Tim, Gene, Kathryn

A new class of measurement error models, 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 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 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 a bias-variance tradeoff as a function of the selected threshold, and we recover this phenomenon in precise mathematical terms in an 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 data generating process. 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 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.
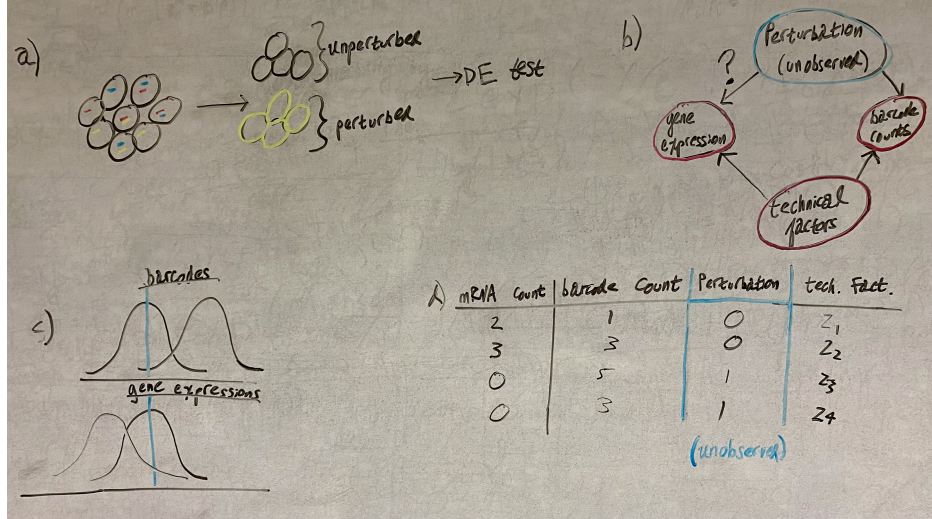

Figure 1: **Overview of experimental design and analysis challenges**: **a,** Experimental design. For a given perturbation (e.g., the perturbation represented in yellow), we partition the cells into two groups: those that received the perturbation, and those that did not receive the perturbation. For 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.

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 broad class of exponential family distributions. For example, Lin, Lei, and Roeder [7] developed eSVD, an extension of SVD to exponential family and curved Gaussian response distributions. Unlike SVD, eSVD captures the relationship between the mean and variance of a gene's expression level, a phenomenon induced by the countedness and sparsity of single-cell data [8]. Along similar lines, Townes et al. [9] proposed GLM-PCA, an extension of PCA that directly models gene expression counts while regressing out technical factors. We see our work as a continuation of this broad 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 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 consider estimation and inference in a k-component mixture of GLMs. While we prefer to view the GLM-EIV model as a kind of errors-in-variables model, the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM densities. Ibrahim derives 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 helpful reference, our estimation and inference tasks are more complex than his.

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 change in gene expression. Additionally, two of these methods (MIMOSCA and scMAGeCK) use linear models to model gene expressions, disregarding the countedness and sparsity of the data.

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 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.