A class of exponential family measurement error models for CRISPR genome engineering and single-cell sequencing

Timothy Barry, Eugene Katsevich, Kathryn Roeder

Abstract

CRISPR genome engineering and single-cell sequencing have transformed 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 substantial 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 tuning parameter. To overcome these limitations, 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 response distributions and sources of measurement error that are (i) exponential family-distributed and (ii) potentially impacted by the same set of confounding variables. We develop a computational infrastructure to deploy GLM-EIV across tens or hundreds of nodes 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, demonstrating improved performance in challenging problem settings.

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 [22, 26] and transforming basic biology research [23]. Recently, scientists have paired CRISPR genome engineering with single-cell sequencing [4, 7]. The resulting assays, known as a "single-cell CRISPR screens," link genetic perturbations in individual cells to changes in gene expression, illuminating regulatory networks underlying human diseases and other traits [21].

Despite their promise, single-cell CRISPR screens present substantial statistical challenges. A major difficulty is that CRISPR perturbations are assigned stochastically to cells and cannot be observed directly. As a consequence, one cannot know with certainty which cells were perturbed.

Instead, one must leverage an indirect, noisy proxy of perturbation presence or absence – namely, transcribed guide RNA counts – to "guess" which cells were perturbed. Using these imputed perturbation assignments, one can attempt to estimate the effect of the perturbation on gene expression. The standard approach, which we call "thresholded regression" or the "thresholding method," is to assign perturbation identities to cells by simply thresholding the guide RNA counts.

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 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 to response distributions and sources of measurement error that are (i) exponential family-distributed and (ii) 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 or another heuristic. We implement several statistical accelerations (that possibly are of independent utility) to bring the cost of GLM-EIV down to within about an order of magnitude of the thresholding method.

Finally, we develop a Docker-containerized application to deploy GLM-EIV at-scale across tens or hundreds of nodes on clouds (e.g., Microsoft Azure) and high-performance clusters. Leveraging this application, we apply GLM-EIV to analyze two recent, large-scale, single-cell CRISPR screen datasets. We find that in some settings, GLM-EIV outperforms thresholded regression by a considerable margin; in other settings the two methods work best in conjunction, with GLM-EIV providing a statistically principled and empirically effective procedure for selecting the threshold.

2 Background and analysis challenges

2.1 Related work Motivated by the challenges of single-cell data, several authors recently have extended statistical models that (implicitly or explicitly) assume Gaussianity and homoscedasticity to a broader class of exponential family distributions. For example, Lin et al. [16] and Townes et al. [32] (separately) developed eSVD and GLM-PCA, generalizations of SVD and PCA, respectively, to exponential family response distributions. Unlike their vanilla counterparts, eSVD and GLM-PCA can model gene expression counts directly, thereby capturing the relationship between the mean and variance of a gene's expression level [15] and improving performance on dimension reduction tasks. We see our work as a continuation of this effort to "port" common statistical methods and models to single-cell count data. Our focus, however, is on regression rather than dimension reduction. We extend the classical errors-in-variables model in several directions, most importantly generalizing to exponential family-distributed response distributions and sources of measurement error.

The closest parallels to our work in the statistical methodology literature are Grün & Leisch [11] and Ibrahim [14]. Grün & 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. Other relevant references include Aigner [1] and Savoca [28].

The genomics literature has produced several applied methods for single-cell CRISPR screen analysis. In a prior work we developed SCEPTRE [2], a custom implementation of the conditional randomization test [3, 17] tailored to single-cell CRISPR screen data. SCEPTRE tests whether a given perturbation is associated with the change in expression of a given gene, adjusting for sources of confounding and ensuring robustness to expression model misspecification. In this work we tackle a set of analysis challenges that are complimentary to those addressed by SCEPTRE.

Most importantly, we account for the fact that the perturbation is measured with noise, an issue that SCEPTRE circumvents via thresholding. Additionally, we seek to *estimate* (with confidence) the effect size of a perturbation on gene expression change, an objective that is challenging to attain within the nonparametric hypothesis testing framework of SCEPTRE.

2.2 Assay overview There are several broad classes of single-cell CRISPR screen assays, each suited to answer a different set of biological questions [5, 10, 19]. In this work we focus on high-multiplicity of infection (MOI) single-cell CRISPR screens. We expect the ideas that we develop for this assay to apply (with some effort) to other classes of single-cell CRISPR screens as well. In this section we motivate high MOI single-cell screens, overview the experimental protocol, and present relevant analysis challenges.

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 importance to the current work). Genome-wide association studies (GWAS) have revealed that the majority ($> 90\%$) of variants associated with diseases lie outside genes and (very likely) inside enhancers [8]. 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 [9].

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 transcrip-

tome (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.

2.3 Analysis challenges 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, and most subtly, sequenced gRNAs sometimes map to cells that have not received a perturbation. This phenomenon, which we "background contamination," results from errors in the sequencing and alignment processes [24]. 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.

In this work we analyze two large-scale, high MOI, single-cell CRISPR screen datasets published by Gasperini et al. [10] and Xie et al. [33]. 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 non-targeting, negative control perturbations to assess method sensitivity and specificity.

3 Thresholding method

We study thresholding from empirical and theoretical perspectives, highlighting several inherent limitations of the approach. Gasperini and Xie imputed perturbation assignments onto the cells via thresholding, but they carried out the subsequent differential expression analysis in different ways: Gasperini fitted negative binomial regression models to the data, whereas Xie applied nonparametric tests of independence. We study Gasperini's variant of the thresholding method, as it relates more closely to GLM-EIV and (in our view) is the more natural method.

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 $m_i \in \mathbb{N}$ 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); 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. The thresholding 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

$$\hat{p}_i = 0 \text{ if } g_i < c; \quad \hat{p}_i = 1 \text{ if } g_i \geq c.$$

2. Assume that m_i is related to $\hat{p}_i, d_i^m$, and z_i through the following GLM:

$$m_i | (\hat{p}_i, z_i, d_i^m) \sim \text{NB}_{\theta^m}(\mu_i); \quad \log(\mu_i) = \beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i + \log\left(d_i^m\right), \tag{1}$$

where (i) $NB_{\theta^m}(\mu_i)$ is a negative binomial distribution with mean μ_i and known size parameter θ^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.

3. Fit a GLM to obtain an estimate and confidence interval for the target of inference β_1^m.

The sequencing depth d_i^m is included as an offset term in (1) so that $\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i$ can be interpreted as a relative expression. Exponentiating both sides of (1) yields

$$\mu_i = \exp\left(\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i\right) d_i^m.$$

We see that $\exp\left(\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i\right)$ is the *fraction* of all transcripts sequenced in the cell produced by the gene under consideration.

Fold change in this context is defined as the ratio of mean gene expression in perturbed cells to mean gene expression in unperturbed cells. The target of inference β_1^m is the *log* fold change in expression in response to the perturbation, controlling for the technical factors. Therefore, $\exp(\beta_1^m) = 1$ (eq., $\beta_1^m = 0$) indicates no change in expression, whereas $\exp(\beta_1^m) > 1$ (eq., $\beta_1^m > 0$) and $\exp(\beta_1^m) < 1$ (eq., $\beta_1^m < 0$) indicate an increase and decrease in expression, respectively.

3.1 Empirical challenges of the thresholding method We examined the behavior of the thresholding method on real data and uncovered attenuation bias and bias-variance tradeoff effects. We applied the thresholding method to analyze the set of 322 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 the selected threshold, we deployed the thresholding method to the positive control data using three different choices for the threshold: 1, 5, and 20. We found that the chosen threshold substantially impacted the results (Figure 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. (Figure 2a.) Estimates produced by threshold = 5 and threshold = 20 were more concordant (Figure 2b).

We reasoned that thresholded regression systematically underestimated true effect sizes, especially for small thresholds. 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 (Figures 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 distributions and checked for bimodality. Figures 2e and 2f display the empirical distribution of a randomly-selected gRNA from the Gasperini and Xie datasets, respectively (counts of 0 omitted). The distributions peak at 1 and then taper off gradually; there does not exist a sharp boundary that cleanly separates 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 bias-variance tradeoff; and (iii) the gRNA count distributions do not imply a clear threshold selection strategy.

3.2 Theoretical challenges of the thresholding method Next, we studied the thresholding method from a theoretical perspective, recovering in precise mathematical terms phenomena revealed in the empirical analysis (plus several others). We worked in a simplified Gaussian setting. Suppose we observe gRNA expression and gene expression data $(g_1, m_1), \ldots, (g_n, m_n)$ on n cells

from the following linear model:

$$m_i = \beta_0^m + \beta_1^m p_i + \epsilon_i; \quad g_i = \beta_0^g + \beta_1^g p_i + \tau_i; \quad p_i \sim \text{Bern}(\pi); \quad \epsilon_i, \tau_i \sim N(0,1), \tag{2}$$

where p_i, τ_i, and ϵ_i are independent. For a given threshold $c \in \mathbb{R}$, the imputed perturbation assignment $\hat{p}_i$ is $\hat{p}_i = \mathbb{I}(g_i \geq c)$. The thresholding estimator $\hat{\beta}_1^m$ is the OLS solution, i.e. $\hat{\beta}_1^m = \left[\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})(m_i - \overline{m})\right] \left[\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})^2\right]^{-1}$. We derive the almost sure limit of $\hat{\beta}_1^m$:

Proposition 1. *The almost sure limit (as $n \to \infty$) of $\hat{\beta}_1^m$ is*

$$\hat{\beta}_1^m \xrightarrow{a.s.} \beta_1^m \left(\frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \right) \equiv \beta_1^m \gamma(\beta_1^g, \pi, c, \beta_0^g), \tag{3}$$

where

$$\mathbb{E}[\hat{p}_i] = \zeta(1-\pi) + \omega\pi; \quad \omega = \Phi\left(\beta_1^g + \beta_0^g - c\right); \quad \zeta = \Phi\left(\beta_0^g - c\right).$$

The function $\gamma : \mathbb{R}^4 \to \mathbb{R}$ does not depend on the gene expression parameters β_1^m or β_0^m. The asymptotic relative bias $b : \mathbb{R}^4 \to \mathbb{R}$ of $\hat{\beta}_1^m$ is given by

$$b(\beta_1^g, \pi, c, \beta_0^g) \equiv \frac{1}{\beta_1^m} \left(\beta_1^m - \lim_{\text{a.s.}} \hat{\beta}_1^m \right) = 1 - \gamma(\beta_1^g, \pi, c, \beta_0^g).$$

Having derived an exact expression for the asymptotic relative bias of $\hat{\beta}_1^m$, we can prove several results about this quantity. For the sake of simplicity, we consider a slightly specialized setting in which we set π to to a specific value (see Appendix A for details; proofs and detailed statements likewise deferred to Appendix A).

First, the thresholding estimator strictly *underestimates* (in absolute value) the true value of β_1^m over all choices of the threshold and over all possible values of the model parameters. This phenomenon, called attenuation bias, is a common attribute of estimators that ignore measurement in errors-in-variables models [30]. Second, 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. Third, the Bayes-optimal decision boundary $c_{\text{bayes}} \in \mathbb{R}$ (i.e., the most accurate decision boundary for classifying cells as perturbed or unperturbed) is a critical value of the bias function. Finally, and most subtly, there is no universally applicable rule

for selecting a threshold that produces minimal bias: when β_1^g is small, setting the threshold to an arbitrarily large number yields smaller bias than setting the threshold to the Bayes decision boundary; when β_1^g is large, the reverse is true.

Next, we studied the variance of the thresholding estimator, considering a slightly simpler model for this purpose. Suppose the intercepts in (2) are fixed at 0 (i.e., $\beta_0^m = \beta_0^g = 0$). For notational simplicity write $\beta_m = \beta_1^m$ and $\beta_g = \beta_1^g$. The thresholding estimator $\hat{\beta}_m$ is the no-intercept OLS solution $\hat{\beta}_m = \left[\sum_{i=1}^n \hat{p}_i m_i\right] \left[\sum_{i=1}^n \hat{p}_i^2\right]^{-1}$. We have the following limiting distributional result.

Proposition 2. *The limiting distribution of* $\hat{\beta}_m$ *is*

$$\sqrt{n}(\hat{\beta}_m - l) \xrightarrow{d} N\left(0, \frac{\beta_m \omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2)}{\left(\mathbb{E}[\hat{p}_i]\right)^2}\right),$$

where

$$l = \beta_m \omega\pi / [\zeta(1-\pi) + \omega\pi]; \quad \mathbb{E}[\hat{p}_i] = \pi\omega + (1-\pi)\zeta; \quad \omega = \Phi(\beta_g - c); \quad \zeta = \Phi(-c).$$

This proposition yields an asymptotically exact bias-variance decomposition for $\hat{\beta}_m$ (Figure 3): as the threshold tends to infinity, the bias decreases and the variance increases, consistent with empirical observations that we made in Section 3.1. Overall, our empirical and theoretical analyses reveal that thresholded regression poses several difficulties: thresholding incurs strict attenuation bias (that in some settings is large), and selecting a good threshold is challenging. These considerations motivate our core research question: *does modeling the gRNA counts directly, thereby circumventing thresholding, facilitate estimation and inference in single-cell CRISPR screen analysis*?

4 GLM-EIV

To answer this question, we introduce GLM-EIV (GLM-based errors-in-variables), a method that models the entire data-generating process in a single-cell CRISPR screen experiment.

4.1 Model Building on the work of several previous authors [12, 31, 32], Sarkar and Stephens [27] proposed a simple strategy for modeling single-cell gene expression data, which, in the frame-

work of negative binomial GLMs, is equivalent to using the log-transformed library size as an offset term (as in (1)). We generalize Sarkar and Stephens' approach to model *both* gene and gRNA modalities. To this end, let the latent variable $p_i \in \{0, 1\}$ indicate whether cell $i \in \{1, \ldots, n\}$ was perturbed. We model the gene expression counts according to

$$m_i | (p_i, z_i, d_i^m) \sim \text{NB}_{\theta^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + \log(d_i^m), \tag{4}$$

where $\theta^m > 0$ is a known negative binomial size parameter, and $\beta_0^m \in \mathbb{R}, \beta_1^m \in \mathbb{R}$, and $\gamma_m \in \mathbb{R}^{d-2}$ are unknown constants. The model (4) is identical to the thresholding model (1), but the imputed perturbation indicator $\hat{p}_i$ is replaced by the latent perturbation indicator p_i. Next, let $d_i^g \in \mathbb{N}$ be the number of gRNA transcripts sequenced across *all* gRNAs in cell i (i.e., the gRNA library size). The model for the gRNA counts is

$$g_i | (p_i, z_i, d_i^g) \sim \text{NB}_{\theta^g}(\mu_i^g); \quad \log(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + \log(d_i^g), \tag{5}$$

where, similar to above, $\theta^g > 0$ is a known negative binomial size parameter, and $\beta_0^g \in \mathbb{R}, \beta_1^g \in \mathbb{R}, \gamma_g \in \mathbb{R}^{d-2}$ are unknown constants. We use a negative binomial GLM to model the gRNA counts because the gRNA molecules are transcribed in the cell in the same way as gene transcripts [4, 13]. Finally, we model the marginal perturbation probability as $p_i \sim \text{Bern}(\pi)$, where $\pi \in (0, 1/2]$. Together, (4), (5), and the marginal distribution of p_i define the negative binomial GLM-EIV model. The terms $(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i)$ and $(\beta_0^g + \beta_1^g p_i + \gamma_g^T z_i)$ can be interpreted as relative gene and gRNA expressions, similar to the analogous term in the thresholding model. Likewise, the target of inference β_1^m is the log fold change in gene expression in response to the perturbation, accounting for technical factors. See Appendix C for a parallel methodology that we developed in which we model the gRNA distribution as (in a certain sense) zero-inflated counts that may be applicable in some settings.

General model To provide greater modeling flexibility, we generalize the GLM-EIV model to arbitrary exponential family response distributions and link functions. To increase notational compactness, let $\tilde{x}_i = [1, p_i, z_i]^T \in \mathbb{R}^d$ be the vector of covariates (including an intercept term) for the ith cell. (We use the tilde as a reminder that the vector is partially unobserved.) Let

$\beta_m = [\beta_0^m, \beta_1^m, \gamma_m]^T \in \mathbb{R}^d$ and $\beta_g = [\beta_0^g, \beta_1^g, \gamma_g]^T \in \mathbb{R}^d$ be the unknown coefficient vectors corresponding to the gene and gRNA expression models, respectively. Finally, let o_i^m and o_i^g be 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 $\log(d_i^g)$, respectively).

We use GLMs to model the gene and gRNA expressions. Considering first the gene expression model, let the ith linear component l_i^m of the model be $l_i^m = \langle \tilde{x}_i, \beta_m \rangle + o_i^m$. Next, let the mean μ_i^m of the ith observation be $r_m(\mu_i^m) = l_i^m$, where $r_m : \mathbb{R} \to \mathbb{R}$ is a strictly increasing, differentiable link function. Let $\psi_m : \mathbb{R} \to \mathbb{R}$ be the differentiable, cumulant-generating function of the selected exponential family distribution. We can express the canonical parameter η_i^m in terms of ψ_m and r_m by $\eta_i^m = \left([\psi_m']^{-1} \circ r_m^{-1}\right)(l_i^m) := h_m(l_i^m)$. Finally, let $c_m : \mathbb{R} \to \mathbb{R}$ be the carrying density of the selected exponential family distribution. The density f_m of m_i conditional on the the canonical parameter η_i is $f_m(m_i; \eta_i^m) = \exp\left\{m_i \eta_i^m - \psi_m(\eta_i^m) + c_m(m_i)\right\}$. The function c_m does not appear in the log likelihood of m_i; therefore, the only functions relevant to inference are ψ_m and r_m.

Let the terms $l_i^g, o_i^g, \mu_i^g, \eta_i^g, \psi_g, r_g, h_g$ and c_g be defined in an analogous way for the gRNA model:

$$l_i^g = \langle \tilde{x}_i, \beta_g \rangle + o_i^g; \quad r_g(\mu_i^g) = l_i^g; \quad \eta_i^g = \left([\psi_g']^{-1} \circ r_g^{-1}\right)(l_i^g) := h_g(l_i^g).$$

The density f_g of g_i given the canonical parameter is $f_g(m_i; \eta_i^g) = \exp\left\{g_i \eta_i^g - \psi_g(\eta_i^g) + c_g(g_i)\right\}$. Finally, the unobserved variable p_i is assumed to follow a Bernoulli distribution with mean $\pi \in (0, 1/2]$. Its marginal density f_p is given by $f_p(p_i) = \pi^{p_i}(1-\pi)^{1-p_i}$. The unknown parameters in the model are $\theta = [\beta_m, \beta_g, \pi]^T \in \mathbb{R}^{2d+1}$.

Notation We briefly introduce notation that we will use throughout. For $j \in \{0, 1\}$, let $\tilde{x}_i(j) := [1, j, z_i]^T$ denote the value of $\tilde{x}_i$ that results from setting p_i to j. Next, let $l_i^m(j)$, $\eta_i^m(j)$, and $\mu_i^m(j)$ be the values of l_i^m, η_i^m, and μ_i^m, respectively, that result from setting p_i to j, i.e.,

$$l_i^m(j) := \langle \tilde{x}_i(j), \beta_m \rangle + o_i^m; \quad \eta_i^m(j) := h_m(l_i^m(j)); \quad \mu_i^m(j) = r_m^{-1}(l_i^m(j)).$$

Let the corresponding gRNA quantities $l_i^g(j)$, $\eta_i^g(j)$, and $\mu_i^g(j)$ be defined analogously. Let $X \in \mathbb{R}^{n \times d-1}$ be the observed design matrix, and let $\tilde{X} \in \mathbb{R}^{n \times d}$ be the augmented design matrix that

results from concatenating the column of (unobserved) p_is to X, i.e.

$$X := \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix}; \quad \tilde{X} := \begin{bmatrix} 1 & p_1 & z_1 \\ 1 & p_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & p_n & z_n \end{bmatrix} = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}.$$

Furthermore, for $j \in \{0, 1\}$, let $\tilde{X}(j) \in \mathbb{R}^{n \times d}$ be the matrix that results from setting p_i to j for all $i \in \{1, \ldots, n\}$ in $\tilde{X}$, and let $[\tilde{X}(0)^T, \tilde{X}(1)^T]^T$ denote the $\mathbb{R}^{2n \times d}$ matrix that results from vertically concatenating $\tilde{X}(0)$ and $\tilde{X}(1)$. Furthermore, define $m := [m_1, \ldots, m_n]$, and let g, p, o^m, and o^g be defined analogously. Finally, let $[m, m]^T \in \mathbb{R}^{2n}$ be the vector that results from concatenating m to itself, i.e.

$$[m, m]^T := [\underbrace{m_1, m_2 \ldots, m_{n-1}, m_n}_{\text{first copy of } m}, \underbrace{m_1, m_2, \ldots, m_{n-1}, m_n}_{\text{second copy of } m}],$$

and let $[g, g]^T$, $[o^g, o^g]^T$, and $[o^m, o^m]^T$ be defined similarly.

Log likelihood and model properties We derive the log-likelihood of the GLM-EIV model. We conduct estimation and inference conditional on the library sizes and technical factors l_i^m, l_i^g, and z_i; therefore, we treat these quantities as fixed constants. We assume that the gene expression m_i and gRNA expression g_i are conditionally independent given the perturbation p_i. The the joint density f of (m_i, g_i, p_i) given θ is

$$f(m_i, g_i, p_i; \theta) = f_m(m_i|p_i) f_g(g_i|p_i) f_p(p_i) = \pi^{p_i}(1-\pi)^{1-p_i} f_m(m_i; \eta_i^m) f_g(g_i; \eta^g). \tag{6}$$

Integrating over the unobserved variable p_i, we can write the density f of (m_i, g_i) as

$$f(m_i, g_i; \theta) = (1-\pi) f_m(m_i; \eta_i^m(0)) f_g(g_i; \eta_i^g(0)) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^g(1)). \tag{7}$$

Therefore, the log-likelihood is

$$\mathcal{L}(\theta; m, g) = \sum_{i=1}^{n} \log \left[(1-\pi) f_m(m_i; \eta_i^m(0)) f_g(g_i; \eta_i^g(0)) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^g(1)) \right]. \tag{8}$$

We see from (7) that the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM densities. Additionally, the GLM-EIV model is a generalization the simple errors-in-variables model (when the predictor is binary); the latter is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i; \quad x_i = x_i^* + \tau_i, \tag{9}$$

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 (9) 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 linear *or* nonlinear link functions; and finally, GLM-EIV allows confounders z_i to impact both x_i and y_i. Therefore, x_i and y_i are dependent *even after conditioning on* x_i^*, enabling GLM-EIV to capture more complex dependence relationships between x_i and y_i than is possible in (9) or other standard measurement error models.

4.2 Estimation and inference We derive an EM algorithm (Algorithm 1) to estimate the parameters of the GLM-EIV model. 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 $(\beta_m^{(\text{t})}, \beta_g^{(\text{t})}, \pi^{(\text{t})})$ and observed data (m_i, g_i) is

$$T_i(1) = \mathbb{P}(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{(\text{t})}, \beta_g^{(\text{t})}, \pi^{(\text{t})}).$$

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 $\pi^{(\text{t}+1)}$, $\beta_g^{(\text{t}+1)}$, and $\beta_m^{(\text{t}+1)}$ of the parameters by maximizing the M step objective function. It turns out that maximizing this objective function is equivalent to setting $\pi^{(\text{t}+1)}$ to the mean of the current membership probabilities and setting $\beta_g^{(\text{t}+1)}$ and $\beta_m^{(\text{t}+1)}$ 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 (8) converges (see Appendix B). Our EM algorithm is reminiscent of (but distinct from) that of Ibrahim [14], who also applied weighted GLM solvers to carry out the M step.

After fitting the model, we perform inference on the estimated parameters. The easiest ap-

Algorithm 1 EM algorithm for GLM-EIV model.

Input: Pilot estimates $\beta_m^{\text{curr}}, \beta_g^{\text{curr}}$, and π^{curr}; data m, g, o^m, o^g, and X; gene expression distribution f_m and link function r_m^{-1}; gRNA expression distribution f_g and link function r_g^{-1}.

while Not converged **do**

  for $i \in \{1, \ldots, n\}$ **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

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

  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^{-1}.

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

proach, 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 [18] (Appendix B). Leveraging this analytic formula, we can calculate standard errors quickly, enabling us to perform inference in practice on real, large-scale data.

4.3 Statistical accelerations A downside of the 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, 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 devised a strategy to produce a highly accurate pilot estimate $(\pi^{\text{pilot}}, \beta_m^{\text{pilot}}, \beta_g^{\text{pilot}})$ 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" (some new, some old) on top of one another. We expect these tricks to be independently useful for accelerating other single-cell methods.

The first step (Algorithm 2) is to obtain good parameter estimates for $[\beta_0^m, \gamma_m]^T$ and $[\beta_0^g, \gamma_g]^T$ via regression. Recall that the underlying gene expression parameter vector β_m is $\beta_m = [\beta_0^m, \beta_1^m, \gamma_m]^T \in$

Algorithm 2 Computing $[\beta_0^m]^{\text{pilot}}$, $[\gamma_m^T]^{\text{pilot}}$, $[\beta_0^g]^{\text{pilot}}$, and $[\gamma_g^T]^{\text{pilot}}$.

Input: Data m, g, o^m, o^g, and X; gene expression distribution f_m and link function r_m^{-1}; gRNA expression distribution f_g and link function r_g^{-1}; number of EM starts B.

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

Fit a GLM GLM_k with responses k, offsets o^k, design matrix X, distribution f_k, and link function r_k^{-1}.

Set $[\beta_0^k]^{\text{pilot}}$ and $[\gamma_k^T]^{\text{pilot}}$ to the fitted coefficients of GLM_k.

for $i \in \{1, \ldots, n\}$ **do**

$\hat{f}_i^k \leftarrow [\beta_0^k]^{\text{pilot}} + \langle [\gamma_k^T]^{\text{pilot}}, z_i \rangle + o_i^k$ ▷ untransformed fitted values

end for

end for

return ($[\beta_0^m]^{\text{pilot}}$, $\hat{f}^m$, $[\gamma_m^T]^{\text{pilot}}$, $[\beta_0^g]^{\text{pilot}}$, $[\gamma_g^T]^{\text{pilot}}$, $\hat{f}^g$)

$\mathbb{R}^d$, where β_0^m is the intercept, β_1^m is the effect of the perturbation, and γ_m^T is the effect of the technical factors. To produce estimates $[\beta_0^m]^{\text{pilot}}$ and $[\gamma_m^T]^{\text{pilot}}$, we regress the gene expressions m onto the technical factors X. The intuition for this procedure is as follows: the probability of perturbation π is very small. Therefore, the true log likelihood is approximately equal to the log likelihood that results from omitting p_i from the model:

$$\sum_{i=1}^n f_m(m_i; \eta_i^m) = \underbrace{\sum_{i:p_i=1} f_m(m_i; h_m(\beta_0 + \beta_1 + \gamma^T z_i + o_i^m))}_{\text{few terms}} + \underbrace{\sum_{i:p_i=0} f_m(m_i; h_m(\beta_0 + \gamma^T z_i + o_i^m))}_{\text{many terms}}$$
$$\approx \sum_{i=1}^n f_m(m_i; h_m(\beta_0 + \gamma^T z_i + o_i^m)).$$

We similarly can obtain pilot estimates $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$ by regressing the gRNA counts g onto the technical factors X. We extract the fitted values (on the scale of the linear component) for use in a subsequent step: $\hat{f}_i^k = [\beta_0^k]^{\text{pilot}} + \langle [\gamma_k^T]^{\text{pilot}}, z_i \rangle + o_i^k$, for $k \in \{m, g\}$.

Next, we obtain estimates $[\beta_1^m]^{\text{pilot}}$, $[\beta_1^g]^{\text{pilot}}$, and π^{pilot} for β_1^m, β_1^g, and π by fitting a "reduced" GLM-EIV. We describe the reduced GLM-EIV procedure in text; the algorithm box (Algorithm 5) is deferred to Appendix D. The log likelihood of the no-intercept, univariate GLM with predictor p_i and offset $\hat{f}_i^m$ is approximately equal to the true log likelihood:

$$\sum_{i=1}^n f_m(m_i; \eta_i^m) = \sum_{i=1}^n f_m(m_i; h_m(\beta_0 + \beta_1 p_i + \gamma^T z_i + o_i^m)) \approx \sum_{i=1}^n f_m(m_i; h_m(\beta_1 p_i + \hat{f}_i^m)).$$

Therefore, to estimate β_1^m, β_1^g, and π, we fit a GLM-EIV model with gene expressions m, gRNA

counts g, gene offsets $\hat{f}^m := [\hat{f}_1^m, \ldots, \hat{f}_n^m]^T$, gRNA offsets $\hat{f}^g := [\hat{f}_1^g, \ldots, \hat{f}_n^g]^T$, and *no* intercept or covariate terms. Intuitively, we "encode" all information about technical factors, library sizes, and baseline expression levels into $\hat{f}^m$ and $\hat{f}^g$. We run the algorithm $B \approx 15$ times over randomly-selected starting values for β^m, β^g, and π and select the solution with greatest the log likelihood.

The M step of the reduced GLM-EIV algorithm requires fitting two no-intercept, univariate GLMs with offsets. We derive analytic formulas for the MLEs of these GLMs in the three most important cases: Gaussian response with identity link, Poisson response with log link, and negative binomial response with log link (see Appendix D; the latter formula is asymptotically exact). Consequently, we do not need to run the relatively slow IRLS procedure to carry out the M step of the reduced GLM-EIV algorithm. Overall, the proposed method for obtaining the full set of pilot parameter estimates requires fitting only two GLMs (via IRLS).

4.4 Computing 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 [6], and ondisc, an R package that we developed (in a separate project) 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 tens or hundreds of nodes on a cloud (e.g., Microsoft Azure, Google Cloud) or high-performance cluster.

Algorithm 3 Applying GLM-EIV at scale.

$G \leftarrow \{\text{gene}_1, \ldots, \text{gene}_{d_g}\}$; $P \leftarrow \{\text{perturbation}_1, \ldots, \text{perturbation}_{d_p}\}$
for gene $\in G$ **do**
  Run precomputation (Algorithm 2) on gene; save $\hat{f}^m$, $[\beta_0^m]^{\text{pilot}}$ and $[\gamma_m^T]^{\text{pilot}}$.
end for
for perturbation $\in P$ **do**
  Run precomputation (Algorithm 2) on perturbation; save $\hat{f}^g$, $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$.
end for
for (gene, perturbation) $\in G \times P$ **do**
  Load $\hat{f}^m, \hat{f}^g$, $[\beta_0^m]^{\text{pilot}}$ $[\gamma_m^T]^{\text{pilot}}$, $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$.
  Compute $[\beta_1^m]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}}, \pi^{\text{pilot}}$ by fitting a reduced GLM-EIV (Algorithm 5).
  Run GLM-EIV using the pilot parameters (Algorithm 1).
end for

Leveraging these technologies, we develop a scalable and efficient pipeline for GLM-EIV (Algorithm 3). First, we run a round of "precomputations" on all d_g genes and d_p perturbations.

The precomputations involve regressing the gene expressions (or gRNA counts) onto the technical factors, thereby "factoring out" Algorithm 2. Next, we run differential expression analyses on the full set of gene-perturbation pairs; for a given pair, this amounts to obtaining the complete set of pilot parameters (by running a reduced GLM-EIV), fitting the GLM-EIV model (Algorithm 1), and performing inference. The three loops in Algorithm 3 are embarassingly parallel and therefore can be massively parallelized.

5 Simulation studies

We conducted a simulation study to compare the empirical performance of GLM-EIV to that of the thresholding method. We generated data on $n = 150,000$ cells from the GLM-EIV model using realistic parameter values, setting the 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 problems. Finally, we generated the gene expression and gRNA count data from two response distributions: Poisson and negative binomial (size parameter fixed at $\theta = 20$ for the latter). For each parameter setting (defined by a β_1^g-distribution pair), we synthesized $n_{\text{sim}} = 500$ i.i.d. datasets. Section E presents additional simulation results on Gaussian response distributions.

We applied three methods to the simulated data: "vanilla" GLM-EIV, accelerated GLM-EIV, and thresholded regression. We used the Bayes-optimal decision boundary for classification as the threshold for the thresholding method. We ran all methods on the negative binomial data twice: once treating the size parameter θ as a known constant and once treating θ as unknown. In the latter case we used the `glm.nb` function from the `MASS` package to estimate θ before applying the methods [25]. We display the results of the simulation study in Figure 4. Columns correspond to distributions (i.e., Poisson, NB with known θ, and NB with unknown θ), rows correspond to performance metrics (i.e., bias, mean squared error, CI coverage rate (nominal rate 95%), CI width,

and method execution time). The problem difficulty parameter β_1^g is plotted on the horizontal axis, and the methods are depicted in different colors (GLM-EIV masked by accelerated GLM-EIV in several panels).

First, we observed that GLM-EIV dominated thresholded regression on all statistical metrics: GLM-EIV exhibited lower bias (row 1) and mean squared error (row 2) than thresholded regression; additionally, GLM-EIV had superior confidence interval coverage (row 3) despite having produced generally narrower confidence intervals (row 4). Intuitively, GLM-EIV outperformed the thresholding 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 found that 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 execution time of accelerated GLM-EIV was almost within an order of magnitude of that of the thresholding method (bottom row).

Interestingly, thresholded regression exhibited better confidence interval coverage under estimated θ than under known θ (row 3). Estimating θ leads to slight inflation bias (i.e., overestimating the true effect size), whereas, as we showed previously, thresholding leads to attenuation bias (i.e., underestimating the true effect size). These phenomena partially cancel, yielding less biased estimates. GLM-EIV exhibited worse performance under unknown θ than known θ, likely due to poor θ estimation. We note that GLM-EIV and the thresholding method in principal are compatible with *any* θ estimation procedure, including those based on more sophisticated techniques, such as regularization [12]. We defer rigorous investigation of the impact of different θ estimation strategies on these methods to future work.

6 Data analysis

Leveraging our computational infrastructure (Section 4.4), we applied GLM-EIV and the thresholding method to analyze the entire Gasperini and Xie datasets. (We report only the most important aspects of the analysis and results in the main text; full details are available in Appendix F.) We set the threshold to the approximate Bayes-optimal decision boundary, as our theoretical analyses

and simulation studies revealed that the Bayes-optimal decision boundary is a good choice for the threshold when the gRNA count 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. We repeated this procedure on both datasets, yielding a threshold of 3 for Gasperini and 7 for Xie. These thresholds were close to the original thresholds, which were selected in a more heuristic way.

We compared GLM-EIV to thresholded regression on the real data, focusing specifically on the negative control pairs (i.e., gene-perturbation pairs that, by design, are expected to exhibit a fold change of 1, or no association). We found that GLM-EIV and the thresholding method produced similar results (Figure 5a-b): estimates, CI coverage rates, and CI widths were concordant. The estimated effect of the perturbation on gene expression $\exp(\hat{\beta}_1^g)$ was unexpectedly large: the 95% CI for this parameter was $[4306, 5186]$ and $[300, 316]$ on the Gasperini and Xie data, respectively. We reasoned that the datasets lay in an "easy" region of the parameter space, making thresholding 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 (likely due to non-global EM convergence) on a small ($< 2.5\%$ on Gasperini, $< 0.05\%$ on Xie) number of pairs consisting of a handful of genes (not plotted).

We artificially 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* 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 Figures 5c-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 (Figure 5d). Next, we computed the median REC across all positive control pairs (Figure 5e; 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, taking the estimate obtained on the raw data as "ground truth," we computed CI coverage across pairs as a function of excess contamination. GLM-EIV exhibited significantly higher CI coverage than thresholded regression as the data became increasingly contaminated (Figure 5f; bands indicate 95% pointwise CIs for population coverage).

7 Discussion

In this work we introduced 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 response distributions and sources of measurement error that are exponential family-distributed and potentially impacted by the same set of confounding variables. These extensions enable GLM-EIV to resolve novel analysis challenges posed by single-cell CRISPR screens. We demonstrated through simulation studies, real data analyses, and theory that GLM-EIV outperforms thresholded regression by a considerable margin in high background contamination settings. GLM-EIV intuitively

achieves this performance gain by leveraging information from *both* modalities (rather than the gRNA modality alone) to assign perturbation identities to cells. On the other hand, in low background contamination settings GLM-EIV and thresholded regression work best in conjunction, with GLM-EIV providing a statistically principled and empirically effective procedure for selecting the threshold. GLM-EIV thereby neutralizes a tuning parameter that, until this point, has been selected using heuristic procedures, with little confidence that the choice is near optimal. Figure 6 summarizes how we anticipate GLM-EIV being used in practice.

To our knowledge this is the first single-cell CRISPR screen paper oriented toward a statistical audience. We hope that this work helps to introduce the broader statistics community to an emerging class functional genomics assays that likely will exert a major impact on biology research in the coming years [23]. Additionally, this is the first work to leverage the ondisc-Nextflow-cloud/HPC technology stack, a tightly-integrated, user-friendly, and powerful set of computational tools for large-scale single-cell analysis that we expect to be of interest to other researchers.

We anticipate that GLM-EIV could be applied to other types of single-cell CRISPR screen and multimodal single-cell data. For example, we likely (with some effort) could extend GLM-EIV to "low-multiplicity of infection" screens [29] in which each cell receives one or two perturbations rather than dozens (as is the case in "high multiplicity screens," studied in this work). We also likely could apply GLM-EIV to analyze multimodal 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) [20]. We do not directly observe the chromatin state of a cell; instead, we observe tagged DNA fragments that serve as proxies for whether a given region of chromatin is open or closed. These tagged DNA fragments come in the form of discrete counts; therefore, GLM-EIV might be applied in such experiments to aid in the selection of thresholds or to analyze whole datasets.

8 Acknowledgments and code availability

We thank Xuran Wang helping to process the Xie dataset. We also thank Songcheng Dai for helping to deploy the GLM-EIV pipeline on Azure. 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. Code for the analyses is available at `/timothy-barry/glmeiv-manuscript`.

References

[1] Dennis J Aigner. "Regression with a binary independent variable subject to errors of observation". In: *Journal of Econometrics* 1.1 (1973), pp. 49–59 (cit. on p. 3).

[2] Timothy Barry et al. "SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis". In: *Genome Biology, to appear* (2021) (cit. on p. 3).

[3] Emmanuel Candès et al. "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 (2018), pp. 551–577 (cit. on p. 3).

[4] Paul Datlinger et al. "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature Methods* 14.3 (2017), pp. 297–301 (cit. on pp. 1, 11).

[5] Paul Datlinger et al. "Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing". In: *Nature Methods* 18.6 (2021), pp. 635–642 (cit. on p. 4).

[6] Paolo DI Tommaso et al. "Nextflow enables reproducible computational workflows". In: *Nature Biotechnology* 35.4 (2017), pp. 316–319 (cit. on p. 17).

[7] Atray Dixit et al. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". In: *Cell* 167.7 (2016), 1853–1866.e17 (cit. on p. 1).

[8] Michael D. Gallagher and Alice S. Chen-Plotkin. "The Post-GWAS Era: From Association to Function". In: *American Journal of Human Genetics* 102.5 (2018), pp. 717–730 (cit. on p. 4).

[9] Molly Gasperini, Jacob M. Tome, and Jay Shendure. "Towards a comprehensive catalogue of validated and target-linked human enhancers". In: *Nature Reviews Genetics* 21.5 (2020), pp. 292–310 (cit. on p. 4).

[10] Molly Gasperini et al. "A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens". In: *Cell* 176.1-2 (2019), 377–390.e19 (cit. on pp. 4, 6).

[11] Bettina Grün and Friedrich Leisch. "Finite Mixtures of Generalized Linear Regression Models". In: *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*. Heidelberg: Physica-Verlag HD, 2008, pp. 205–230 (cit. on p. 3).

[12] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1 (2019), pp. 1–15 (cit. on pp. 10, 19).

[13] Andrew J. Hill et al. "On the design of CRISPR-based single-cell molecular screens". In: *Nature Methods* 15.4 (2018), pp. 271–274 (cit. on p. 11).

[14] Joseph G. Ibrahim. "Incomplete Data in Generalized Linear Models". In: *Journal of the American Statistical Association* 85.411 (1990), pp. 765–769 (cit. on pp. 3, 14).

[15] Jan Lause, Philipp Berens, and Dmitry Kobak. "Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data". In: *Genome Biology* 22.1 (2021), pp. 1–20 (cit. on p. 3).

[16] Kevin Z. Lin, Jing Lei, and Kathryn Roeder. "Exponential-Family Embedding With Application to Cell Developmental Trajectories for Single-Cell RNA-Seq Data". In: *Journal of the American Statistical Association* 0.0 (2021), pp. 1–32 (cit. on p. 3).

[17] Molei Liu et al. "Fast and Powerful Conditional Randomization Testing via Distillation". In: *Biometrika* (2021), pp. 1–25 (cit. on p. 3).

[18] By Thomas A Louis. "Finding the Observed Information Matrix when Using the EM Algorithm". In: *Society* 44.2 (1982), pp. 226–233 (cit. on p. 15).

[19] Eleni P. Mimitou et al. "Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells". In: *Nature Methods* 16.5 (2019), pp. 409–412 (cit. on p. 4).

[20] Eleni P. Mimitou et al. *Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells*. Vol. 39. 10. Springer US, 2021, pp. 1246–1258 (cit. on p. 22).

[21] John A. Morris et al. "Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing". In: *bioRxiv* (2021), p. 2021.04.07.438882 (cit. on p. 1).

[22] Kiran Musunuru et al. “In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates”. In: *Nature* 593.7859 (2021), pp. 429–434 (cit. on p. 1).

[23] Laralynne Przybyla and Luke A. Gilbert. “A new era in functional genomics screens”. In: *Nature Reviews Genetics* 0123456789 (2021) (cit. on pp. 1, 22).

[24] Joseph M. Replogle et al. “Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing”. In: *Nature Biotechnology* (2020) (cit. on p. 5).

[25] Brian Ripley et al. “Package ‘mass’”. In: *Cran r* 538 (2013), pp. 113–120 (cit. on p. 18).

[26] Tanja Rothgangl et al. “In vivo adenine base editing of PCSK9 in macaques reduces LDL cholesterol levels”. In: *Nature Biotechnology* 39.8 (2021), pp. 949–957 (cit. on p. 1).

[27] Abhishek Sarkar and Matthew Stephens. “Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis”. In: *Nature Genetics* 53.6 (2021), pp. 770–777 (cit. on p. 10).

[28] E. Savoca. “Measurement errors in binary regressors: An application to measuring the effects of specific psychiatric diseases on earnings”. In: *Health Services and Outcomes Research Methodology* 1.2 (2000), pp. 149–164 (cit. on p. 3).

[29] Daniel Schraivogel et al. “Targeted Perturb-seq enables genome-scale genetic screens in single cells”. In: *Nature Methods* 17.6 (2020), pp. 629–635 (cit. on p. 22).

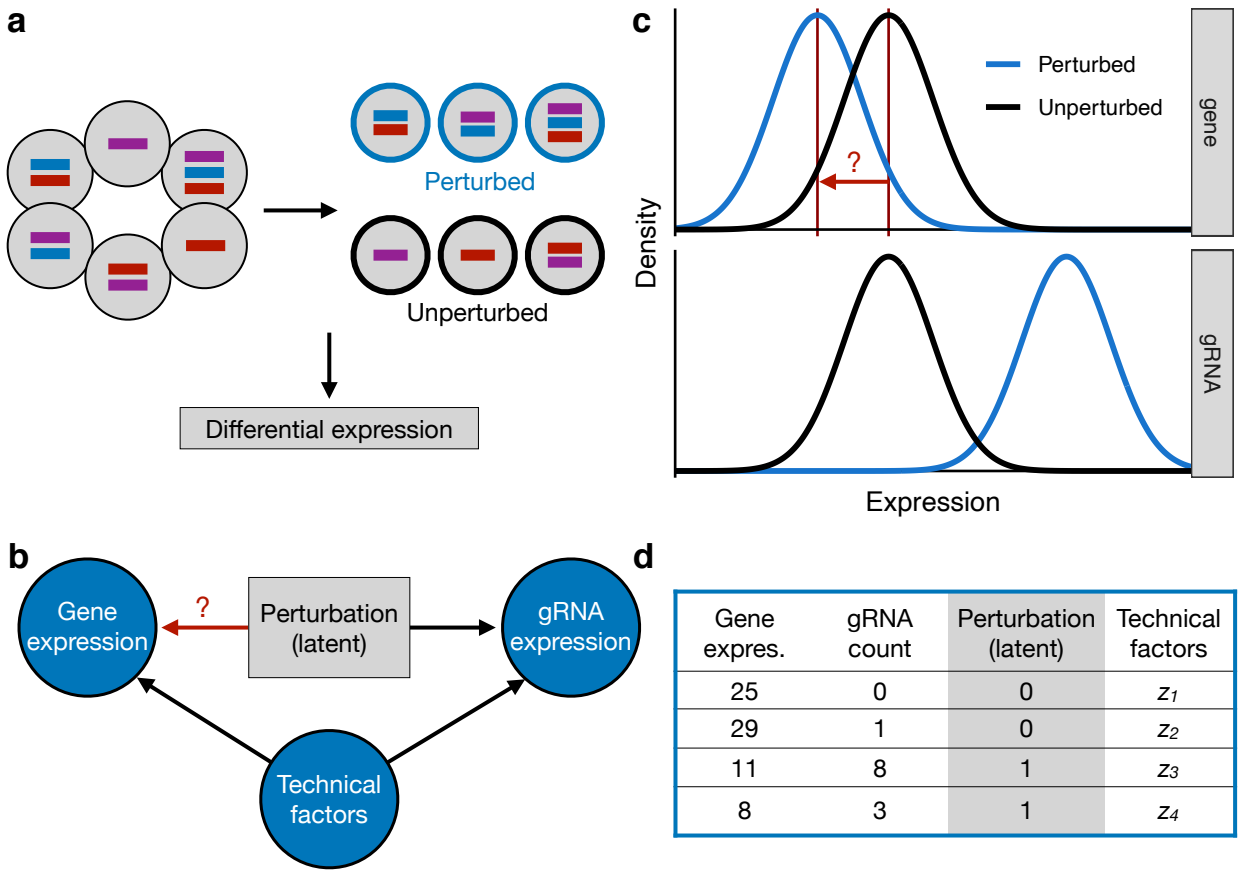
[30] L. A. Stefanski. “Measurement Error Models”. In: *Journal of the American Statistical Association* 95.452 (2000), pp. 1353–1358 (cit. on p. 9).

[31] Valentine Svensson. “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38 (2020), pp. 142–150 (cit. on p. 10).

[32] F. William Townes et al. “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome Biology* 20.1 (2019), pp. 1–16 (cit. on pp. 3, 10).

[33] Shiqi Xie et al. “Global Analysis of Enhancer Targets Reveals Convergent Enhancer-Driven Regulatory Modules”. In: *Cell Reports* 29.9 (2019), 2570–2578.e5 (cit. on p. 6).

Figures



d

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.

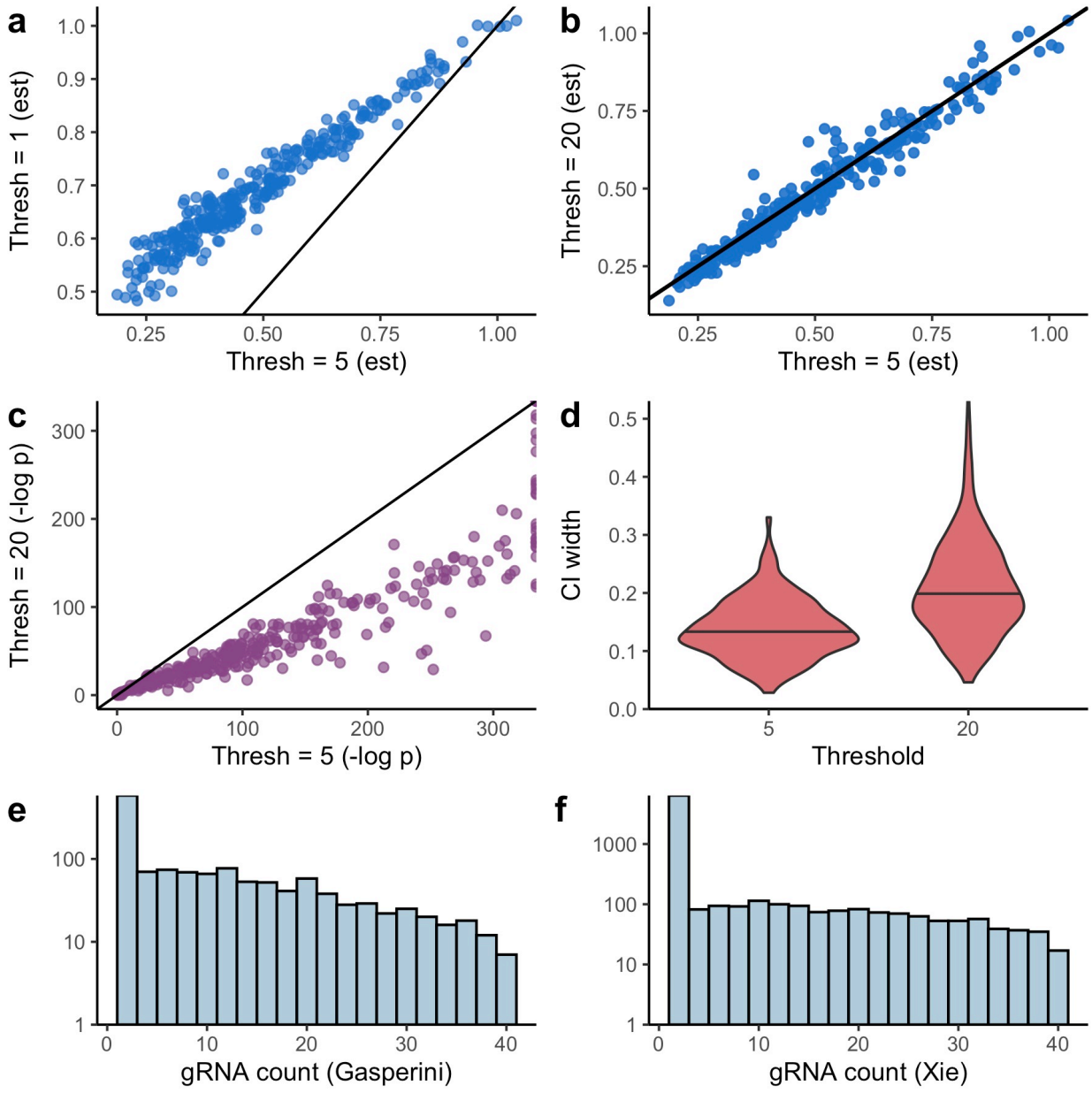


Figure 2: **Empirical challenges of thresholded regression. a-b,** Fold change estimates produced by threshold = 1 versus threshold = 5 (a) and threshold = 20 versus threshold = 5 (b). The selected threshold substantially impacts the results. **c-d,** p-values (c) and CI widths (d) produced by threshold = 20 versus threshold = 5. The latter threshold yields more confident estimates. **e-f**, Empirical distribution of randomly-selected gRNA from Gasperini (e) and Xie (f) data (0 counts not shown). The gRNA data do not appear to imply an obvious threshold selection strategy.

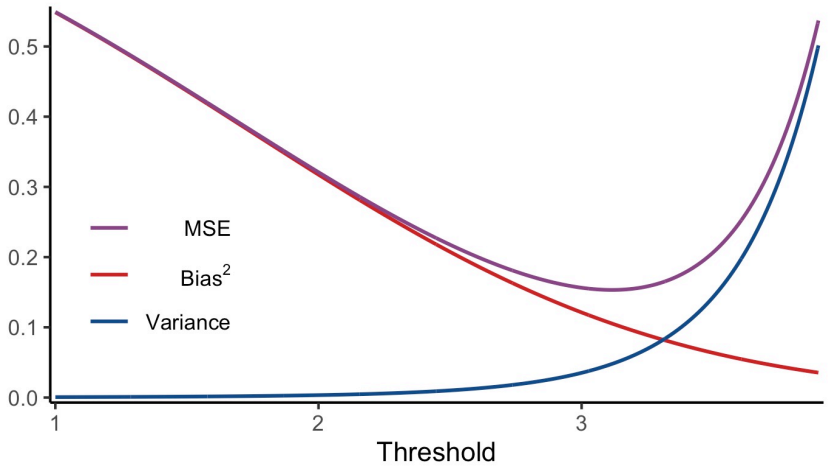


Figure 3: **Thresholding method bias-variance decomposition.** Bias decreases and variance increases as the threshold tends to infinity. $\beta_1^g = 1, \beta_1^m = 1$, and $\pi = 0.1$ in this plot.

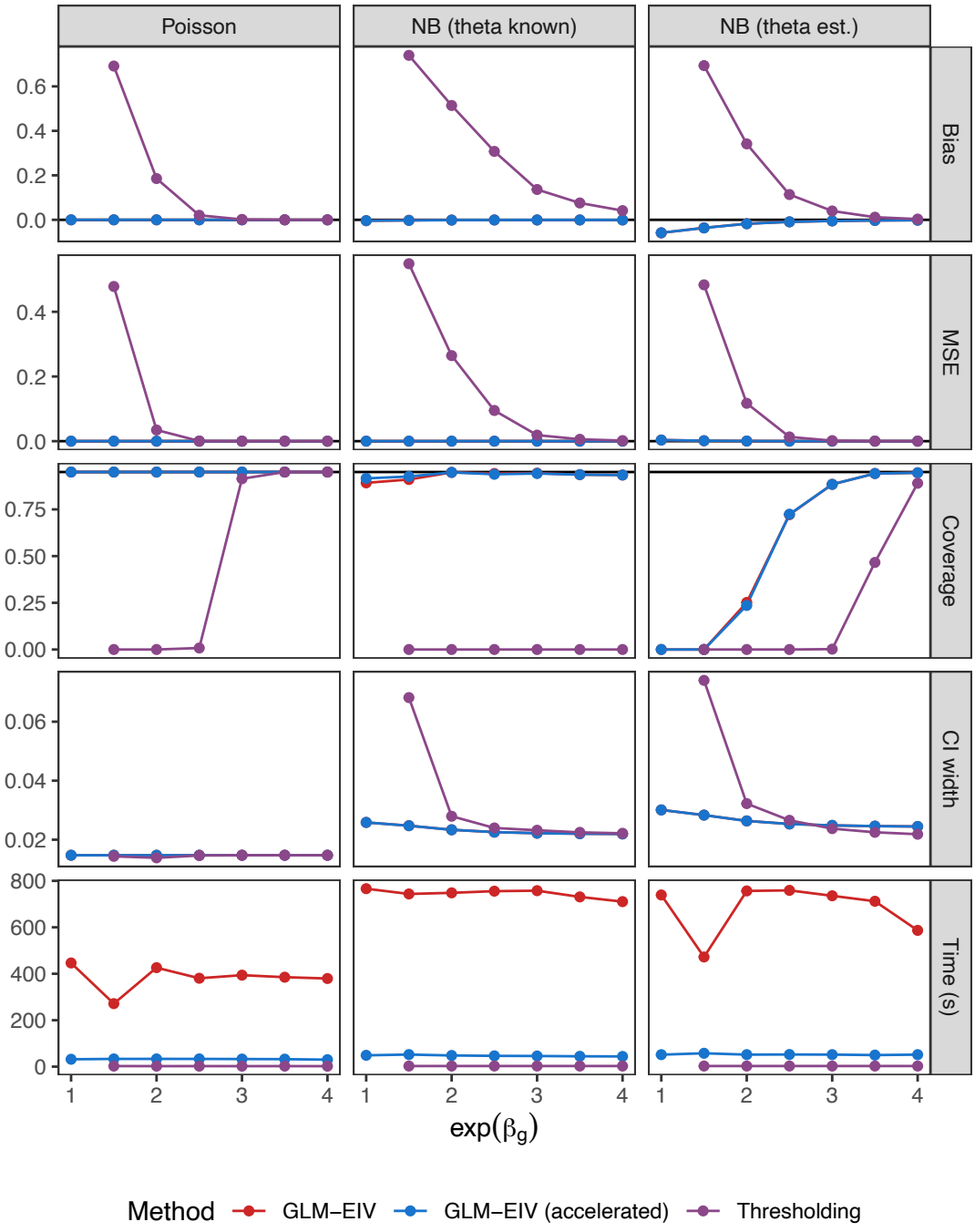


Figure 4: **Simulation study**. Columns correspond to distributions (Poisson, NB with known θ, NB with estimated θ), and rows correspond to metrics (bias, MSE, coverage, CI width, and time). Methods are shown in different colors; GLM-EIV (red) is masked by accelerated GLM-EIV (blue) in several panels. GLM-EIV demonstrated superior statistical performance to the thresholding method on all metrics (rows 1-4). Accelerated GLM-EIV had substantially lower computational cost than "vanilla" GLM-EIV (bottom row) despite demonstrating identical statistical performance (rows 1-4).

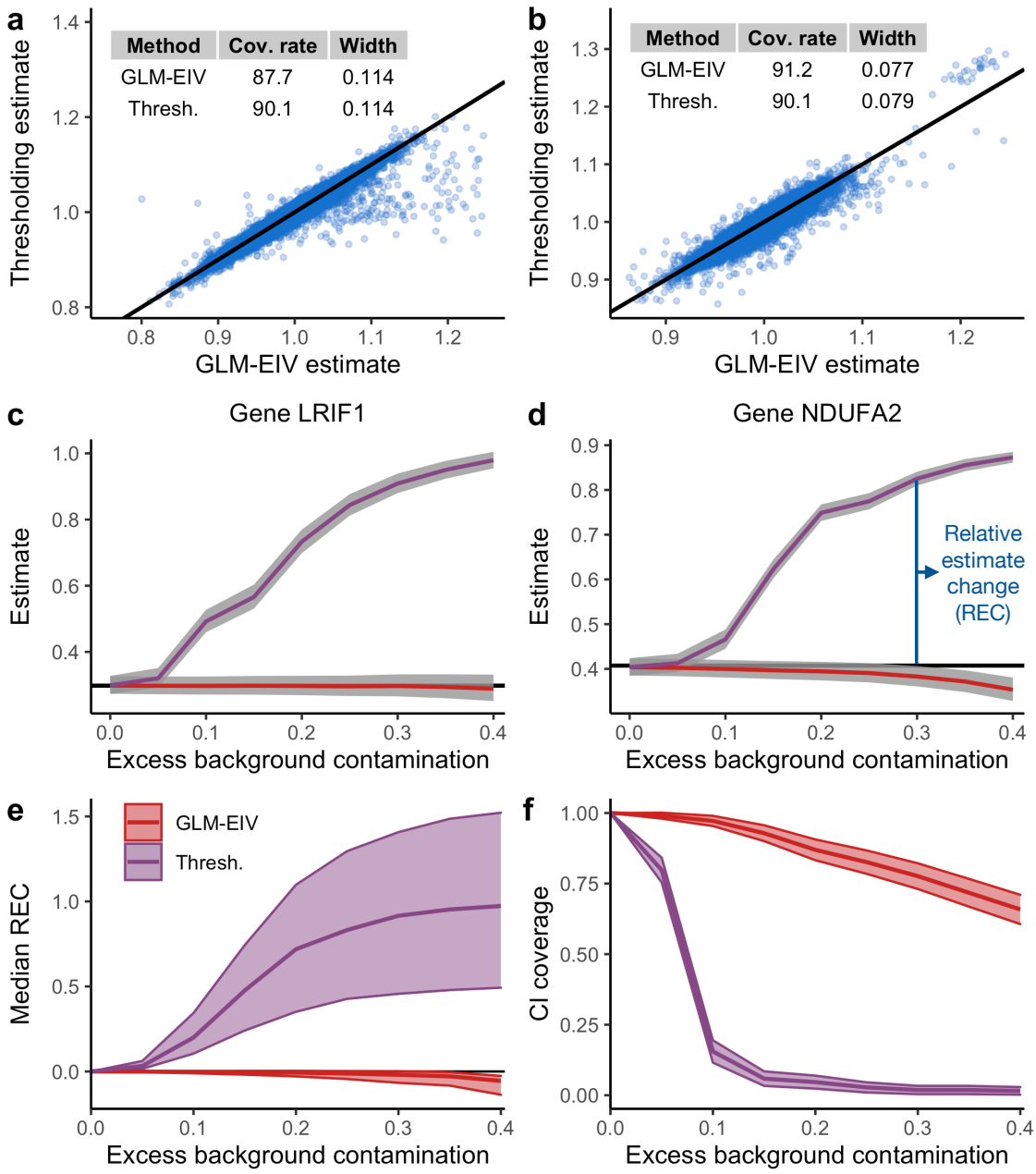


Figure 5: **Data analysis**. **a-b**, Estimates for fold change produced by GLM-EIV and thresholded regression on Gasperini (**a**) and Xie (**b**) negative control pairs. **c-d**, Estimates produced by GLM-EIV and thresholded regression on two positive control pairs – *LRIF1* (**a**) and *NDUFA2* (**b**) – 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. Panels (**c-f**) together illustrate that GLM-EIV demonstrated greater stability than thresholded regression as background contamination increased.

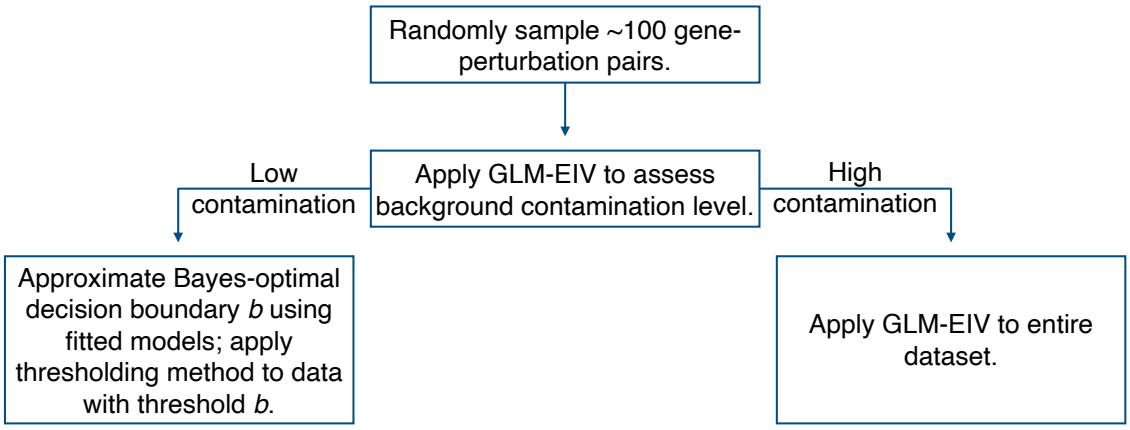


Figure 6: **Use of GLM-EIV in practice**. The decision tree above illustrates how we anticipate GLM-EIV could be used in practice. First, apply GLM-EIV to a set of randomly-sampled gene-perturbation pairs to assess background contamination level (positive control pairs work best for this purpose). If GLM-EIV indicates that background contamination is high (e.g., $\exp(\beta_1^g) \lessapprox 10$), apply GLM-EIV to analyze the entire dataset; otherwise, approximate the Bayes-optimal decision boundary using the fitted GLM-EIV models. Next, apply a thresholding method (e.g., SCEPTRE or thresholded negative binomial regression) to analyze the data, setting the threshold to the estimated Bayes-optimal decision boundary.