

Tim, Gene, Kathryn

**A generalized errors-in-variables model, with application to
CRISPR genome editing and single-cell sequencing**

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 (i) unobserved and (ii) assigned randomly to cells. The analyst cannot know with certainty which cells were perturbed and instead 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, the analyst can attempt to estimate the effect of the perturbation on gene expression. The standard approach, which we call the “thresholding method,” is to assign perturbation identities to cells by thresholding the barcode counts.

We study estimation and inference in single-cell CRISPR screens from a statistical perspective, formulating the data generating mechanism as 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.

Within this framework we make two contributions. First, we study the thresholding method from theoretical and empirical perspectives. We characterize the properties of the estimator and uncover settings in which

GLM-EIV (GLM-based errors in variables). GLM-EIV implicitly estimates the probability that each cell was perturbed, obviating the need to

explicitly impute perturbation identities via

2 Background and analysis challenges

-

3 Related work

4 Thresholding method

-
-

4.1	Theoretical analysis
4.2	Empirical analysis
5	GLM-EIV
6	Simulation studies
7	Real data analysis
8	Discussion
9	Appendix
9.1	Proofs of 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 Bortel, 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.