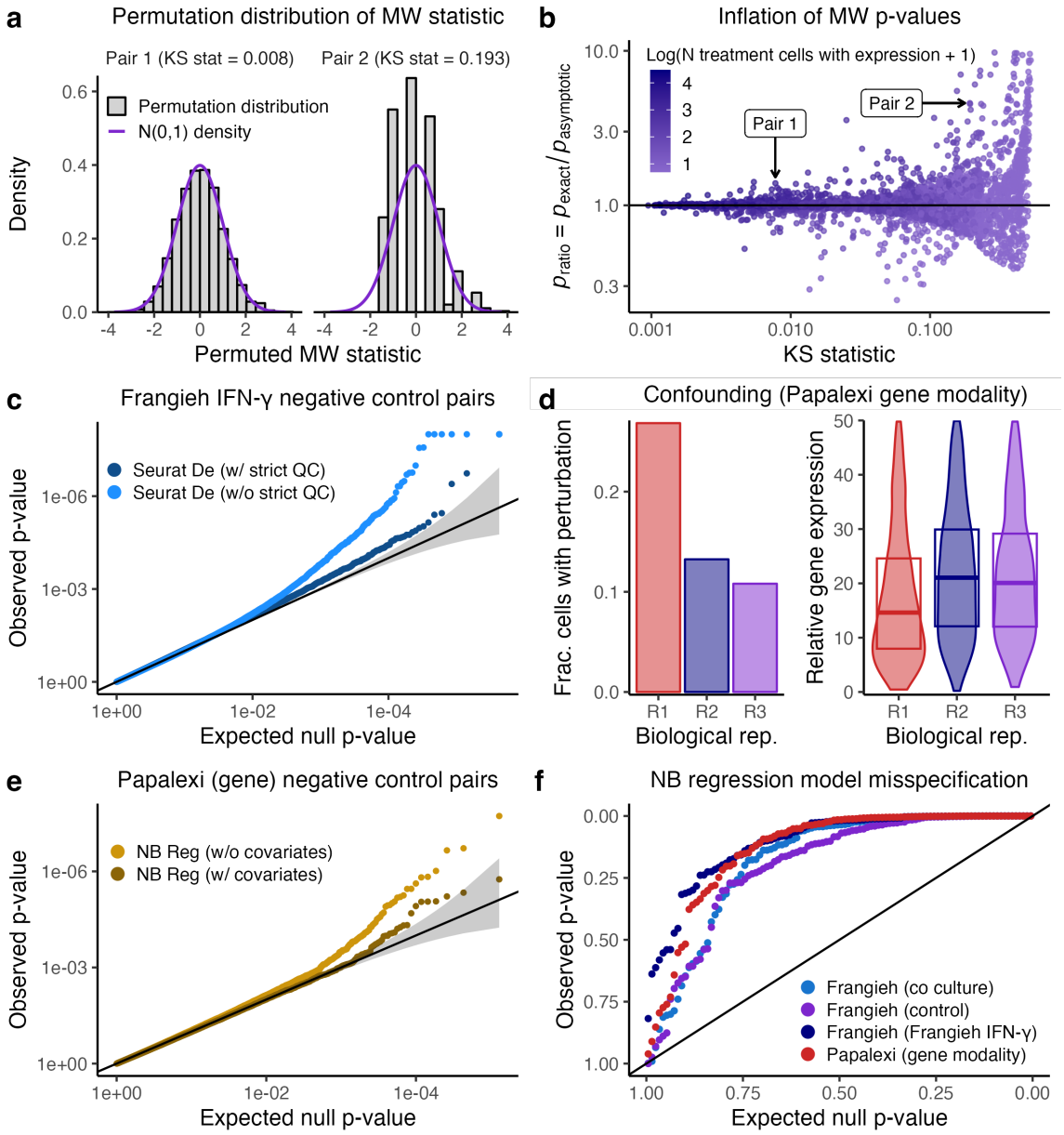

Figure 2: **Sparsity, confounding, and model misspecification are core analysis challenges in single-cell CRISPR screen analysis**. **a**, The exact null distribution of the Mann–Whitney (MW) test statistic (obtained via permutations; grey) on two pairs from the Frangieh IFN-γ data. The MW test (and thus Seurat DE) approximates the exact null distribution using a standard Gaussian density (purple). For pair 1 (left), the Gaussian

approximation to the exact null distribution is good (KS statistic = 0.008); for pair 2, by contrast (right), the approximation is poor (KS statistic = 0.193). **b**, A plot of p_{ratio} (defined as the ratio of the exact MW p-value, p_{exact}, to the asymptotic MW p-value, $p_{\text{asymptotic}}$) vs. goodness of fit of the Gaussian distribution to the exact null distribution (as quantified by the KS statistic). Each point represents a gene-gRNA pair; pairs 1 and 2 (from panel **a**) are annotated. As the KS statistic increases (indicating worse fit of the Gaussian distribution to the exact MW null distribution), p_{ratio} deviates more from one, indicating miscalibration. Points are colored according to the effective sample size (as quantified by the number of treatment cells with nonzero expression) of the corresponding pair. **c**, An application Seurat DE to the IFN-γ negative control data with and without stringent QC; applying stringent QC in this context amounts to filtering for pairs with with a very large effective sample size. **d**, An example of confounding on the Papalexi data. Left (resp. right), the fraction of cells that received a given NT gRNA (resp., the relative expression of a given gene) across biological replicates "R1," "R2," and "R3." If we failed to account for biological replicate, we would conclude (incorrectly) the the NT gRNA *decreases* the relative expression of the gene. **e**, Application of NB regression to the Papalexi data. Inclusion of confounders (such as biological replicate) in the regression model improves calibration (although further improvements are possible). **f**, A QQ-plot of p-values obtained from testing for goodness of fit of the NB regression model to the gene expression data (points colored by dataset). The p-values are inflated, indicating that the NB regression model provides a poor fit to some subset of the genes.

SCEPTRE: robust single-cell CRISPR screen DE testing

To resolve the analysis challenges of sparsity, confounding, and model misspecification, we introduce SCEPTRE, a method for robust single-cell CRISPR screen DE testing (Figure 3). SCEPTRE at its core is a permutation that uses a test statistic with very appealing statistical and computational properties. First, SCEPTRE regresses the gene expressions onto the technical factors via a negative binomial GLM, ignoring (in this initial step) the gRNA indicators. Second, SCEPTRE permutes the vector of gRNA indicators B times (where B is a large number, for example $25,000$). Third, SCEPTRE tests the original (i.e., unpermuted) gRNA indicator vector for inclusion in the fitted GLM via a score test[13,14], resulting in a z-score z_{orig}. SCEPTRE likewise carries out score tests for each of the permuted gRNA indicator vectors, yielding B "null" z-scores $z_1, \ldots, z_B$. Finally, SCEPTRE computes a permutation p-value by comparing the original z-score z_{orig} to the null z-scores.

SCEPTRE in effect layers a GLM on top of a permutation test, imbuing it with a special robustness property that enables it to address all three core analysis challenges (Table S3). A key observation is that the strength of confounding due to technical factors varies across

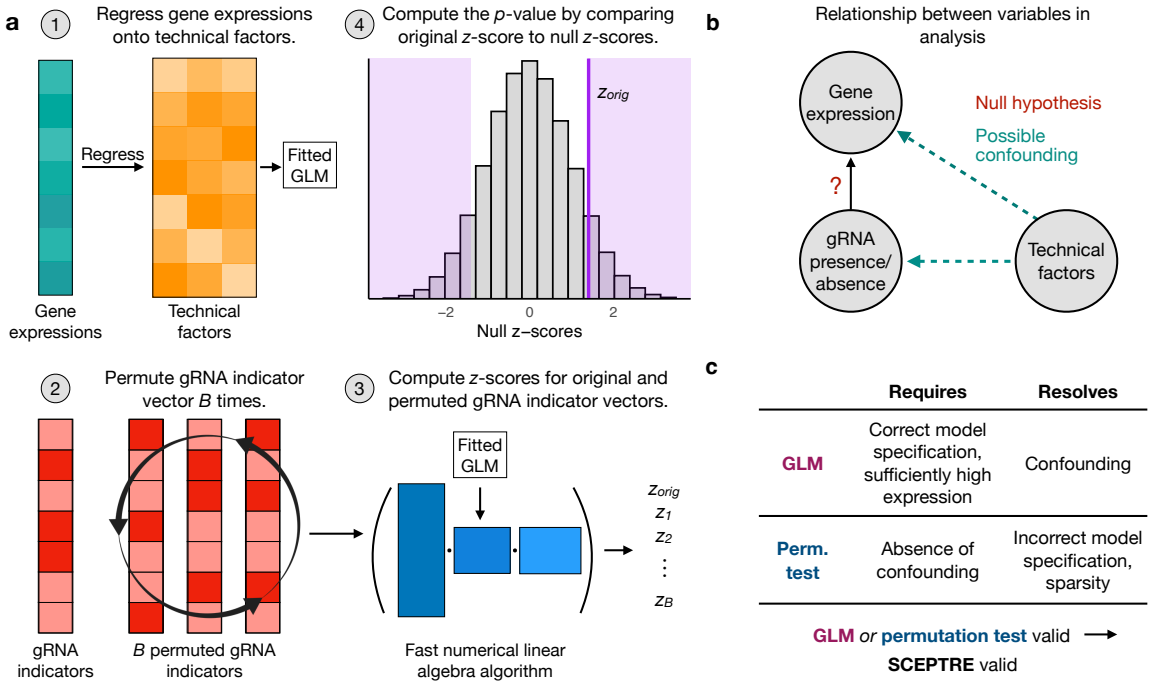

Figure 3: **SCEPTRE: robust single-cell CRISPR screen differential expression analysis**. **a**, A schematic of the SCEPTRE algorithm. SCEPTRE (i) fits a negative binomial GLM of gene expressions onto technical factors, (ii) permutes the gRNA indicator vector B times, (iii) computes the "original" z-score (resp., "null z-scores") by testing for inclusion of the original gRNA indicator vector (resp., permuted gRNA indicator vectors) in the fitted GLM via a score test, and (iv) calculates the final p-value by comparing the original z-score to the null z-scores. **b**, A directed acyclic graph (DAG) representing the variables in the analysis. The technical factors often (but not always) exert a confounding effect on the gRNA indicator and gene expressions. **c**, SCEPTRE in effect layers a GLM on top of a permutation test, imbuing it with a key robustness property: if either the requirements of the GLM or those of the permutation test are satisfied, then SCEPTRE is valid. This robustness property, which we call CAMP ("confounder adjustment via marginal permutations"), enables SCEPTRE to resolve the core analysis challenges.

gene-gRNA pairs: for some pairs, confounding is negligible; for other pairs, by contrast, confounding is considerable (Figure 3b). SCEPTRE exploits this variability in the strength of confounding, automatically leveraging permutation-based inference when the problem is negligibly confounded and GLM-based inference when the confounding is more potent. Formally, SCEPTRE produces a valid p-value if either (i) the gene-gRNA pair is negligibly confounded (in which case the permutation test powers inference, overcoming sparsity and model misspecification), or (ii) the negative binomial GLM is correctly specified and the effective sample size is sufficiently large (in which case the GLM powers inference, overcoming confounding; Figure 3c). In fact, for (ii) to hold, the negative binomial GLM only needs to be specified correctly up to the size parameter, sidestepping the notoriously difficult problem of NB size parameter estimation[15,16]. We refer to this statistical phenomenon as "confounder adjustment via marginal permutations" (or CAMP), and we demonstrate the existence and utility of CAMP a brief simulation study (Figure S5).

In addition to enjoying the above appealing statistical property, SCEPTRE is computationally efficient. We derive a fast numerical linear algebra algorithm to compute the set of original and null z-scores for a given gene-gRNA pair. (The algorithm is parameterized by quantities extracted from the fitted GLM; Figure 3a, step 3). Computing $B = 25{,}000$ z-scores on a gene-gRNA pair from the Papalexi data, for example, takes about 0.08 seconds. Further accelerations, including the use of an adaptive permutation testing scheme and the sharing of fitted GLMs across gene-gRNA pairs, are detailed in the Methods section.

SCEPTRE calibration and power results

Leveraging the undercover gRNA procedure, we apply SCEPTRE to analyze the negative control data and find that SCEPTRE is better calibrated than its main competitor, Seurat DE, across datasets (Figure 4; comparisons to other methods in Figures S1-S3). For example, on the negative control Papalexi (gene modality) data, SCEPTRE makes zero Bonferroni rejections and yields p-values that lie almost exclusively within the grey 95% confidence band; Seurat DE, by contrast, makes eight rejections and produces p-values that fall far outside this region. Importantly, SCEPTRE matches or exceeds the performance of both the permutation test and NB regression on the Papalexi and Frangieh data (Figure S6), highlighting the CAMP robustness property that SCEPTRE exploits (Figure 3c).

We additionally interrogate the power of the methods by applying them to analyze positive control data. We construct positive control pairs for each dataset by coupling gRNAs that target TSSs or known enhancers to the genes (or proteins) regulated by these elements. We plot the number of "highly significant" discoveries — operationally defined as rejections made at level $\alpha = 10^{-5}$ — made by each method on each dataset (Figure 5; larger

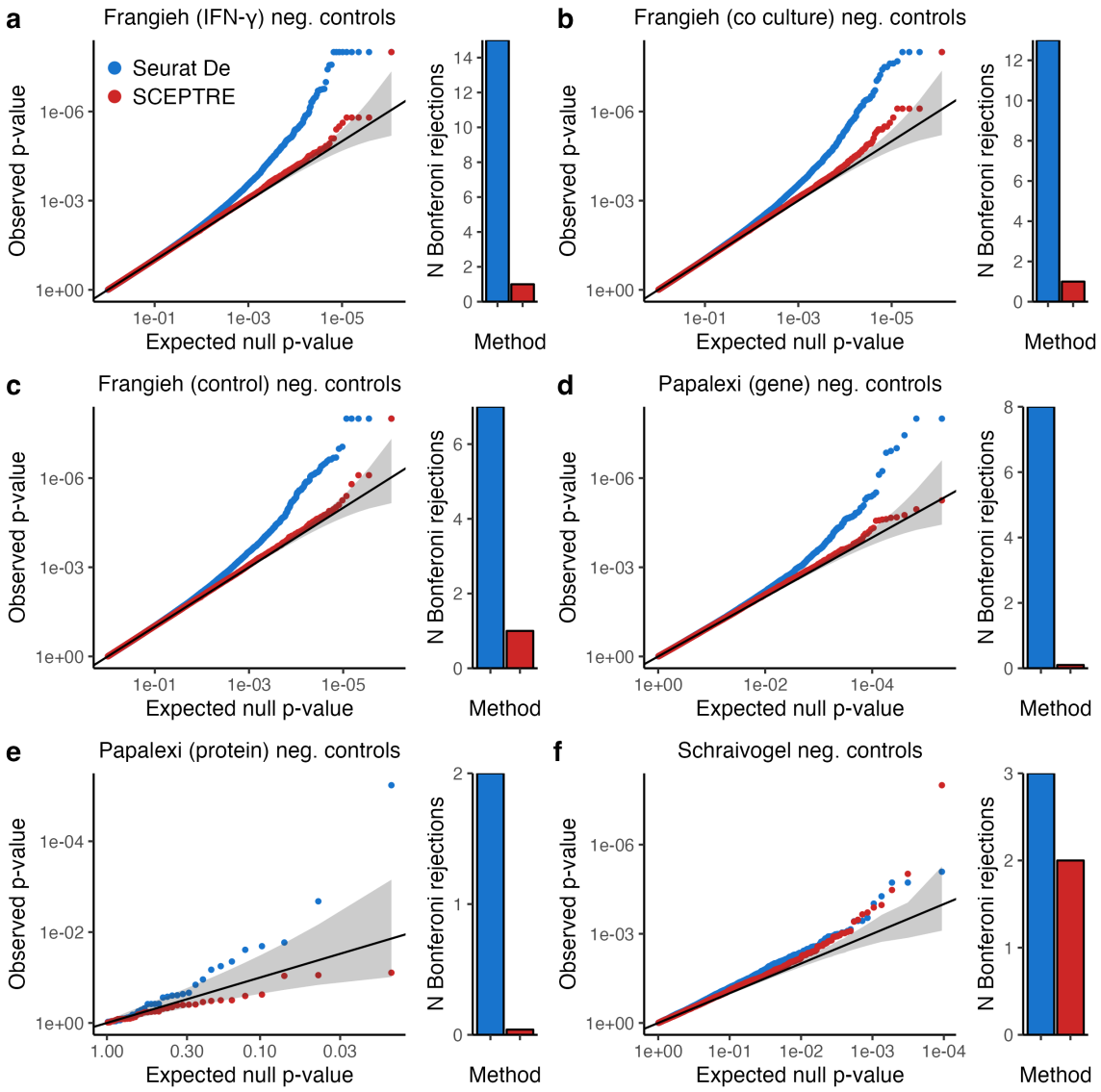

Figure 4: **SCEPTRE is better calibrated than Seurat DE, the current state-of-the-art, across all datasets**. **a**-**f**, Calibration results for SCEPTRE and Seurat DE across all six real datasets (treating the gene and protein modalities of the Papalexi data separately). In each panel the left display is a QQ-plot of the null p-values (colored by method) on a negative log transformed scale; the right display, meanwhile, is the number of Bonferroni rejections that either method makes on the negative control data.

values are better). Methods that exhibit extreme miscalibration on a given dataset (defined as > 50 Bonferroni rejections on the negative control pairs of that dataset) are excluded from the positive control analysis, as assessing the power of these methods is challenging. An inspection of the results reveals that SCEPTRE matches or outperforms the other methods with respect to power on nearly every dataset (Figure 5a-f). For example, SCEPTRE makes ten (or 13.5%) more discoveries than Seurat DE on the Frangieh control data while maintaining proper calibration. It is worth noting that MIMOSCA exhibits aberrantly low power, likely because it outputs p-values that are truncated.

Quality control

Carrying out QC at the level of the gene-gRNA pair (in addition to standard QC at the level of the gene, gRNA, and cell) helps filter out low-quality pairs in single-cell CRISPR screen analysis. The most important indicator of pair quality is effective sample size: pairs with an extremely small effective sample size cannot be rejected by a well-calibrated method and thus needlessly aggravate the multiple testing burden. A reasonable strategy to ameliorate this issue is to (i) determine the minimum effective sample size such that SCEPTRE rejects positive control pairs with this effective sample size and then (ii) to filter for pairs whose effective sample size is equal to or greater than this threshold. (A reasonable metric for "effective sample size" is the number of treatment cells with nonzero expression). Applying the above criterion to the data analyzed in this work yields a QC threshold of about seven treatment cells with nonzero expression per pair (Figure 5).

Discussion

Before concluding we briefly highlight relevant connections to other works. First, Zhou et al.[17] introduce guided sparse factor analysis (GSFA), a method that combines factor analysis and DE analysis to infer the effects of perturbations on individual genes and gene modules in single-cell CRISPR screens. In contrast to the methods studied in this work, GSFA is Bayesian, returning "posterior inclusion probabilities" instead of p-values for tests of association. Next, Wang[18] proposes Normalisr, a very general method for single-cell data analysis that includes a single-cell CRISPR screen DE module. Normalisr transforms the expression data in a special way, models the transformed data using a linear model, and then tests for association between a treatment (e.g., a CRISPR perturbation) and a gene using this linear model. We discuss Normalisr in greater detail in Section Existing methods details. Finally, several works in the neuroscience literature[19–21] have proposed coupling resampling-based tests (e.g., permutation tests) to regression models (e.g., GLMs) in various ways. To the best of our knowledge, SCEPTRE's use of a GLM score statistic within

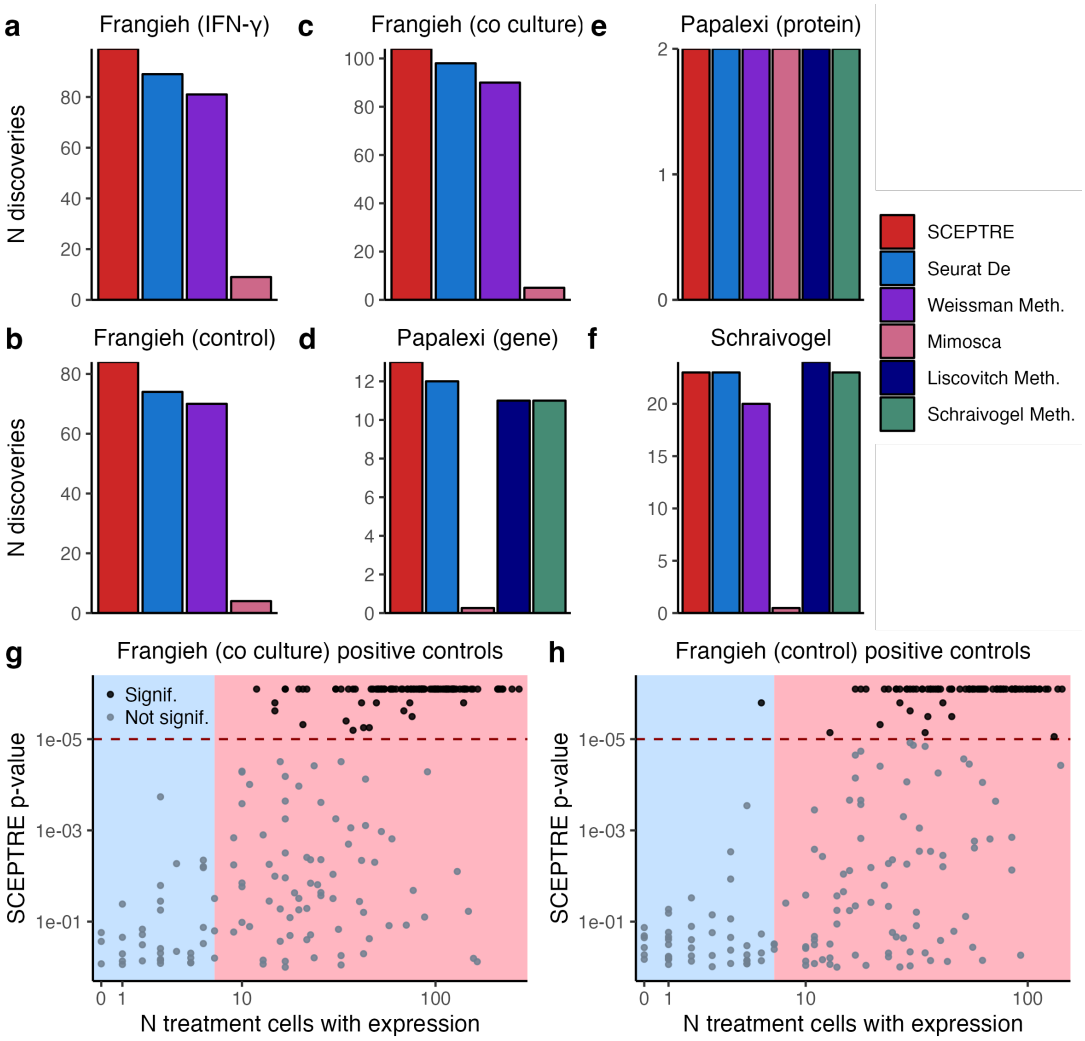

Figure 5: **SCEPTRE is more powerful than competing methods**. **a**-**f**, The number of rejections that each method makes on the positive control pairs of each dataset (at level 10^{-5}). Methods that demonstrate extreme miscalibration on a given dataset (defined as > 50 Bonferroni rejections on the corresponding negative control paris) are excluded. **g** (resp., **h**), SCEPTRE p-value versus effective sample size for each positive control pair on the Frangieh co-culture (resp., control) data. Significant pairs (black) are defined as those with p-values of less than 10^{-5} (horizontal dotted line). The pairwise QC threshold is seven treatment cells with nonzero expression; thus, pairs in the blue region would be filtered out, whereas those in the red region would be retained. (This criterion is used to filter all pairs, not just positive controls.)

permutation test — and the algorithm that SCEPTRE leverages for this purpose — are novel.

More broadly, SCEPTRE takes steps toward resolving a fundamental tension between parametric and nonparametric approaches to two-sample testing. Nonparametric approaches, such as the permutation test, make minimal assumptions but are unable to adjust for confounders; parametric approaches, conversely, adjust for confounders at the expense of making strong assumptions. SCEPTRE leverages a statistically principled and computationally efficient algorithm to couple GLMs to permutation tests, thereby inheriting the best properties of both approaches for two-sample testing (with respect to calibration). We anticipate that the statistical and algorithmic ideas that underly SCEPTRE could be applied more broadly to other differential expression testing applications in genetics and genomics.

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. Schraivogel, D. *et al.* Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nature methods* **17**, 629–635 (2020).

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

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

5. Papalexi, E. *et al.* Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics* **53**, 322–331 (2021).

6. Frangieh, C. J. *et al.* Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics* **53**, 332–341 (2021).

7. Liscovitch-Brauer, N. *et al.* Profiling the genetic determinants of chromatin accessibility with scalable single-cell crispr screens. *Nature biotechnology* **39**, 1270–1277 (2021).

8. Replogle, J. M. *et al.* Combinatorial single-cell crispr screens by direct guide rna capture and targeted sequencing. *Nature biotechnology* **38**, 954–961 (2020).

9. Barry, T., Wang, X., Morris, J. A., Roeder, K. & Katsevich, E. Sceptre improves calibration and sensitivity in single-cell crispr screen analysis. *Genome biology* **22**, 1–19 (2021).

10. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).

11. Finak, G. *et al.* Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology* **16**, 1–13 (2015).

12. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome biology* **23**, 1–13 (2022).

13. Dunn, P. K. & Smyth, G. K. *Generalized linear models with examples in R*, chap. 7, 286–287 (Springer, 2018).

14. Lehmann, E. & Romano, J. P. *Testing statistical hypotheses*, chap. 14, 685–689 (Springer, 2020), 4 edn.

15. Lause, J., Berens, P. & Kobak, D. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *Genome biology* **22**, 1–20 (2021).

16. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21 (2014).

17. Zhou, Y., Luo, K., Chen, M. & He, X. A novel bayesian factor analysis method improves detection of genes and biological processes affected by perturbations in single-cell crispr screening. *bioRxiv* (2022).

18. Wang, L. Single-cell normalization and association testing unifying crispr screen and gene co-expression analyses with normalisr. *Nature communications* **12**, 1–13 (2021).

19. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *Neuroimage* **92**, 381–397 (2014).

20. Eklund, A., Andersson, M. & Knutsson, H. Fast random permutation tests enable objective evaluation of methods for single-subject fmri analysis. *International journal of biomedical imaging* **2011** (2011).

21. Helwig, N. E. Robust nonparametric tests of general linear model coefficients: a comparison of permutation methods and test statistics. *NeuroImage* **201**, 116030 (2019).

22. McGinnis, C. S. *et al.* Multi-seq: sample multiplexing for single-cell rna sequencing using lipid-tagged indices. *Nature methods* **16**, 619–626 (2019).

23. 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).

24. Barry, T., Katsevich, E. & Roeder, K. Exponential family measurement error models for single-cell crispr screens. *arXiv preprint arXiv:2201.01879* (2022).

25. Ripley, B. *et al.* Package 'mass'. *Cran r* **538**, 113–120 (2013).

26. DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316–319 (2017).

27. Chung, E. & Romano, J. P. Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference* **168**, 97–105 (2016).

28. Eddelbuettel, D. & François, R. Rcpp: Seamless r and c++ integration. *Journal of statistical software* **40**, 1–18 (2011).

29. Zhang, M., Zou, J. & Tse, D. Adaptive monte carlo multiple testing via multi-armed bandits. In *International Conference on Machine Learning*, 7512–7522 (PMLR, 2019).

30. Besag, J. & Clifford, P. Sequential monte carlo p-values. *Biometrika* **78**, 301–304 (1991).

Methods

Dataset details

We download, process, and harmonize five single-cell CRISPR screen datasets (Table S1), inheriting several data-related analysis decisions made by the original authors. First, we use the gRNA-to-cell assignments that the original authors used, thereby circumventing the need to assign gRNAs to cells using gRNA UMI and/or read count matrices. Papalexi and Schraivogel employed a simple strategy for this purpose: Papalexi identified the gRNA with the greatest UMI count in a given cell and assigned that gRNA to the cell, while Schraivogel assigned gRNAs by thresholding gRNA UMI counts. Frangieh, meanwhile,

assigned gRNAs to cells via a more complex approach involving a separate dial-out PCR procedure. We find the gRNA-to-cell assignments adequate and thus use them without modification. Next, we inherit the cell-wise QC that the original authors implemented. For example, Papalexi removed likely duplets (as determined by the `Seurat` function `MULTIseqDemux`[10,22]) as well as cells with excessive mitochondrial content and low gene expression.

We generate a synthetic single-cell CRISPR screen dataset to use for benchmarking purposes (alongside the other datasets). The synthetic dataset contains 5,000 genes, 25 gRNAs, and 10,000 cells. We generate the matrix of gene expressions by sampling counts from a negative binomial distribution, allowing each gene to have its own mean and size parameter. (We draw gene-wise means and sizes i.i.d. from a Gamma$(0.5, 2)$ distribution and a Unif$(1, 25)$ distribution, respectively.) We randomly insert gRNAs into cells such that the expected number of cells per gRNA is equal across gRNAs. The dataset is entirely devoid of signal and confounding: no gRNA affects the expression of any gene, and there do not exist technical factors that impact the gRNA assignments or gene expressions.

We apply our own minimal gene-wise, gRNA-wise, and cell-wise QC uniformly to the datasets. We filter for genes expressed in at least 0.005 of cells, gRNAs expressed in at least 10 cells, and cells with exactly one gRNA, respectively. Table S2 summarizes the statistical attributes (e.g., number of genes, number of cells, etc.) of each dataset. Finally, we obtain the set of cell-specific covariates (or technical factors) that each dataset possesses, which we list below. Frangieh Co-culture, control, and IFN-γ datasets: number of gene UMIs, number of genes expressed, cell cycle phase, and sequencing batch; Papalexi (gene modality): number of gene UMIs, number of genes expressed, biological replicate, cell cycle phase, and percent of gene transcripts that map to mitochondrial genes; Papalexi (protein modality): number of protein UMIs, cell cycle phase, biological replicate, and percent of gene transcripts that map to mitochondrial genes; Schraivogel: number of gene UMIs, number of genes expressed, and sequencing batch.

Existing methods details

The methods we examine in this work — Seurat DE, Liscovitch Method, Schraivogel Method, Weissman Method, and MIMOSCA — have software implementations that differ from one another considerably. In fact, one method — Liscovitch Method — has no software implementation at all, while two others — Schraivogel Method and Weissman Method — are written in Python rather than R. Furthermore, no two methods have similar APIs (application programming interfaces). Thus, to facilitate comprehensive benchmarking of the methods, we implement all methods ourselves in an R package (`lowmoi`[1]). To

[1] github.com/Katsevich-Lab/lowmoi

this end, we carefully read the Methods section of the paper in which each method was proposed, copy and paste relevant source code from the associated Github repository (filling in any missing pieces on our own), communicate with the original authors to ensure accurate implementation of the method (when necessary), and then test for agreement between the output of our implementation and the original implementation (when possible). To ensure consistency of the API across methods, we implement the methods in such a way that each takes the same inputs — namely, (i) the gene expression matrix, (ii) the gRNA-to-cell assignment matrix, (iii) the mapping of each gRNA to its type (targeting versus NT) and target site (if applicable), and (iv) the set of gene-gRNA pairs to analyze — and returns the same output — namely, a set of p-values for each test of association conducted.

Although negative binomial (NB) regression has not yet (to our knowledge) been used to analyze low-MOI single-cell CRISPR screen data, we additionally implement NB regression in the `lowmoi` package. NB regression is a staple tool in single-cell data science[23] and has been used to analyze high-MOI single-cell CRISPR screens[3,24]; as such, NB regression serves as a useful point of comparison to the other methods. We implement NB regression as follows. First, we regress the gene expression vector onto the matrix of technical factors via the `MASS`[25] package's implementation of NB regression. (This step involves estimation of the NB size parameter θ.) We then test for inclusion of the vector of gRNA indicators in the fitted model using a GLM score test, implemented in the `statmod`[13] package. The score test yields a p-value for the test of association between the gRNA indicator and the expression of the gene (controlling for the technical factors). Regressing the gene expressions onto the technical factors *and* the gRNA indicator and then computing a Wald p-value for the gRNA indicator is asymptotically equivalent to the score test. Although the Wald test is more popular, we use the score test to make our implementation of NB regression comparable to SCEPTRE; the latter method uses a permutation test built upon an NB regression score test statistic.

We do not apply the methods directly "out of the box" and instead adjust them in two small ways to enable their more meaningful application to the data. First, some methods have an internal QC step in which gene-gRNA pairs that are unpromising or low-quality (as determined by the method itslef) are removed. For example, Seurat DE by default filters out gene-gRNA pairs for which the log-fold change of the expression of the gene (across the treatment and control cells) falls below a certain threshold. We disable such method-specific pairwise QC, allowing us to apply competing methods to the exact same set of gene-gRNA pairs on each dataset, greatly facilitating head-to-head comparisons across methods. Second, a few methods — namely MIMOSCA, NB regression, and SCEPTRE — provide the option of adjusting for cell-specific technical factors (e.g., batch, biological replicate, etc.). When using one of these methods to analyze a given dataset, we

deploy the method such that it adjusts for the cell-specific technical factors of that dataset (as listed in the Section Dataset details).

We note that the methods analyzed in this work vary with respect to their "control group," or the set of cells used as the comparison group in the differential expression test (Table 1). For a given targeting gRNA, we call the cells that received the gRNA the "treatment cells." Some single-cell CRISPR screen DE methods carry out the test of association by comparing the treatment cells to the set of cells that received a non-targeting gRNA (i.e., the NT cells), while others compare the treatment cells to the set of *all other cells*, including those that received another targeting gRNA (i.e., the "compliment set"). Every method that we examine but one — MIMOSCA — falls into the former category.

We note that we do not prioritize Normalisr[18] for benchmarking for the following reasons. Normalisr is a very general method with modules for conventional single-cell DE, low MOI CRISPR screen DE, gene co-expression analysis, and high MOI CRISPR screen DE. Normalisr appears to be primarily geared toward the latter two tasks. Indeed, we locate code examples of gene co-expression analysis and high MOI CRISPR screen DE in the Normalisr Github repository, but we are unable to find an example of low MOI CRISPR screen DE. Given the complexity of Normalisr relative to the other methods (i.e., those listed in Table 1), and given that the other methods currently are more widely used in the primary analysis of single-cell CRISPR screen data, we decide to prioritize the other methods for benchmarking.

Details of the undercover gRNA procedure, negative control benchmarking results

An important attribute of the undercover gRNA procedure is that it is highly flexible. In particular, the procedure applies both to methods that take as their control group the NT cells (e.g., Seurat DE, Schraivogel Method) and those that take as their control group the compliment set (e.g., MIMOSCA; Table 1). In the former case the "undercover" cells (i.e., the cells that received an "undercover" gRNA) are compared to the other NT cells; in the latter case, by contrast, the "undercover" cells are compared to all other cells. Although not our focus in this work, high-MOI single-cell CRISPR screen data also are compatible with the undercover gRNA procedure. We design a Nextflow[26] pipeline to deploy the undercover gRNA procedure at scale on all dataset-method pairs[2]. We filter for gene-gRNA pairs with at least seven treatment cells and seven control cells with nonzero gene (or protein) expression (as explained in Section Quality control). Figures 1 and S1-S3 present the results; Table S2 presents the number of negative control pairs in each dataset.

[2] `github.com/Katsevich-Lab/undercover-grna-pipeline`

Details of the investigation into the core analysis challenges

We explicate in greater detail our empirical investigations into the core analysis challenges of sparsity, confounding, and model misspecification (as described in Section Core analysis challenges).

Sparsity. To explore the impact of sparsity on calibration, we deploy the Mann-Whitney (MW) test to a randomly-selected subset of 5,400 negative control gene-gRNA pairs from the Frangieh IFN-γ data. (The pairs are selected so that each has at least one treatment cell with nonzero gene expression.) Following Seurat DE, we deploy the MW test as follows: first, we normalize the gene expressions by dividing the raw counts by the cell-specific library sizes; then, we apply the MW test (as implemented by the `wilcox.test` function from the `stats` package in R) to the normalized data, comparing the treatment cells to the control cells. We compute the MW p-value in two ways. First, we calculate the asymptotic p-value $p_{\text{asymptotic}}$ by comparing the MW test statistic to the standard Gaussian distribution. This approach implicitly assumes that the number of cells with nonzero expression (across both groups) is large enough for the null distribution of the MW test statistic to be approximately Gaussian. Next, we calculate the exact p-value p_{exact} by (i) computing the MW statistic on the original data; (ii) permuting the gRNA indicator vector $B = 200,000$ times (while holding fixed the vector of normalized gene expressions), resulting in B permuted datasets; (iii) computing the MW test statistic on each of these B permuted datasets, yielding a permutation (or "null") distribution of MW statistics; and then (iv) calculating the p-value p_{exact} by comparing the original MW statistic to the null MW statistics[27]. The latter approach, though computationally expensive (due to the slowness of computing the MW statistic), yields a much more accurate p-value than the asymptotic approach for lowly expressed genes. Seurat DE returns the asymptotic p-value $p_{\text{asymptotic}}$ instead of the exact p-value p_{exact} in virtually all cases.[3]

To study the impact making the above approximation, we plot the asymptotic null distribution of the MW statistic (i.e., the standard Gaussian distribution) superimposed on top of the exact null distribution of the MW statistic (i.e., the permutation distribution) for two pairs from the Frangieh IFN-γ negative control data (Figure 2a). The asymptotic and exact distributions must be highly similar for the asymptotic p-value $p_{\text{asymptotic}}$ to be reliably accurate. We meaure goodness of fit of the Gaussian distribution to the exact null distribution by calculating the Kolmogorov–Smirnov (KS) statistic; this statistic ranges from zero to one, with smaller values indicating better fit of the Gaussian distribution to the exact null distribution. We report the KS statistic for both example pairs in the panels

[3]The `wilcox.test` function on which Seurat DE relies returns p_{exact} only if (i) there are fewer than 50 cells across both treatment and control groups and (ii) no two cells (in either the treatment or the control group) have the same normalized expression level. This condition is expected to hold rarely, if ever.

of the plot.

Next, we calculate p_{ratio}, defined as the ratio of the exact p-value p_{exact} to the asymptotic p-value $p_{\text{asymptotic}}$, for each of the the 5,400 negative control pairs sampled from the Frangieh IFN-γ data. A p_{ratio} value of one indicates that the asymptotic and exact p-values coincide; a p_{ratio} value of greater than one (resp., less than one), on the other hand, indicates inflation (resp., deflation) of the asymptotic p-value relative to the exact p-value. We seek to explore visually how a small effective sample size can lead to degradation of the Gaussian approximation, thereby resulting in p-value miscalibration (as reflected by p_{ratio} values that deviate from one). To this end, we plot p_{ratio} versus goodness of fit of the the Gaussian distribution to the exact null distribution (as quantified by the KS statistic) for each pair (Figure 2b). We color the points according their effective sample size (defined as the number of treatment cells with nonzero expression). Pairs 1 and 2 from Figure 2a are annotated in Figure 2b.

Finally, to directly assess the impact of sparsity on calibration, we apply Seurat DE to the IFN-γ negative control data, partitioning the pairs into two categories: those with an effective sample size less than 30, and those with an effective sample size greater than or equal to 30. We create of QQ-plot of the negative control p-values, coloring the points by category (effective sample size < 30, light blue; effective sample size ≥ 30, dark blue). Pairs in the former category outnumber those in the latter category; thus, to facilitate comparisons across the two groups, we downsample the p-values so that the number of p-values is equal across categories. We note that filtering for pairs with an effective sample size of 30 or greater leads to the loss of many interesting pairs (Section Quality control) and is therefore not a viable strategy for resolving the analysis challenge of sparsity.

Confounding. We explore how the variable of biological replicate confounds the Papalexi (gene modality) data. The Papalexi data were generated and sequenced across three independent experimental replicates (which we label "R1," "R2," and "R3")[4]. We visually examine the relationship between biological replicate and a given NT gRNA ("NTg4") and a given gene (*FTH1*). We plot the fraction of cells in each biological replicate that received the NT gRNA (Figure 2d, left); additionally, we create a violin plot of the relative expression of the gene across biological replicate. (The relative expression r_i of the gene in cell i is defined as $r_i = 1000 \cdot \log\left(u_i/l_i + 1\right)$, where u_i is the UMI count of the gene in cell i, and l_i is the library size of cell i. The violin plots are truncated at a relative expression level of 50). We superimpose boxplots indicating the 25th, 50th, and 75th percentiles of the empirical relative expression distributions on top of the violin plots (Figure 2d, right).

Observing clear visual evidence that biological replicate impacts both NTg4 and *FTH1*,

[4]The original data contained a fourth biological replicate as well, but this replicate was removed by the original authors, as it was deemed to be low quality.

we extend the above analysis to investigate the entire set of NT gRNAs and genes. First, we test for association between each NT gRNA and biological replicate. To this end, we construct a contingency table of gRNA presences and absences across biological replicate, testing for significance of the contingency table using a using a Fisher exact test (as implemented in the R function `fisher.test`). Next, we test for association between the relative expression of each gene and biological replicate. To do so, we fit two NB regression models to each gene; the first contains only library size as a covariate, while the second contains both library size *and* biological replicate as covariates. We compare these two models via a likelihood ratio test, yielding a p-value for the test of association between relative gene expression and biological replicate. Finally, we create QQ-plots of the resulting p-values (Figure S4; gRNA p-values, left; gene p-values, right). An inflation of the p-values across modalities suggests that the bulk of gene-NT gRNA pairs is confounded by biological replicate.

Finally, we directly assess the impact of adjusting for biological replicate (alongside other potential confounders) by applying two variants of NB regression to the Papalexi (gene modality) negative control data: (i) NB regression with library size (only) included as a covariate, and (ii) NB regression with library size as well as all potential confounders included as covariates. We plot the negative control p-values on a QQ-plot (Figure 2e); superior calibration of the latter method suggests that adjusting for confounders is useful. To neutralize the effect of sparsity (i.e., the first analysis challenge), we restrict our attention in this plot to gene-gRNA pairs with an effective sample size greater than 30.

Model misspecification. To illuminate the analysis challenge of parametric model specification, we evaluate goodness of fit of the NB regression model to the Frangieh and Papalexi gene expression data. (The Schraivogel data and the protein modality of the Papalexi data contain too few features to interrogate this question.) We restrict our attention to the NT cells (i.e., the cells that received an NT gRNA), circumventing the need to account for the effect of targeting gRNAs on expression. Unfortunately, using the same set of cells to estimate the NB size parameter θ and assess goodness of fit of the NB regression model is not statistically valid. Thus, we randomly split the NT cells into two groups. On the first group of NT cells, we regress the gene expressions onto the matrix of technical factors, yielding gene-wise estimates for the NB size parameter θ (alongside gene-wise estimates for the model coefficients, which we do not use). Next, on the second set of cells, we again regress the gene expressions onto the matrix of technical factors, this time treating the NB size parameter (as estimated in the first step) as fixed and known. We test the resulting fitted NB regression models for goodness of fit to the gene expression data (via a GLM deviance test), yielding gene-wise goodness-of-fit p-values. We display the goodness-of-fit p-values on a QQ-plot, coloring each p-value according to the dataset on which it was computed (Figure 2f). Inflation of the p-values indicates misspecification

of the NB regression model on some subset of the genes. We plot only p-values derived from highly expressed genes (operationally defined as genes with a UMI count of three or greater in at least 95% of cells), as the goodness-of-fit test can produce misleading p-values for lowly expressed genes. Additionally, we downsample the p-values so that the number of p-values per dataset is equal across datasets.

SCEPTRE overview

Consider a given gene and a given targeting gRNA. We call the cells that receive the targeting gRNA the "treatment cells" and those that receive an NT gRNA the "control cells." (We ignore the cells that receive another targeting gRNA.) Suppose there are n cells across treatment and control groups. Let $Y = [Y_1, \ldots, Y_n]^T$ be the vector of raw gene (or protein) expressions, and let $X = [X_1, \ldots, X_n]^T$ be the vector of gRNA indicators, where an entry of one (resp., zero) indicates presence of the targeting (resp. NT) gRNA. Finally, for cell $i \in \{1, \ldots, n\}$, let Z_i be the p-dimensional vector of technical factors for cell i (containing library size, batch, etc.). We include an entry of one in each Z_i to serve as an intercept term. Let Z be the $n \times p$ matrix formed by concatenating the Z_is, and let $[X, Z]$ be the $n \times (p + 1)$ matrix formed by concatenating X and Z.

We model Y_i as a function of X_i and Z_i via an NB generalized linear model (GLM):

$$Y_i \sim \text{NB}_\theta(\mu_i); \quad \log(\mu_i) = \gamma X_i + \beta^T Z_i, \tag{1}$$

where $\text{NB}_\theta(\mu_i)$ denotes a negative binomial distribution with mean μ_i and size parameter θ, and $\gamma \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are unkown constants. (In fact, SCEPTRE is compatible with arbitrary GLMs, including Poisson GLMs, which may be more appropriate for highly sparse data.) To convey intuition for SCEPTRE, we first describe a slightly modified version SCEPTRE that we call "SCEPTRE-Wald." SCEPTRE-Wald is a permutation test that uses a GLM-based Wald z-score as its test statistic. SCEPTRE-Wald is computationally infeasible to apply in practice but is useful for illustration purposes (Algorithm 1).

SCEPTRE-Wald posesses the key robustness property of "CAMP" (confounder adjustment via marginal permutations), which we state in a slightly more formal way here. If at least one of the following conditions holds, then the left-, right-, and two-tailed p-values are valid: (i) the gene-gRNA pair is unconfounded (i.e., the vector of technical factors Z_i contains all possible confounders, and Z_i is independent of at least one of Y_i and X_i); (ii) the NB GLM (1) is correctly specified up to the size parameter θ, and the effective sample size is sufficiently large (Figure 2, Figure S5). CAMP imbues SCEPTRE-Wald with two considerable advantages relative the vanilla NB GLM. First, SCEPTRE-Wald always yields valid inference when confounding is negligible, even if the NB model is

Algorithm 1: SCEPTRE-Wald algorithm.

1. Regress Y onto the matrix $[X, Z]$ by fitting the GLM (1). Compute a z-score z_{orig} for a test of the null hypothesis $H_0 : \gamma = 0$ via a Wald test.
2. Permute the X vector B (e.g., $B = 25,000$) times, resulting in permuted vectors $\tilde{X}_1, \ldots, \tilde{X}_B$.
3. For each $i \in \{1, \ldots, B\}$, repeat step 1, substituting the matrix $[\tilde{X}_i, Z]$ for $[X, Z]$. Label the resulting z-scores $z_1, \ldots, z_B$.
4. Compute a left-tailed (p_{left}), right-tailed (p_{right}), or two-tailed (p_{both}) p-value as follows:

$$\begin{cases} p_{\text{right}} = 1/(B+1) \sum_{i=1}^{B} \mathbb{I}(z_{\text{orig}} \geq z_i) \\ p_{\text{left}} = 1/(B+1) \sum_{i=1}^{B} \mathbb{I}(z_{\text{orig}} \leq z_i) \\ p_{\text{both}} = 2 \cdot \min\left\{p_{\text{right}}, p_{\text{left}}\right\}. \end{cases}$$

misspecificed or the effective sample size is small. Second, when confounding is non-negligible, SCEPTRE-Wald yields valid inference if the NB GLM is correctly specified up to the size parameter (and the effective sample size is sufficiently large), sidestepping the notoriously difficult problem of NB size parameter estimation[15,16]. These two improvements enable SCEPTRE-Wald to address all core single-cell CRISPR screen analysis challenges (namely, sparsity, confounding, and model misspecification; Table S3). However, SCEPTRE-Wald is not practical to deploy on real data, as it requires fitting $B + 1$ GLMs per gene-gRNA pair.

SCEPTRE Accelerations

To overcome the computational infeasibility of SCEPTRE-Wald, we introduce SCEPTRE. SCEPTRE possesses the exact same statistical attributes as SCEPTRE-Wald, including the critical robustness property of CAMP (described above). However, SCEPTRE is orders of magnitude faster than SCEPTRE-Wald: the former typically takes a fraction of a second to compute a p-value for a gene-gRNA pair, while the latter can take minutes or longer. We obtain this substantial speedup by implementing several accelerations, described below.

Acceleration 1: Score test. First, we replace the Wald test statistic of SCEPTRE-Wald (steps 1 and 3, Algorithm 1) with a score test statistic, enabling us to compute the z-scores for X and $\tilde{X}_1, \ldots, \tilde{X}_B$ using a single fitted GLM. Consider the following simplified NB GLM in which the gene expression Y_i is modelled as a function of the technical factor vector Z_i only:

$$Y_i \sim NB_\theta(\mu_i); \quad \log(\mu_i) = \beta^T Z_i. \tag{2}$$

Regressing Y onto Z by fitting the GLM (2) produces estimates $\hat{\beta}$ and $\hat{\theta}$ of the coefficient vector β and the size parameter θ, respectively, under the null hypothesis of no relationship between the gRNA indicator and the gene expression. We fit this GLM via maximum likelihood estimation (as implemented by the `MASS` package[25]), but other estimation procedures, including penalized maximum likelihood estimation, also are valid[14]. Denote the ith fitted mean of the model by $\hat{\mu}_i = \exp(\hat{\beta}^T Z_i)$, and let $\hat{\mu} = [\hat{\mu}_1, \ldots, \hat{\mu}_n]^T$ be the vector of fitted means. We can test the gRNA indicator vector X for inclusion in the fitted model by computing a score statistic z_{score}, as follows:

$$z_{\text{score}} = \frac{X^T W M (Y - \hat{\mu})}{X^T W X - X^T W Z (Z^T W Z)^{-1} Z^T W X}. \tag{3}$$

Here, W and M are a matrix and vector, respectively, that depend on the fitted means $\hat{\mu}$, gene expressions Y, and estimated size $\hat{\theta}$:

$$W = \text{diag}\left\{ \frac{\hat{\mu}_1}{1 + \hat{\mu}_1/\hat{\theta}}, \ldots, \frac{\hat{\mu}_n}{1 + \hat{\mu}_n/\hat{\theta}} \right\}; \quad M(Y - \hat{\mu}) = \left[\frac{Y_1}{\hat{\mu}_1} - 1, \ldots, \frac{Y_n}{\hat{\mu}_n} - 1 \right]^T.$$

(We display $M(Y - \hat{\mu})$ instead of M for convenience.) The score statistic (3) is asymptotically equivalent to the Wald statistic that one obtains by testing $H_0 : \gamma = 0$ in the full model (1). However, unlike the Wald statistic, the score statistic only depends on a fit of the model under the null hypothesis. SCEPTRE (Algorithm 2) exploits this useful property of the score statistic to accelerate SCEPTRE-Wald, reducing the number of GLMs that must be fit from $B + 1$ to one.

Algorithm 2: SCEPTRE algorithm.

1. Regress Y onto the matrix Z by fitting the GLM (2).
2. Compute the score statistic for X using the formula (3), yielding z_{orig}.
3. Permute the X vector B times, generating $\tilde{X}_1, \ldots, \tilde{X}_B$.
4. For each $i \in \{1, \ldots, B\}$, repeat step 2, substituting the vector $\tilde{X}_i$ for X. Label the resulting z-scores $z_1, \ldots, z_B$.
5. Compute a p-value using the formula from step 4 of Algorithm 1.

Acceleration 2: A fast score test for binary treatments. Calculating the score statistic (3) is not trivial. The quadratic form $X^T W Z (Z^T W Z)^{-1} Z^T W X$ in the denominator of (3) is hard to compute, as the matrix $W Z (Z^T W Z)^{-1} Z^T W$ is a large, dense matrix. The standard solution, employed for example by the package `statmod`[13], is to algebraically manipulate the score statistic so that it can be evaluated via a QR decomposition. However, the QR decomposition approach fails to leverage the structure in X when X is binary

and sparse (as is the case in single-cell CRISPR screen analysis). We therefore introduce an alternate strategy for computing the score statistic that instead is based on a spectral decomposition; the proposed strategy is hundreds of times faster than the standard QR decomposition approach in the single-cell CRISPR screen setting.

First, observe that $Z^T W Z$ is a symmetric matrix. Thus, $Z^T W Z$ can be spectrally decomposed as $Z^T W Z = U^T \Lambda U$, where U is an orthonormal matrix and Λ is a diagonal matrix of eigenvalues. Exploiting this decomposition, we can express the quadratic form in the denominator of (3) as follows:

$$X^T W Z (Z^T W Z)^{-1} Z^T W X = X^T W Z U \Lambda^{-1/2} \Lambda^{-1/2} U^T Z^T W X = L^T L = ||L||^2,$$

where $L = \Lambda^{-1/2} U^T Z^T W X$ is a p-dimensional vector. Evaluating the above expression reduces to computing the vector L and then summing over the squared entries of L, which is fast and easy. This insight motivates Algorithm 3, which computes the score statistics for $X, \tilde{X}_1, \ldots, \tilde{X}_B$ via a spectral decomposition.[5] The inner product and matrix-vector multiplication operations of step 3 are extremely fast because X_{curr} is sparse and binary. Furthermore, we program step 3 in C++ (via Rcpp[28]) for maximum speed.

Algorithm 3: Computing the GLM score statistics for $X, \tilde{X}_1, \ldots, \tilde{X}_B$ via spectral decomposition. Below, w is the n-dimensional vector constructed from the diagonal entries of W.

---

1. Spectrally decompose the matrix $Z^T W Z$, yielding diagonal matrix of eignvalues Λ and an orthonormal matrix U.
2. Compute the matrix $B = \Lambda^{-1/2} U^T Z^T W$ and the vector $a = WM(Y - \hat{\mu})$.

for $X_{\text{curr}} \in \{X, \tilde{X}_1, \ldots, \tilde{X}_B\}$ **do**

3. Compute

$$\begin{cases} \texttt{top} = a^T X_{\text{curr}} \\ \texttt{bottom\_right} = B X_{\text{curr}} \\ \texttt{bottom\_left} = w^T X_{\text{curr}}. \end{cases}$$

4. Compute $z = \texttt{top} / (\texttt{bottom\_left} + ||\texttt{bottom\_right}||^2)$

end

Acceleration 3: Sharing information across genes and gRNAs. The SCEPTRE algorithm introduced above (Algorithm 2) involves fitting one GLM per gene-gRNA pair.

[5] A Cholesky decomposition of $Z^T W Z$ could be used in place of the spectral decomposition, but the spectral decomposition is slightly more general, as it applies to matrices with eigenvalues equal to zero, which can occur (for example) when columns of Z are highly correlated.

We devise a strategy to share GLMs across gene-gRNA pairs containing the same gene, reducing the number of GLMs that must be fit per dataset to one per gene. Let n_{total} denote the total number of cells in the dataset. Index the cells by $\{1, \ldots, n_C, n_C + 1, \ldots, n_{\text{total}}\}$, where the first n_{C} cells are the negative control cells. Let Y denote the expression vector of a given gene. We proceed as follows. First, we fit the reduced GLM (2) using the control cells only. In other words, letting $Y_{[n_{\text{C}}]}$ and $Z_{[n_{\text{C}}]}$ denote the first n_{C} entries of Y and Z, respectively, we regress $Y_{[n_{\text{C}}]}$ onto $Z_{[n_{\text{C}}]}$, producing an estimate $\hat{\beta}_{\text{C}}$ of β. Next, we impute the mean expression of each cell (under the null hypothesis of $\gamma = 0$) by setting $\hat{\mu}_i = \exp(Z_i^T \hat{\beta}_{\text{C}})$ for all $i \in \{1, \ldots, n_{\text{total}}\}$. Using these imputed means, we can compute the score statistic (3) to test for association between the gene and any given targeting gRNA. Next, we seek to reduce the compute associated with permuting the gRNA indicator vector B times, which can become expensive when B is large. To this end, for a given gRNA, we share the permutations $\tilde{X}_1, \ldots, \tilde{X}_B$ of its indicator vector X across all genes that we test for association with the gRNA. This technique is common in permutation-based genome-wide association (GWA) analysis[29], for example.

Acceleration 4: Adaptive permutation testing. Computing a large number of permutation resamples (e.g., $B = 250,000$) for a gene-gRNA pair that yields an unpromising p-value after only a few thousand resamples is wasteful. To reduce this inefficiency, we implement a two-step adaptive permutation testing scheme. First, we conduct an initial "screen" of all gene-gRNA pairs, computing the p-value of each out to a small number (B_1) of resamples. If this initial p-value is unpromising (i.e., if it exceeds some pre-selected threshold of p_{thresh}, where $p_{\text{thresh}} \approx 0.01$), then we return this p-value to the user. Otherwise, we draw a large number (B_2) of fresh resamples and compute the p-value using this second set of resamples. The tuning parameters of this procedure are B_1, B_2, and p_{thresh}; reasonable default values are $B_1 = 5,000$, $B_2 = 250,000$, and $p_{\text{thresh}} = 0.01$. As most pairs are expected to be null (and thus yield unpromising p-values), this procedure eliminates most of the compute associated with carrying out the permutation tests. Other, more complex adaptive permutation testing algorithms have been developed[29,30] (and in theory are compatible with SCEPTRE), but we favor the two-step strategy outlined above for its simplicity.

Positive control analysis and QC details

Positive control analysis. We group together gRNAs that target the same genomic location. We call these grouped gRNAs "gRNA groups[3]". We construct positive control pairs by coupling a given gRNA group to the gene or protein that the gRNA group targets. We develop a Nextflow pipeline to apply all methods to analyze the positive control pairs of

all datasets[6].

QC. We have found that SCEPTRE produces mildly inflated p-values when applied to pairs with an effective sample size of zero. To avoid this pathological corner case, we recommend restricting the QC threshold to one treatment cell with nonzero expression or greater.

Simulation study

We conduct a simulation study (Figure S5) to demonstrate the existence and utility of the CAMP ("confounder adjustment via marginal permutations") phenomenon. We base the simulation study on a gene (namely, *CXCL10*) and gRNA (namely, "CUL3") from the Papalexi data. Following the notation introduced in Section SCEPTRE overview, let $Y = [Y_1, \ldots, Y_n]^T$ denote the vector of gene expressions of *CXCL10* and $X = [X_1, \ldots, X_n]^T$ the vector of gRNA indicators of "CUL3." Next, let $Z_i \in \mathbb{R}^p$ denote the vector of technical factors of the ith cell (for $i \in \{1, \ldots, n\}$), and let Z denote the $n \times p$ matrix formed by assembling the Z_is into a matrix. We regress Y onto Z by fitting the GLM (2), yielding estimates $\hat{\beta}$ for β and θ^* for θ under the null hypothesis of no association between the gRNA and gene. An examination of $\hat{\beta}$ reveals that the gene expressions Y are moderately associated with the technical factors Z. Letting $\hat{\mu}_i = \exp(\hat{\beta}^T Z_i)$ denote the fitted mean of cell i, we sample B i.i.d. synthetic expressions $\tilde{Y}_i^1, \ldots, \tilde{Y}_i^B$ from an NB model with mean $\hat{\mu}_i$ and size parameter θ^*. We then construct B synthetic gene expression vectors $\tilde{Y}^j = [\tilde{Y}_1^j, \ldots, \tilde{Y}_n^j]^T \in \mathbb{R}^n$ for $j \in \{1, \ldots, B\}$. Next, we generate a synthetic gRNA indicator vector $\tilde{X} \in \mathbb{R}^n$ such that $\tilde{X}$ is independent of Z. To this end, we marginally sample synthetic gRNA indicators $\tilde{X}_1, \ldots, \tilde{X}_n$ i.i.d. from a Bernoulli model with mean $\hat{\pi}$, where $\hat{\pi} = (1/n) \sum_{i=1}^n X_i$ is the fraction of cells that received the targeting gRNA. (The observed gRNA indicator vector X is moderately associated with Z.)

We assess three methods in the simulation study: NB regression, SCEPTRE, and the vanilla permutation test. We deploy NB regression and SCEPTRE in a slightly different way than usual: we set the NB size parameter θ upon which these methods rely to a fixed value. (Typically, NB regression and SCEPTRE estimate θ using the data.) This enables us to assess the impact of misspecification of the size parameter on the calibration of NB regression and SCEPTRE. We set the test statistic of the vanilla permutation test to the sum of the gene expressions in the treatment cells. We then generate B confounded (resp., unconfoundeed) datasets by pairing the synthetic response vectors $\tilde{Y}_1, \ldots, \tilde{Y}_B$ to the design matrix $[X, Z]$ (resp., $[\tilde{X}, Z]$). We apply the methods to the datasets twice: once setting the SCEPTRE/NB regression size parameter to the correct value of θ^*, and once setting this parameter to the incorrect value of $5 \cdot \theta^*$. We display the results produced by

[6] github.com/Katsevich-Lab/pc-grna-pipeline

the methods in each of the four settings (i.e., confounded versus unconfounded, correct versus incorrect specification of the size parameter; Figure S5) on a QQ-plot. We seek to show that SCEPTRE maintains calibration in all settings, while the vanilla permutation test and NB regression break down under confounding and incorrect specification of the size parameter, respectively.

Author contributions

GK identified the research problem. TB and GK performed the analyses and developed the method, with assistance from KR. GK and KR supervised the project. TB wrote the manuscript. All authors edited the manuscript and approved the final version.

Supplementary Tables

Paper	**Datasets**	**CRISPR modality**	**Tech. platform**	**Target**	**Modality measured**	**Cell type**
Frangieh 2021	co-culture, control, IFN-γ (3)	CRISPRko	Perturb-CITE seq	Gene TSSs	Gene expressions*	TIL
Papalexi 2021	ECCITE screen (1)	CRISPRko	ECCITE-seq	Gene TSSs	Gene and protein expressions	K562
Schraivogel 2020	Enhancer screen (1)	CRISPRi	Targeted perturb-seq	Enhancers	Gene expressions	THP1
-	Simulated dataset (1)	-	-	Gene TSSs	Gene expressions	-

Table S1: **Datasets analyzed in this work**. The first column indicates the name of a recent low-MOI single-cell CRISPR screen paper; the second column indicates the datasets that we obtained from that paper; and the subsequent columns indicate the (paper-specific) biological attributes of the data, including CRISPR modality, technology platform, target type, cellular modality measured, and cell type. Tech., technology. *The Frangieh data also contain protein measurements. However, unlike the Papalexi protein modality, the Frangieh protein modality requires a complicated normalization procedure to use, and so we focus exclusively on the Frangieh gene modality in this work.

Dataset	**N genes (or proteins)**	**N cells**	**N targeting gRNAs**	**N NT gRNAs**	**N neg. control pairs**	**N pos. control pairs**
Frangieh Co-culture	14,438	46,427	744	74	596,344	181
Frangieh control	15,449	30,486	744	74	528,239	170
Frangieh IFN-γ	14,654	50,053	744	74	565,502	181
Papalexi (gene)	14,559	20,729	101	9	100,458	25
Papalexi (protein)	4	20,729	101	9	36	2
Schraivogel*	82 (Chr11), 71 (Chr8)	99,884 (Chr11), 88,715 (Chr8)	3,073 (Chr11), 4,089 (Chr8)	30 (Chr11), 30 (Chr8)	4,693 (pooled)	26 (pooled)
Simulated	4439	10,000	-	25	108,510	-

Table S2: **Statistical attributes of the datasets**. The number of genes, cells, targeting gRNAs, NT gRNAs, negative control pairs, and positive control pairs for each dataset. Neg., negative; pos., positive.

*Schraivogel separately assayed two chromosomes: Chr11 and Chr8. Given the similarity of these assays, we pool together the negative and positive control pairs across assays.

Method	**Sparsity**	**Confounding**	**Model misspecification**
Seurat DE	No	No	Yes
Liscovitch Method	Yes	No	No
Schraivogel Method	No	No	No
Weissman Method	No	No	Yes
NB regression	No	Yes	No
Vanilla permutation test	Yes	No	Yes
SCEPTRE	Yes	Yes	Yes

Table S3: **Analysis challenges addressed by each method**. The cells indicate whether the method in the row addresses the analysis challenge in the column. SCEPTRE (bottom row) is the only method that addresses all three analysis challenges. Note: MIMOSCA is excluded from this table, as it is unclear which (of any) of the analysis challenges MIMOSCA addresses.

Supplementary Figures

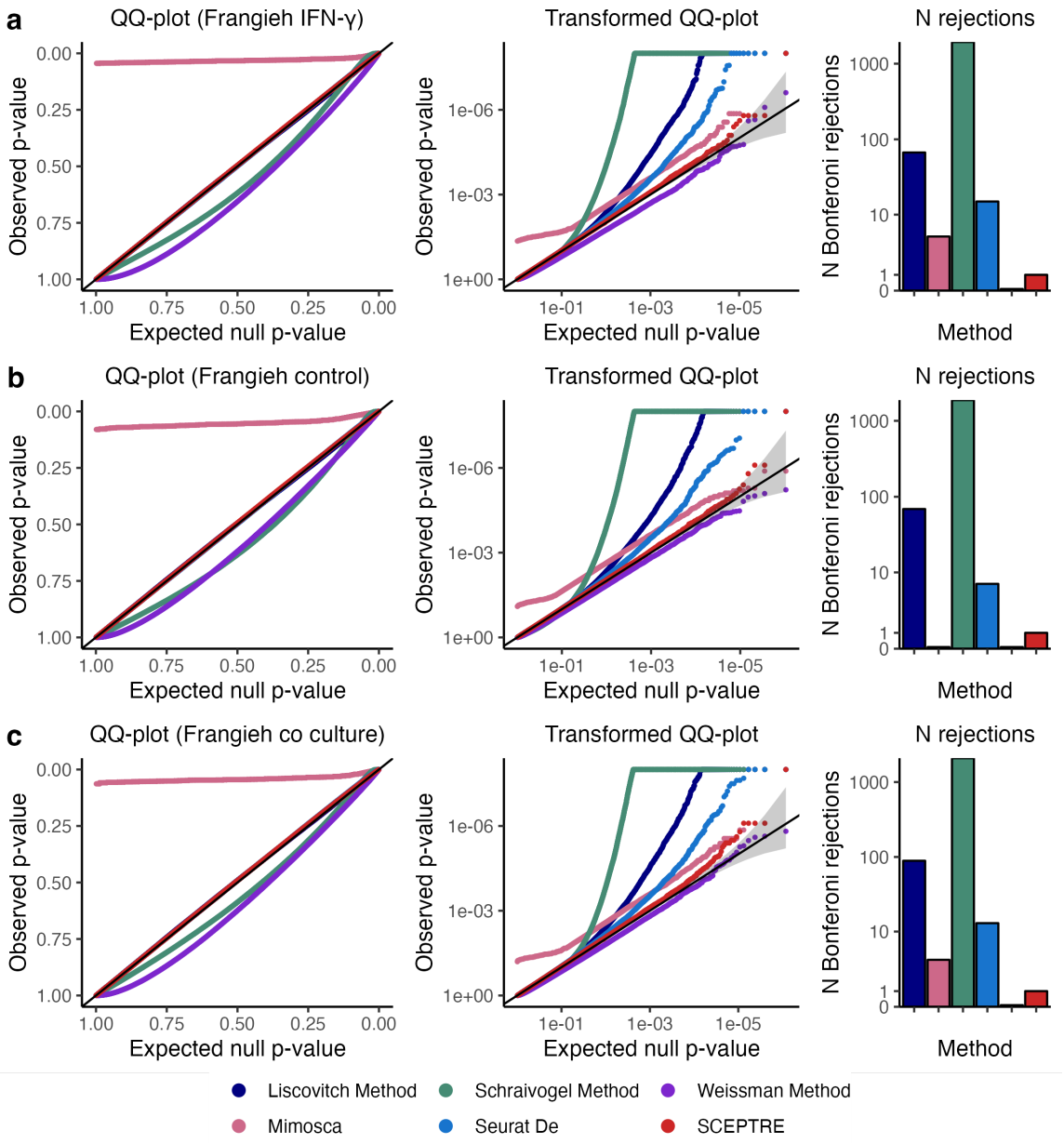

Figure S1: **Calibration results for all methods on Frangieh IFN-γ, Frangieh control, and Frangieh co-culture negative control data.** The interpretation of these plots (as well as those in Figures S1 - S3) is similar to the interpretation of the plots in Figure 1, panels b-c.

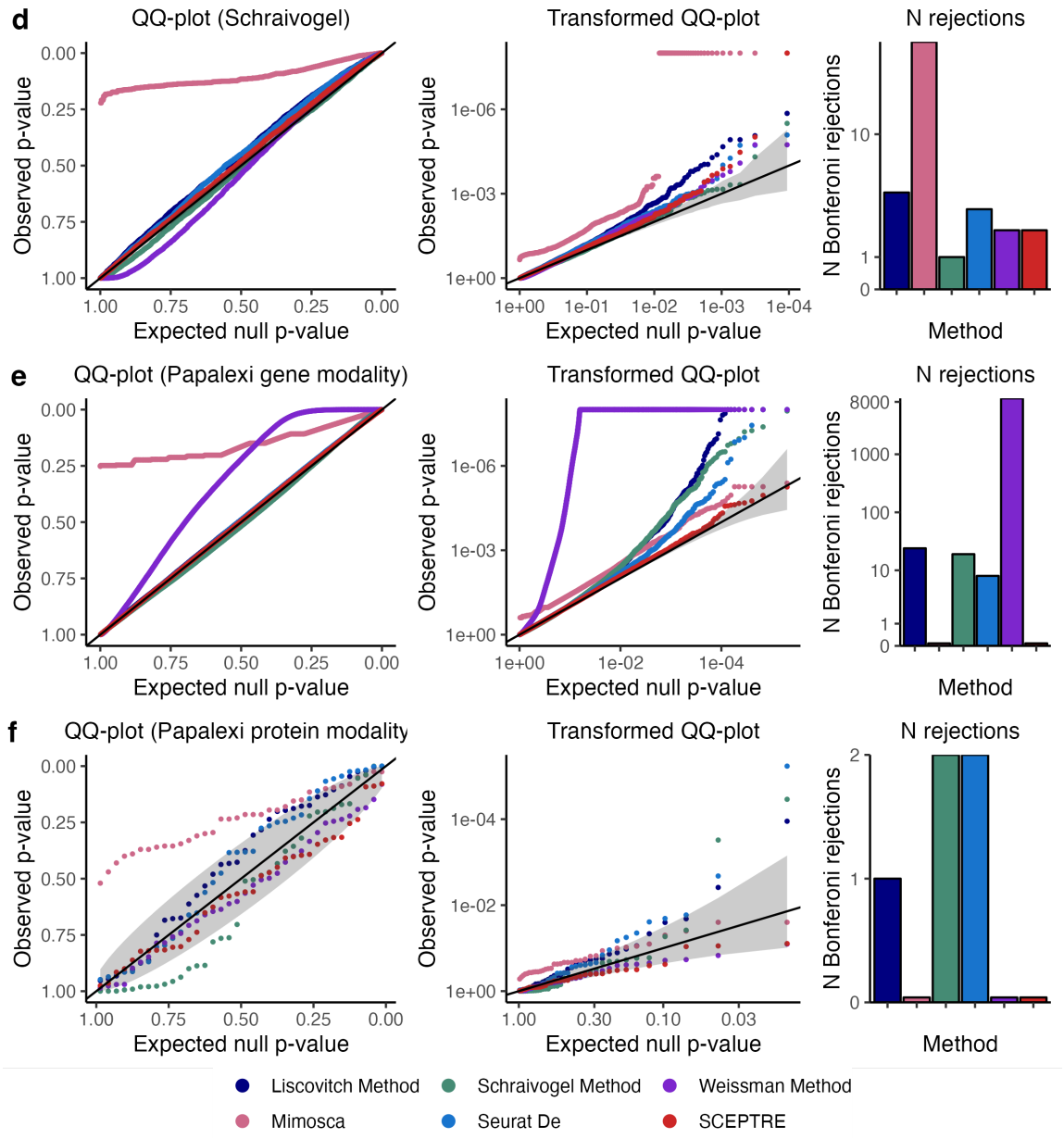

Figure S2: **Calibration results for all methods on Schraivogel, Papalexi (gene modality), and Papalexi (protein modality) negative control data.**

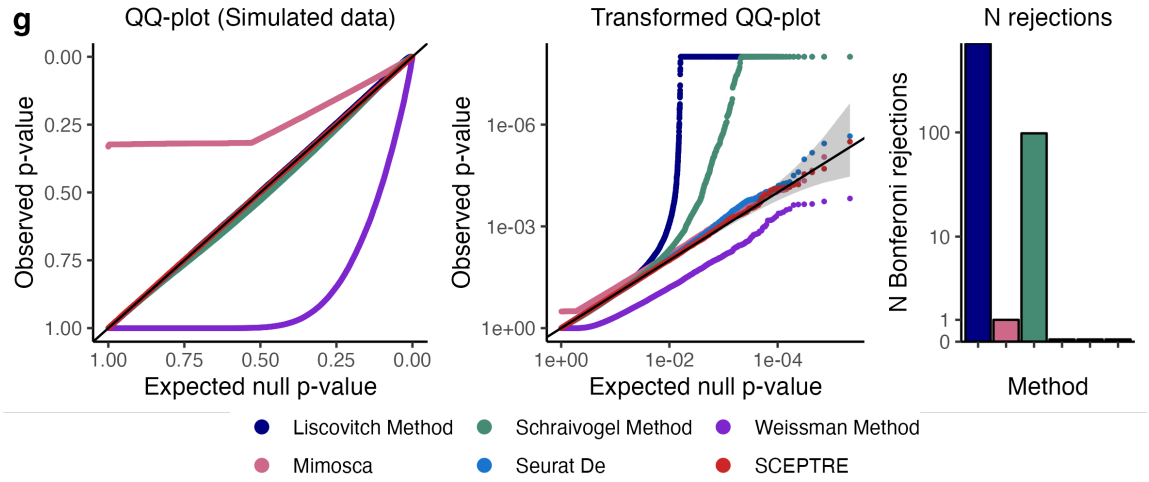

Figure S3: **Calibration results for all methods on simulated data.**

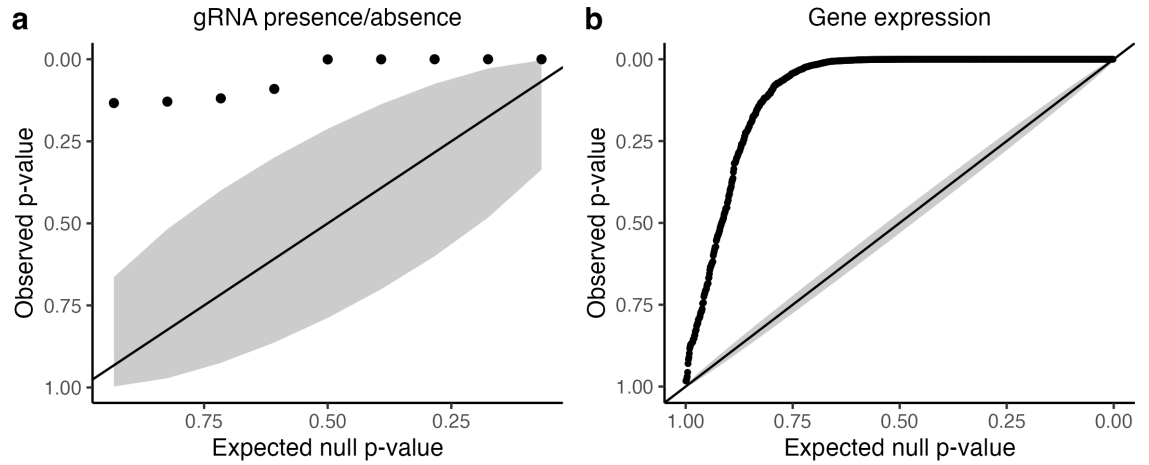

Figure S4: **Confounding due to biological replicate on the Papalexi (gene modality) data.** Left, QQ-plot of p-values for tests of association between the gRNA indicator and biological replicate for each NT gRNA (tests carried out using Fisher's exact test). Right, p-values for tests of association between (relative) gene expression and biological replicate for each gene (tests carried out using NB GLM likelihood ratio test). The inflation of the p-values indicates that the bulk of NT gRNAs and genes is impacted by biological replicate, creating a confounding effect.

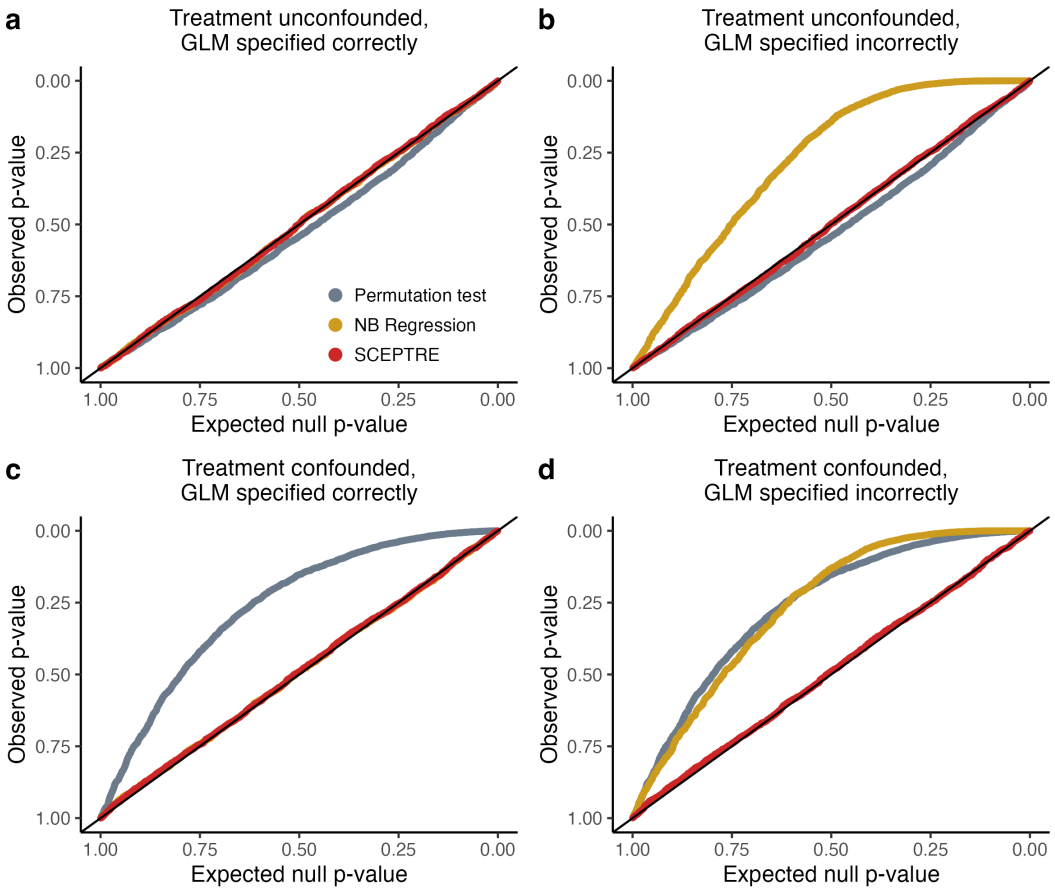

Figure S5: **Demonstration of the CAMP ("confounder adjustment via marginal permutations") phenomenon on realistic semi-synthetic data**. Application of a vanilla permutation test, NB regression, and SCEPTRE to realistic semi-synthetic data generated under two conditions: confounded and unconfounded. Panels **a** and **b** (resp., **c** and **d**) show the results on the unconfounded (resp., confounded) data; meanwhile, panels **a** and **c** (resp., **b** and **d**) show the results under correct (resp. incorrect) specification of the negative binomial size parameter. The permutation test (grey) works well when the data are unconfounded (panels **a** and **b**) but breaks down in the presence of confounding (panels **c** and **d**). On the other hand, NB regression is well-calibrated when the size parameter is correctly specified (panels **a** and **c**) but fails when the size parameter is misspecified (panels **b** and **d**). SCEPTRE is well-calibrated in all settings. We note that SCEPTRE is expected to break down when the (i) problem is confounded and (ii) the NB regression model is arbitrarily misspecified. Details of the simulation study are given in Section Simulation study.

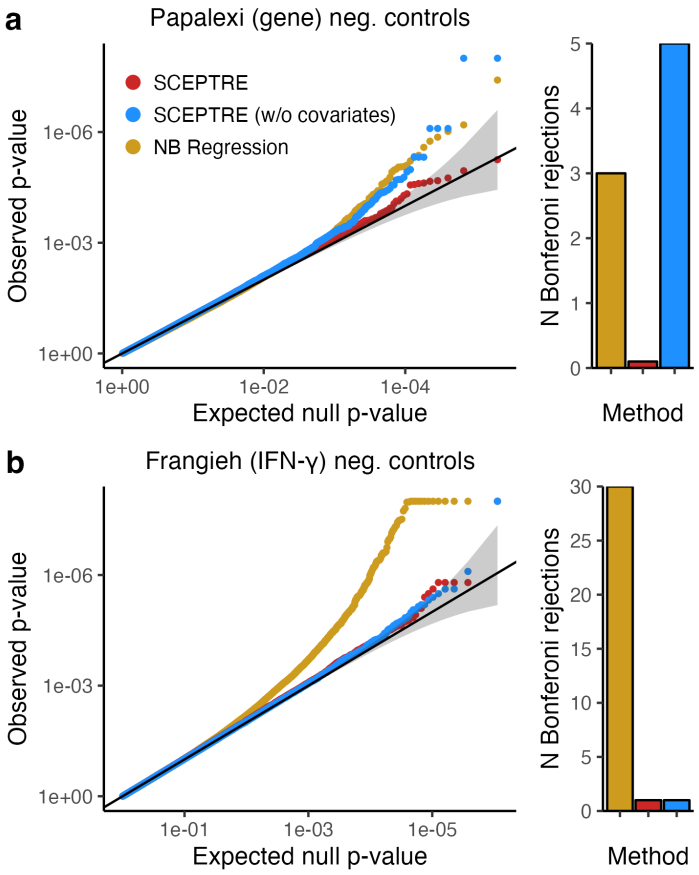

Figure S6: **Demonstration of the CAMP phenomenon on real data.** Application of SCEPTRE, SCEPTRE (without covariates), and NB regression to the Papalexi (gene modality) and Frangieh IFN-γ negative control data. SCEPTRE (without covariates) can be thought of as a vanilla permutation test that adjusts for library size only. **a**, Results on the Papalexi negative control data. SCEPTRE (without covariates) fails, as it does not adjust for confounding. NB regression, by contrast, does adjust for confounding, and so it outperforms SCEPTRE (without covariates). However, due to likely misspecification of the size parameter, NB regression exhibits mild inflation. Finally, SCEPTRE is well-calibrated, as the NB GLM underlying SCEPTRE lkely is correctly specified up to the size parameter. These results mirror those of panel **d** in Figure S5. **b**, Results on the Frangieh IFN-γ negative control data. NB regression fails due to a combination of model misspecification and sparsity. SCEPTRE (without covariates), on the other hand, is well-calibrated, as confounding is negligible. Finally, SCEPTRE matches the performance of SCEPTRE (without covariates). These results mirror those of panel **b** in Figure S5.