01/05/2022
Pittsburgh PA, United States

Dear Dr. Zhang,

We are writing to submit our manuscript “Exponential family measurement error models for single-cell CRISPR screens” for consideration in JASA (Applications and Case Studies section). CRISPR is a genome engineering tool that has enabled scientists to precisely edit human and nonhuman genomes, transforming biological discovery. Recently, scientists have paired CRISPR genome engineering with single-cell RNA sequencing. The resulting assays, called “single-cell CRISPR screens,” enable scientists to causally map GWAS variants to their target genes at genome-wide scale, potentially solving a long-standing problem in genetics.

Despite their promise, single-cell CRISPR screens pose significant statistical challenges. We demonstrate on real data that a commonly-used method exhibits attenuation bias and a bias-variance tradeoff as a function of a challenging-to-select tuning parameter. We recover these phenomena in precise mathematical terms in an idealized Gaussian setting. Next, to overcome these limitations, we introduce “GLM-EIV” (“GLM-based errors-in-variables”), a new measurement error model that extends the classical measurement error model in several key directions. Most importantly, GLM-EIV (i) accommodates responses and noisy predictors that are arbitrarily exponential family-distributed and (ii) accounts for a common source of measurement error between the response and predictor. These properties, which are not shared by classical measurement error models, enable GLM-EIV to resolve novel analysis challenges posed by single-cell CRISPR screens. 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, demonstrating improved performance in challenging problem settings.

In conclusion, we couple a novel statistical methodology to several of the most powerful technologies from biology and computer science: CRISPR genome engineering, single-cell RNA sequencing, and cloud and high-performance computing. This union yields new analytic insights and adds a new tool to the working biologist’s toolkit. We also anticipate that the proposed methodology is sufficiently general to apply to areas beyond genomics, such as psychology (see final paragraph of manuscript).

We confirm that all authors approve the manuscript and that the manuscript is not under consideration at any other journal.

Sincerely,
Timothy Barry (Carnegie Mellon University)
Eugene Katsevich (Wharton School, University of Pennsylvania)
Kathryn Roeder (Carnegie Mellon University)