Timothy Barry
Technologies for single-cell CRISPR screen data analysis: a proposal
Advisors: Eugene Katsevich, Kathryn Roeder
Committee members: Jing Lei, Jian Ma, Aaditya Ramdas

Abstract

CRISPR is a genome engineering technology that has enabled scientists to precisely edit human and nonhuman genomes. Single-cell CRISPR screens couple CRISPR genome engineering to single-cell sequencing, linking genetic perturbations in individual cells to changes in gene expression and other cellular phenotypes. Despite their enormous potential for unraveling mechanisms underlying disease, single-cell CRISPR screens pose considerable statistical and computational challenges. The broad objective of this thesis is to develop statistically rigorous and computationally efficient tools for the analysis of single-cell CRISPR screen data. This is a multidisciplinary line of work that spans statistics, computer science, and biotechnology.

First, we introduce a new statistical method — "GLM-EIV" (GLM-based errors-in-variables) — for estimating the effect sizes of CRISPR perturbations on gene expressions in single-cell CRISPR screen experiments. GLM-EIV extends the classical Gaussian measurement error model to responses and noisy predictors that are exponential family-distributed and potentially impacted by the same set of confounding variables. Second, we propose "SCEPTRE" (pronounced "scepter"), a custom implementation of the conditional randomization test tailored to single-cell CRISPR screen data. SCEPTRE simultaneously adjusts for confounder presence and ensures robustness to expression model misspecification, significantly outperforming existing methods on sensitivity and specificity metrics. Third, we develop a computational framework for out-of-core and cloud-scale distributed computing on single-cell (including single-cell CRISPR screen) data. The framework is powered by several new, highly efficient algorithms and functional data structures. For example, one of the processing steps involves transposing a large, sparse matrix out-of-core; we introduce a generalized radix sort algorithm for this purpose and prove that the algorithm is asymptotically optimal in running time, memory space, and disk space.

We will advance this line of work with two additional projects, one applied and one methodological. First, we will extend our SCEPTRE methodology to apply to a much broader class of single-cell CRISPR screen datasets and also improve its computational performance. Second, we will introduce a general selective inference framework for boosting the sensitivity and specificity of hypothesis testing procedures on "multi-axis high-throughput data," i.e. data on which many variables are measured along multiple distinct axes (e.g., single-cell CRISPR screen data, eQTL data, etc.).

Background and motivation

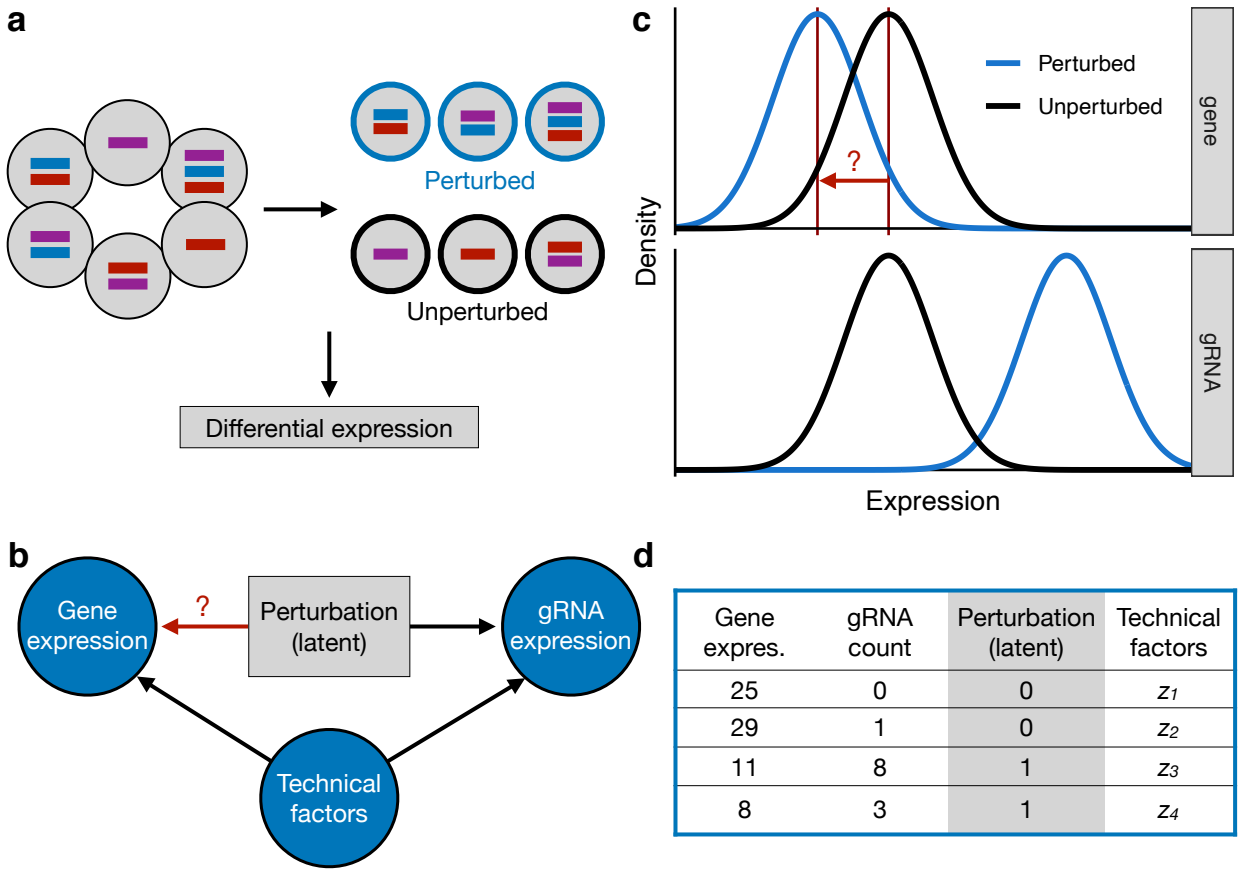
There are several broad classes of single-cell CRISPR screen assays, each suited to answer a different set of biological questions. Most of our completed work has focused on so-called

high-multiplicity of infection (MOI) single-cell CRISPR screens, which we motivate and describe here. The human genome consists of genes, enhancers (segments of DNA that regulate the expression of one or more genes), and other genomic elements (that are not of relevance to the current work). GWAS have revealed that the majority ($> 90\%$) of variants associated with diseases lie outside genes and inside enhancers (Gallagher et al. 2018). These noncoding variants are thought to contribute to disease by modulating the expression of one or more disease-relevant genes. Scientists do not know the gene (or genes) through which most noncoding variants exert their effect, limiting the interpretability of GWAS results. A central open challenge in genetics, therefore, is to link enhancers that harbor GWAS variants to the genes that they target at genome-wide scale (Morris et al. 2021).

High MOI single-cell CRISPR screens are the most promising biotechnology for solving this challenge. High MOI single-cell CRISPR screens combine CRISPR interference (CRISPRi) — a version of CRISPR that represses a targeted region of the genome — with single-cell sequencing. The experimental protocol is as follows. First, the scientist develops a library of several hundred to several thousand CRISPRi perturbations, each designed to target a candidate enhancer for repression. The scientist then cultures tens or hundreds of thousands of cells and delivers the CRISPRi perturbations to these cells. The perturbations assort into the cells randomly, with each cell receiving on average 10-40 distinct perturbations. Conversely, a given perturbation enters about 0.1-2% of cells.

After waiting several days for CRISPRi to take effect, the scientist profiles each cell's transcriptome (i.e., its gene expressions) and the set of perturbations that it received. Finally, the scientist conducts perturbation-to-gene association analyses. Figure 1a depicts this process schematically, with colored bars (blue, red, and purple) representing distinct perturbations. For a given perturbation (e.g., the perturbation represented in blue), the scientist partitions the cells into two groups: those that received the perturbation (top) and those that did not (bottom). Next, for a given gene, the scientist runs a differential expression analysis across the two groups of cells, producing an estimate for the magnitude of the gene expression change in response to the perturbation. If the estimated change in expression is large, the scientist can conclude that the enhancer *targeted* by the perturbation exerts a strong regulatory effect on the gene. This procedure is repeated for a large set of preselected perturbation-gene pairs. The enhancer-by-enhancer approach is valid because the perturbations assort into cells approximately independently of one another.

The genomics literature has produced a few applied methods for single-cell CRISPR screen analysis (Gasperini et al. 2019; Xie et al. 2019). Gasperini et al. applied negative binomial GLMs (as implemented in the Monocle software; (Trapnell et al. 2014)) to carry out the differential expression analysis described above. Xie et al., by contrast, applied chi-squared-like tests of independence for this purpose. Both of these approaches have limitations: the former is not robust to misspecification of the gene expression model, and the latter is unable to correct for the presence of technical confounders. Our goal, broadly, is to develop statistically rigorous and computationally efficient methods for differential expression testing and estimation in single-cell CRISPR screen analysis.



Gene expres.	gRNA count	Perturbation (latent)	Technical factors
25	0	0	z_1
29	1	0	z_2
11	8	1	z_3
8	3	1	z_4

Figure 1: **Experimental design and analysis challenges**: **a,** Experimental design. For a given perturbation (e.g., the perturbation indicated in blue), we partition the cells into two groups: perturbed and unperturbed. Next, for a given gene, we conduct a differential expression analysis across the two groups, yielding an estimate of the impact of the given perturbation on the given gene. **b,** DAG representing all variables in the system. The perturbation (latent) impacts both gene expression and gRNA expression; technical factors act as confounders, also impacting gene and gRNA expression. The target of estimation is the effect of the perturbation on gene expression. **c,** Schematic illustrating the "background read" phenomenon. Due to errors in the sequencing and alignment processes, unperturbed cells exhibit a nonzero gRNA count distribution (bottom). The target of estimation is the change in mean gene expression in response to the perturbation (top). **d**, Example data on four cells for a given perturbation-gene pair. Note that (i) the perturbation is unobserved, and (ii) the gene and gRNA data are discrete counts.

Analysis challenges and statistical model

High MOI single-cell CRISPR screens present several statistical challenges, four of which we highlight here. Throughout, we consider a single perturbation-gene pair. First, the "treatment" variable — i.e., the presence or absence of a perturbation — cannot be directly observed. Instead, perturbed cells transcribe molecules called *guide RNAs* (or *gRNAs*) that serve as indirect proxies of perturbation presence. We must leverage these gRNAs to

impute (explicitly or implicitly) perturbation assignments onto the cells (Figure 1b). Second, "technical factors" — sources of variation that are experimental rather than biological in origin — impact the measurement of both gene and gRNA expressions and therefore act as confounders (Figure 1b). Third, the gene and gRNA data are sparse, discrete counts. Consequently, classical statistical approaches that assume Gaussianity or homoscedasticity are inapplicable. Finally, sequenced gRNAs sometimes map to cells that have not received a perturbation. This phenomenon, which we call "background contamination," results from errors in the sequencing and alignment processes. The marginal distribution of the gRNA counts is best conceptualized as a mixture model (Figure 1c; Gaussian distributions used for illustration purposes only). Unperturbed and perturbed cells both exhibit nonzero gRNA count distributions, but this distribution is shifted upward for perturbed cells. Figure 1d shows example data on four (of possibly tens or hundreds of thousands of) cells. The analysis objective is to leverage the gene expressions and gRNA counts to estimate the effect of the (latent) perturbation on gene expression, accounting for the technical factors.

We propose to model the single-cell CRISPR screen data-generating process using a pair of GLMs. Let $n \in \mathbb{N}$ be the number of cells assayed in the experiment. Consider a single perturbation and a single gene. For cell $i \in \{1, \ldots, n\}$, let $p_i \in \{0, 1\}$ indicate perturbation presence or absence; let $m_i \in \mathbb{N}$ be the number of gene transcripts sequenced; let $g_i \in \mathbb{N}$ be the number of gRNA transcripts sequenced; let $d_i^m \in \mathbb{N}$ be the number of gene transcripts sequenced across *all* genes (i.e., the library size or sequencing depth); let d_i^g be the gRNA library size; and finally, let $z_i \in \mathbb{R}^{d-1}$ be the cell-specific technical factors (e.g., sequencing batch, percent mitochondrial reads, etc.) The letters "m," "g", and "d" stand for "mRNA," "gRNA," and "depth," respectively.

Building on the work of several previous authors (Townes et al. 2019; Hafemeister et al. 2019), Sarkar et al. 2021 proposed a simple strategy for modeling single-cell gene expression data, which, in the framework of negative binomial GLMs, is equivalent to using the log-transformed library size as an offset term. Sarkar and Stephens' framework enjoys strong theoretical and empirical support; therefore, we generalize their approach to model *both* gene and gRNA modalities in single-cell CRISPR screen experiments. To this end, we assume that the gene expression counts are given by

$$m_i | (p_i, z_i, d_i^m) \sim \mathrm{NB}_{s^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + \log(d_i^m), \tag{1}$$

where (i) $\mathrm{NB}_{s^m}(\mu_i^m)$ is a negative binomial distribution with mean μ_i^m and known size parameter s^m; (ii) $\beta_0^m \in \mathbb{R}, \beta_1^m \in \mathbb{R}$, and $\gamma_m \in \mathbb{R}^{d-1}$ are unknown parameters; and (iii) $\log(d_i^m)$ is an offset term. Similarly, we model the gRNA counts by

$$g_i | (p_i, z_i, d_i^g) \sim \mathrm{NB}_{s^g}(\mu_i^g); \quad \log(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + \log(d_i^g), \tag{2}$$

where μ_i^g, s^g, β_0^g, β_1^g, γ_g, and d_i^g are analogous. We use a negative binomial GLM to model the gRNA counts as well as the gene expressions because the gRNA transcripts are generated via the same biological mechanism as the gene transcripts (Datlinger et al. 2017; Hill et al. 2018). Finally, we model the marginal perturbation probability as $p_i \sim \mathrm{Bern}(\pi)$, where $\pi \in (0, 1/2]$, and p_i is unobserved.

The log-transformed sequencing depth $\log(d_i^m)$ is included as an offset term in (1) so that $\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i$ can be interpreted as a relative expression. Exponentiating both sides of

(1) reveals that the mean gene expression μ_i^m of the ith cell is $\exp\left(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i\right) d_i^m$. Because d_i^m is the sequencing depth, $\exp\left(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i\right)$ is the *fraction* of all transcripts sequenced in the cell produced by the gene under consideration. The target of inference β_1^m is the log fold change in expression in response to the perturbation, controlling for the technical factors. Fold change in this context is the ratio of the mean gene expression in perturbed cells to the mean gene expression in unperturbed cells. Hence, $\exp(\beta_1^m) = 1$ (i.e., $\beta_1^m = 0$) indicates no change in expression, whereas $\exp(\beta_1^m) > 1$ (i.e., $\beta_1^m > 0$) and $\exp(\beta_1^m) < 1$ (i.e., $\beta_1^m < 0$) indicate an increase and decrease in expression, respectively.

Completed (or mostly completed) works

Having introduced the problem, we turn to a brief summary of three completed (or mostly completed) projects on statistical and computational methods for single-cell CRISPR screen analysis (and more generally single-cell analysis). The following discussion is based partly on the works Barry et al. 2021 and Barry et al. 2022.

Exponential family measurement error models for single-cell CRISPR screens

In Barry et al. 2022 we introduce a model and associated method — "GLM-EIV" (generalized linear model with errors-in-variables) — for estimating (with confidence) the log fold change parameter β_1^m introduced above. First, we generalize the the model $(1-2)$ to arbitrary exponential family response distributions and link functions:

$$\begin{cases} m_i | (p_i, z_i, d_i^m) \sim \text{GLM}\left(p_i, z_i, d_i^m\right) \\ g_i | (p_i, z_i, d_i^g) \sim \text{GLM}\left(p_i, z_i, d_i^g\right) \\ p_i \sim \text{Bern}(\pi). \end{cases} \tag{3}$$

This model can be seen as an extension of the classical Gaussian measurement error model. Next, we develop methods for rapid estimation and inference in the model (3). We derive an EM algorithm to estimate the parameters of (3); interestingly, the M-step involves fitting two weighted GLMs and thus does not require direct maximization of a complicated objective function. After fitting the model, we perform inference on the estimated parameters. The easiest approach, given the complexity of the log likelihood, would be to run a bootstrap. This strategy, however, is prohibitively slow, as the data are large and the EM algorithm is iterative. Therefore, we derive an analytic formula for the asymptotic observed information matrix using Louis's Theorem (Louis 1982). Leveraging this analytic formula, we can calculate standard errors quickly, enabling us to perform inference in practice on real, large-scale data.

We additionally devise a strategy to produce a highly accurate pilot estimate of the true parameters, enabling us to run the algorithm once and converge upon the MLE within a few iterations. The strategy involves layering several statistical tricks — that likely are of independent utility — on top of one another. Finally, we develop a computational infrastructure

to apply GLM-EIV to large-scale, single-cell CRISPR screen data. The infrastructure leverages `Nextflow`, a programming language that facilitates building data-intensive pipelines, and `ondisc`, an `R/C++` package that we developed (in a separate project) to facilitate large-scale computing on single-cell data. Overall, the statistical accelerations and computational infrastructure make the deployment of GLM-EIV to large-scale single-cell CRISPR screen quite feasible.

We compare GLM-EIV to the so-called "thresholding method," which imputes perturbations assignments onto the cells simply by thresholding the gRNA counts. Specifically, let $c \in \mathbb{N}$ be a given threshold. The thresholding method sets the imputed perturbation assignment $\hat{p}_i \in \{0, 1\}$ to 0 if $g_i < c$ and to 1 if $g_i \geq c$. Next, the thresholding method fits the GLM (1) using $\hat{p}_i$ in place of the latent variable p_i. We show through empirical and theoretical analyses that the thresholding method poses several fundamental difficulties. Most importantly, (i) the gRNA count distributions do not imply a clear threshold selection strategy, and (ii) the choice of threshold exerts a major impact on the estimate $\hat{\beta}_1^m$ of β_1^m.

We compare GLM-EIV to the thresholding method on real and simulated data. GLM-EIV outperformed the thresholding method on simulated data by a substantial margin. On real data the two methods produced more concordant results. When we increased the difficulty of the problem by generating partially synthetic datasets with increased background contamination, the two methods diverged, with GLM-EIV demonstrating superior performance. Our conclusion therefore is as follows: in low background contamination settings either GLM-EIV or the thresholding method can be used. An advantage to GLM-EIV, however, is that it obviates the need to tune a threshold. In high background contamination settings, by contrast, GLM-EIV (or some similar method that accounts for measurement error) is required.

SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis

Next, in Barry et al. 2021 we tackle the problem of *hypothesis testing* (as opposed to *estimation*) in single-cell CRISPR screen experiments. *Testing* for an association between a CRISPR perturbation and the expression of a gene (as opposed to *estimating* the strength of such an association) poses a somewhat different set of challenges. Importantly, we typically test tens or hundreds of thousands of associations with the objective of producing a discovery set that controls the false discovery rate (FDR). The rejected hypotheses have p-values in the extreme tail of the p-value distribution. To avoid making excess false positive discoveries, we must ensure that our testing procedure is excellently calibrated. Our focus in the hypothesis testing setting, therefore, is largely on p-value calibration (and secondarily on power). Crucially, single-cell CRISPR screen datasets typically come equipped with "non-targeting" or "negative" control gRNAs; these are gRNAs designed to have *no* impact on the expression of any gene. Non-targeting gRNAs thereby put us under the null hypothesis of "no regulatory relationship" and thus can be used to assess the calibration of a hypothesis testing method.

We consider two recent single-cell CRISPR screen datasets – one published by Gasperini et al. 2019 and the other by Xie et al. 2019 – to illustrate these ideas. Both Gasperini et al. and Xie et al. encountered substantial difficulties in calibrating tests of association between

candidate enhancers and genes. Gasperini et al. computed p-values using a DESeq2-inspired negative binomial regression analysis implemented in Monocle2 (Trapnell et al. 2014), and Xie et al. computed p-values using Virtual FACS, a nonparametric, chi-squared test-like method proposed by these authors. Gasperini et al. assessed calibration by pairing each of 50 non-targeting (or negative) control gRNAs with each protein-coding gene. These "null" p-values exhibited inflation, deviating substantially from the expected uniform distribution. To assess the calibration of Virtual FACS in a similar manner, we constructed a set of *in silico* negative control pairs of genes and gRNAs on the Xie et al. data. The resulting p-values were likewise miscalibrated, with some pairs exhibiting strong conservative bias and others strong liberal bias (Figure 2).

Technical covariates, such as sequencing depth and batch, induce a correlation between gRNA detection probability and gene expression, even in the absence of a regulatory relationship (Figure 2b-c). This confounding effect can lead to severe test miscalibration and is especially problematic for traditional nonparametric approaches (such as Virtual FACS), which implicitly (and incorrectly) treat cells symmetrically with respect to confounders. Parametric regression approaches, such as the negative binomial regression approach of Gasperini et al., are the most straightforward way to adjust for confounders. However, parametric methods rely heavily on correct model specification, a challenge in single-cell analysis given the heterogeneity and complexity of the count data.

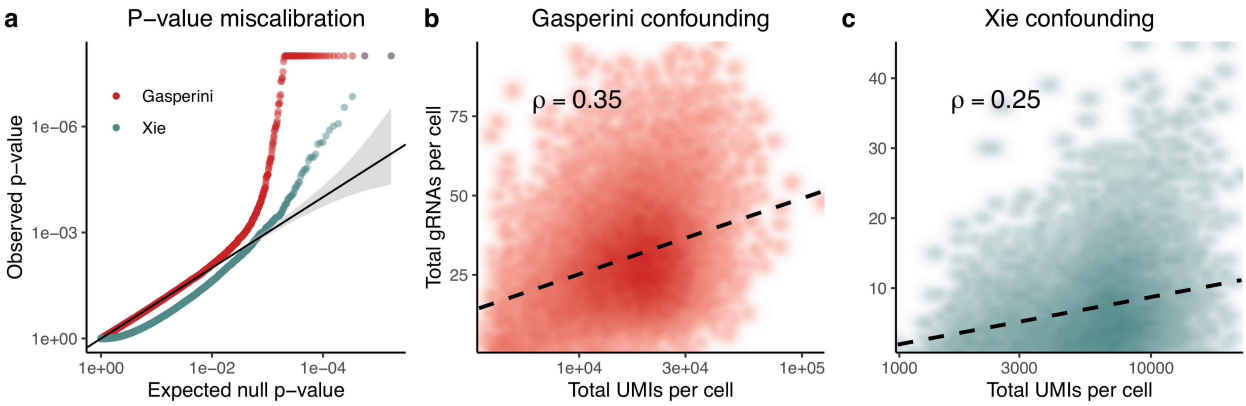


Figure 2: ***P*-value miscalibration**. **a**, QQ-plot of negative control p-values produced by Gasperini et al. (red; downsampled for visualization) and Xie et al. (gray-green). These p-values deviate substantially from the expected uniform distribution, indicating test miscalibration. **b-c**, Technical factors, such as sequencing depth (i.e., total UMIs per cell) and batch, impact gRNA detection probability and observed gene expression levels in both Gasperini et al. (b) and Xie et al. (c) data. Thus, technical factors act as confounders, differentiating CRISPR screens from traditional differential expression applications.

To overcome these challenges, we propose SCEPTRE (analysis of Single-CEll PerTurbation screens via conditional REsampling; pronounced "scepter"). SCEPTRE is based on the conditional randomization test (Candès et al. 2018), a powerful and intuitive statistical methodology that, like parametric methods, enables simple confounder adjustment, and like nonparametric methods, is robust to expression model misspecification. To test the association between a given gRNA and gene, we first fit a negative binomial GLM. This yields

a z-value, which typically would be compared to a standard normal null distribution based on the parametric negative binomial model. Instead, we build a null distribution for this statistic via conditional resampling. First, we estimate the probability that the gRNA will be detected in a given cell based on the cell's technical factors, such as sequencing depth and batch. Next, we resample a large number of "null" datasets, holding gene expression and technical factors constant while redrawing gRNA assignment independently for each cell based on its fitted probability. We compute a negative binomial z-value for each resampled dataset, resulting in an empirical null distribution. Finally, we compute a left-, right-, or two-tailed probability of the original z-value under the empirical null distribution, yielding a well-calibrated p-value. This p-value can deviate substantially from that obtained based on the standard normal.

To assess the calibration of SCEPTRE on real data, we applied SCEPTRE to test the association between negative control gRNAs and genes in the Gasperini et al. data (Figure 3b) and Xie et al. data (Figure 3c). We compared SCEPTRE to Monocle regression and the improved negative binomial method. For the Xie et al. data, we also compared to Virtual FACS, the method originally applied to the data. SCEPTRE showed good calibration on both datasets; by contrast, Monocle regression and improved negative binomial regression demonstrated signs of severe p-value inflation, while Virtual FACS exhibited a bimodal p-value distribution peaked at 0 and 1.

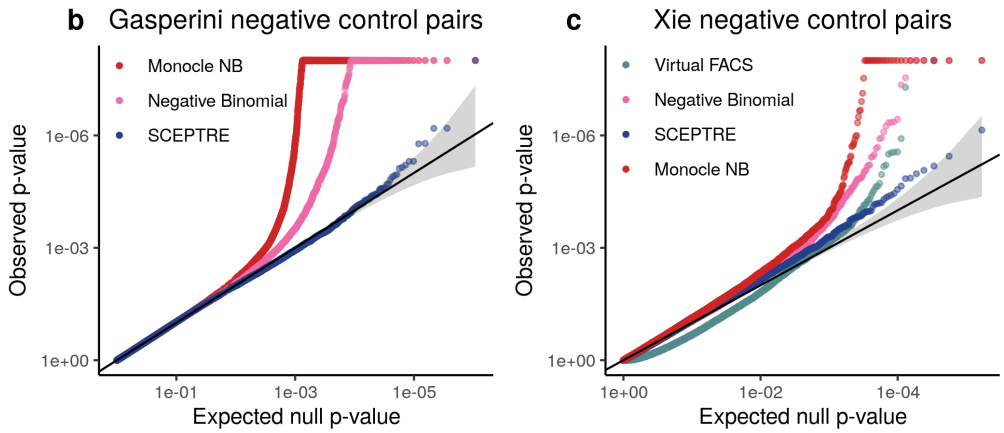


Figure 3: **SCEPTRE demonstrates good calibration on negative control data. b-c**, Application of SCEPTRE, improved negative binomial regression, Monocle regression, and Virtual FACS to pairs of negative control gRNAs and genes in (b) the Gasperini et al. data, and (c) the Xie et al. data. Compared to the other methods, SCEPTRE showed good calibration.

Finally, we applied SCEPTRE to test all candidate *cis*-regulatory pairs in the Gasperini et al. ($n = 84,595$) and Xie et al. ($n = 5,209$) data. A given gene and gRNA were considered a "candidate pair" if the gRNA targeted a site within one Mb the gene's TSS. SCEPTRE discovered 563 and 139 gene-enhancer links at an FDR of 0.1 on the Gasperini et al. and Xie et al. data, respectively. We used several orthogonal assays (e.g., ChIP-seq, topologically associating domain, and physical distance data) to quantify the enrichment of SCEPTRE's discovery set for regulatory biological signals, and we compared the SCEPTRE results to

those of other methods. We found that SCEPTRE's discovery set was far more biologically plausible than that of other methods.

Ondisc: efficient algorithms and functional data structures for out-of-core and distributed single-cell analysis

Single-cell datasets are growing in size, posing challenges as well as opportunities to biology researchers. We develop a computational framework for out-of-core and cluster/cloud-scale distributed computing on single-cell (including single-cell CRISPR screen) data. The framework is powered by several new, highly efficient algorithms and data structures. We implement the framework in an R/C++/Nextflow package called `ondisc`. Due to space constraints, and because `ondisc` is largely a software-based project, we touch only on its key features here.

A challenge of large-scale single-cell data is twofold: first, the data might not fit into memory. Second, the data are sparse ($> 99\%$ zeros) and therefore must be stored in compressed sparse matrix format. Because single-cell data are sequenced in batches, the data arrive in out-of-core *cell-accessible* sparse matrix format. This, however, poses a difficulty: most single-cell analyses require access to genes rather than (or in addition to) cells. For example, state-of-the-art normalization methods entail regressing genes onto a matrix of technical factors (including sequencing depth, sequencing batch, etc.) upstream of the dimension reduction step. Additionally, differential expression, gene-gene coexpression, and multimodal (including single-cell CRISPR screen) association analyses entail regressing genes onto other genes or across experimental conditions. We introduce a generalized radix sort algorithm to transpose the cell-accessible, sparse gene expression matrix out-of-core, yielding a *gene-accessible*, sparse expression matrix. The algorithm is optimal in disk space (i.e., it uses only as much disk space as is required to store the gene-accessible matrix) and asymptotically optimal in running time ($\mathcal{O}(n)$) and memory space ($\mathcal{O}(1)$).

The core class implemented by `ondisc` is `ondisc_matrix`, which provides an API for reading from and operating on large single-cell expression matrices. `ondisc_matrix` takes up very little space in memory, storing only essential metadata and a file path to the so-called "backing .odm file", which is HDF5 that file stores the expression data on disk. We implement several functions to operate on `ondisc_matrices`, including "subset by cell" (useful for quality control), "subset by gene" (useful for feature selection), and "normalize by regression" (useful for normalizing data ahead of dimension reduction). Importantly, `ondisc_matrices` are purely functional; that is, an `ondisc_matrix` remains accessible after applying an arbitrary sequence of functions to it. This pure functionality trivializes parallel computing on `ondisc_matrices`. Different "versions" of an `ondisc_matrix` (e.g., normalized or unnormalized, subset or unsubset, etc.) share the same backing .odm file, making `ondisc_matrices` space-efficient.

Future works

We will advance this line of work with two future projects: one applied and one methodological. Our exposition aims to convey intuition and high-level ideas rather than in-the-weeds

mathematical and biological details.

SCEPTRE version 2

Our SCEPTRE method demonstrates state-of-the-art calibration and power on high multiplicity-of-infection (MOI) single-cell CRISPR screen data. However, much remains to be done. The followup to our SCEPTRE project — currently called "SCEPTRE version 2" (or SCEPTREv2) — will expand SCEPTRE in several major directions. Our primary goal is to develop a powerful and well-calibrated test of association for *low MOI* (i.e., low multiplicity-of-infection) single-cell CRISPR screen data. Low MOI data are data in which each cell receives a *single* perturbation instead of a handful of perturbations (as in the high MOI setting, discussed above). The perturbation can target an enhancer or a gene. The statistical objective is to assess the impact of the perturbation on some cellular phenotype, typically the expression of a gene.

The low MOI setting, while related to the high MOI setting, poses a unique set of analysis challenges. In particular, the differential expression analysis protocol is slightly different in the low MOI setting. Recall that in the high MOI setting, to test the effect of a given perturbation on a given gene, we compare all cells that received the perturbation against all cells that did not receive the perturbation. In the low MOI setting, by contrast, we compare all cells that received the perturbation against the set of cells that received a non-targeting perturbation. Given this (and several other) differences, we expect the statistical challenges in the low MOI setting to differ from those in the high MOI setting. However, we currently are not sure what the exact statistical challenges in the low MOI setting will be.

Our second goal is to improve the computational efficiency of SCEPTRE. Our current SCEPTRE implementation is reasonably fast: analyzing a large (e.g., ≈200,000 cell) single-cell CRISPR screen dataset takes about one day. However, the implementation requires considerable disk space. Furthermore, the compute time could be reduced substantially through statistical accelerations. Therefore, our secondary goal is to improve the computational performance of our method, both in low and high MOI settings. We additionally plan to fully integrate SCEPTREv2 with `ondisc`, facilitating out-of-core and/or distributed single-cell CRISPR screen analysis.

Scope

This section is somewhat technical. SCEPTREv2 should apply broadly to many different kinds of single-cell CRISPR screen data. In particular, SCEPTREv2 should be handle datasets that vary along the following axes: (i) cell type (K562 cells, THP1 cells, etc.), (ii) perturbation modality (CRISPRi, CRISPRko, CRISPRa, etc.), (iii) sequencing assay (ECCITE-seq, direct-capture Perturb-seq, etc.), (iv) perturbation target (genes, enhancers), (v) readout (gene expression, chromatin accessibility, protein abundance, etc.), and most importantly (vi) MOI (low vs. high). SCEPTREv2 additionally must address the following challenges: (i) assigning perturbation identities to cells using gRNA counts, (ii) adjusting for possible perturbation inefficacy (especially relevant for CRISPRko screens), (iii) controlling for unwanted sources of technical variation, and (iv) combining information across gRNAs that target the same site.

Datasets and competitor methods

We will evaluate our method on five low MOI single-cell CRISPR screen datasets published across four papers. The papers are as follows:

1. Efthymia Papalexi et al. (2021). "Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens". In: *Nature Genetics* 53.3, pp. 322–331.

2. Daniel Schraivogel et al. (2020). "Targeted Perturb-seq enables genome-scale genetic screens in single cells". In: *Nature Methods* 17.6, pp. 629–635. (This paper presents two datasets.)

3. Noa Liscovitch-Brauer et al. (2021). "Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens". In: *Nature Biotechnology*.

4. Chris J Frangieh et al. (2021). "Multi-modal pooled Perturb-CITE-Seq screens in patient models define novel mechanisms of cancer immune evasion". In: *Nature Genetics* 53, pp. 332–341.

We additionally will evaluate the methods that the authors of the above papers used to analyze their data; we call these methods "competitor methods," as they will compete against SCEPTREv2.

Evaluation

We will evaluate all methods (i.e., SCEPTREv2 and the competitor methods) on all datasets and assess their performance on both statistical and computational metrics. Statistically, we will assess calibration (on the negative control data) and power (on the positive control data); computationally, we will assess running time, RAM usage, and disk usage.

To assess calibration using the negative control data, we will introduce a permutation test-inspired strategy called the "undercover gRNA procedure." Assume that there are n cells, p genes, and d gRNAs. Let $X \in \mathbb{R}^{n \times p}$ be the cell-by-gene expression matrix, $Y \in \mathbb{R}^{n \times d}$ be the cell-by-gRNA expression matrix, and

$$L \in \{\text{"targeting,"} \ \text{"negative control,"} \ \text{"positive control"}\}^d$$

be the vector of gRNA labels (where each gRNA is assigned a label of "targeting," "negative control," and "positive control"). Also, let P be a list of gRNA-gene pairs to analyze. Finally, let M be a single-cell CRISPR screen method (e.g., SCEPTREv2 or one of the competitor methods); M takes as arguments the gene expression matrix X, the gRNA expression matrix Y, the gRNA labels L, and the list of pairs to analyze P and outputs a p-value for each gRNA-gene pair within P.

The undercover gRNA calibration check works as follows. First, for a given negative control gRNA g, we swap its label from "negative control" to "targeting," resulting in a modified label vector L'. We call g the "undercover gRNA," as g is a negative control gRNA that has gone "undercover" as a targeting gRNA. Next, we set P' to the set of gRNA-gene pairs that results from pairing the undercover gRNA g to every gene. Finally, we

run the method M on the gene expression matrix Y, the gRNA expression matrix X, the modified gRNA labels L', and the undercover gRNA-gene pairs P'. We iterate through the set of $d_{\text{nc}} \in \mathbb{N}$ negative control gRNAs, setting each to the role of "undercover gRNA" and repeating the above steps. This procedure yields a vector of $d_{\text{nc}} \cdot p$ (dependent) p-values. If the method M is correctly calibrated, then this vector of p-values should be uniformly distributed. The "undercover gRNA calibration check" therefore can be used to assess the calibration of a method.

Preliminary results

We applied two of the competitor methods ("Schraivogel Method," a negative binomial GLM-based approach, and "Seurat DE", a two-sample Wilcoxin test-based approach) to three of the datasets ("Schraivogel TAP," "Schraivogel Perturb," and "Papalexi Gene"). We evaluated the calibration of the methods using the undercover gRNA procedure. Both methods exhibited severe miscalibration on all datasets, as indicated by QQ-plots (Figure 4). These preliminary results suggest that obtaining a well-calibrated test for low MOI single-cell CRISPR screen data is challenging.

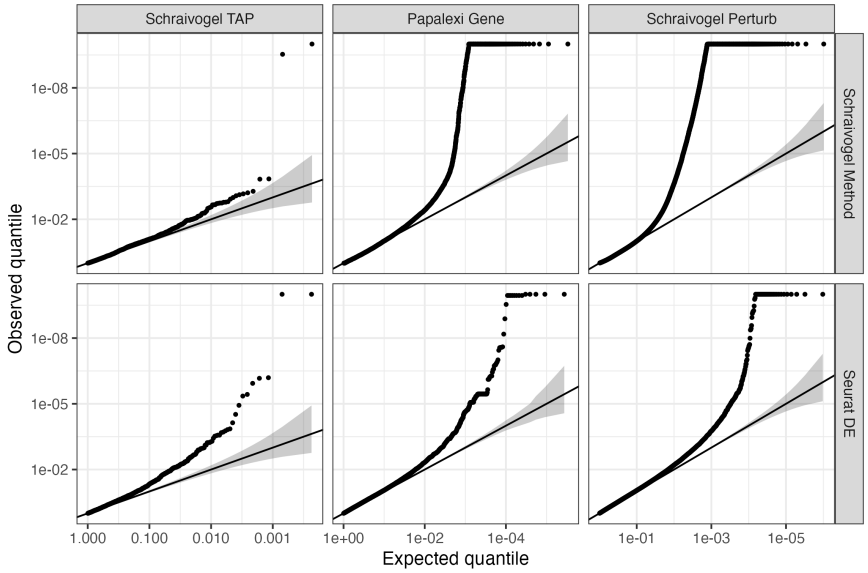


Figure 4: Results of the undercover gRNA procedure applied to two competitor methods and three datasets.

MATHS: Multi-axis testing of hypotheses via feature splitting

A transformation occurred in statistics at the turn of the century: high-throughput data, born of technological advances in biology, astronomy, and other sciences, burst onto the scene. High-throughput data are data on which tens of thousands (or more) of predictors or features are measured; examples include microarrays, genome-wide assocation studies (GWAS), and bulk RNA-seq. High-throughput data spurred the development of a flurry of methods in multiple testing, empirical Bayes, and high-dimensional statistics.

Two decades later another transformation is underway. Current and emerging technologies capture tens of thousands of features along *multiple* axes, yielding what we call "multi-axis high-throughput data." Examples include eQTL, biobank, multimodal single-cell, and single-cell CRISPR screen data. The transition from traditional to multi-axis high-throughput data, like the transition from low-throughput to high-throughout data, is fundamental: multi-axis high-throughput data pose novel analysis challenges and opportunities, especially in the context of hypothesis testing. We suggest an approach to analyzing this new class of data called "**m**ulti-**a**xis **t**esting of **h**ypotheses by feature **s**plitting" (MATHS). MATHS is analogous to ML-driven prediction: a hypothesis testing procedure is tuned on "labeled" training data so as to maximize its power while preserving type-I error control; unbiased estimates of the procedure's power and type-I error are obtained on validation data; and finally, the procedure is deployed at-scale on unlabeled discovery data.

The dawn of multi-axis high-throughput data

Traditional (or "single-axis") high-throughput data consist either of many predictors or many responses. We call a dataset "tall" if it contains many predictors and a single or only a few responses; by contrast, we call dataset "wide" if it contains many responses and a single or only a few predictors. GWAS data are tall (Figure 5a): hundreds of thousands of genetic variants are measured and tested for association with a phenotype of interest (e.g., height). Microarrays and bulk RNA-seq, on the other hand, are wide (Figure 5b): tens of thousands of genes are measured across an experimental condition (e.g., the presence or absence of a disease) and tested for differential expression. Our use of the terms "tall" and "wide" in this context refer only to the number of responses and/or predictors in the data, not the dimension.

Many current and emerging technologies measure a large number of features along *multiple* axes, yielding data that are both tall and wide (Figure 5c). We call such data "multi-axis high-throughput data." For example, eQTL (respectively, single-cell CRISPR screen) studies measure hundreds of thousands of genetic variants (respectively, thousands of CRISPR perturbations) alongside the expression levels of tens of thousands of genes. The analysis objective is to assess the association between some subset of the predictors (i.e., the genetic variants or the CRISPR perturbations) and the responses (i.e., the gene expression levels), typically controlling for a set of technical confounders.

Mutli-axis high-throughput data differ from single-axis high-throughput data in important ways. First, assuming we measure roughly p features along both axes, we potentially could test for p^2 associations, considerably more than the p possible associations in the single-axis (with univariate response) case. Testing all p^2 hypotheses would incur an enormous (and likely unacceptable) multiple testing penalty. In practice, therefore, it is common to restrict the search space to a pre-selected, interesting set of associations. In eQTL studies, for example, scientists often test for associations between GWAS-nominated variants and nearby genes to uncover *cis*-regulatory relationships; single-cell CRISPR screens are similar.

As a consequence of this restriction of the search space, many (and in fact most) associations are not tested as part of the discovery set. The "leftover" associations — which are of lesser scientific interest — can be leveraged to assess calibration of the hypothesis testing procedure. Multi-axis high-throughput data often possess some sort of spatial structure that

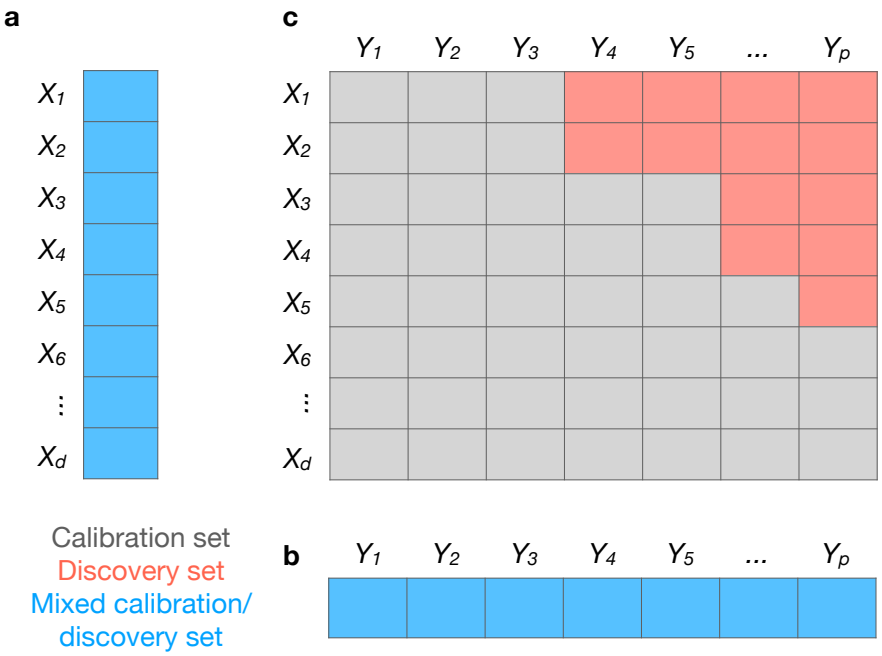


Figure 5: Three types of high-throughput data: (**a**) tall (GWAS); (**b**) wide (microarrays, bulk RNA-seq); and (**c**) tall and wide (eQTL, biobank, multimodal single-cell, single-cell CRISPR screen). Types (**a**) and (**b**) are single-axis high-throughput, while type (**c**) is multi-axis high throughput.

can be exploited to construct "natural negative controls," predictor-response pairs that *a priori* are expected to exhibit no association. For example, in eQTL studies, one can map genetic variants to genes located on different chromosomes to construct natural negative control pairs. These natural negative controls — which are not part of the discovery set — can provide valuable information about the calibration (or lack thereof) of the testing procedure. Additionally, single-cell CRISPR screens typically come equipped with *experimental* negative and positive controls. Experimental negative controls are inactive perturbations designed to have no effect on *any* gene; conversely, experimental positive controls are designed to target a specific gene for up- or down-regulation.

On single-axis high-throughput data, one must use the same data *both* to test for associations *and* to assess calibration; genomic control (in the context of GWAS) and empirical Bayes methods (in the context of microarrays) are based on this idea (Devlin et al. 1999; Efron et al. 2001). The calibration and discovery sets therefore coincide in the single-axis setting (Figure 5). By contrast, on multi-axis high-throughput data, the calibration and discovery sets are separate.

MATHS

The intuitive idea behind MATHS is as follows: for $i \in [B]$, let ϕ_i be a candidate hypothesis testing procedure. We partition the calibration set (which consists of known negative and positive control pairs) into two pieces: training set and validation set. We "audition" $\{\phi\}_{i=1}^{B}$ on the training set, selecting the testing procedure ϕ^* with the highest power (on the positive

control pairs) subject to the constraint that it controls type-I error (on the negative control pairs). Next, we obtain an unbiased estimate of the type-I error and power of ϕ^* on the validation set. Finally, we apply ϕ^* to the discovery set to make a set of discoveries.

MATHS opens the door to fresh, practical, and empirical ways of addressing several fundamental (and not fully resolved) statistical challenges in the multi-axis, high-throughput setting, including the following:

i. **Assumption checking in conditional independence testing**. Valid conditional independence tests (either parametric or nonparametric) must make an assumption or set of assumptions about the data-generating process (Shah et al. 2020). MATHS can verify the assumptions of conditional independence testing procedures, including those based on complicated machine learning models (e.g., neural networks paired with the generalized covariance measure; Shah et al. 2020).

ii. **Model selection**. Controlling for too many variables in a regression analysis can result in loss of power, while controlling for too few can result in model misspecification (and thus miscalibration). MATHS can help determine the "correct" set of confounders to adjust for.

iii. **Nuisance parameter estimation in parametric models**. Negative binomial GLMs contain a nuissance parameter (called the "dispersion parameter") that is challenging to estimate but crucial for valid inference. MATHS suggets a new strategy for estimating this nuissance parameter via "estimation by calibration."

We will mathematically formalize the MATHS framework, establish several theoretical results, implement MATHS in software, and apply MATHS to analyze single-cell CRISPR screen and eQTL data.

References

Barry, Timothy et al. (2021). "SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis". In: *Genome Biology*, pp. 1–19 (cit. on pp. 5, 6).

Barry, Timothy, Eugene Katsevich, and Kathryn Roeder (2022). "Exponential family measurement error models for single-cell CRISPR screens". In: *arXiv preprint arXiv:2201.01879* (cit. on p. 5).

Candès, Emmanuel et al. (2018). "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3, pp. 551–577 (cit. on p. 7).

Datlinger, Paul et al. (2017). "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature Methods* 14.3, pp. 297–301 (cit. on p. 4).

Devlin, B. and Kathryn Roeder (1999). "Genomic control for association studies". In: *Biometrics* 55.4, pp. 997–1004 (cit. on p. 14).

Efron, Bradley et al. (2001). "Empirical bayes analysis of a microarray experiment". In: *Journal of the American Statistical Association* 96.456, pp. 1151–1160 (cit. on p. 14).

Frangieh, Chris J et al. (2021). "Multi-modal pooled Perturb-CITE-Seq screens in patient models define novel mechanisms of cancer immune evasion". In: *Nature Genetics* 53, pp. 332–341 (cit. on p. 11).

Gallagher, Michael D. and Alice S. Chen-Plotkin (2018). "The Post-GWAS Era: From Association to Function". In: *American Journal of Human Genetics* 102.5, pp. 717–730 (cit. on p. 2).

Gasperini, Molly et al. (2019). "A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens". In: *Cell* 176.1-2, 377–390.e19 (cit. on pp. 2, 6).

Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1, pp. 1–15 (cit. on p. 4).

Hill, Andrew J. et al. (2018). "On the design of CRISPR-based single-cell molecular screens". In: *Nature Methods* 15.4, pp. 271–274 (cit. on p. 4).

Liscovitch-Brauer, Noa et al. (2021). "Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens". In: *Nature Biotechnology* (cit. on p. 11).

Louis, By Thomas A (1982). "Finding the Observed Information Matrix when Using the EM Algorithm". In: *Society* 44.2, pp. 226–233 (cit. on p. 5).

Morris, John A. et al. (2021). "Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing". In: *bioRxiv*, p. 2021.04.07.438882 (cit. on p. 2).

Papalexi, Efthymia et al. (2021). "Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens". In: *Nature Genetics* 53.3, pp. 322–331 (cit. on p. 11).

Sarkar, Abhishek and Matthew Stephens (2021). "Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis". In: *Nature Genetics* 53.6, pp. 770–777 (cit. on p. 4).

Schraivogel, Daniel et al. (2020). "Targeted Perturb-seq enables genome-scale genetic screens in single cells". In: *Nature Methods* 17.6, pp. 629–635 (cit. on p. 11).

Shah, Rajen D. and Jonas Peters (2020). "The hardness of conditional independence testing and the generalised covariance measure". In: *Annals of Statistics* 48.3, pp. 1514–1538 (cit. on p. 15).

Townes, F. William et al. (2019). "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model". In: *Genome Biology* 20.1, pp. 1–16 (cit. on p. 4).

Trapnell, Cole et al. (2014). "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32.4, pp. 381–386 (cit. on pp. 2, 7).

Xie, Shiqi et al. (2019). "Global Analysis of Enhancer Targets Reveals Convergent Enhancer-Driven Regulatory Modules". In: *Cell Reports* 29.9, 2570–2578.e5 (cit. on pp. 2, 6).