Tim, Gene, Kathryn

CRISPR genome editing, single-cell sequencing, and exponential family measurement error models

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. In this work we study single-cell CRISPR screens from a statistical perspective. First, we demonstrate on real data 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. We recover these phenomena in precise theoretical terms in an idealized Gaussian setting. Next, we introduce GLM-EIV ("generalized linear model with errors-in-variables"), a new method for single-cell CRISPR screen analysis. GLM-EIV generalizes the classical errors-in-variables model to response distributions and sources of measurement error that are exponential family-distributed, overcoming limitations of thresholded regression. We develop a computational infrastructure to deploy GLM-EIV across hundreds or thousands 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 new biological insights.

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 [1, 2] and transforming basic biology research [3]. Recently, scientists have paired CRISPR genome engineering with single-cell sequencing [4, 5]. 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 [6].

Despite their promise, single-cell CRISPR screens present substantial statistical challenges. A major difficulty is that CRISPR perturbations are unobservable and assigned stochastically to cells. 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 barcode 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 barcode 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 errors-in-variables (or measurement error) models. We assume that the response variable y is a GLM of an underlying predictor variable x^*. We do not observe x^* directly; rather, we observe a noisy version x of x^* that itself is a GLM of x^*. The goal of the analysis is to estimate the effect of x^* on y using the observed data (x, y) only. In the context of the biological application, x^*, y, and x are CRISPR perturbations, gene expressions, and barcode counts, respectively.

Our work makes two main contributions. First, we study the thresholding method from empirical and theoretical perspectives. 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 an idealized Gaussian model. Second, we introduce a new method for estimation and inference in single-cell CRISPR screens that accounts for the measurement error inherent in the experiment. The method, called *GLM-EIV* (generalized linear model with errors-in-variables), implicitly estimates the probability that each cell was perturbed, obviating the need to explicitly impute perturbation assignments via thresholding or another heuristic. Theoretical analyses and simulation studies indicate that GLM-EIV outperforms the thresholding method in large regions of the parameter space.

We implement several statistical accelerations (that possibly are of independent utility) to bring the cost of GLM-EIV down to within an order of magnitude of the thresholding method. Finally, we develop a computational infrastructure to deploy GLM-EIV at-scale across hundreds or thousands 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, high multiplicity-of-infection single-cell CRISPR screen datasets, yielding new biological and statistical insights.

2 Background and analysis challenges

Our focus in this work is on high multiplicity-of-infection (MOI), enhancer-targeting, single-cell CRISPR screens. In this section we cover relevant biological background and motivation.

P1: The human genome consists of genes (segments of DNA that code for proteins), enhancers (segments of DNA that regulate the expression of one or more genes), and other genomic regions. Genome-wide association studies have revealed that the majority ($> 95\%$) of variants associated with diseases lie outside genes and (very likely) inside enhancers. These noncoding variants contribute to disease by modulating the expression one or more genes, which in turn encode proteins that affect the phenotype. A central open challenge in genetics, therefore, is to link enhancers that harbor disease-associated variants to the genes that they target at genome-wide scale.

P2: High MOI single-cell CRISPR screens are the most promising biotechnology for solving this problem. Describe the experimental protocol here. Explain that we use the terms "barcodes" and "gRNAs" interchangeably, as polyadenylated gRNAs serve as barcodes in CROP-seq. Link to Figure 1.

P3: Single-cell CRISPR screens pose several core analysis challenges. Describe the analysis challenges here: (i) unobserved perturbation; (ii) existence of background reads; (iii) highly discrete count data; (iv) nuisance variables.

3 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, Lei, and Roeder [7] developed eSVD, an extension of SVD to exponential family and curved Gaussian responses. Unlike SVD, eSVD models the relationship between the mean and variance of a gene's expression level, a phenomenon induced by the countedness of single-cell data [8]. Similarly, Townes et al. [9] proposed GLM-PCA, an extension of PCA that directly models Poisson- or negative binomially-distributed gene expression counts. We see our work as a continuation of this broad effort to "port"

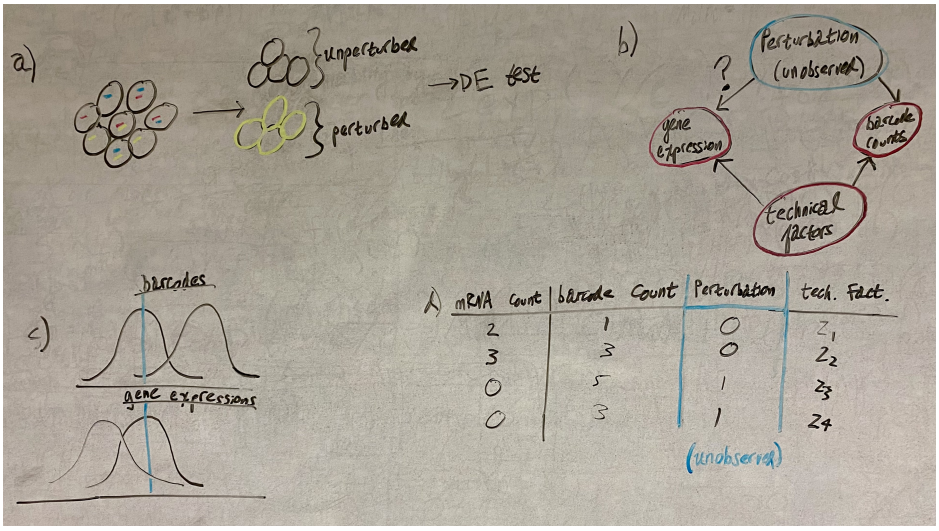

Figure 1: **Overview of experimental design and analysis challenges**: **a,** Experimental design. For a given perturbation (e.g., the perturbation represented in yellow), we partition the cells into two groups: those that received the perturbation, and those that did not receive the perturbation. For a given gene, we conduct a differential expression analysis across the two groups of cells, yielding an estimate of the impact of the given perturbation on the given gene. **b,** DAG representing the variables in the analysis. The perturbation (unobserved) affects both gene expression and gRNA expression; technical factors (e.g., batch, sequencing depth, etc.) act as nuisance variables. The target of inference is the effect of the perturbation on gene expression (denoted with question mark). **c,** Schematic illustrating "background reads." The gRNA modality has a nonzero, "background read" distribution even in the absence of a perturbation, complicating the assignment of perturbations to cells. **d**, Example data for a given perturbation-gene pair. Notice that (i) the perturbations are unobserved, and (ii) the gene and gRNA expression data take the form of discrete counts.

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 to response distributions and sources of measurement error that are exponential family-distributed.

The closest parallels to our work in the statistical methodology literature are Grün & Leisch [10] and Ibrahim [11]. Grün & Leisch considered

estimation and inference in a k-component mixture of GLMs. While we prefer to view the GLM-EIV model as an errors-in-variables model, the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM densities. Ibrahim derived a method 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 is a helpful reference, our estimation and inference tasks are more complex.

The genomics literature has produced several applied methods for linking perturbations to changes in gene expression in single-cell CRISPR screens: SCEPTRE [12], MIMOSCA [4], and scMAGeCK [13]. These methods in general are focused on hypothesis testing rather than estimation; none, for instance, produces a confidence interval for the effect size of a perturbation on gene expression change. Additionally, two of these methods (MIMOSCA and scMAGeCK) use (possibly penalized) linear models to model gene expressions, thereby disregarding the countedness and sparsity of the data.

4 Thresholding method

In this section we study the thresholding method from empirical and theoretical perspectives. First, we define some notation.

Let $n \approx 100,000 - 250,000$ be the number of cells in the experiment. Consider a single perturbation-gene pair. For cell $i \in \{1, \ldots, n\}$, let $p_i \in \{0, 1\}$ indicate whether the cell was perturbed, $m_i \in \mathbb{N}$ be the number of observed gene UMIs, $g_i \in \mathbb{N}$ be the number of observed gRNA UMIs, $l_i^m \in \mathbb{N}$ be the gene library size, and $z_i \in \mathbb{R}^{d-1}$ be a vector of technical factors (e.g., batch, percent mitochondrial reads, etc.). The thresholding method is defined as follows:

1. For given threshold $c \in \mathbb{N}$, calculate the imputed value $\hat{p}_i$ of p_i by

$$\begin{cases} \hat{p}_i = 0 \text{ if } g_i \geq c, \\ \hat{p}_i = 1 \text{ if } g_i < c. \end{cases}$$

2. Fit the regression model [14]

$$m_i | \, (\hat{p}_i, z_i, l_i^m) \sim \mathrm{NB}_\theta(\mu_i),$$

where $\theta > 0$ is the NB size parameter, and

$$\log(\mu_i) = \beta_m^0 + \beta_m \hat{p}_i + \gamma_m^T z_i + \log(l_i^m).$$

3. Fit a GLM to obtain an estimate $\hat{\beta}_m$ of β_m. Compute a confidence interval and p-value for β_m.

4.1 Empirical analysis

To investigate the impact of threshold selection on the thresholding method, we applied the thresholding method to the set of positive control (i.e., gene-targeting) perturbation-gene pairs in the Gasperini data using three different choices for the threshold: 1, 5, and 20.

Description of panels a-d; threshold = 1 leads to considerable attenuation bias (a); threshold = 5 and threshold = 20 produce similar estimates, though the effect sizes are slightly greater for threshold = 20, suggesting threshold = 5 yields mild attenuation bias (b); the threshold = 20 estimates are more variable than than those of threshold = 5, as evidenced by the smaller p-values and wider confidence intervals.

Description of panels e-f: these are the empirical gRNA count distributions for randomly selected gRNAs from the Gasperini and Xie datasets. There does not appear to be a clear location at which to draw the threshold; aside from the initial spike at one (zero not shown), the histograms gradually decrease.

Take-home message: (i) the threshold is a tuning parameter that substantially affects the result; (ii) as the threshold increases, bias seems to decrease and variance seems to increase; (iii) it is not clear where to draw the threshold from gRNA counts alone.

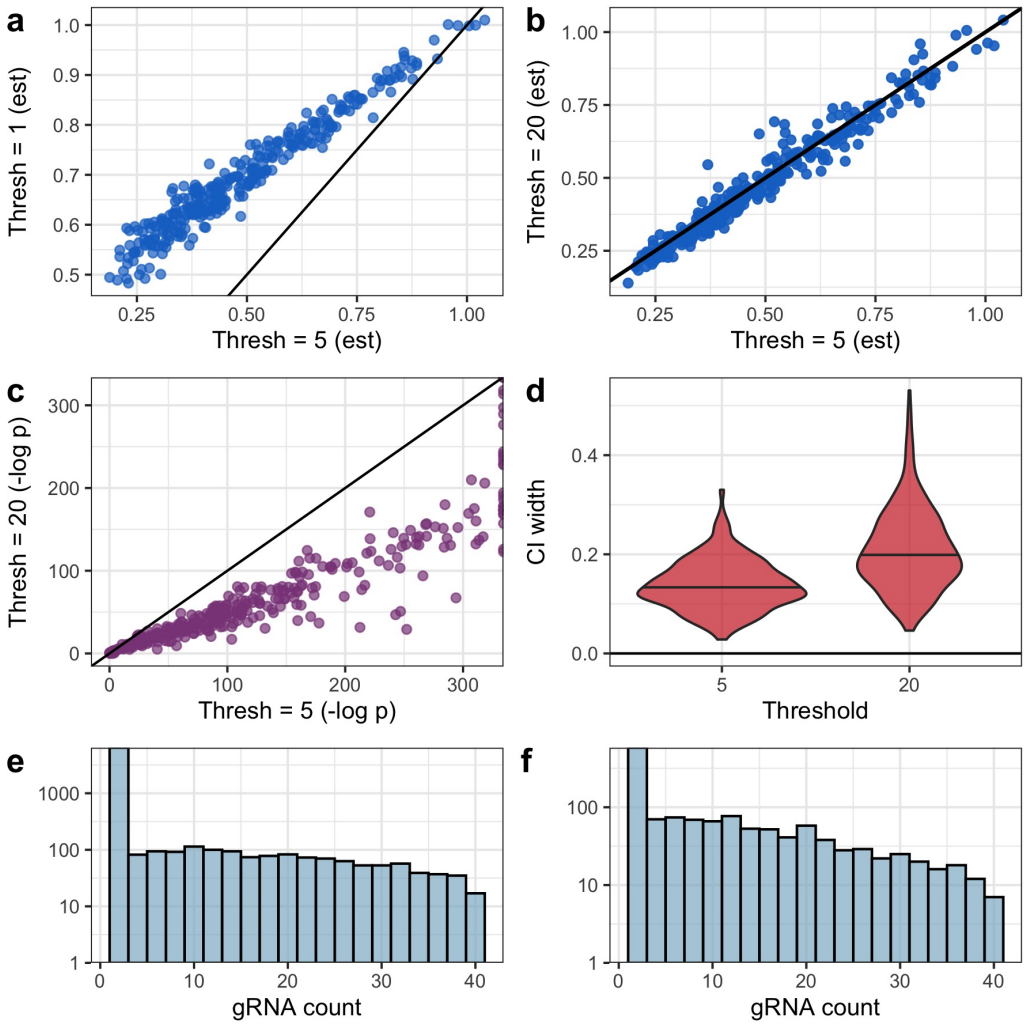

Figure 2: **Empirical challenges of thresholded regression.**

4.2 Theoretical analysis

Bias: We use the following theoretical model to study bias of the thresholding method:

$$\begin{cases} m_i = \beta_0^m + \beta_1^m p_i + \epsilon_i \\ g_i = \beta_0^g + \beta_1^g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i \end{cases}.$$

We can solve for the limit in probability of $\hat{\beta}_m$. We have that

$$\hat{\beta}_1^m \xrightarrow{P} \beta_1^m \left(\frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \right),$$

where

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

Let $\gamma : \mathbb{R}^4 \to \mathbb{R}$ be defined by

$$\gamma(\beta_0^g, \beta_1^g, c, \pi) = \frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])}.$$

We call γ the attenuation function. Observe that (i) γ does not depend on β_1^m or β_0^m, and (ii) $\hat{\beta}_1^m \xrightarrow{P} [\gamma(\beta_0^g, \beta_1^g, c, \pi)]\beta_1^m$. Let $b : \mathbb{R}^4 \to \mathbb{R}$ be the asymptotic bias of $\hat{\beta}_1^m$. We can express b in terms of γ as

$$b\left(\beta_0^g, \beta_1^g, c, \pi\right) = \beta_1^m - \gamma(\beta_0^g, \beta_1^g, c, \pi)\beta_1^m = \beta_1^m \left[1 - \gamma(\beta_0^g, \beta_1^g, c, \pi)\right].$$

The asymptotic bias vanishes when the attenuation function equals 1.

Bias as a function of threshold (Panel a)

To investigate the basic question of "What is a good threshold selection strategy?", we studied the relationship between the asymptotic bias b of $\hat{\beta}_1^m$ and the selected threshold c. For simplicity, we set the perturbation probability π to 1/2. Additionally, without loss of generality, we set the target of inference β_1^m to 1. Let $c_{\text{bayes}} \in \mathbb{R}$ be the Bayes-optimal decision boundary for classifying cells as perturbed or unperturbed, i.e.

$$c_{\text{bayes}} = \arg\min_{c \in \mathbb{R}} \mathbb{P}(\hat{p}_i \neq p_i).$$

Simple algebra shows that $c_{\text{bayes}} = \beta_0^g + (1/2)\beta_1^g$. Below, we state several results relating the limiting bias b of $\hat{\beta}_1^m$ to c_{bayes} and β_1^g, deferring proofs to the appendix.

1. The Bayes-optimal threshold c_{bayes} is a critical value of the bias function b, i.e.

$$\frac{\partial b(\beta_0^g, \beta_1^g, c, 1/2)}{\partial c}\Big|_{c=c_{\text{bayes}}} = 0.$$

2. For certain values of β_0^g and β_1^g, c_{bayes} is a maximizer of the bias (Figure 3a, left); for other values of β_0^g and β_1^g, c_{bayes} is a minimizer of the bias (Figure 3a, right).

3. The limit of the bias in c is 1/2, i.e.

$$\lim_{c \to \infty} b(\beta_0^g, \beta_1^g, c, 1/2) = 1/2.$$

   In other words, we always can set the threshold to a large number and achieve a bias of 1/2 (Figure 3a, all panels). This fact establishes an upper bound on the bias of thresholded regression (under optimal threshold selection).

4. For $\beta_1^g < 2\Phi(3/4) \approx 1.35$, $c = c_{\text{bayes}}$ yields a *larger* bias than $c = \infty$ (Figure 3a, left); for $\beta_1^g > 2\Phi(3/4)$, c_{bayes} yields a *smaller* bias than $c = \infty$ (Figure 3a, right); finally, for $\beta_1^g = 2\Phi(3/4)$, $c = c_{\text{bayes}}$ and $c = \infty$ both yield a bias of 1/2 (Figure 3a, middle).

5. The bias decreases monotonically as β_1^g increases (Figure 3a; fix a threshold and scan the panels from left to right). This is consistent with the intuition that the problem becomes easier as the gRNA mixture distribution becomes increasingly well-separated.

6. For all β_0^g, β_1^g, and $c \in \mathbb{R}$, the bias is strictly greater than 0 and less than 2. Therefore, the thresholding method suffers from attenuation bias, a common phenomenon in errors-in-variables models.

These results are subtle, but we can distill them as follows. First, selecting a threshold that minimizes the bias is deceptively challenging, as there is no rule of thumb that we can apply universally for this purpose (e.g., "always choose the Bayes-optimal decision boundary" or "always choose a large number"). Furthermore, the optimal threshold depends on unknown parameters. Finally, even *if* we have selected a good threshold, the thresholding method incurs a strict attenuation bias.

Bias of large threshold strategy vs. π (Panel b)

Intrigued by the finding that selecting a large threshold when $\pi = 1/2$ yields a bias of 1/2, we investigated the relationship between bias and other values of π under the large threshold selection strategy.

Bias-variance tradeoff (Panel c)

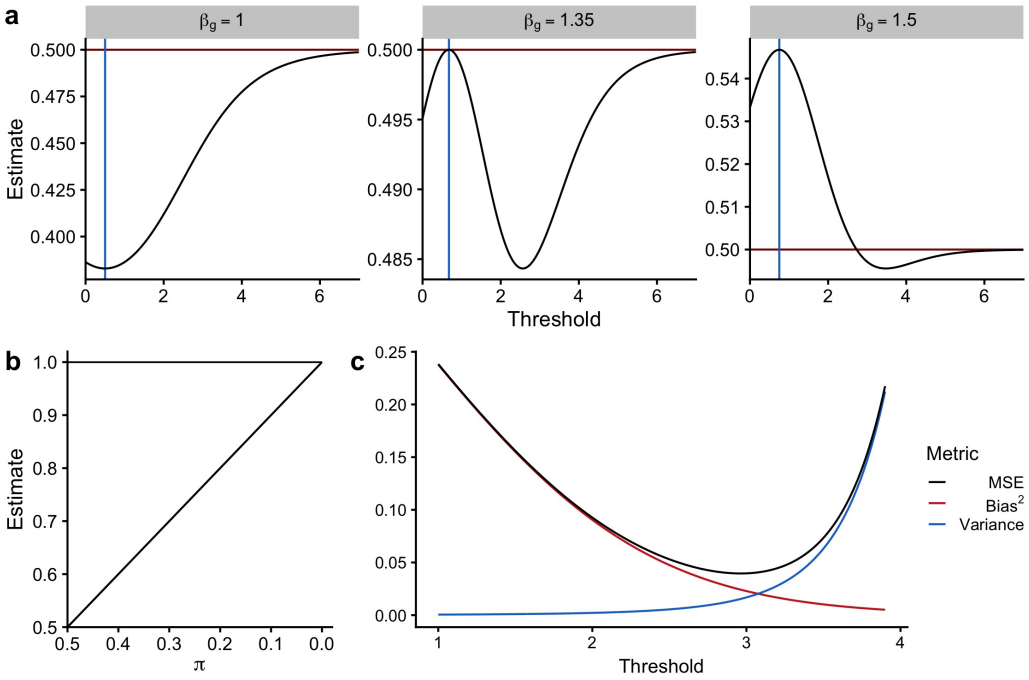

Figure 3: **Theoretical challenges of thresholded regression.**

5 GLM-EIV

6 Simulation studies

7 Real data analysis

8 Discussion

9 Appendix

9.1 Theoretical results for thresholding estimator

9.2 Derivation of EM algorithm

9.3 Derivation of observed information matrix

9.4 Implementation using R family objects

9.5 Statistical accelerations to GLM-EIV

9.6 Additional simulation results

References

[1] Tanja Rothgangl, Melissa K. Dennis, Paulo J.C. Lin, Rurika Oka, Dominik Witzigmann, Lukas Villiger, Weihong Qi, Martina Hruzova, Lucas Kissling, Daniela Lenggenhager, Costanza Borrelli, Sabina Egli, Nina Frey, Noëlle Bakker, John A. Walker, Anastasia P. Kadina, Denis V. Victorov, Martin Pacesa, Susanne Kreutzer, Zacharias Kontarakis, Andreas Moor, Martin Jinek, Drew Weissman, Markus Stoffel, Ruben van Boxtel, Kevin Holden, Norbert Pardi, Beat Thöny, Johannes Häberle, Ying K. Tam, Sean C. Semple, and Gerald Schwank. In vivo adenine base editing of PCSK9 in macaques reduces LDL cholesterol levels. *Nature Biotechnology*, 39(8):949–957, 2021.

[2] Kiran Musunuru, Alexandra C. Chadwick, Taiji Mizoguchi, Sara P. Garcia, Jamie E. DeNizio, Caroline W. Reiss, Kui Wang, Sowmya Iyer, Chaitali Dutta, Victoria Clendaniel, Michael Amaonye, Aaron Beach, Kathleen Berth, Souvik Biswas, Maurine C. Braun, Huei Mei Chen, Thomas V. Colace, John D. Ganey, Soumyashree A. Gangopadhyay, Ryan Garrity, Lisa N. Kasiewicz, Jennifer Lavoie, James A. Madsen, Yuri Matsumoto, Anne Marie Mazzola, Yusuf S. Nasrullah, Joseph Nneji, Huilan Ren, Athul Sanjeev, Madeleine Shay, Mary R. Stahley, Steven H.Y. Fan, Ying K. Tam, Nicole M. Gaudelli, Giuseppe Ciaramella, Leslie E. Stolz, Padma Malyala, Christopher J. Cheng, Kallanthottathil G. Rajeev, Ellen Rohde, Andrew M. Bellinger, and Sekar Kathiresan. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. *Nature*, 593(7859):429–434, 2021.

[3] Laralynne Przybyla and Luke A. Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 0123456789, 2021.

[4] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, 2016.

[5] Paul Datlinger, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie

Kuchler, Donat Alpar, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017.

[6] John A. Morris, Zharko Daniloski, Júlia Domingo, Timothy Barry, Marcello Ziosi, Dafni A. Glinos, Stephanie Hao, Eleni P. Mimitou, Peter Smibert, Kathryn Roeder, Eugene Katsevich, Tuuli Lappalainen, and Neville E. Sanjana. Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing. *bioRxiv*, page 2021.04.07.438882, 2021.

[7] Kevin Z. Lin, Jing Lei, and Kathryn Roeder. Exponential-Family Embedding With Application to Cell Developmental Trajectories for Single-Cell RNA-Seq Data. *Journal of the American Statistical Association*, 0(0):1–32, 2021.

[8] Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, 22(1):1–20, 2021.

[9] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):1–16, 2019.

[10] Bettina Grün and Friedrich Leisch. *Finite Mixtures of Generalized Linear Regression Models*, pages 205–230. Physica-Verlag HD, Heidelberg, 2008.

[11] Joseph G. Ibrahim. Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

[12] Timothy Barry, Xuran Wang, John A. Morris, Kathryn Roeder, and Eugene Katsevich. Conditional resampling improves calibration and sensitivity in single-cell CRISPR screen analysis. *bioRxiv*, page 2020.08.13.250092, 2020.

[13] Lin Yang, Yuqing Zhu, Hua Yu, Sitong Chen, Yulan Chu, He Huang, Jin Zhang, and Wei Li. Linking genotypes with multiple phenotypes in single-cell CRISPR screens. *bioRxiv*, page 658146, 2019.

[14] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021.