Robust differential expression testing for single-cell CRISPR screens at low multiplicty of infection

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

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

Single-cell CRISPR screens have emerged as a critical method for linking genetic perturbations to phenotypic changes in individual cells. The most 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 conducted the first-ever comprehensive benchmarking study of association testing methods for single-cell CRISPR screens, applying six leading methods to analyze six diverse datasets. We found that existing methods exhibit varying degrees of miscalibration, suggesting that results obtained using these methods may contain excess false positives. Next, to understand why existing methods demonstrate miscalibration, we conducted an extensive empirical investigation of the data. We identified three core analysis challenges: sparsity, confounding, and model misspecification. Finally, we developed SCEPTRE (`katsevich-lab.github.io/sceptre`), an association testing method based on the novel and statistically principled technique of permuting negative binomial score statistics. SCEPTRE addresses the core analysis challenges both in theory and in practice, demonstrating markedly improved calibration and power across datasets.

Pooled CRISPR screens with single-cell readout (e.g. Perturb-seq[1]) have emerged as a scalable, flexible, and powerful technique for connecting genetic perturbations to molecular phenotypes, with applications ranging from fundamental molecular biology to medical genetics and cancer research.[2] In such screens, a library of genetic perturbations is transfected into a population of cells via CRISPR guide RNAs (gRNAs), followed by single-cell sequencing to identify the perturbations present and measure a rich molecular phenotype for each cell. The perturbations can target either genes[1] or non-coding regulatory elements,[3–5] either repressing[1] or activating[6] these targets; the molecular readouts can include gene expression[1], protein expression,[7–9] or epigenetic phenotypes like chromatin accessibility.[10] Typically, perturbations are introduced at low multiplicity of infection (MOI), with one perturbation per cell. In cases where perturbations are expected to have weak effects (like regulatory-element-targeting screens), perturbations also can be introduced at high-MOI (with many perturbations per cell) to increase scalability.[3–5,11,12]

The most fundamental statistical task involved in the analysis of single-cell CRISPR screen data is to test for association between a perturbation and a univariate, count-based molecular phenotype, like the expression of a gene or protein. In our previous work on high-MOI single-cell CRISPR screen analysis,[13] we discovered that existing methods for association testing are prone to an excess of false positive hits.[13] In that work we proposed SCEPTRE, a well-calibrated method for association testing on high-MOI data. Although important, the high-MOI association testing problem is less pressing than its low-MOI counterpart, as low-MOI screens far outnumber high-MOI screens. A variety of methods has been deployed for association testing in low-MOI.[1,8–10,14–17] Currently, there is no consensus as to which of these methods represents the "state of the art;" these methods have not undergone rigorous statistical validation and comparison; and in fact there is no commonly accepted framework for quantifying the statistical validity of single-cell CRISPR screen association testing methods. Resolving these fundamental issues is essential to ensuring the reliability of biological conclusions made on the basis single-cell CRISPR screen experiments.

We aimed to address the aforementioned challenges by making three contributions. First, we developed a simple framework for evaluating the calibration of association testing methods for single-cell CRISPR screens. We then leveraged this framework to conduct the first-ever comprehensive benchmarking study of association methods on low-MOI data, applying six leading methods to analyze six diverse datasets. We found that all existing methods exhibit varying degrees of miscalibration, indicating that results obtained using these methods may be contaminated by excess false positive discoveries. Second, to shed light on why existing methods might demonstrate miscalibration, we conducted 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. Finally, we developed SCEPTRE (low-MOI), a substantial extension of the original SCEPTRE[13] tailored to the analysis of low-MOI single-cell CRISPR screens. SCEPTRE (low-MOI) is based on the novel and statistically principled technique of permuting negative binomial score statistics. (We often will refer to the low-MOI version of SCEPTRE simply as "SCEPTRE" for the sake of brevity). SCEPTRE addresses all three core analysis challenges both in theory and in practice, demonstrating markedly improved calibration and power relative to existing methods across datasets.

Results

Comprehensive benchmarking study of leading analysis methods

Association testing on low-MOI single-cell CRISPR screen data is a variation on the classical single-cell differential expression testing problem (Figure 1a). To test for association between a given targeting CRISPR perturbation and gene, one first divides the cells into two groups: those that received the targeting perturbation, and those that received a non-targeting (NT) perturbation. (All other cells typically are ignored.) One then tests for differential expression of the given gene across these two groups of cells, yielding a fold change estimate and p-value. One repeats this procedure for a (typically) large, preselected set of perturbation-gene pairs. Finally, one computes the discovery set by subjecting the tested pairs to a multiple comparison correction procedure (e.g., Benjamini-Hochberg).

For a given targeting perturbation-gene pair, we refer to the cells that received the targeting perturbation as the "treatment cells," and we refer to the cells against which the treatment cells are compared as the "control group." As indicated above, the control group typically is the set of cells that received an NT perturbation (i.e., the "NT cells"). Certain single-cell CRISPR screen methods, however, take as their control group the set of cells that did *not* receive the targeting perturbation (i.e., the "complement set"). The NT cells generally constitute a more natural control group than the complement set, as we seek to compare the effect of the targeting perturbation to that of a "null" perturbation rather than to the average of the effects of all other perturbations introduced in the pooled screen.

We surveyed recent analyses of single-cell CRISPR screen data and identified five methods commonly in use: the default Seurat[18] `FindMarkers()` function based on the Wilcoxon test (Seurat-Wilcox); MIMOSCA;[1] a t-test on the library-size-normalized expressions;[10] MAST;[19] and a Kolmogorov-Smirnov (KS) test on the library-size-normalized expressions.[20] We also considered applying `FindMarkers()` with negative binomial (NB) regression rather than the Wilcoxon test (Seurat-NB). These methods vary along several dimensions (Table 1; Existing methods details), including their testing paradigm (two-sample test versus regression-based test), how they normalize the data, whether they make parametric assumptions, and whether they use the NT cells or the complement set as their control group. Most of these methods are popular single-cell differential expression procedures that have been adapted to single-cell CRISPR screen data.

We sought to assess whether these methods are correctly calibrated (i.e., whether they yield uniformly distributed p-values under the null hypothesis of no association between the perturbation and gene). Methods that are not correctly calibrated can produce discovery sets that are contaminated by excess false positives or false negatives. Unfortunately, there does not exist a standard protocol for assessing the calibration of single-cell CRISPR

Method	Paradigm	Parametric assumption	Null distribution	Normalization/ Adjustments	Control group
Seurat-Wilcox [7,9,21]	Two-sample test	No	Asymptotic	Library size	NT cells
MIMOSCA[1] [1,8,22–25]	Regression-based	No	Permutation	Library size, other covariates	Complement set
t-test [10]	Two-sample test	Yes	Asymptotic	Library size	NT cells
MAST[19] [15]	Regression-based	Yes	Asymptotic	Library size, expressed genes	NT cells
KS test [16,20,26]	Two-sample test	No	Asymptotic	Library size, batch	NT cells
Seurat-NB (single-cell DE)	Two-sample test	Yes	Asymptotic	Library size	NT cells

Table 1: **A summary of low-MOI single-cell CRISPR screen DE methods currently in use**. The applications of each method to single-cell CRISPR screens are cited below the method name. The methods vary along several key axes, including the use (or lack thereof) of parametric assumptions, the construction of the null distribution, the variables adjusted for, and the control group. NT, non-targeting.

screen association methods. The closest existing analysis[15] proceeds by applying methods to analyze gene-perturbations pairs for perturbations with known targets. Any pair where the gene is not the known target of the perturbation is considered null. As acknowledged by the original authors, this approach underestimates precision because downstream effects of perturbations are not taken into account.

To help fill this methodological gap, we designed a simple procedure to check the calibration of a single-cell CRISPR screen association method (Figure 1b). We constructed a set of "null" or "negative control" perturbation-gene pairs by pairing each NT gRNA to each gene. We then deployed a given method to analyze these null pairs. (For methods that use the NT cells as their control group — the majority of methods — this check consists of comparing cells containing a given NT gRNA to cells containing *any other* NT gRNA.) The output of this check is 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. Since the null perturbation-gene pairs are devoid of signal, a well-calibrated association method should output uniformly distributed p-values on these pairs. Deviations from uniformity — and thus miscalibration of the method — can be detected by inspecting a QQ plot of the p-values. Quantitatively,

the number of null pairs passing a Bonferroni correction measures the extent of the miscalibration; well-calibrated methods should have roughly zero such pairs.

We employed the above framework to systematically benchmark the performance of the existing methods on six single-cell CRISPR screen datasets, five real and one simulated (Tables S1-S2). The five real datasets come from three recent papers: Frangieh 2021[8] (three datasets), Papalexi 2021[9] (one dataset), and Schraivogel 2020[15] (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 TSS). Notably, the Papalexi data are multimodal, containing both gene and protein expression measurements. For simplicity, we analyzed the gene and protein modalities separately throughout.

Surprisingly, the results of our analyses (Figures 1c, S1, S2, S3) revealed substantial miscalibration for many dataset-method pairs. On the Papalexi data, for example, the KS test produced inflated p-values, yielding over 9,000 false Bonferroni discoveries. MAST was similarly inflated on the Frangieh IFN-γ data, falsely rejecting nearly 2,000 null perturbation-gene pairs. MIMOSCA, meanwhile, exhibited noticeably non-uniform behavior on both datasets, outputting p-values strictly less than 0.26 across all pairs. Overall, the two best methods were Seurat-Wilcox and Seurat-NB. Still, Seurat-Wilcox and Seurat-NB demonstrated clear signs of miscalibration, suggesting that space for improvement is available.

Systematic identification of core analysis challenges

We conducted an extensive empirical investigation of the data to search for possible sources of miscalibration, uncovering three core analysis challenges: sparsity, confounding, and model misspecification. No method that we examined addressed more than one of these analysis challenges (Table S3), explaining their collective lack of calibration.

Single-cell CRISPR screen data typically are sparse, both in terms of gene expression and perturbation presence. Many genes have nonzero expression in only a small fraction of cells. On the other hand, due to the pooling of a large number of perturbations in a single experiment, the perturbation presence data are also sparse: most perturbations are present in only a small fraction of cells. The latter sparsity distinguishes single-cell CRISPR screens from other single-cell applications and is particularly pronounced in low-MOI. To summarize both sources of sparsity in a single number, we defined the "effective sample size" for a given perturbation-gene pair as the number of cells containing both the perturbation and nonzero gene expression.

We found that effective sample size had a substantial effect on the calibration of many methods under consideration (Figure S4), especially those based on asymptotic approx-

imations, such as Seurat-Wilcox. Asymptotic approximations tend to break down when the effective sample size is too low. For example, we compared the exact null distribution of the Wilcoxon test statistic (obtained via permutations) to the asymptotic Gaussian distribution used by Seurat-Wilcox; the latter is a computationally tractable approximation to the former in large samples. The Gaussian distribution provided a reasonable approximation to the exact null distribution for some pairs (Figure 2a, left) but not others (2a, right). Furthermore, as the effective sample size decreased and the Gaussian approximation degraded in accuracy, the p-value obtained via the Gaussian approximation likewise degraded in accuracy (Figure 2b). Finally, stratifying the Seurat-Wilcox null p-values by effective sample size on the Frangieh IFN-γ data revealed that pairs with small effective sample sizes yielded more inflated p-values than pairs with large effective sample sizes (Figure 2c).

Second, technical factors, such as biological replicate, batch, and library size, impact not only a cell's expression level, but also its probability of receiving a perturbation, thereby creating a confounding effect that, if not accounted for, can lead to spurious associations[13] (Figure 2d, Figure S5). All existing methods adjust for library size, but few adjust for other technical factors (Table 1). To assess the utility of adjusting for technical factors beyond library size, we applied negative binomial (NB) regression — both with and without biological replicate included as a covariate — to the Papalexi negative control data (Figure 2e). The variant of NB regression with biological replicate, though not perfectly calibrated, outperformed its counterpart without biological replicate. Methods not adjusting for biological replicate on the Papalexi data (such as Seurat-Wilcox) exhibited *worse* calibration for large effective sample sizes (Figure S4), where there is more power to detect the spurious confounding-driven associations.

Third, methods that rely upon parametric models for the gene expression distribution, such as NB regression and MAST, can yield miscalibrated p-values when those models are misspecified.[27] To assess this effect, we monitored p-value calibration of the NB regression method on the Frangieh IFN-γ data while gradually increasing the effective sample size (Figure 2f). We found that the calibration quality improved until a point before plateauing; even for large effective sample sizes, noticeable miscalibration remained. (The non-parametric Seurat-Wilcox method, by contrast, attained good calibration for large effective sample sizes on this dataset.) This pattern was consistent with poor fit of the NB regression model, potentially due to inadequate estimation of the NB size parameter.

SCEPTRE (low-MOI) addresses the analysis challenges

We next developed SCEPTRE (low-MOI), a method for robust single-cell CRISPR screen association testing on low-MOI data (Figure 3a). For a given targeting perturbation-gene

pair, SCEPTRE first regresses the vector of gene expressions onto the vector of perturbation indicators and matrix of technical factors via an NB GLM. (A given entry of the perturbation indicator vector is set to "1" if the corresponding cell contains a targeting perturbation and "0" if it contains a non-targeting perturbation.) SCEPTRE then computes the z-score z_{obs} corresponding to a test of the null hypothesis that the the coefficient corresponding to the perturbation indicator in the fitted GLM is zero. Next, SCEPTRE permutes the perturbation indicator vector B times (while holding fixed the gene expression vector and technical factor matrix) and recomputes a z-score for each of the permuted indicator vectors, yielding B "null" z-scores. Finally, SCEPTRE fits a smooth (skew-normal) density to the histogram of null z-scores and computes a p-value by evaluating the tail probability of the fitted density based on the original test statistic z_{obs}.

SCEPTRE possesses several appealing theoretical and computational properties. Theoretically, SCEPTRE is robust to the calibration threats of sparsity, confounding, and model misspecification. A key observation is that the technical factors (e.g., biological replicate) may or may not exert a confounding effect on the perturbation indicator and gene expression (Figure 3b). If confounding is absent for a given perturbation-gene pair, then SCEPTRE is valid regardless of misspecification of the NB model or the level of sparsity. On the other hand, if confounding is present, then SCEPTRE retains validity if the NB model is correctly specified and the problem is not too sparse (Figure 3c). (In fact, in the latter case, the NB model need only be specified correctly up to the dispersion parameter, sidestepping the difficult problem of NB dispersion parameter estimation.[28,29]) In this sense SCEPTRE is the only method that addresses all three core analysis challenges (Table S3). We explored the above key robustness property of SCEPTRE in a brief simulation experiment (Figure S6).

SCEPTRE also is performant, capable of analyzing hundreds of perturbation-gene pairs per second. We attained this efficiency by implementing several computational accelerations. First, we elected to use a *score* test (as opposed to a more standard *Wald* or *likelihood ratio* test) to compute the NB z-scores; the score test enabled us to fit a single NB GLM per perturbation-gene pair and share this fitted GLM across all permuted perturbation indicator vectors. Second, we derived a new algorithm for computing GLM score tests; this new algorithm is hundreds of times faster than the classical algorithm when the perturbation indicator vector is sparse, as is the case in single-cell CRISPR screen analysis. Finally, we developed several strategies for recycling compute across distinct perturbation-gene pairs. We note that SCEPTRE (low-MOI) is inspired by, but not at all identical to, SCEPTRE (high-MOI).[13] We clarify similarities and differences between these two methods in Existing methods details.

Application of SCEPTRE to negative and positive control data

We added SCEPTRE to the calibration benchmarking analysis presented before. An inspection of the QQ plots revealed that SCEPTRE markedly improved on the calibration of the two best existing methods, namely Seurat-Wilcox and Seurat-NB (Figure 4a-b). For example, on the Frangieh IFN-γ data, SCEPTRE made one Bonferroni rejection and yielded p-values that lay mostly within the grey 95% confidence band. The Seurat methods, by contrast, made fifteen false rejections each and produced p-values that fell considerably outside the confidence band. Next, we tabulated the number of Bonferroni-significant false positives for each dataset-method pair (Figure 4c; smaller values are better). SCEPTRE generally made the fewest number of false discoveries among all methods. On average over datasets, SCEPTRE made only 0.7 false discoveries, a roughly tenfold improvement over the Seurat methods.

Next, we assessed the power of the methods by applying them to positive control data. We constructed positive control pairs for each dataset by coupling perturbations targeting TSSs or known enhancers to the genes (or proteins) regulated by these elements. We examined the number of "highly significant" discoveries — operationally defined as rejections made at level $\alpha = 10^{-5}$ — made by each method on each dataset (Figure 4c; larger values are better). Methods that exhibited extreme miscalibration on a given dataset (defined as > 50 Bonferroni rejections on the negative control pairs of that dataset) were excluded from the positive control analysis, as assessing the power of such methods is challenging. We found that SCEPTRE matched or outperformed the other methods with respect to power on nearly every dataset (while at the same time achieving better calibration on negative control data).

Pairwise quality control

Quality control (QC) — the removal of low-quality genes and cells — is a key step in the analysis of single-cell data. In the context of single-cell CRISPR screens, it is useful not only only to remove low quality genes and cells but also low-quality perturbation-gene pairs. We term this latter type of QC "pairwise QC." As discussed previously, "effective sample size" — the number of cells containing both the perturbation and nonzero gene expression — affects the calibration of several methods considered. It also affects power, as small effective sample sizes yield low power and therefore needlessly increase the multiplicity burden. We found that SCEPTRE rarely rejected positive control pairs with an effective sample size below seven (Figure S7); moreover, SCEPTRE maintained calibration for negative control pairs with an effective sample size of seven and above. For this reason, our pairwise QC strategy consisted of filtering for pairs with an effective sample size of seven or greater. We applied this pairwise QC throughout.

Application of SCEPTRE for discovery analyses

The standard workflow in applying SCEPTRE to analyze a new single-cell CRISPR screen dataset consists of three main steps. First, the user prepares the data to pass to SCEPTRE and defines the "discovery set," which is the set of perturbation-gene pairs that the user seeks to test for association. (A reasonable default choice is the set of all possible pairs.) Second, the user runs the "calibration check" to verify that SCEPTRE is adequately calibrated on the dataset under analysis. The calibration check involves applying SCEPTRE to analyze a set of automatically-constructed negative control pairs. These negative control pairs are "matched" to the discovery pairs in several respects. For example, the negative control pairs and discovery pairs are subjected to the exact same pairwise QC, and the number of negative control pairs is set equal to the number of discovery pairs. If the calibration check fails, the user can take steps to improve calibration, such as adding covariates or increasing the pairwise QC threshold. After verifying adequate calibration, the user runs the "discovery analysis," which entails applying SCEPTRE to analyze the pairs contained in the discovery set (Figure 5a).

To illustrate the above workflow, we applied SCEPTRE to carry out a complete *trans* analysis of the Papalexi (gene expression) and Frangieh (control) datasets. Many of the genes targeted for knockout in these datasets were transcription factors (TFs); thus, our main biological objective was to map the TFs to their target genes. We carried out a calibration check and discovery analysis on both datasets (Figure 5b). These fairly large analyses completed within a matter of hours on a single laptop processor and required a few gigabytes of memory. We used publicly-available ChIP-seq data to validate the SCEPTRE-discovered targets of IRF1 and STAT1 on the Papalexi data. (IRF1 and STAT1 were the only TFs for which cell-type-relevant ChIP-seq data were available.) We found significant enrichment for both TFs (IRF1: odds ratio = 3.94, $p = 8 \times 10^{-76}$; STAT1: odds ratio = 1.37, $p = 5 \times 10^{-16}$), increasing our confidence in the discovery results.

Discussion

Single-cell CRISPR screens have emerged as a powerful method for linking genetic perturbations to rich phenotypic profiles in individual cells. Although poised to impact a variety of research directions, single-cell CRISPR screens will play an especially important role in dissecting the regulatory logic of the noncoding genome. The bulk of genetic risk for diseases lies in noncoding regions, implicating dysregulation of gene expression.[30–32] A major challenge in genetics, therefore, is to map noncoding disease variants to the genes that they target, target genes to the molecular programs that they regulate, and — ultimately — molecular programs to disease.[33] Single-cell screens have enabled breakthrough

progress on these tasks. For example, two recent studies leveraged high-MOI single-cell screens to perturb blood disease[11] and cancer[34] GWAS variants (in some cases at single nucleotide resolution) and link these variants to target genes in *cis* and *trans*. (Both studies used SCEPTRE (high MOI) to analyze their data.) Another recent study leveraged low-MOI single-cell screens to knock down genes regulated by heart disease GWAS variants and map these genes to the molecular programs that they regulate.[33] Given the promise that single-cell screens have demonstrated in understanding noncoding variation, we anticipate that a wave of single-cell screens aiming to link noncoding variants to genes and genes to molecular programs will emerge over the coming decade.

It is therefore crucial that reliable methods for single-cell CRISPR screen data analysis become available. The broad objective of this work was to put single-cell CRISPR screen analysis onto a solid statistical foundation. To this end we devised a simple framework for assessing the calibration and power of competing methods; applied this framework to conduct the first-ever comprehensive benchmarking study of existing methods; identified and dissected core statistical challenges; and developed a method, SCEPTRE, that combines careful modeling with a resampling procedure to produce a well-calibrated, powerful, fast, and memory-efficient test of association. Taken together, these contributions help bring statistical clarity and rigor to the practice of single-cell CRISPR screen data analysis. Furthermore, given the appealing theoretical, empirical, and computational properties of the proposed method, we anticipate that the method could be extended (with appropriate modifications) to applications beyond single-cell CRISPR screens, such as single-cell eQTL analysis and single-cell case-control differential expression analysis.

We identified sparsity, confounding, and model misspecification as key challenges in single-cell CRISPR screen analysis. However, several challenges beyond these impact the data, and SCEPTRE does not currently address these other challenges. First, some NT gRNAs may have off-targeting effects. In such cases testing for association by comparing cells that contain a targeting perturbation to those that contain an NT perturbation can result in a loss of error control. At least one prior work has attempted to address this problem.[26] Second, some targeting gRNAs are ineffective, i.e., they fail to perturb their target. Including such defective gRNAs in the analysis can result in a loss of power. Several methods, including MIMOSCA,[1] MUSIC,[35] and Mixscape[9], have attempted to resolve this issue. Third, it is challenging to distinguish between direct and indirect effects, in the sense that perturbations can be associated with their direct targets or with targets further downstream. Disentangling direct and indirect effects likely admits a statistical solution, but to our knowledge, this problem remains unaddressed. Finally, genes often are co-expressed alongside other genes in the same gene "module." An exciting opportunity is to pool information across genes within the same module to increase the power of perturbation-to-gene association tests; the recent method GSFA attempts to do just this.[36]

In summary single-cell CRISPR screens, though promising, present a variety of statistical challenges, demanding robust analytic tools. The `sceptre` software, which now includes functionality for both low- and high-MOI CRISPR screens, aims to provide practitioners with a unified solution for reliable, efficient, and user-friendly single-cell CRISPR screen differential expression analysis.

Figures

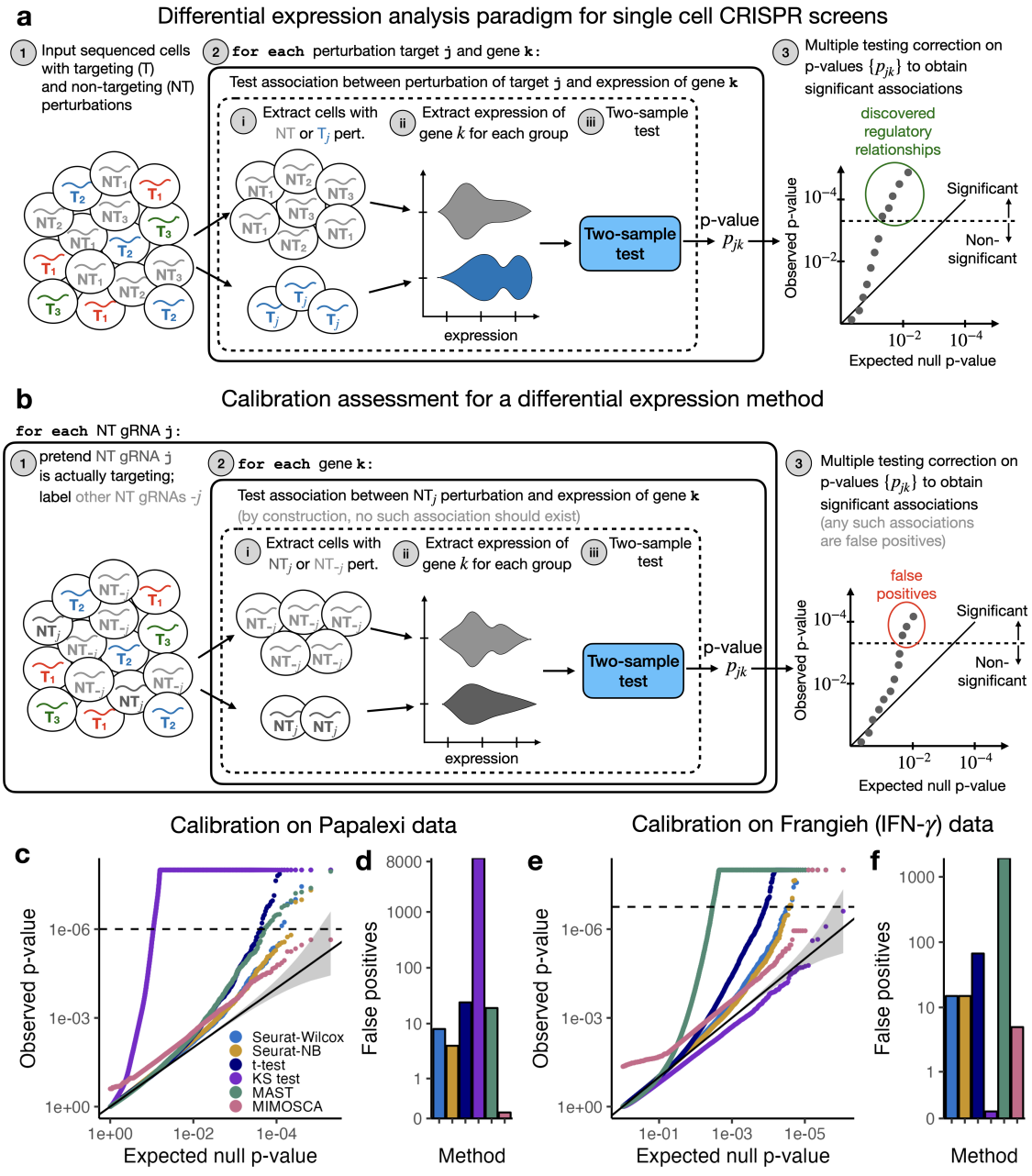


Figure 1: (Caption on next page.)

Figure 1: **Comprehensive benchmarking study of single-cell CRISPR screen association testing methods on low-MOI data**. **a**, The standard paradigm for association testing on low-MOI single-cell CRISPR screen data. To test for association between a given targeting perturbation and gene, one tests for differential expression of the gene across two groups of cells: those containing the given targeting perturbation, and those containing a non-targeting (NT) perturbation. One typically repeats this procedure for a large, preselected set of targeting-perturbation gene pairs, obtaining a discovery set by subjecting the resulting p-values to a multiple comparison correction procedure (e.g., Benjamini-Hochberg). **b**, The calibration check paradigm. One constructs "null" or "negative control" perturbation-gene pairs by coupling each individual NT gRNA to the entire set of genes. One then assesses the calibration of a method by deploying the method to analyze these null pairs. Any p-values that survive the multiple testing correction procedure correspond to false positive discoveries. **c-d**, Results of the calibration check benchmarking analysis on the Papalexi gene expression data. **c**, QQ plot of the null p-values (colored by method) plotted on a negative log transformed scale. Gray region, 95% confidence band. **d**, Number of false discoveries that each method makes on the null pairs after a Bonferroni correction at level 0.1. **e-f**, Similar to panels **c-d**, but for the Frangieh IFN-γ data.

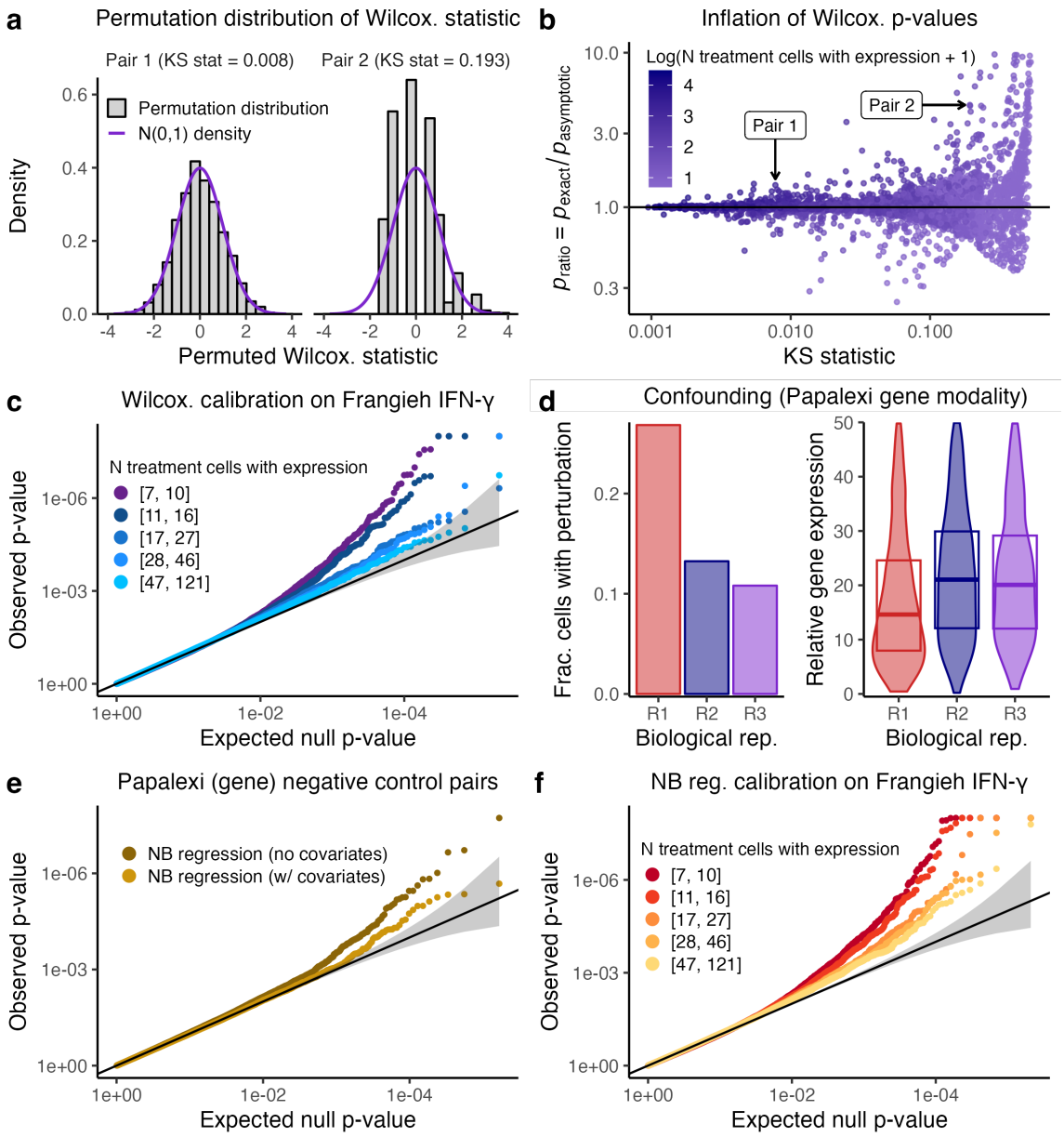


Figure 2: (Caption on next page.)

Figure 2: **Sparsity, confounding, and model misspecification are core analysis challenges in single-cell CRISPR screen analysis**. **a**, The exact null distribution of the Wilcoxin test statistic (obtained via permutations; grey) on two pairs from the Frangieh IFN-γ data. The Wilcoxin test (and thus Seurat-Wilcox) 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), while for pair 2 (right) the approximation is poor (KS statistic = 0.193). **b**, A plot of p_{ratio} (defined as the ratio of the exact Wilcoxin p-value, p_{exact}, to the asymptotic Wilcoxin 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 Wilcoxin null distribution), p_{ratio} deviates more from one, indicating miscalibration. Points are colored according to the effective sample size of the corresponding pair. **c**, Stratification of the Seurat-Wilcox p-values on the Frangieh IFN-γ negative control data by 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." **e**, Application of NB regression with and without covariates to the Papalexi data. **f**, Stratification of the NB regression p-values on the Papalexi (gene expression) negative control data by effective sample size.

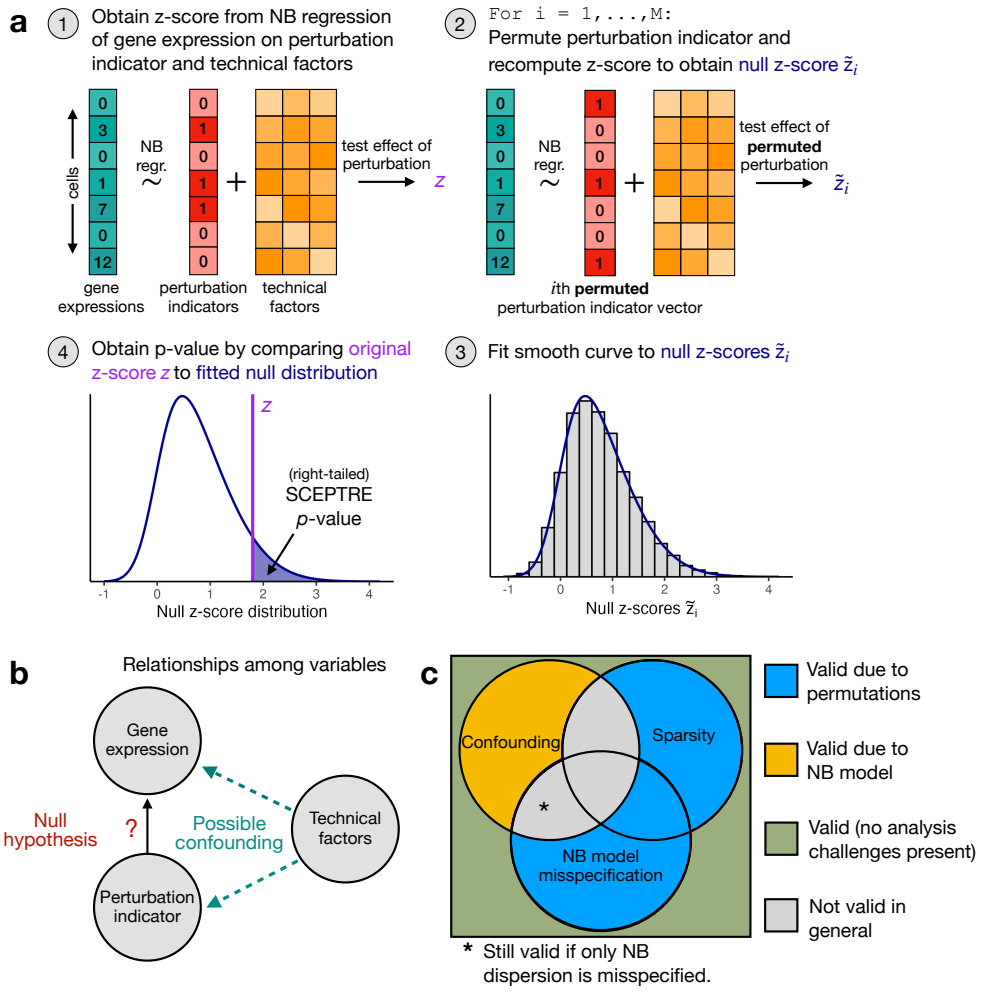
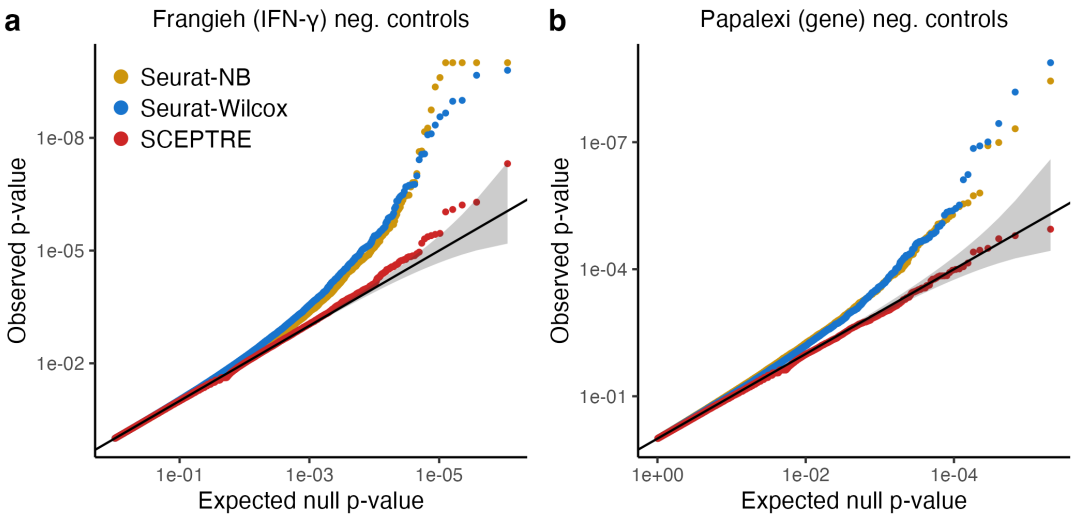


Figure 3: (Caption on next page.)

Figure 3: **SCEPTRE addresses the core analysis challenges of sparsity, confounding, and model misspecification in theory**. **a**, The SCEPTRE algorithm. First, the gene expressions are regressed onto the perturbation indicators and technical factors, and the z-score z_{obs} corresponding to the perturbation indicator is computed. Second, the perturbation indicators are permuted (while the gene expressions and technical factors are held fixed) and the z-score is recomputed, yielding B "null" z-values. Third, a smooth density is fit to the histogram of the null z-values. Fourth, a p-value is computed by evaluating the tail probability of the fitted density at z_{obs}. **b**, A diagram representing the relationship between the variables in the analysis. The technical factors often (but not always) exert a confounding effect on the perturbation indicator and gene expression. **c**, A diagram illustrating the robustness property of SCEPTRE. The circles represent analysis challenges. A perturbation-gene pair can be affected any subset of the analysis challenges. The color in each region of the diagram indicates whether SCEPTRE is valid on pairs affected by that subset of analysis challenges (blue, yellow, or green = valid; grey = not valid in general). For regions in which SCEPTRE is valid, the color of the region indicates *why* SCEPTRE is valid (yellow = NB model, blue = permutations). The validity of SCEPTRE is overdetermined on pairs unaffected by *any* analysis challenge (green region) due to the combination of the NB model and permutations.



c Number of false positives

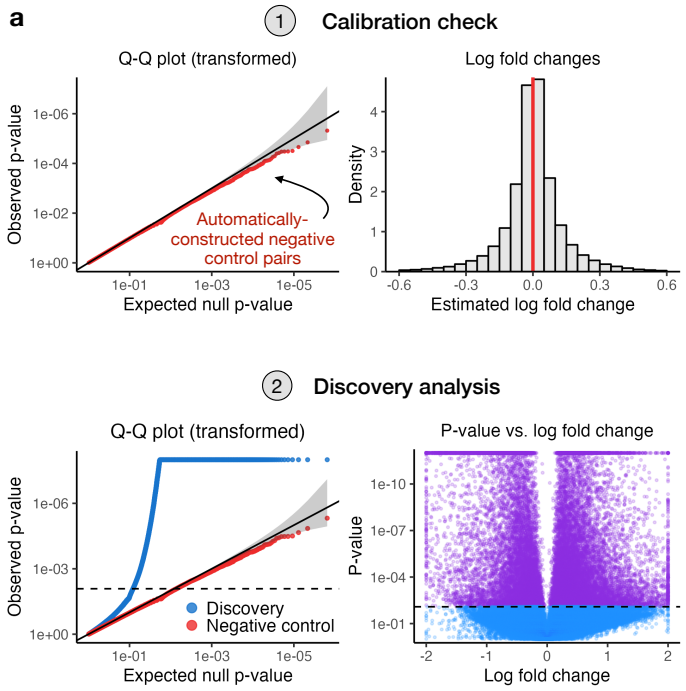
Dataset	SCEPTRE	Seurat-Wilcox	Seurat-NB	t-test	MAST	KS test	MIMOSCA	NT pairs
Frangieh (Co Culture)	1	13	10	89	2083	**0**	4	596344
Frangieh (Control)	**0**	7	16	69	1873	**0**	**0**	528239
Frangieh (IFN-γ)	1	15	15	67	1933	**0**	5	565502
Papalexi (Gene)	**0**	8	4	24	19	9191	**0**	100458
Papalexi (Protein)	**0**	2	**0**	1	2	**0**	**0**	36
Schraivogel	3	2	3	4	**1**	**1**	8	4357
Simulated	**0**	**0**	**0**	7	16	**0**	1	96944
Average	**0.7**	6.7	6.9	37.3	846.7	1313.1	2.6	

d Number of true positives

Dataset	SCEPTRE	Seurat-Wilcox	Seurat-NB	t-test	MAST	KS test	MIMOSCA	PC pairs
Frangieh (Co Culture)	**103**	98	94	-	-	90	5	181
Frangieh (Control)	**77**	74	72	-	-	70	4	170
Frangieh (IFN-γ)	**94**	89	81	-	-	81	8	181
Papalexi (Gene)	**13**	12	**13**	11	11	-	0	25
Papalexi (Protein)	**2**	**2**	**2**	**2**	**2**	**2**	**2**	2
Schraivogel	22	22	21	**23**	22	19	0	25

Figure 4: (Caption on next page.)

Figure 4: **SCEPTRE demonstrates improved calibration and power relative to existing methods across datasets**. **a** (resp. **b**), QQ plot of the p-values outputted by Seurat-NB, Seurat-Wilcox, and SCEPTRE on the Frangieh IFN-γ (resp., Papalexi gene expression) negative control data. Gray band, 95% confidence region. **c**, Number of false discoveries (at Bonferroni correction level 0.1) on the negative control data for each method-dataset pair. **d**, Number of true discoveries on the positive control data for each each method-dataset pair.



b Computational performance metrics* of SCEPTRE

Dataset	Frangieh (Control)	Papalexi (gene)
Number of pairs	2.2 million	323,906
Number of cells	30,486	20,729
Calibration check run time	45 min (794 pair/s)	32 min (167 pair/s)
Calibration check peak memory	3.4 GB	2.4 GB
Discovery analysis run time	8.6 h (70 pair/s)	1.8 h (51 pair/s)
Discovery analysis peak memory	4.0 GB	2.8 GB

*Benchmarking conducted on standard laptop using one processor

Figure 5: **Applying SCEPTRE to make biological discoveries**. **a**, The standard workflow involved in applying SCEPTRE to a new dataset, using the Papalexi gene expression data as a running example. First, SCEPTRE is applied to analyze a set of automatically-constructed negative control pairs (the "calibration check"). The resulting negative control p-values are plotted on a QQ plot to ensure uniformity (upper left), and the negative control log-fold changes are plotted on a histogram to ensure symmetry about zero (upper right). Second, SCEPTRE is applied to analyze the discovery pairs (the "discovery analysis"). The discovery p-values are superimposed over the negative control p-values to ensure that signal is present in the discovery set (lower left), and a volcano plot is created (lower right). **b**, Computational performance metrics of SCEPTRE on the Frangieh (control) and Papalexi (gene expression) data. A complete *trans* analysis was conducted on both datasets. Several metrics are reported, including calibration check run time, calibration check peak memory usage, discovery analysis run time, and discovery analysis peak memory usage.

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. Bock, C. *et al.* High-content crispr screening. *Nature Reviews Methods Primers* **2**, 1–23 (2022).

3. 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).

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

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

6. Alda-Catalinas, C. *et al.* A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program. *Cell Systems* **11**, 25–41.e9 (2020).

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

8. 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).

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

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

11. Morris, J. A. *et al.* Discovery of target genes and pathways at gwas loci by pooled single-cell crispr screens. *Science* eadh7699 (2023).

12. Yao, D. *et al.* Compressed Perturb-seq: highly efficient screens for regulatory circuits using random composite perturbations. *bioRxiv* (2023).

13. 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).

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

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

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

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

18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018).

19. 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).

20. 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).

21. Wessels, H.-h. *et al.* Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. *Nature Methods* (2022).

22. Genga, R. M. *et al.* Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development. *Cell Reports* **27**, 708–718.e10 (2019).

23. Jin, X. *et al.* In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with Autism risk genes. *Science* 791525 (2020).

24. Lalli, M. A., Avey, D., Dougherty, J. D., Milbrandt, J. & Mitra, R. D. High-throughput single-cell functional elucidation of neurodevelopmental disease-associated genes reveals convergent mechanisms altering neuronal differentiation. *Genome Research* **30**, 1317–1331 (2020).

25. Ursu, O. *et al.* Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nature Biotechnology* (2022).

26. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e28 (2022).

27. 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).

28. 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).

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

30. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675–8 (2012).

31. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory dna. *Science* **337**, 1190–5 (2012).

32. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–35 (2015).

33. Schnitzler, G. R. *et al.* Mapping the convergence of genes for coronary artery disease onto endothelial cell programs. *bioRxiv* 2022–11 (2022).

34. Tuano, N. K. *et al.* Crispr screens identify gene targets at breast cancer risk loci. *Genome biology* **24**, 1–23 (2023).

35. Duan, B. *et al.* Model-based understanding of single-cell CRISPR screening. *Nature Communications* **10** (2019). URL http://dx.doi.org/10.1038/s41467-019-10216-x.

36. 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).

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

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

39. 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).

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

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`[37,38]) as well as cells with excessive mitochondrial content and low gene expression.

We additionally generate a synthetic single-cell CRISPR screen dataset to use for benchmarking purposes. 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 assign the set of cell-specific covariates (or technical factors) to each dataset, which we list below. Frangieh co-culture, control, and IFN-γ datasets: number of gene UMIs, number of genes expressed; Papalexi (gene modality): number of gene UMIs, number of genes expressed, biological replicate, and percent of gene transcripts that map to mitochondrial genes; Papalexi (protein modality): number of protein UMIs, biological replicate, and percent of gene transcripts that map to mitochondrial genes; Schraivogel: number of gene UMIs, number of genes expressed, sequencing lane.

Existing methods details

The latter uses the complement set as control cells, since in high-MOI settings there are few cells containing only NT perturbations. On the other hand, SCEPTRE (low-MOI) uses the NT cells as control cells, for reasons discussed above. Aside from this distinction, the high-MOI problem suffers stronger confounding by sequencing depth than the low-MOI problem, but the low-MOI problem suffers from stronger sparsity. Due to these differences, we reasoned that permutations rather than conditional resampling would give better calibration in the low-MOI setting. Finally, SCEPTRE (low-MOI) has built into it a number of novel computational accelerations not present in the originally proposed SCEPTRE (high-MOI).

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 Systematic identification of core analysis challenges).

Sparsity. To explore the impact of sparsity on calibration, we deploy the two sample Wilcoxin 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-Wilcox, we deploy the Wilcoxin test as follows: first, we normalize the gene expressions by dividing the raw counts by the cell-specific library sizes; then, we apply the Wilcoxin 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. Finally, we compute the Wilcoxin p-value in two ways. First, we calculate the asymptotic p-value $p_{\text{asymptotic}}$ by comparing the Wilcoxin 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 Wilcoxin test statistic to be approximately Gaussian. Next, we calculate the exact p-value p_{exact} by (i) computing the Wilcoxin 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 Wilcoxin test statistic on each of these B permuted datasets, yielding a permutation (or "null") distribution of Wilcoxin statistics; and then (iv) calculating the p-value p_{exact} by comparing the original Wilcoxin statistic to the null Wilcoxin statistics[39]. The latter approach, though computationally expensive (due to the slowness of computing the Wilcoxin statistic), yields a much more accurate p-value than the asymptotic approach for lowly expressed genes. Seurat-Wilcox returns the asymptotic p-value $p_{\text{asymptotic}}$ instead

of the exact p-value p_{exact} in virtually all cases.[1]

To study the impact making the above approximation, we plot the asymptotic null distribution of the Wilcoxin statistic (i.e., the standard Gaussian distribution) superimposed on top of the exact null distribution of the Wilcoxin 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 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 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 to their number of treatment cells with nonzero expression. Pairs 1 and 2 from Figure 2a are annotated in Figure 2b.

Finally, to assess directly the impact of sparsity on calibration, we apply Seurat-Wilcox to the IFN-γ negative control data, binning the pairs into five categories based on the number of treatment cells with nonzero expression per pair. (The categories are defined by the intervals [7,10], [11,16], [17,27], [28,46], and [47,121], where the left and right endpoints of an interval indicate the minimum and maximum number of nonzero treatment cells, respectively, for pairs in that interval. The intervals are constructed such that an approximately equal number of pairs falls into each interval.) We observe that as the number of nonzero treatment cells increases, the Seurat-Wilcox p-values converge to uniformity, illustrating that sparsity is a cause of miscalibration of Seurat-Wilcox. We emphasize that filtering for pairs with extremely high gene expression levels leads to the loss of many interesting pairs and thus is not a viable strategy for addressing 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

[1] The `wilcox.test` function on which Seurat-Wilcox 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.

independent experimental replicates (which we label "R1," "R2," and "R3")[2]. 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). We observe clear visual evidence that biological replicate impacts both NTg4 and *FTH1*, creating a confounding effect.

Next, 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 S5; 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 (including biological replicate) included as covariates. We plot the negative control p-values on a QQ plot (Figure 2e). The variant of NB regression with confounders included as covariates demonstrates superior calibration, demonstrating that confounding is an analysis challenge. To reduce 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 15.

Model misspecification. To explore the analysis challenge of model misspecification, we apply NB regression to the Frangieh IFN-γ negative control data. As in Figure 2c (in which we apply Seurat-Wilcox to the Frangieh IFN-γ negative control data), we parti-

[2]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.

tion the pairs into five categories based on the number of nonzero treatment cells per pair. As the number of nonzero treatment cells increases, the NB regression p-values fail to converge to uniformity (in contrast to the Seurat-Wilcox p-values). The key difference between Seurat-Wilcox and NB regression is that the former is a nonparametric method while the latter is parametric method. Thus, we reason that miscalibration of the NB regression p-values likely is due to misspecification of the NB regression model. (We note that miscalibration of the NB regression p-values is not due to confounding, as Seurat-Wilcox, which does not adjust for confounding, is well-calibrated for pairs with high expression levels.)

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." 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.) SCEPTRE is a permutation test that uses as its test statistic the z-score that results from testing the hypothesis $\gamma = 0$ in the model (1). We present the basic SCEPTRE algorithm in Algorithm 1. Several key accelerations speed Algorithm 1 by several orders of magnitude.

Acceleration 1: Score test. First, we use a GLM score test to compute the test statistics $z_{\text{orig}}, z_1, \ldots, z_B$. 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

Algorithm 1: Basic SCEPTRE 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 $H_0 : \gamma = 0$.
2. Permute the X vector B (e.g., $B = 5,000$) times, resulting in permuted vectors $\tilde{X}_1, \ldots, \tilde{X}_B$.
3. For each $i \in \{1, \ldots, B\}$, regress Y onto the matrix $[\tilde{X}_i, Z]$. Test the null hypothesis $H_0 : \gamma = 0$, and label the resulting z-score z_i.
4. Compute a left-tailed (p_{left}), right-tailed (p_{right}), or two-tailed (p_{both}) p-value using the standard permutation test p-value formula:

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

between the gRNA indicator and the gene expression. 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 $M(Y - m\hat{u})$ are a matrix and vector, respectively, that depend on the fitted means $\hat{\mu}$, gene expressions Y, and estimated size $\hat{\theta}$:

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

The score statistic (3) is asymptotically equivalent to the Wald or likelihood ratio 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 (with score statistic; Algorithm 2) exploits this useful property of the score statistic to accelerate the basic SCEPTRE algorithm.

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

Algorithm 2: SCEPTRE (with score statistic) 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 standard permutation test p-value formula from step 4 of Algorithm 1.

classical solution 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 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.[3] 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[40]) for maximum speed.

Acceleration 3: Adaptive permutation testing. Computing a large number of permutation resamples 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 compute the p-value of a given gene-gRNA pair out to a small number (e.g., $B_1 = 500$) 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 larger number ($B_2 = 5,000$) of fresh

[3] 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.

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

resamples and compute the p-value using this second set of resamples. 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.

Acceleration 4: Skew-normal fit. The null distribution of the test statistics $z_1, \ldots, z_B$ converges to a standard Gaussian distribution. Thus, to compute a more precise p-value using a small number of permutations, we fit a skew-normal distribution to the set of null statistics. We then compute a p-value by evaluating the tail probability of the fitted skew-normal distribution at the observed test staitstic z_{obs}. If the skew-normal fit to the null statistics is poor (an event that happens rarely), we instead return the standard permutation test p-value. We fit the skew-normal distribution via a method of moments estimator and evaluate the skew-normal tail probability via the C++ Boost library. Thus, all operations involving the skew-normal distribution are fast.

Statistical robustness property of SCEPTRE

SCEPTRE posesses a robustness property that we term "CAMP," or "confounder adjustment via marginal permutations." We state CAMP in a slightly more formal way here. If at least one of the following conditions holds, then the left-, right-, and two-tailed SCEPTRE p-values are valid: (i) the gRNA is unconfounded (i.e., the vector of technical factors Z_i contains all possible confounders, and Z_i is independent of X_i); (ii) the NB GLM (1) is correctly specified up to the size parameter θ, and the effective sample size is sufficiently large. CAMP imbues SCEPTRE with two considerable advantages relative a standard

NB GLM. First, SCEPTRE always yields valid inference when the perturbation is unconfounded, even if the NB model is arbitrarily misspecificed or the effective sample size is small. Second, when confounding is non-negligible, SCEPTRE 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 difficult problem of NB size parameter estimation[28,29]. These two improvements enable SCEPTRE to address the core single-cell CRISPR screen analysis challenges in theory.

Simulation study details

We conduct a simulation study (Figure S6) 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 standard 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 standard 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 the methods in each of the four settings (i.e., confounded versus unconfounded, correct versus incorrect specification of the size parameter; Figure S6) on a QQ plot. We seek to show that SCEPTRE maintains calibration in all settings, while the standard permutation test and NB regression break down under confounding and incorrect specification of the size parameter, respectively.

Positive control analysis

We group together gRNAs that target the same genomic location and refer to these grouped gRNAs as "gRNA groups[5]". 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.

ChIP-seq enrichment analysis

We obtained ChIP-seq data for CD14+ monocyte cultures with MCSF (10ng/ml) and stimulated with IFN-gamma (100U/ml) for 24 hours (DOI: 10.1016/j.immuni.2013.08.009). We filtered peaks by calling the top 25% by enrichment score. We defined the promoter region of a gene as 5kb upstream or downstream of the TSS. After converting the ChIP-seq data and the promoter regions for all genes into `Granges` objects, we found those peaks that intersected with any part of a promoter region via the `join_overlap_left` function from the `plyranges` package in R. To determine if SCEPTRE's discoveries were consistent with the ChIP-seq data, we computed odds ratios and their corresponding p-values via a fisher exact test on the contingency table consisting of the downstream genes SCEPTRE found to be associated with the perturbed gene and genes whose promoter regions overlapped with a ChIP-seq peak. We carried out this analysis for the STAT1 and IRF1 genes.

Methods not included in the benchmarking analysis

Several methods that recently have been proposed for single-cell CRISPR screen analysis are not included in our benchmarking study. First, guided sparse factor analysis (GSFA; introduced by Zhou et al.[36]) couples factor analysis to differential expression analysis to infer the effects of perturbations on gene modules and individual genes. GSFA is a Bayesian method, returning a posterior inclusion probability instead of a p-value for each test of association. Given that the methods that we study in this work are frequentist (and thus return a p-value), we deprioritize GSFA for benchmarking. Next, Normalisr (proposed by

Wang[17]) is a method for single-cell differential expression, co-expression, and CRISPR screen analysis. Normalisr attempts to non-linearly transform the gene expression counts to Gaussianity and then model the transformed counts via a linear model. We were unable to locate an example low-MOI single-cell CRISPR screen analysis in the Normalisr Github repository (although gene co-expression, case-control differential expression, and high-MOI CRISPR screen examples are available). Given this, and given the complexity of the Normalisr codebase, we deprioritize Normalisr for benchmarking.

Code, software, data, and results repositories

The code for this paper is spread across nine Github repositories. Navigate to `https://github.com/Katsevich-Lab/sceptre2-manuscript` (i.e., the second Github repository of those listed) for instructions on reproducing the analyses reported in this manuscript.

1. The `sceptre` package implements the SCEPTRE method. The repository contains detailed tutorials and examples.

   `https://katsevich-lab.github.io/sceptre/`

2. The `sceptre2-manuscript` repository contains code to reproduce all analyses reported in this paper.

   `https://github.com/Katsevich-Lab/sceptre2-manuscript`

3. The `lowmoi` package implements the existing single-cell CRISPR screen analysis methods. (Methods originally written in Python are implemented via `reticulate`).

   `https://github.com/Katsevich-Lab/lowmoi`

4. The `undercover-grna-pipeline` repository contains the Nextflow pipeline to carry out negative control benchmarking analysis.

   `https://github.com/Katsevich-Lab/undercover-grna-pipeline`

5. The `pc-grna-pipeline` repository contains the Nextflow pipeline to carry out the positive control benchmarking analysis.

   `https://github.com/Katsevich-Lab/pc-grna-pipeline`

6. The `ondisc` package implements data structures that we use to store the single-cell expression data.

   `https://github.com/timothy-barry/ondisc`

7. The `import-frangieh-2021` repository imports and processes the Frangieh data.

   https://github.com/Katsevich-Lab/import-frangieh-2021

8. The `import-papalexi-2021` repository imports and processes the Papalexi data.

   https://github.com/Katsevich-Lab/import-papalexi-2021

9. The `import-schraivogel-2020` repository imports and processes the Schraivogel data.

   https://github.com/Katsevich-Lab/import-schraivogel-2020

Next, the uniformly processed single-cell CRISPR screen data (stored in `ondisc` format) are available in the following directory: https://www.dropbox.com/sh/jekmk1v4mr4kj3b/AAAhznGqk-TIZKhW40xiU6ORa?dl=0. Finally, all results are stored in the following directory: https://www.dropbox.com/sh/yuubaoro3k75c61/AACPi9LDjY0B1pxwMxUMSoTca?dl=0.

Author contributions

EK identified the research problem. TB, KM, and EK performed the analyses. TB and EK developed the method. TB implemented the low-MOI `sceptre` software. KR and EK supervised the project. TB and EK wrote the manuscript with assistance from KM and KR.

Acknowledgements

We thank Hugh MacMullan and Gavin Burris for extensive support in using the Wharton high performance computing cluster (HPCC). We thank Sophia Lu for preliminary work on benchmarking existing methodologies. We thank Ziang Niu for carrying out analyses related to modeling the SCEPTRE resampling distributions. We thank John Morris for help in designing the computational experiments and providing feedback on an early draft. We thank Stephanie Hicks, Kasper Hansen, and the Hicks and Hansen Labs at Johns Hopkins University for detailed feedback on the exposition and content of the paper. We thank Chris Frangieh for providing instructions on downloading and processing the Frangieh data and using the MIMOSCA method. Finally, we thank Rahul Satija for clarifying several points about the Papalexi data.

Supplementary Tables and Figures

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 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, but we focus exclusively on the 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)	25 (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-Wilcox	No	No	Yes
Seurat-NB	No	No	No
t-test	Yes	No	No
MAST	No	No	No
KS test	No	No	Yes
NB regression (with covariates)	No	Yes	No
Standard permutation test	Yes	No	Yes
SCEPTRE	**Yes**	**Yes**	**Yes**

Table S3: **Analysis challenges addressed by each method**. Each cell indicates 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 we could not determine which analysis challenge(s) MIMOSCA addresses.

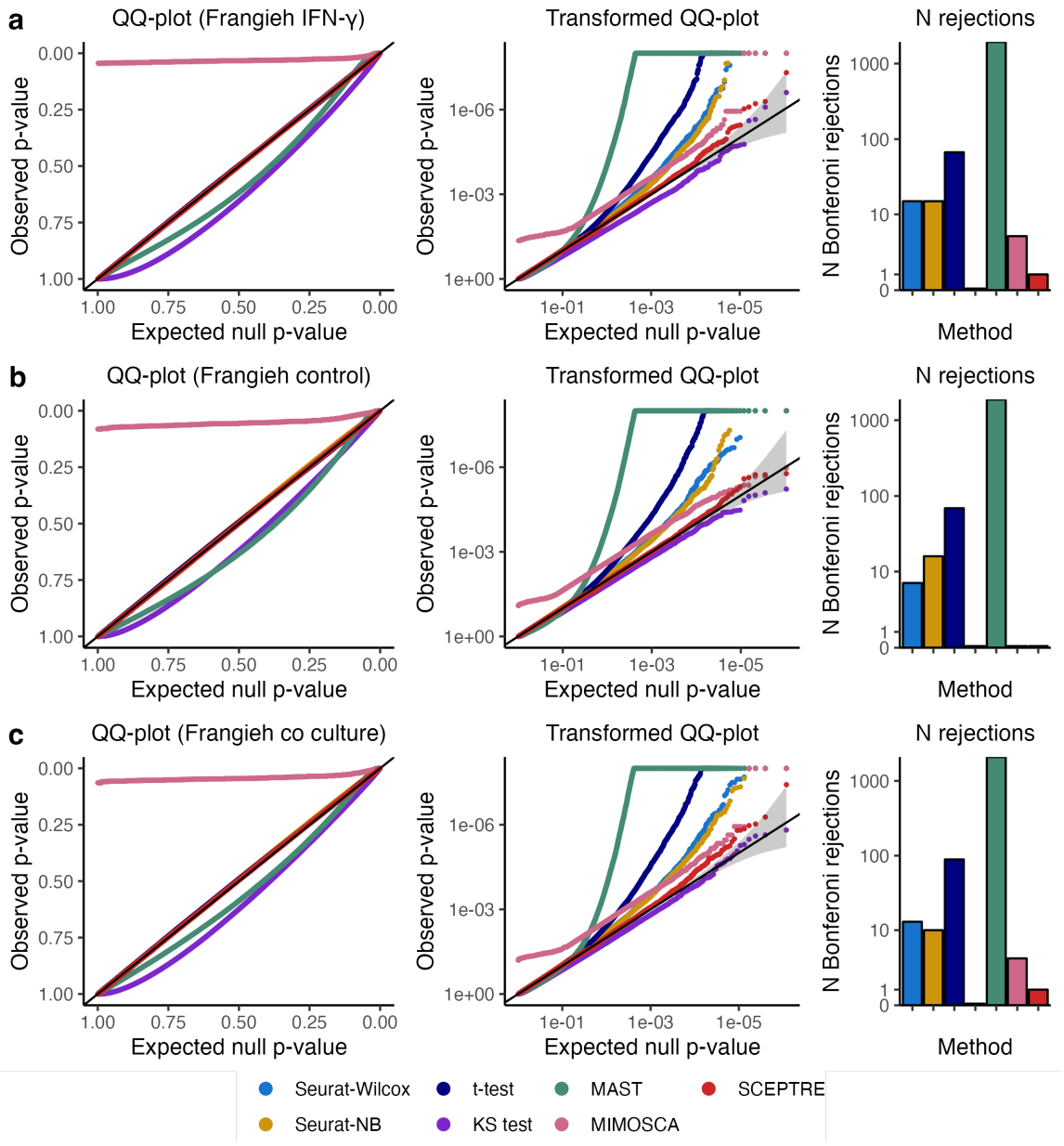


Figure S1: **Calibration results for all methods on Frangieh IFN-γ, Frangieh control, and Frangieh co-culture negative control data.** Left, untransformed QQ plots; middle, negative log-10 transformed QQ plots; right; number of false rejections after a Bonferroni correction at level 0.1.

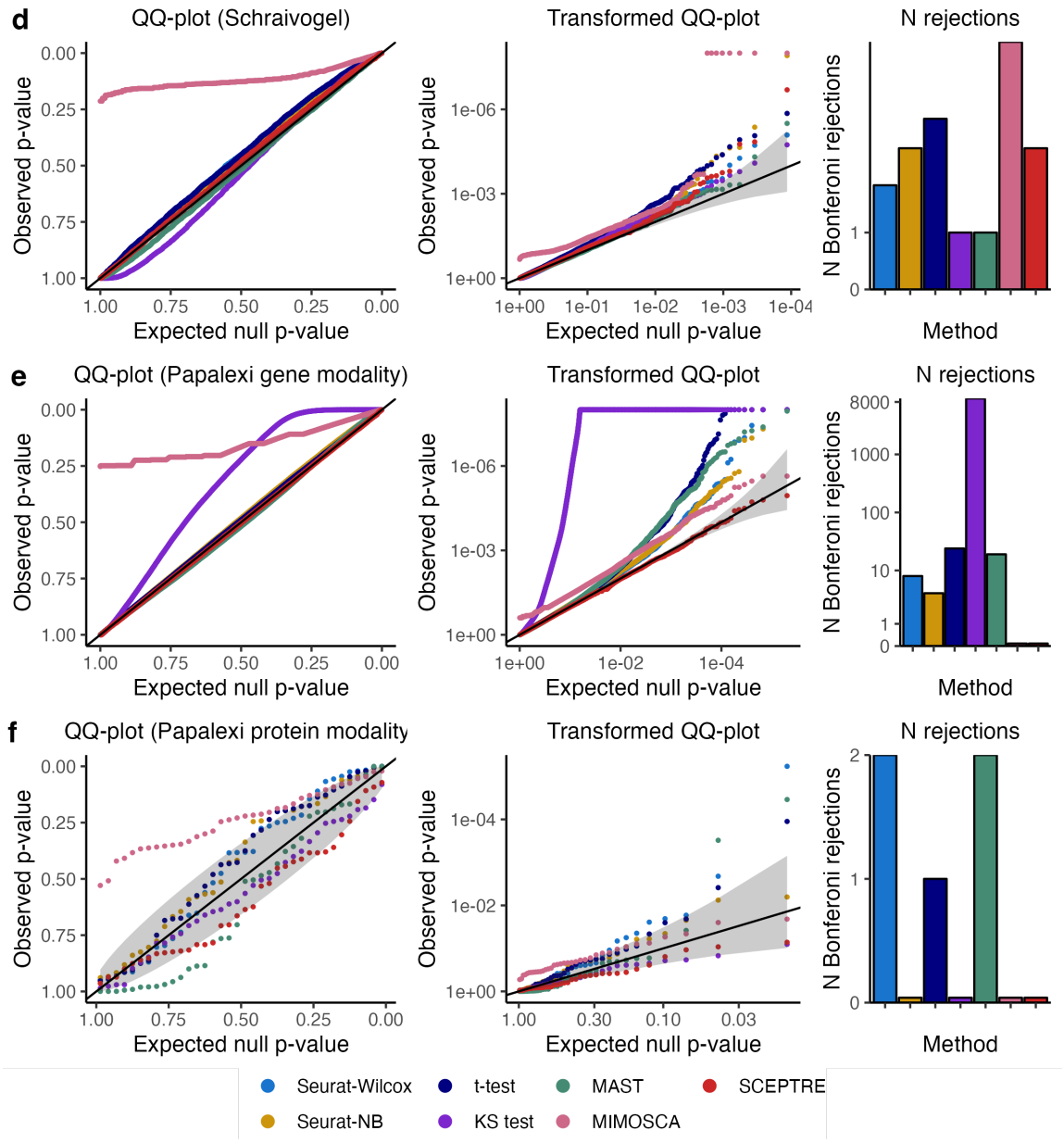


Figure S2: **Calibration results for all methods on Schraivogel, Papalexi (gene modality), and Papalexi (protein modality) negative control data.** Interpretation is the same as in Figure S1.

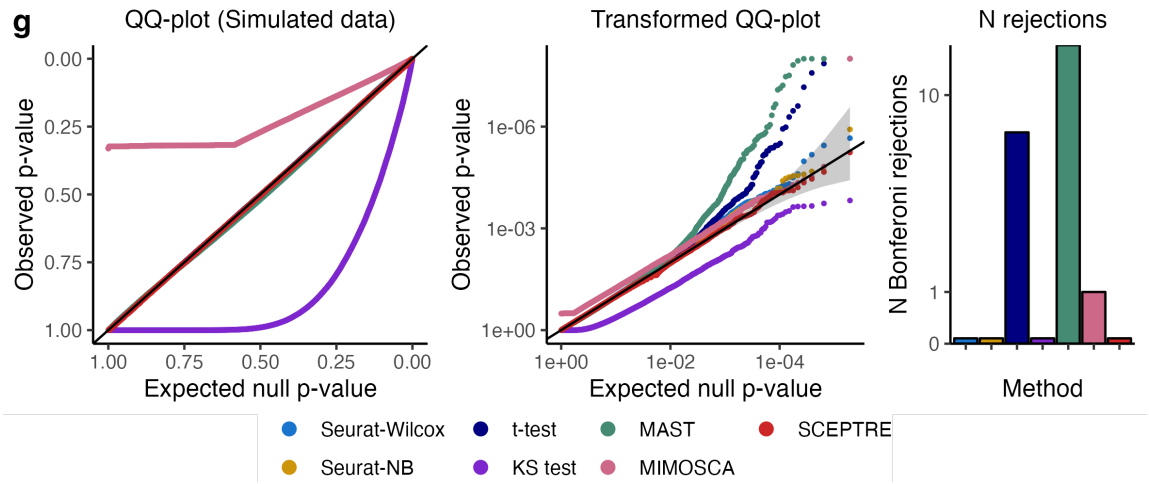


Figure S3: **Calibration results for all methods on simulated data.** Interpretation is the same as in Figure S1.

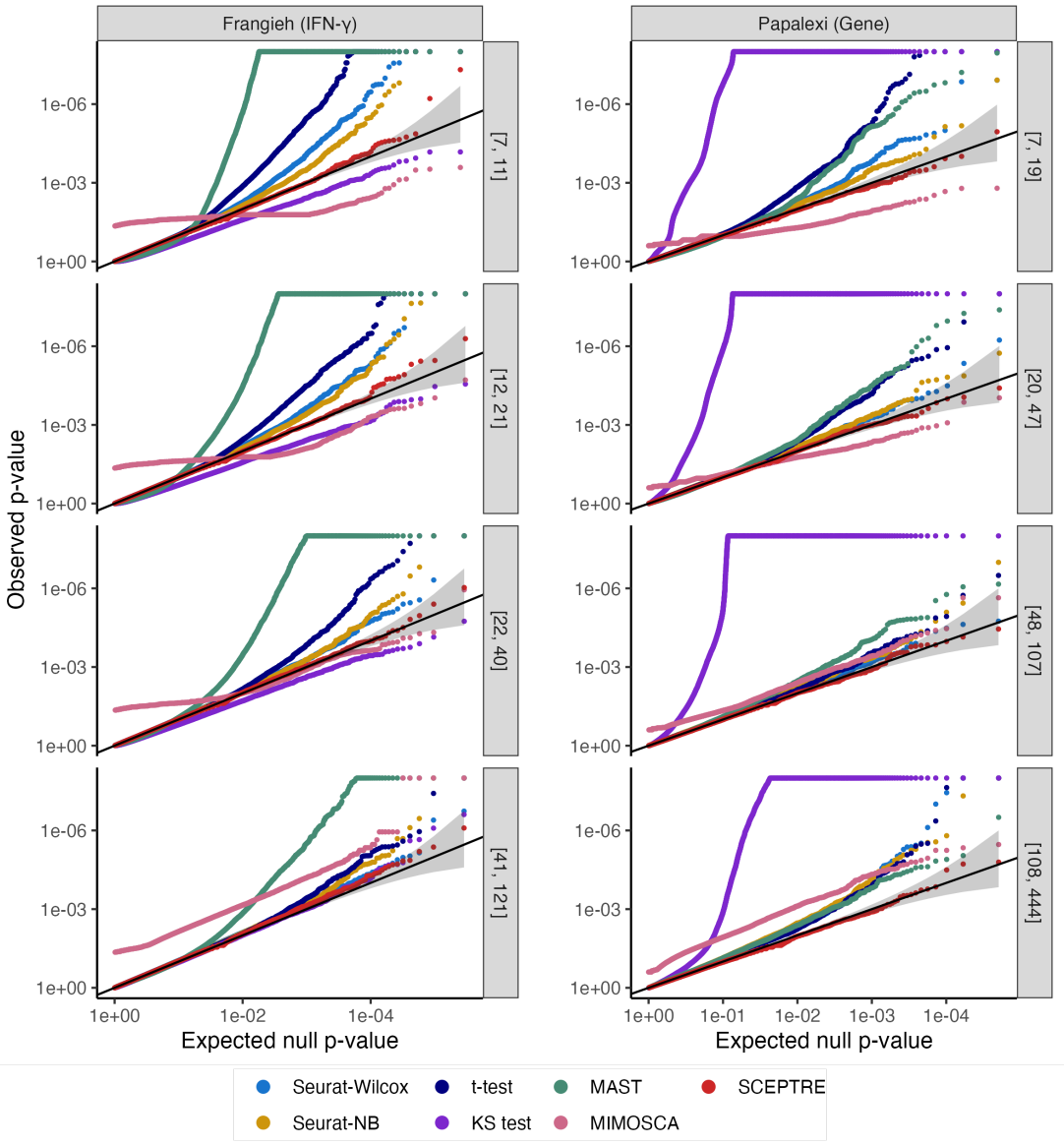


Figure S4: **Calibration results for all methods on Frangieh IFN-γ and Papalexi (gene modality) datasets, stratified by effective sample size**. Negative control gene-gRNA pairs are partitioned into four bins of approximately equal size based on the number of treatment cells with nonzero expression in a given pair. The interval on the right-hand side of each panel indicates the minimum and maximum number of treatment cells with nonzero gene expression for pairs in that bin. Some methods (e.g., Seurat-Wilcox on the Frangieh IFN-γ data) exhibit better calibration as the number of treatment cells with nonzero expression increases (i.e., as sparsity decreases).

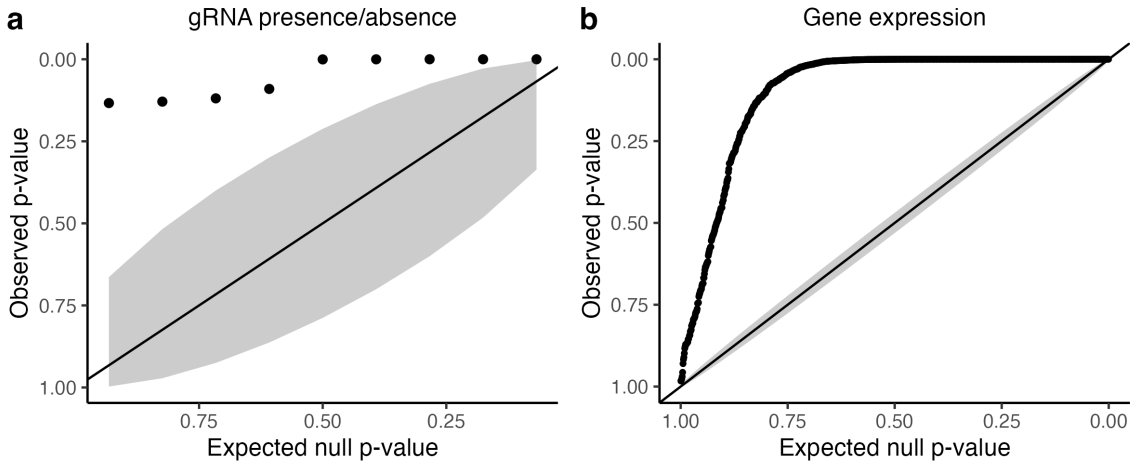


Figure S5: **Confounding due to biological replicate on the Papalexi (gene modality) data.** Left, QQ plot of p-values for tests of association between the gRNA indicator and biological replicate for each NT gRNA (tests carried out using Fisher's exact test). Right, p-values for tests of association 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.

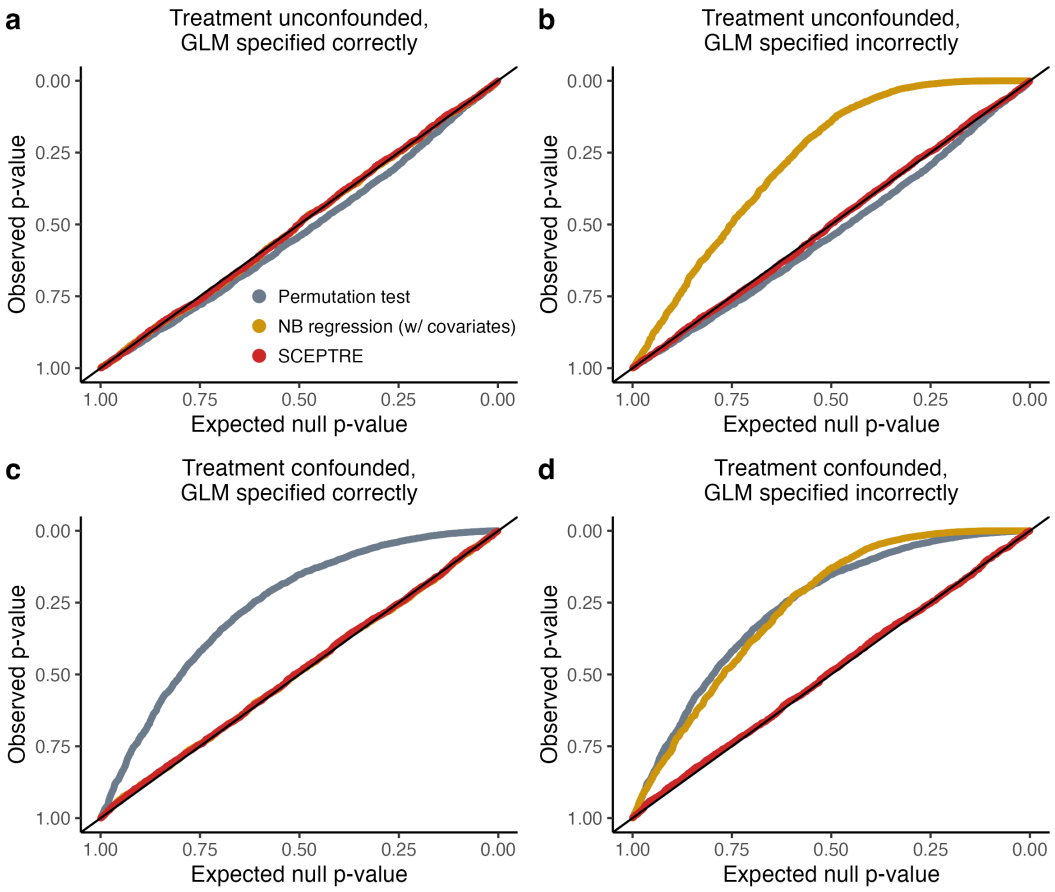


Figure S6: **Demonstration of the CAMP ("confounder adjustment via marginal permutations") phenomenon on realistic semi-synthetic data**. Application of a standard permutation test, NB regression, and SCEPTRE to realistic semi-synthetic data generated under two conditions: confounded and unconfounded. 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 the NB regression model is arbitrarily misspecified or (ii) the problem is confounded and the sparsity is high. Details of the simulation study are given in Section Simulation study details.

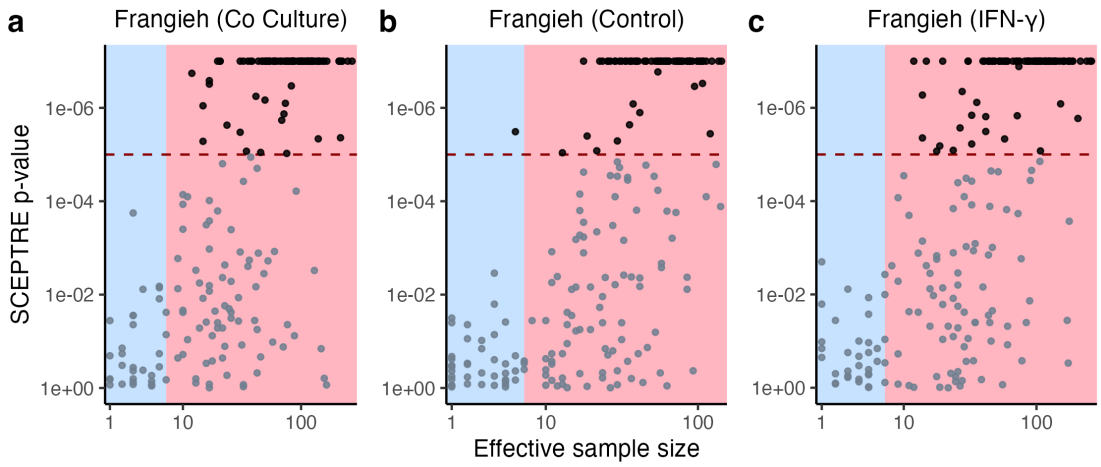


Figure S7: **SCEPTRE's power to detect associations increases as effective sample size increases**. **a-c**, SCEPTRE p-value (truncated at 10^{-6}) versus effective sample size for each pair on the Frangieh co-culture (**a**), control (**b**), and IFN-γ (**c**) positive control data. The horizontal dashed line is drawn at 10^{-5}, demarcating a highly significant discovery. SCEPTRE makes only one rejection at a highly significant level on pairs for which the effective sample size less than seven (blue region).