Biostatistics, 2024, **25**, 1254–1272
https://doi.org/10.1093/biostatistics/kxae010
Advance Access publication April 22, 2024

Exponential family measurement error models for single-cell CRISPR screens

Timothy Barry[1,*], Kathryn Roeder[2], Eugene Katsevich[3]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Building 2 435, 655 Huntington Ave, Boston, MA 02115, United States
[2]Department of Statistics and Data Science, Carnegie Mellon University, Baker Hall 228B, 4909 Frew St, Pittsburgh, PA 15213, United States
[3]Department of Statistics and Data Science, University of Pennsylvania, Academic Research Building 311, 265 South 37th Street Philadelphia, PA 19104, United States

*Corresponding author: Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States. Email: tbarry@hsph.harvard.edu

SUMMARY

CRISPR genome engineering and single-cell RNA sequencing have accelerated biological discovery. Single-cell CRISPR screens unite these two technologies, linking genetic perturbations in individual cells to changes in gene expression and illuminating regulatory networks underlying diseases. Despite their promise, single-cell CRISPR screens present considerable statistical challenges. We demonstrate through theoretical and real data analyses that a standard method for estimation and inference in single-cell CRISPR screens—"thresholded regression"—exhibits attenuation bias and a bias-variance tradeoff as a function of an intrinsic, challenging-to-select tuning parameter. To overcome these difficulties, we introduce GLM-EIV ("GLM-based errors-in-variables"), a new method for single-cell CRISPR screen analysis. GLM-EIV extends the classical errors-in-variables model to responses and noisy predictors that are exponential family-distributed and potentially impacted by the same set of confounding variables. We develop a computational infrastructure to deploy GLM-EIV across hundreds of processors on clouds (e.g. Microsoft Azure) and high-performance clusters. Leveraging this infrastructure, we apply GLM-EIV to analyze two recent, large-scale, single-cell CRISPR screen datasets, yielding several new insights.

KEYWORDS: CRISPR; GLM; mixture model; parallel computing; single cell.

1. INTRODUCTION

CRISPR is a genome engineering tool that has enabled scientists to precisely edit human and nonhuman genomes, opening the door to new medical therapies (Musunuru et al. 2021) and accelerating biological discovery (Przybyla and Gilbert 2022). Recently, scientists have paired CRISPR genome engineering with single-cell RNA sequencing (Datlinger et al. 2017). The resulting assays, known as "single-cell CRISPR screens," link genetic perturbations in individual cells to changes in gene expression. Single-cell CRISPR screens have enabled breakthrough progress on longstanding challenges in genetics, such as causally mapping genome wide association study (GWAS) variants to target genes at genome-wide scale (Morris et al. 2023).

Received: March 24, 2022. **Revised:** January 10, 2024. **Accepted:** March 11, 2024

Despite their promise, single-cell CRISPR screens present considerable statistical challenges. One difficulty is that the "treatment"—i.e. the presence or absence of a CRISPR perturbation—is assigned randomly to cells and is not directly observable. As a consequence, one cannot know with certainty which cells were perturbed. Instead, one must leverage an indirect, quantitative proxy of perturbation presence or absence to "guess" which cells received a perturbation. This indirect proxy takes the form of a so-called guide RNA count, with higher counts indicating that a cell is more likely to have been perturbed. A standard approach to single-cell CRISPR screen analysis is to impute perturbation assignments onto the cells by simply thresholding the guide RNA counts; using these imputations, one can attempt to estimate the effect of the perturbation on gene expression. We call this standard approach "thresholded regression" or the "thresholding method."

We study estimation and inference in single-cell CRISPR screens from a statistical perspective, formulating the data-generating mechanism using a new class of measurement error models. We assume that the response variable y is a GLM of an underlying predictor variable x^* and vector of confounders z. We do not observe x^* directly; rather, we observe a noisy version x of x^* that itself is a GLM of x^* and the same set of confounders z. The goal of the analysis is to estimate the effect of x^* on y using the observed data (x, y, z) only. In the context of the biological application, x^*, x, y, and z are CRISPR perturbations, guide RNA counts, gene expressions, and technical confounders, respectively.

Our work makes two main contributions. First, we conduct a detailed study of the thresholding method. Notably, we demonstrate on real data that the thresholding method exhibits attenuation bias and a bias-variance tradeoff as a function of the selected threshold, and we recover these phenomena in precise mathematical terms in a simplified Gaussian setting. Second, we introduce a new method, GLM-EIV ("GLM-based errors-in-variables"), for single-cell CRISPR screen analysis. GLM-EIV extends the classical errors-in-variables model (Carroll et al. 2006) to responses and noisy predictors that are exponential family-distributed and potentially impacted by the same set of confounding variables. GLM-EIV thereby implicitly estimates the probability that each cell was perturbed, obviating the need to explicitly impute perturbation assignments via thresholding. We implement several statistical accelerations to bring the cost of GLM-EIV down to within about an order of magnitude of the thresholding method. We additionally develop a Docker-containerized application to deploy GLM-EIV at-scale across tens or hundreds of processors on clouds (e.g. Microsoft Azure) and high-performance clusters.

Our analyses indicate that single-cell CRISPR screens fall into two main problem settings: the more challenging "high background contamination" setting and the easier "low background contamination" setting. GLM-EIV outperforms thresholded regression by a considerable margin in the high background contamination setting; in the low background contamination setting, by contrast, GLM-EIV and thresholded regression perform similarly, provided that accurate guide RNA-to-cell assignments are used within the thresholded regression model. We show that a simplified version of GLM-EIV can be used to obtain these guide RNA-to-cell assignments in the low background contamination setting, thereby neutralizing a tuning parameter that until this point has been challenging to select.

2. ASSAY BACKGROUND

There are several classes of single-cell CRISPR screen assays, each suited to answer a different set of biological questions. In this work we mostly focus on 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. GWAS have revealed that the majority (>90%) of variants associated with diseases lie outside genes and inside enhancers (Gallagher and Chen-Plotkin 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

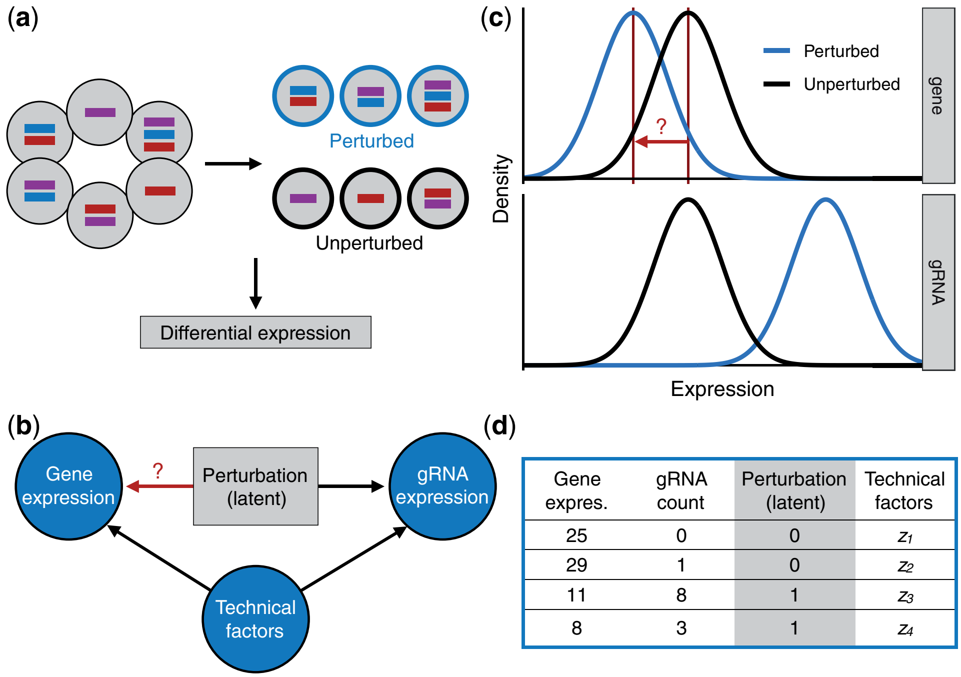

Figure 1. Experimental design and analysis challenges: a) Experimental design. For a given perturbation (e.g. the perturbation indicated in blue), we partition the cells into two groups: perturbed and unperturbed. Next, for a given gene, we conduct a differential expression analysis across the two groups, 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.

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. 2023).

High-MOI single-cell CRISPR screens are a promising emerging technology for resolving this challenge (Morris et al. 2023; Mostafavi et al. 2023). 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 (this work).

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 several methods for high-MOI single-cell CRISPR screen analysis (Gasperini et al. 2019; Xie et al. 2019; Barry et al. 2021; Wang 2021). For example, 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. Moreover, Xie et al. applied chi-squared-like tests of independence for this purpose. Unfortunately, both of these approaches have limitations: the former can break down when the gene expression model is misspecified, and the latter does not adjust for the presence of technical confounders. In a prior work we introduced SCEPTRE, a custom implementation of the conditional randomization test (Candès et al. 2018; Liu et al. 2022) tailored to single-cell CRISPR screen data. SCEPTRE simultaneously adjusts for confounder presence and ensures robustness to expression model misspecification, thereby overcoming limitations of previous approaches and demonstrating improved sensitivity and specificity on single-cell CRISPR screen data. In this work we tackle a set of analysis challenges complimentary to those addressed by SCEPTRE. Most importantly, we seek to account for the fact that the perturbation is measured with noise. Additionally, we seek to *estimate* (with confidence) the effect size of a perturbation on gene expression change, an objective that we did not consider in the original SCEPTRE study.

3. ANALYSIS CHALLENGES AND PROPOSED 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 (Fig. 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 (Fig. 1b). Third, the gene and gRNA data are sparse, discrete counts. Consequently, classical statistical approaches that assume Gaussianity or homoscedasticity are not directly applicable. 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 (Fig. 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 $z_i \in \mathbb{R}^{d-2}$ be the cell-specific covariates, including sequencing batch, percent mitochondrial reads, etc. (We note that most single-cell CRISPR screens have been carried out on cell lines consisting of a uniform cell type; however, if multiple cell types are present in the data, then cell type could be included as a covariate in the model.) The letters "m," "g", and "d" stand for "mRNA," "gRNA," and "depth," respectively.

Building on the work of several previous authors (Robinson and Smyth 2008; Townes et al. 2019; Hafemeister and Satija 2019), Sarkar and Stephens (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

$$m_i|(p_i, z_i, d_i^m) \sim \mathrm{NB}_{s^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + \log(d_i^m), \tag{3.1}$$

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 $\gamma_m \in \mathbb{R}^{d-2}$ are unknown parameters; and (iii) $\log(d_i^m)$ is an offset term. (We note that the "size parameter" is simply the inverse of the negative binomial dispersion parameter; "size parameter" does not refer to library size in this context.) Similarly, we model the gRNA counts by

$$g_i|(p_i, z_i, d_i^g) \sim \mathrm{NB}_{s^g}\left(\mu_i^g\right); \quad \log(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + \log(d_i^g), \tag{3.2}$$

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). We model the marginal perturbation as $p_i \sim \mathrm{Bern}(\pi)$, where p_i is an unobserved binary variable indicating presence ($p_i = 1$) or absence ($p_i = 0$) of the perturbation. We restrict π, the probability of perturbation, to the interval $(0, 1/2]$ to ensure that the model is identifiable; this restriction is reasonable given that each perturbation infects only a small fraction of cells. The gRNA intercept term β_0^g controls the ambient level of gRNA expression, i.e. the rate at which gRNA reads are generated in the absence of the perturbation. The perturbation coefficient β_1^g controls the extent to which perturbed and unperturbed cells differentially express the gRNA; the target of inference β_1^m is challenging to estimate when β_1^g is close to zero, as the gRNA distributions of the perturbed and unperturbed cells are hard to differentiate in this region of the problem space. Together, (3.1), (3.2), and the marginal distribution of p_i define the negative binomial GLM-EIV model.

The log-transformed sequencing depth $\log(d_i^m)$ is included as an offset term in (3.1) 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 (3.1) 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.

In this work we analyzed two large-scale, high-MOI, single-cell CRISPR screen datasets published by Gasperini et al. (2019) and Xie et al. (2019). Gasperini (resp., Xie) targeted approximately 6,000 (resp., 500) candidate enhancers in a population of approximately 200,000 (resp., 100,000) cells. Gasperini additionally designed several hundred positive control, gene-targeting perturbations and 50 nontargeting, negative control perturbations to assess method sensitivity and specificity.

4. ANALYSIS OF THE THRESHOLDING METHOD

We studied thresholding from empirical and theoretical perspectives, highlighting several potential limitations of the approach. In the context of the negative bionomial GLM-EIV model introduced above (3.1–3.2), the thresholding method leverages the gRNA counts (3.2) to impute the latent perturbation indicator (3.2), thereby reducing the full data-generating process to a single, gene expression model (3.1). We studied Gasperini et al.'s variant of the thresholding method (i.e. thresholded negative binomial regression), as this version of the thresholding method is standard and relates most closely to GLM-EIV. The method is defined as follows:

1. For a given threshold $c \in \mathbb{N}$, let the imputed perturbation assignment $\hat{p}_i \in \{0, 1\}$ be given by $\hat{p}_i = 0$ if $g_i < c$ and $\hat{p}_i = 1$ otherwise.
2. Assume that m_i is related to $\hat{p}_i$, d_i^m, and z_i through the following GLM:

$$m_i|(\hat{p}_i, z_i, d_i^m) \sim \mathrm{NB}_{s^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i + \log\left(d_i^m\right). \tag{4.3}$$

   The model (4.3) is equivalent to the model (3.2), but the latent perturbation indicator p_i has been replaced by the imputed perturbation indicator $\hat{p}_i$.
3. Fit a GLM to (4.3) to obtain an estimate and CI for the target of inference β_1^m.

To shed light on empirical challenges of the thresholding method, we applied thresholded negative binomial regression to analyze the set of positive control perturbation–gene pairs in the Gasperini dataset. The positive control pairs consisted of perturbations that targeted gene transcription start sites (TSSs) for inhibition. Repressing the TSS of a given gene decreases its expression; therefore, the positive control pairs a priori are expected to exhibit a strong decrease in expression.

To investigate the sensitivity of the thresholding method to threshold choice, we deployed the method using three different choices for the threshold: 1, 5, and 20. We found that the chosen threshold substantially impacted the results (Fig. 2a, b): estimates for fold change produced by threshold = 1 were smaller in magnitude (i.e. closer to the baseline of 1) than those produced by threshold = 5 (Fig. 2a). On the other hand, estimates produced by threshold = 5 and threshold = 20 were more concordant (Fig. 2b).

We reasoned that thresholded regression systematically underestimated true effect sizes on the positive control pairs, especially for threshold = 1. For a given perturbation, the majority ($> 98\%$) of cells are unperturbed. This imbalance leads to an asymmetry: misclassifying *unperturbed* cells as *perturbed* is intuitively "worse" than misclassifying *perturbed* cells as *unperturbed*. Misclassified unperturbed cells contaminate the set of truly perturbed cells, leading to attenuation bias; by contrast, misclassified perturbed cells are swamped in number and "neutralized" by the truly unperturbed cells. Setting the threshold to a large number reduces the unperturbed-to-perturbed misclassification rate, decreasing bias.

We hypothesized, however, that the reduction in bias obtained by selecting a large threshold causes the variance of the estimator to increase. To investigate, we compared *P*-values and confidence intervals produced by threshold = 5 and threshold = 20 for the target of inference β_1^m. We found that threshold = 5 yielded smaller (i.e. more significant) *P*-values and narrower confidence intervals than did threshold = 20 (Fig. 2c, d). We concluded that the threshold controls a bias-variance tradeoff: as the threshold increases, the bias of the estimator decreases and the variance increases.

Finally, to determine whether there is an "obvious" location at which to draw the threshold, we examined the empirical gRNA count distribution of a gRNA from the Gasperini (Fig. 2e) and Xie (Fig. 2f) dataset (counts of 0 omitted). The distributions peaked at 1 and then tapered off gradually; there did not exist a sharp boundary that cleanly separated the perturbed from the unperturbed cells. Overall, we concluded that the thresholding method faces several challenges: (i) the threshold is a tuning parameter that significantly impacts the results; (ii) the threshold mediates an intrinsic

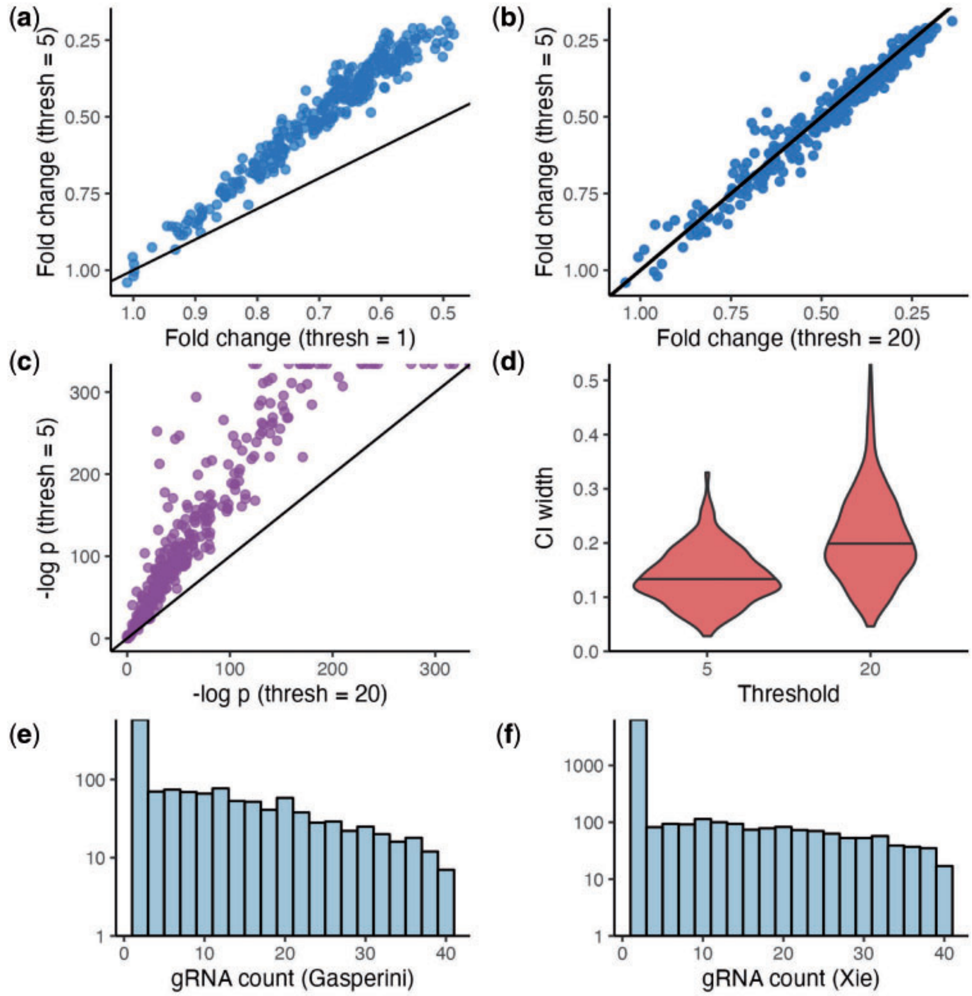

Figure 2. Empirical challenges of thresholded regression. a, b) Estimates for fold change (i.e. $\exp(\beta_1^m)$ in model (4.3)) produced by threshold = 5 versus threshold = 1 (a) and threshold = 5 versus threshold = 20 (b). The selected threshold substantially impacts the results. c, d) *P*-values (c) and CI widths (d) produced by threshold = 5 versus threshold = 20. The *P*-values correspond to a test of the null hypothesis $H_0 : \beta_1^m = 0$, i.e. a log fold change in gene expression of zero. A threshold of five yields more significant *P*-values and more confident estimates. e, f) Empirical distribution of a gRNA from Gasperini (e) and Xie (f) data (0 counts not shown). These gRNA count distributions do not appear to imply an obvious threshold.

bias-variance tradeoff; and (iii) the gRNA count distributions may not imply a clear threshold selection strategy.

Next, we studied the thresholding method from a theoretical perspective, recovering in a simplified Gaussian setting phenomena revealed in the empirical analysis. Due to space constraints, we relegate this analysis to Supplementary Appendix A, but we briefly summarize the main results here. First, we derived an exact expression for the asymptotic relative bias of the thresholding estimator $\hat{\beta}_1^m$. Leveraging this exact expression, we showed that (i) the thresholding estimator strictly underestimates (in absolute value) the true value of β_1^m over all choices of the threshold and over all values of the regression coefficients (an example of *attenuation bias*; Stefanski (2000));

and (ii) the magnitude of the bias decreases monotonically in β_1^g, comporting with the intuition that the problem becomes easier as the gRNA mixture distribution becomes increasingly well-separated. Second, we derived an asymptotically exact bias-variance decomposition for $\hat{\beta}_m$, demonstrating that as the threshold tends to infinity, the bias decreases and the variance increases.

5. GLM-BASED ERRORS-IN-VARIABLES

We introduce the general GLM-EIV (GLM-based errors-in-variables) model, which generalizes the negative binomial GLM-EIV model (3.1–3.2) to arbitrary exponential family response distributions and link functions, thereby providing much greater modeling flexibility. We derive efficient methods for estimation and inference in this model and develop a pipeline to deploy the model at-scale.

5.1. Model and model properties

The general GLM-EIV model uses an arbirary GLM to model the gene and gRNA modalities:

$$m_i|(p_i, z_i, o_i^m) \sim f_m(\mu_i^m); \quad r_m(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + o_i^m, \tag{5.4}$$

$$g_i|(p_i, z_i, o_i^g) \sim f_g(\mu_i^g); \quad r_g(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + o_i^g. \tag{5.5}$$

Here, f_m (resp., f_g) is an exponential family distribution with mean μ_i^m (resp., μ_i^g); r_m and r_g are the link function for the gene and gRNA models, respectively; and o_i^m and o_i^g are the (possibly zero) offset terms for the gene and gRNA models. In practice, we typically set o_i^m and o_i^g to the log-transformed library sizes (i.e. $\log(d_i^m)$ and $\log(d_i^g)$). Again, we assume that the unobserved perturbation indicator p_i is drawn from a $\text{Bern}(\pi)$ distribution. More model details are available in Supplementary Appendix B.

The GLM-EIV model can be seen as a generalization of the simple errors-in-variables model (when the predictor is binary); the latter is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i; \quad x_i = x_i^* + \tau_i, \tag{5.6}$$

where, $x_i^* \sim \text{Bern}(\pi)$, $\epsilon_i, \tau_i \sim N(0, 1)$, and ϵ_i,τ_i, and x_i^* are independent. GLM-EIV extends (5.6) in at least three directions: first, GLM-EIV allows y_i and x_i to follow exponential family (i.e. not just Gaussian) distributions; second, GLM-EIV allows y_i and x_i to be related to x_i^* through arbitrary (i.e. not just linear) link functions; and finally, GLM-EIV allows confounders z_i to impact both x_i and y_i. Therefore, x_i and y_i can be conditionally dependent given x_i^*, enabling GLM-EIV to capture more complex dependence relationships between x_i and y_i than is possible in (5.6) or other standard measurement error models.

5.2. Estimation and inference, and computational infrastructure

We derived an EM algorithm (Algorithm 1) to estimate the parameters of the GLM-EIV model. We briefly introduce some notation. Let $\beta_m = [\beta_0^m, \beta_1^m, \gamma_m]^T$ be the vector of unknown gene model parameters and $\beta_g = [\beta_0^g, \beta_1^g, \gamma_g]^T$ the vector of unknown gRNA model parameters. Let m, g, o^m, and o^g be the vector of gene expressions, gRNA expressions, gene library sizes, and gRNA library sizes. Finally, let X be the observed design matrix; let $\tilde{X}$ be the augmented design matrix that results from concatenating the column of (unobserved) p_is to X; and let $\tilde{X}(0)$ (resp., $\tilde{X}(1)$) be the matrix that results from setting all of the p_is in $\tilde{X}$ to 0 (resp., 1).

The E step entails computing the membership probability (i.e. the probability of perturbation) in each cell. The membership probability $T_i(1)$ of cell $i \in \{1, \ldots, n\}$ given the current parameter estimates $(\beta_m^{(t)}, \beta_g^{(t)}, \pi^{(t)})$ and observed data (m_i, g_i) is $T_i(1) = \mathbb{P}(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{(t)}, \beta_g^{(t)}, \pi^{(t)})$. We can calculate this quantity by applying (i) Bayes rule, (ii) the conditional independence property of M_i and G_i, (iii) the density of M_i and G_i, and (iv) a log-sum-exp-type trick to ensure numerical stability. Next, we produce updated estimates $\pi^{(t+1)}$, $\beta_g^{(t+1)}$, and $\beta_m^{(t+1)}$ of the parameters by maximizing the M step objective function. It turns out that maximizing this objective function is equivalent to setting $\pi^{(t+1)}$ to the mean of the current membership probabilities and setting $\beta_g^{(t+1)}$ and $\beta_m^{(t+1)}$ to the fitted coefficients of a GLM weighted by the current membership probabilities (Algorithm 1). We iterate through the E and M steps until the log likelihood (B.1) converges (Supplementary Appendix B). Our EM algorithm is reminiscent of (but distinct from) that of Ibrahim (1990), who also applied weighted GLM solvers to carry out an M step of an EM algorithm.

Algorithm 1 EM algorithm for GLM-EIV model.

Require: Pilot estimates β_m^{curr}, β_g^{curr}, and π^{curr}; data m, g, o^m, o^g, and X; gene expression distribution f_m and link function r_m; gRNA expression distribution f_g and link function r_g.

while Not converged **do**

  for $i \in \{1, \dots, n\}$ **do** ▷ E step

    $T_i(1) \leftarrow \mathbb{P}\left(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{\text{curr}}, \beta_g^{\text{curr}}, \pi^{\text{curr}}\right)$

    $T_i(0) \leftarrow 1 - T_i(1)$

  end for

  $\pi^{\text{curr}} \leftarrow (1/n) \sum_{i=1}^{n} T_i(1)$ ▷ M step

  $w \leftarrow [T_1(0), T_2(0), \dots, T_n(0), T_1(1), T_2(1), \dots, T_n(1)]^T$

  for $k \in \{g, m\}$ **do**

    Fit a GLM GLM_k with responses $[k, k]^T$, offsets $[o^k, o^k]^T$, weights w, design matrix $[\tilde{X}(0)^T, \tilde{X}(1)^T]^T$, distribution f_k, and link function r_k.

    Set β_k^{curr} to the estimated coefficients of GLM_k.

  end for

  Compute log likelihood using β_m^{curr}, β_g^{curr}, and π^{curr}.

end while

$\hat{\beta}_m \leftarrow \beta_m^{\text{curr}}$; $\hat{\beta}_g \leftarrow \beta_g^{\text{curr}}$; $\hat{\pi} \leftarrow \pi^{\text{curr}}$.

return $(\hat{\beta}_m, \hat{\beta}_g, \hat{\pi})$

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 derived an analytic formula for the asymptotic observed information matrix using Louis's Theorem (Louis (1982); Supplementary Appendix B). Leveraging this analytic formula, we can calculate standard errors quickly, enabling us to perform inference in practice on real, large-scale data.

A downside of the EM algorithm (Algorithm 1) is that it requires fitting many GLMs. Assuming that we run the algorithm 15 times using randomly generated pilot estimates (to improve chances of convergence to the global maximum), and assuming that the algorithm iterates through E and M steps about 10 times per run, we must fit approximately 300 GLMs. (These numbers are based on exploratory applications of the method to real and simulated data.) We instead devised 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" on top of one another. Briefly, we first obtain pilot estimates for the nuisance parameters $\beta_0^m, \gamma_m, \beta_0^g$, and γ_g by regressing the gene and gRNA expression vectors onto the observed design matrix X; the resulting estimates are close to the full GLM-EIV model maximum likelihood estimates because the probability of perturbation is small. Next, we obtain pilot estimates for π and the perturbation effect parameters β_1^m and β_1^g by estimating a simplified, "reduced" GLM-EIV model; this second step does not require fitting any GLMs. (See Appendix C for additional details.) Overall, the statistical accelerations reduce the number of GLMs that must be fit to < 10 in most cases.

Next, we developed 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; preprint forthcoming) to facilitate large-scale computing on single-cell data. `Nextflow` and `ondisc` together enable the construction of highly portable single-cell pipelines: one can analyze data out-of-memory on a laptop or in a distributed fashion across hundreds of processors on a cloud (e.g. Microsoft Azure, Google Cloud) or high-performance cluster. Leveraging these technologies, we built a Docker-containerized pipeline for deploying GLM-EIV at-scale. The pipeline recycles computation when possible, saving a considerable amount of compute; see Supplementary Appendix C.3 for details. Overall, the statistical accelerations and computational infrastructure make the deployment of GLM-EIV to large-scale single-cell CRISPR screen quite feasible.

5.3. The gRNA mixture assignment method

Thus far, we have described two methods for estimating the effect of a perturbation on gene expression: the simple thresholding method and the more complex GLM-EIV method. A third approach of intermediate complexity—which we call the "gRNA mixture assignment" approach—is to (i) fit a mixture model to the gRNA count distribution, (ii) use this fitted mixture model to impute perturbation identities onto cells, and then (iii) regress the gene expressions onto the imputed perturbation indicators (as well as the remaining covariates). The gRNA mixture assignment approach enjoys at least two strengths relative to the simpler thresholding approach: the former negates the threshold tuning parameter and can account for variation across cells due to covariates.

Replogle et al. (2020) proposed a simple gRNA mixture assignment strategy that involves fitting a Poisson–Gaussian mixture model to the log-transformed gRNA counts and then assigning gRNAs to cells using the posterior perturbation probabilities of the fitted model. (We call this method the Nat. Biotech. 2020 method, representing the journal and year in which the method appeared.) Unfortunately, this method poses several conceptual and practical difficulties. First, it is unclear how the method fits the Poisson component of the mixture distribution to the log-transformed gRNA expressions, as the transformed expressions are not integer-valued. Second, due to recent changes in the Python ecosystem, we and others have had difficulty with installing the Python package upon which the Nat. Biotech. 2020 method relies. (See Supplementary Appendix D for further discussion of the Nat. Biotech. 2020 method.)

Following Replogle et al. (2020), we devised an alternate gRNA mixture assignment strategy that is tethered more closely to the data-generating mechanism. For a given gRNA, we regress the gRNA counts onto the (latent) perturbation indicator and covariates (while ignoring the gene expressions; model 5.5). We assign perturbation identities to cells by thresholding the posterior perturbation probabilities of the fitted model at 1/2. The latent variable gRNA model is a subset of the full GLM-EIV model (5.4–5.5). Thus, we used the GLM-EIV EM algorithm to fit the latent variable gRNA model, enabling us to exploit the various techniques that we developed in the context of GLM-EIV for obtaining fast and numerically stable estimates.

6. SIMULATION STUDY

We conducted a comprehensive suite of six simulation studies to compare the empirical performance of GLM-EIV, the thresholding method, and the gRNA mixture assignment method. (We coupled the latter method to standard regression on the imputed perturbation assignments to estimate the perturbation effect size.) We describe one simulation study here and defer the remaining simulation studies to the Supplementary Appendix G. We generated data on $n = 50,000$ cells from the GLM-EIV model, setting the target of inference β_1^m to $\log(0.25)$ and the probability of perturbation π to 0.02. $\beta_1^m = \log(0.25)$ represents a decrease in gene expression by a factor of 4, which is a fairly large effect size on the order of what we might observe for a positive control pair. We included "sequencing batch" (modeled as a Bernoulli-distributed variable) as a covariate and sequencing depth (modeled as a Poisson-distributed variable) as an offset. We varied the log-fold change in gRNA expression, β_1^g, over a grid on the interval $[\log(1), \log(4)]$; β_1^g controls problem difficulty, with higher values corresponding to easier problem settings. We generated the gene expression count data from two response distributions: Poisson and negative binomial (size parameter fixed at $s = 20$ for the latter; see simulation study 3 for an exploration of different values of s). We generated the gRNA count data from a Poisson distribution. For each parameter setting (defined by a β_1^g-distribution pair), we synthesized $n_{\text{sim}} = 500$ i.i.d. datasets. Supplementary Appendix G compares the parameter values used in the simulation study to those estimated from real data.

We applied four methods to the simulated data: "vanilla" GLM-EIV, accelerated GLM-EIV, thresholded regression, and the gRNA mixture assignment method. We used the Bayes-optimal decision boundary for classification as the threshold for the thresholding method (as derived in Supplementary Section A.12). We ran all methods on the negative binomial data twice: once treating the size parameter s as a known constant and once treating s as unknown. In the latter case we used the `glm.nb` function from the `MASS` package to estimate s before applying the methods (Ripley et al. 2013). We note that none of the methods accounts for the error in estimating s when computing coefficient standard errors. We display the results of the simulation study in Fig. 3. Columns correspond to distributions (i.e. Poisson, NB with known s, and NB with unknown s), and rows correspond to performance metrics (i.e. bias, mean squared error, CI coverage rate (nominal rate 95%), CI width, and method run time). The β_1^g parameter is plotted on the horizontal axis, and the methods are depicted in different colors. (GLM-EIV is masked by accelerated GLM-EIV in several panels.)

We found that GLM-EIV outperformed the gRNA mixture method and that the gRNA mixture method outperformed thresholded regression across the metrics of bias, mean squared error, and confidence interval coverage. We reasoned that GLM-EIV outperformed the gRNA mixture method because (i) GLM-EIV leveraged information from *both* modalities (rather than the gRNA modality alone) to assign perturbation identities to cells and (ii) GLM-EIV produced soft rather than hard assignments, capturing the inherent uncertainty in whether a perturbation occurred. We additionally reasoned that the gRNA mixture method outperformed thresholded regression because the gRNA mixture method better accounted for heterogeneity across cells due to the covariates. Notably, accelerated GLM-EIV performed as well as vanilla GLM-EIV on all statistical metrics (rows 1–4) despite having substantially lower computational cost (bottom row). In fact, the running time of accelerated GLM-EIV was almost within an order of magnitude of that of the thresholding method. As expected, the confidence interval coverage of the methods degraded somewhat in the negative binomial case under estimated s as opposed to known s, but this difference was not substantial. Supplementary Appendix G presents additional simulation studies in which we generate data from a Gaussian model, vary β_1^m and s, and assess the performance of the methods on data containing unmeasured covariates and outliers.

7. REAL DATA APPLICATION I: ESTIMATING PERTURBATION EFFECTS ON HIGH-MOI DATA

Leveraging our computational infrastructure, we applied GLM-EIV and the thresholding method to analyze the entire Gasperini and Xie datasets. GLM-EIV ran in under two days on both datasets, using no more than 250 processors and two gigabytes of memory per process. We report only the most important aspects of the analysis and results in the main text; full details are available in Supplementary Appendix E. We set the threshold in the thresholding method to the approximate Bayes-optimal decision boundary, as our theoretical analyses and simulation studies indicated that the Bayes-optimal decision boundary is a good choice for the threshold when the gRNA count

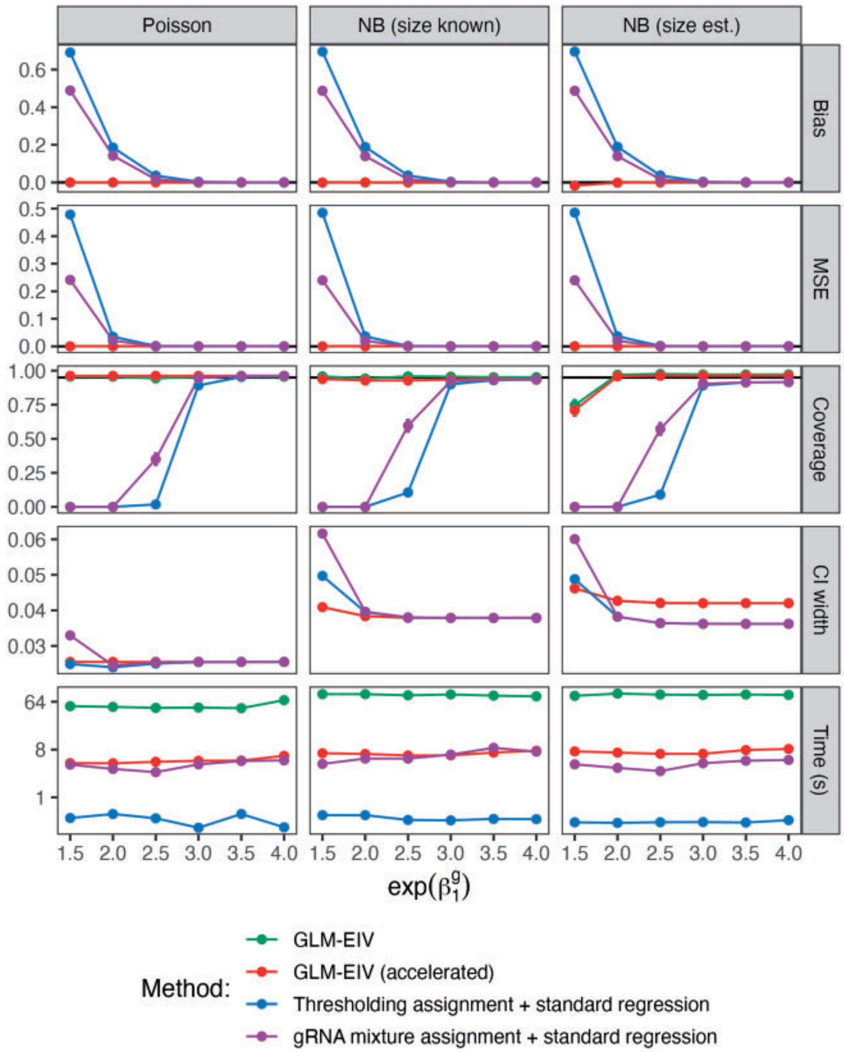

Figure 3. Simulation study. Columns correspond to distributions (Poisson, NB with known s, NB with estimated s), and rows correspond to metrics (bias, MSE, coverage, CI width, and time). Methods are shown in different colors; GLM-EIV (green) is masked by accelerated GLM-EIV (red) in several panels. Generally, GLM-EIV (both accelerated and nonaccelerated versions) outperformed the gRNA-mixture/NB-regression method, which in turn outperformed the thresholding/NB-regression method. The rejection probability (i.e. the probability of rejecting the null hypothesis $H_0 : \beta_1^m = 0$ at level $\alpha = 0.05$) was strictly 1 across methods and parameter settings, likely because the effect size was fairly large.

distribution is well-separated. Operating under the assumption that the effect of the perturbation on gRNA expression is similar across pairs, we leveraged the fitted GLM-EIV models to approximate the Bayes boundary in the following way: we (i) sampled several hundred gene–perturbation pairs, (ii) extracted the fitted values $\hat{\beta}_g$ and $\hat{\pi}$ from the GLM-EIV models fitted to these pairs, (iii) computed the median $\overline{\hat{\beta}_g}$ and $\overline{\hat{\pi}}$ across the $\hat{\beta}_g$s and $\hat{\pi}$s, and (iv) used $\overline{\hat{\beta}_g}$ and $\overline{\hat{\pi}}$ to estimate a dataset-wide Bayes-optimal decision boundary (Section A.12). We repeated this procedure on both datasets, yielding a threshold of 3 for Gasperini and 7 for Xie.

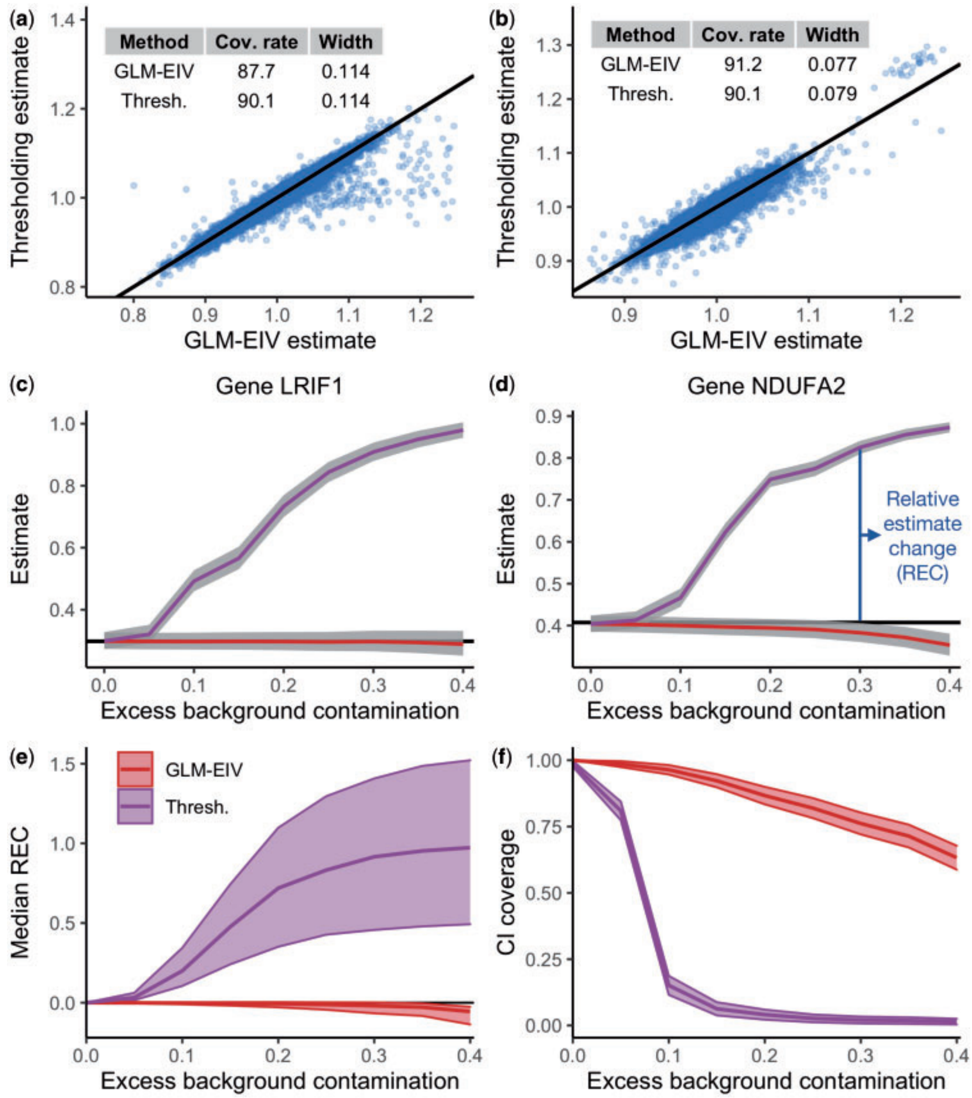

Figure 4. Applying GLM-EIV to analyze large-scale, high-MOI data. a, b) Estimates for fold change produced by GLM-EIV and thresholded regression on Gasperini (a) and Xie (b) negative control pairs. c, d) Estimates produced by GLM-EIV and thresholded regression on two positive control pairs—*LRIF1* (a) and *NDUFA2* (b)—plotted as a function of excess background contamination. Grey bands, 95% CIs for the target of inference outputted by the methods. e, f) Median relative estimate change (REC; e) and confidence interval coverage rate (f) across *all* 322 positive control pairs, plotted as a function of excess background contamination. c–f) Together illustrate that GLM-EIV demonstrated greater stability than thresholded regression as background contamination increased.

We compared GLM-EIV to thresholded regression on the real data, focusing specifically on the negative control pairs (i.e. gene-perturbation pairs for which the ground truth fold change is known to be 1; Supplementary Appendix E). We found that GLM-EIV and the thresholding method produced similar results (Fig. 4a, b): estimates, CI coverage rates, and CI widths were concordant. CI coverage rates, which ranged from 87.7% to 91.2%, were slightly below the nominal rate of 95%, likely due to mild model misspecification. The estimated effect of the perturbation on gRNA expression $\exp(\hat{\beta}_1^g)$ was unexpectedly large: the 95% CI for this parameter (averaged across pairs) was [4306, 5186] and [300, 316] on the Gasperini and Xie data, respectively. We reasoned that the datasets lay in a region of the parameter space in which thresholding is a tenable strategy (provided the threshold is selected well). However, this was not obvious a priori and may not be the case for other datasets. We note that GLM-EIV produced outlier estimates (defined as estimated fold change < 0.75 or > 1.25) on a small ($< 2.5\%$ on Gasperini, $< 0.05\%$ on Xie) number of pairs consisting of a handful of genes, likely due to nonglobal EM convergence. These outliers are not plotted in Fig. 4a, b but were used to compute the CI coverage reported in the inset tables.

To evaluate performance of GLM-EIV versus thresholding in more challenging settings, we increased the difficulty of the perturbation assignment problem by generating partially synthetic datasets. First, for a given pair, we sampled gRNA counts directly from the fitted GLM-EIV model. Next, to simulate elevated background contamination, we sampled gRNA counts from a slightly modified version of the fitted model in which we increased the mean gRNA expression of *unperturbed* cells while holding constant the mean gRNA expression of *perturbed* cells. We defined a parameter called "excess background contamination" (normed to take values in [0, 1]) to quantify the relative distance between the unperturbed and perturbed gRNA count distributions. We held fixed the real-data gene expressions, library sizes, covariates, and fitted perturbation probabilities in all settings.

We generated partially synthetic data in the above manner for each of the 322 positive control pairs in the Gasperini dataset, varying excess background contamination over the interval [0, 0.4]. We then applied GLM-EIV and the thresholding method to analyze the data. We present results on two example pairs (the pair containing gene *LRIF1* and the pair containing gene *NDUFA2*) in Fig. 4c, d. We observed that the estimate produced by the methods on the raw data (depicted as a horizontal black line) coincided almost exactly with the estimate produced by the methods on the partially synthetic data generated by setting excess background contamination to zero (This result replicated across nearly all pairs; average relative difference 0.003.) We additionally observed that as excess background contamination increased, the performance of thresholded regression degraded considerably while that of GLM-EIV remained stable.

We generalized the above analysis to the entire set of positive control pairs. First, for each pair we computed the "relative estimate change" (REC) as a function of excess background contamination, defined as the relative difference between the estimate at a given level of excess contamination and zero excess contamination (Fig. 4d). Next, we computed the median REC across all positive control pairs (Fig. 4e; upper and lower bands indicate the pointwise interquartile range of the REC). As excess background contamination increased, thresholded regression exhibited severe attenuation bias (as reflected by large median REC values); GLM-EIV, by contrast, remained mostly stable. Finally, letting $\hat{\beta}_1^m$ denote the estimate obtained on the raw data, we computed the CI coverage of $\hat{\beta}_1^m$ as a function of excess contamination. Under the assumption that $\hat{\beta}_1^m$ is close to the true parameter β_1^m, the CI coverage of the former is similar to that of the latter. We computed the CI coverage of $\hat{\beta}_1^m$ by calculating each individual pair's coverage of $\hat{\beta}_1^m$ (across the Monte Carlo replicates) and then averaging this quantity across all pairs. GLM-EIV exhibited significantly higher CI coverage than thresholded regression as the data became increasingly contaminated (Fig. 4f; bands indicate 95% pointwise CIs). Coverage rates were slightly above the nominal level of 95% in some settings because we covered an *estimate* of β_1^m rather than β_1^m itself, leading to mild "overfitting." Nonetheless, this experiment was meaningful to assess the stability of both methods to elevated background contamination.

8. REAL DATA APPLICATION II: ASSIGNING PERTURBATIONS TO CELLS ON LOW-MOI DATA

We sought to explore whether the gRNA mixture assignment method that we proposed in Section 5.3—which is in effect a special case of GLM-EIV—might be an independently useful tool for assigning gRNAs to cells on real single-cell CRISPR screen data. We applied the gRNA mixture assignment method to assign gRNAs to cells on a low multiplicity-of-infection (or MOI) single-cell CRISPR screen of immune cells (Papalexi et al. 2021). (A low-MOI dataset, in contrast to a high-MOI dataset, is one in which the experimenter has aimed to insert exactly one perturbation into each cell.) We elected to assess the performance of the gRNA mixture assignment method on low-MOI data because the "ground truth" gRNA-to-cell mapping is easier to ascertain in low MOI than in high MOI. The majority of cells in a low-MOI screen contains a single perturbation, while a fraction of cells contains zero or two or more perturbations. Thus, if a given gRNA constitutes a large fraction (say, $> 25\%$) of the gRNA reads in a given cell, we can confidently map that gRNA to that cell. Athough not foolproof, this strategy yields a reasonable approximation to the ground truth in low MOI. (There is no analogous strategy for obtaining ground truth gRNA assignments in high MOI, as each cell in high MOI contains many gRNAs, and the number of gRNAs per cell is indeterminate and variable.)

We used our proposed gRNA mixture assignment method to obtain gRNA-to-cell assignments for each gRNA in the low-MOI dataset (after restricting our attention to the 95% most highly expressed gRNAs). We included the standard technical factors as covariates, including biological replicate. We compared the mixture-model-based gRNA assignments to the ground truth assignments; the latter were obtained in the manner described above. Encouragingly, we found that these two methods produced near-identical results. For example, the mixture model determined that gRNA "CUL3g2" was present in 141 cells (and absent in the rest), while the ground truth method indicated that "CUL3g2" was present in 137 cells (Fig. 5a). Treating the ground truth assignments as a reference, we constructed a confusion matrix to assess the classification accuracy of the mixture method assignments on CUL3g2 (Fig. 5b). The sensitivity, specificity, and balanced accuracy of the mixture method assignments were high (1.000, 0.9998, and 0.9998, respectively).

We replicated this analysis across the entire set of gRNAs, finding that the mixture method assignments exhibited consistently high concordance with the ground truth assignments as measured by sensitivity, specificity, and balanced accuracy (although there were a few outliers; Fig. 5c). We concluded that the mixture assignment method was a statistically principled, fast, and numerically stable strategy for the recapitulating the ground truth assignments with high fidelity. We sought to compare our gRNA mixture assignment method against the Nat. Biotech. 2020 Poisson-Gaussian mixture method. Unfortunately, as discussed elsewhere (Section 5.3 and Appendix D), we were unable to get the Nat. Biotech. 2020 method (or approximations thereof written in R) working. We note that, in contrast to the Nat. Biotech. 2020 method, the proposed method allows for the inclusion of covariates (e.g. library size and batch) and models the gRNA counts directly.

9. DISCUSSION

In this work, we studied the problem of estimating the effect sizes of perturbations on changes in gene expression in high-MOI single-cell CRISPR screens, focusing specifically on the challenge that the perturbation is unobserved. We showed through empirical, theoretical, and simulation analyses that the commonly used thresholding method poses several difficulties: there exist settings (i.e. high background contamination settings) in which thresholding is not a tenable strategy, and in settings in which thresholding *is* a tenable strategy (i.e. low background contamination settings), selecting a good threshold is challenging and consequential. Next, we developed GLM-EIV, a method that jointly models the gene and gRNA modalities to implicitly assign perturbation identities to cells and estimate perturbation effect sizes, thereby overcoming limitations of the thresholding method. GLM-EIV demonstrated significantly improved performance relative to the thresholding method in high background contamination settings on both synthetic and realistic semi-synthetic data.

However, GLM-EIV and the thresholding method demonstrated roughly similar performance on the two real high-MOI datasets that we examined, as the real data exhibited lower background contamination than anticipated. We believe that this is an interesting finding in itself; moreover,

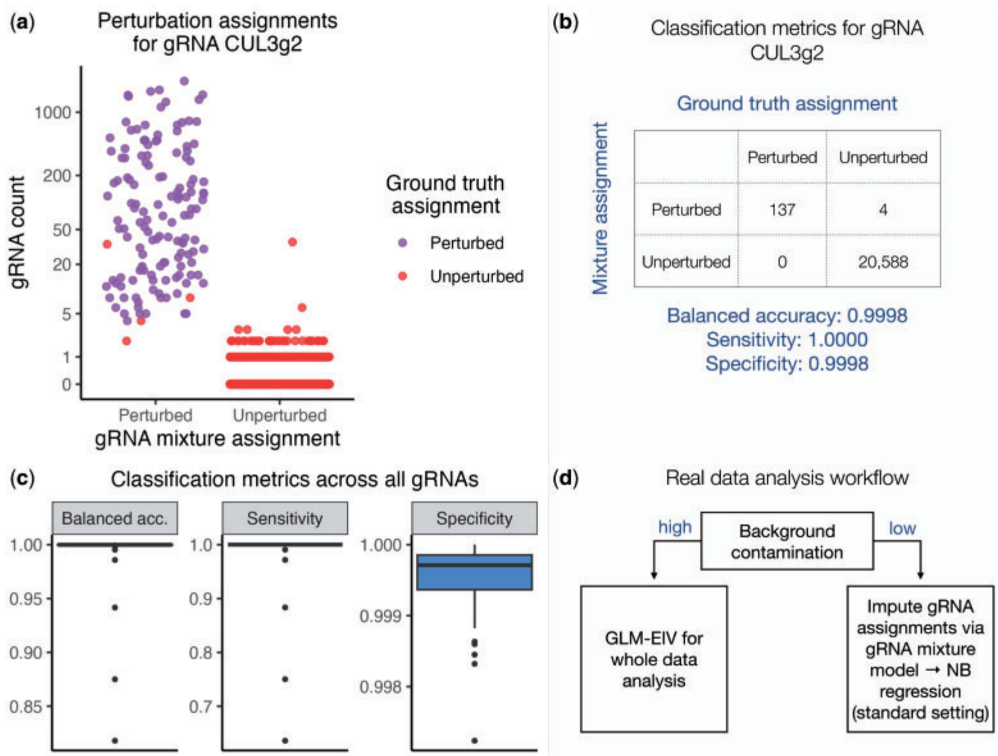

Figure 5. The gRNA-only mixture assignment functionality of GLM-EIV accurately assigns gRNAs to cells on real low-MOI data. a) Each point represents a cell. The position of each cell along the vertical axis indicates the number of gRNA reads (from gRNA "CUL3g2") observed in that cell. Cells in the left column were classified by the gRNA mixture model as perturbed, while those in the right column were classified as unperturbed. Purple (resp., red) cells were classified by the ground truth method as perturbed (resp., unperturbed). b) A confusion matrix comparing the gRNA-to-cell mixture model classifications against the ground truth classifications for gRNA "CUL3g2." The two sets of classifications were highly concordant, as quantified by balanced accuracy, sensitivity, and specificity metrics. c) The balanced accuracy (left), sensitivity (middle), and specificity (right) of the gRNA mixture assignment method across *all* gRNAs. d) The proposed data analysis workflow. If the level of background contamination is low, then the gRNA mixture method can be used to impute perturbation identities onto cells, which can then be plugged into downstream analytic tools, such as negative binomial regression or SCEPTRE. On the other hand, if the level of background contamination is high, then the entire GLM-EIV model can be used to analyze the data.

future datasets may demonstrate higher levels of background contamination, in which case GLM-EIV could serve as an immediately applicable analytic tool. Finally, the gRNA mixture assignment method, which under the hood exploits the estimation machinery of GLM-EIV, is a statistically principled, numerically stable, fast, and accurate strategy for obtaining gRNA-to-cell assignments on real data; these assignments can used as input to downstream methods (e.g. negative binomial regression or SCEPTRE; Fig. 5d).

We anticipate that GLM-EIV could be applied to other types of multimodal single-cell data, such as single-cell chromatin accessibility assays. A question of interest in such experiments is whether chromatin state (i.e. closed or open) is associated with the expression of a gene or abundance of a protein (Mimitou et al. 2021). We do not directly observe the chromatin state of a cell; instead, we observe tagged DNA fragments that serve as count-based proxies for whether a given region of chromatin is open or closed. GLM-EIV might be applied in such experiments to aid in the selection of thresholds or to analyze whole datasets. The full GLM-EIV model potentially could be applied to analyze low-MOI single-cell CRISPR screen data, but we anticipate that the relative ease of assigning gRNAs to cells in low MOI (as described in Section 8) may obviate the need for GLM-EIV in that setting.

The closest parallels to GLM-EIV in the statistical methodology literature are Grün and Leisch (2008) and Ibrahim (1990). Grün and Leisch derived a method for estimation and inference in a *k*-component mixture of GLMs. While we prefer to view GLM-EIV as a generalized errors-in-variables method, the GLM-EIV model is equivalent to a two-component mixture of products of GLM densities. Ibrahim proposed a procedure for fitting GLMs in the presence of missing-at-random covariates. Our method, by contrast, involves fitting two conditionally independent GLMs in the presence of a totally latent covariate. Thus, while Ibrahim and Grün & Leisch are helpful references, our estimation and inference tasks are more complex than theirs. Next, Aigner (1973) and Savoca (2000) proposed measurement error models that consist of unobserved *binary* rather than *continuous* predictors; the latter are more commonly used in measurement error models. GLM-EIV likewise consists of a latent binary predictor, but unlike Aigner and Savoca, GLM-EIV handles a much broader class of exponential family-generated data. Finally, GLM-EIV accounts for a common source of measurement error between the predictor and response, a property not shared by classical measurement error models (Carroll et al. 2006). Additional related work is relayed in Supplementary Appendix F.

GLM-EIV might be applied to areas beyond genomics, such as psychology. Some psychological constructs (e.g. presence or absence of a social media addiction) are latent and can be assessed only through an imperfect proxy (e.g. the number of times one has checked social media). Researchers might use GLM-EIV to regress an outcome variable (e.g. self-reported well-being) onto the latent construct via the imperfect proxy, potentially resolving challenges related to attenuation bias and threshold selection. Applications to psychology and other areas are a topic of future investigation.

SOFTWARE, CODE, AND RESULTS

The gRNA-only mixture assignment functionality of GLM-EIV is implemented in our `sceptre` toolkit for single-cell CRISPR screen analysis (`github.com/Katsevich-Lab/sceptre`). The `sceptre` user manual (`timothy-barry.github.io/sceptre-book/sceptre.html`) presents a detailed guide on analyzing data using the `sceptre` software, including several sections on assigning gRNAs to cells using the mixture assignment method introduced in this work.

Results are deposited at `upenn.box.com/v/glmeiv-files-v1`. Github repositories containing manuscript replication code, the `glmeiv` R package, and the cloud/HPC-scale GLM-EIV pipeline are available at `github.com/timothy-barry/glmeiv-manuscript`, `github.com/timothy-barry/glmeiv`, and `github.com/timothy-barry/glmeiv-pipeline`, respectively. Detailed replication instructions are available in the first repository.

ACKNOWLEDGMENTS

We thank Eric Tchetgen Tchetgen for helpful conversations, Xuran Wang for helping to process the Xie dataset, and Songcheng Dai for helping to deploy the GLM-EIV pipeline on Azure. We additionally thank three anonymous reviewers whose comments considerably improved the manuscript. This work used the Extreme Science and Engineering Discovery Environment (XSEDE; NSF grant ACI-1548562) and the Bridges-2 system (NSF grant ACI-1928147) at the Pittsburgh Supercomputing Center.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biostatistics Journal* online.

FUNDING

This work was funded by National Institute of Mental Health (NIMH) grant R01MH123184 and NSF grant DMS-2113072.

Conflict of interest statement. None declared.

REFERENCES

Aigner DJ. Regression with a binary independent variable subject to errors of observation. J Econ. 1973:1(1):49–59.

Barry T, et al. SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. Genome Biol. 2021:1–19.

Candès E, et al. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J R Stat Soc Ser B. 2018:80(3):551–577.

Carroll RJ, et al. Measurement error in nonlinear models: a modern perspective. New York, New York, USA: Chapman and Hall/CRC; 2006.

Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. 2022:23(1):1–20.

Datlinger P, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017:14(3):297–301.

Gallagher MD, Chen-Plotkin AS. The post-GWAS era: from association to function. Am J Hum Genetics. 2018:102(5):717–730.

Gasperini M, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell 2019:176(1–2):377–390.e19.

Grün B, Leisch F. Finite mixtures of generalized linear regression models. Heidelberg: Physica HD; 2008. p. 205–230.

Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019:20(1):1–15.

Hill AJ, et al. On the design of CRISPR-based single-cell molecular screens. Nat Methods. 2018:15(4):271–274.

Ibrahim JG. Incomplete data in generalized linear models. J Am Stat Assoc. 1990:85(411):765–769.

Lin KZ, Lei J, Roeder K. Exponential-family embedding with application to cell developmental trajectories for single-cell RNA-seq data. J Am Stat Assoc. 2021:116(534):457–470.

Liu M, Katsevich E, Janson L, Ramdas A. Fast and powerful conditional randomization testing via distillation. Biometrika. 2022:109(2):277–293.

Louis TA. Finding the observed information matrix when using the EM algorithm. J R Stat Soc Ser B. 1982:44(2):226–233.

McCullagh P, Nelder JA. Generalized linear models. 2nd ed. 1990. London, UK and New York, New York, USA: Chapman and Hall.

Mimitou EP, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 2021:39(10):1246–1258.

Morris J, et al. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. Science. 2023:380(6646):eadh7699.

Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat Genet. 2023:55(11):1866–1875.

Musunuru K, et al. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. Nature. 2021:593(7859):429–434.

Papalexi E, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genetics*. 2021:53(3):322–331.

Przybyla L, Gilbert LA. A new era in functional genomics screens. *Nat Rev Genetics*. 2022:23(2):89–103.

Replogle JM, et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat Biotechnol. 2020:38(8):954–961.

Ripley B et al.. Package 'mass'. Cran r. 2013:538:113–120.

Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to sage data. Biostatistics. 2008:9(2):321–332.

SARKAR A, STEPHENS M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat Genet. 2021:53(6):770–777.
SAVOCA E. Measurement errors in binary regressors: an application to measuring the effects of specific psychiatric diseases on earnings. Health Serv Outcomes Res Methodol. 2000:1(2):149–164.
STEFANSKI LA. Measurement error models. J Am Stat Assoc. 2000:95(452):1353–1358.
TOWNES FW, ET AL. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 2019:20(1):1–16.
TRAPNELL C ET AL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014:32(4):381–386.
WANG L. Single-cell normalization and association testing unifying CRISPR screen and gene co-expression analyses with Normalisr. Nat Commun. 2021:12(1):6395.
XIE S. ET AL. Global analysis of enhancer targets reveals convergent enhancer-driven regulatory modules. Cell Rep. 2019:29(9):2570–2578.e5.