

Exponential family measurement error models for single-cell CRISPR screens

Tim Barry
Carnegie Mellon University



Joint work with my advisors



Kathryn Roeder (Dept. of Statistics,
Carnegie Mellon University)



Gene Katsevich (Dept. of Statistics,
University of Pennsylvania)

CRISPR is a genome engineering technology that can be used to modify living organisms in incredible ways.

- Fix genes that cause diseases in humans.
- Make crops more tolerant to hot and arid weather.
- Transform elephants into woolly mammoths!?

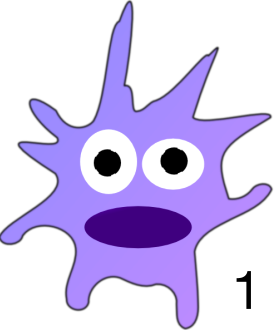
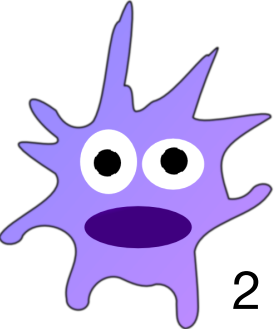
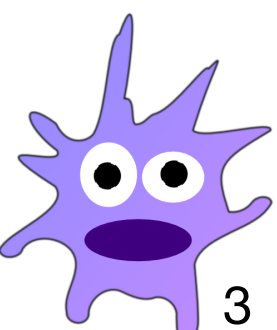
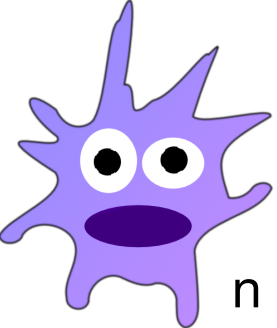


CRISPR
→



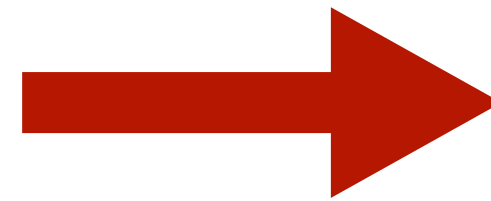
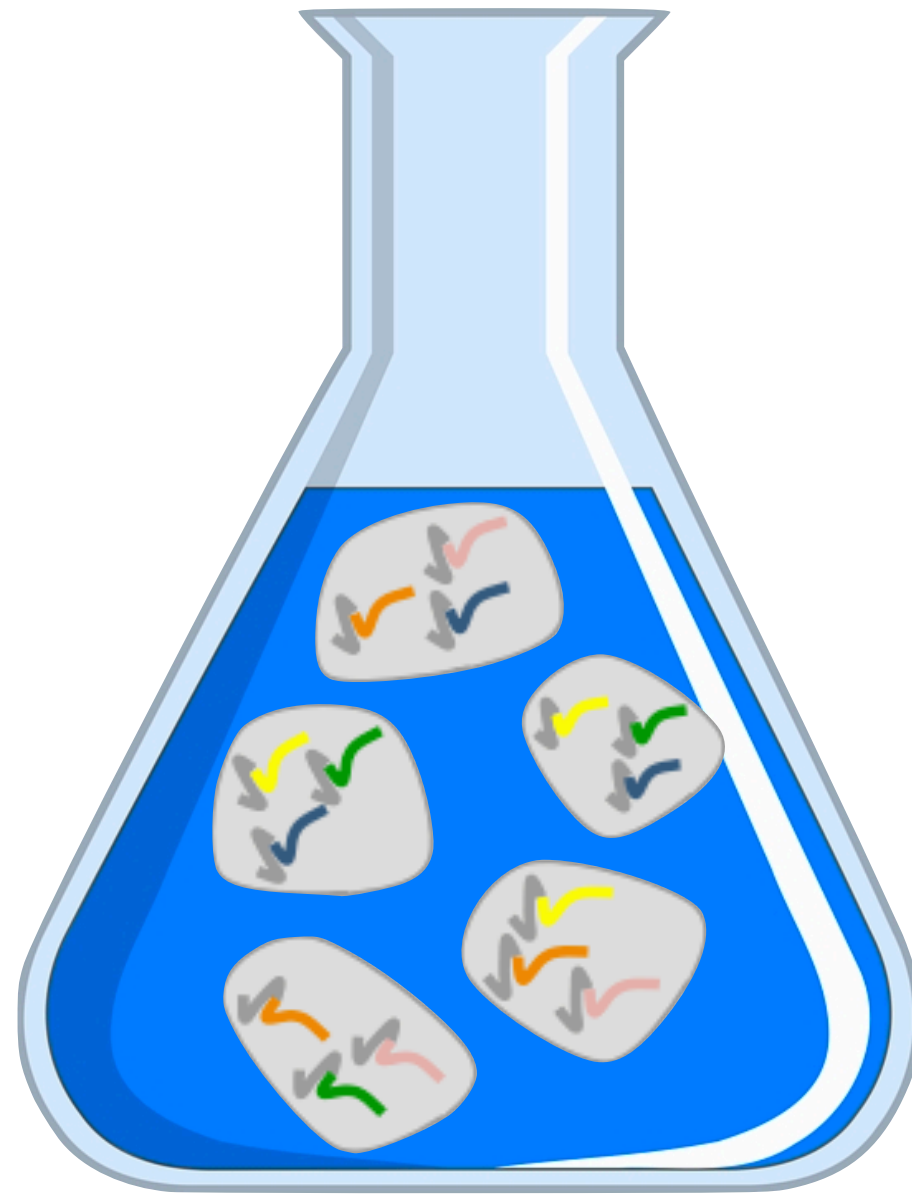
Single-cell RNA sequencing is a technology for measuring gene expressions in individual cells.

i.i.d. cells

	Gene 1	Gene 2	Gene 3		Gene p
 ₁	1	0	4	...	2
 ₂	0	1	0	...	0
 ₃	3	0	0	...	2
⋮	⋮	⋮	⋮		⋮
 _n	0	2	4	...	1

Single-cell CRISPR screens couple CRISPR to single-cell sequencing, enabling scientists to interrogate the effects of perturbations in individual cells.

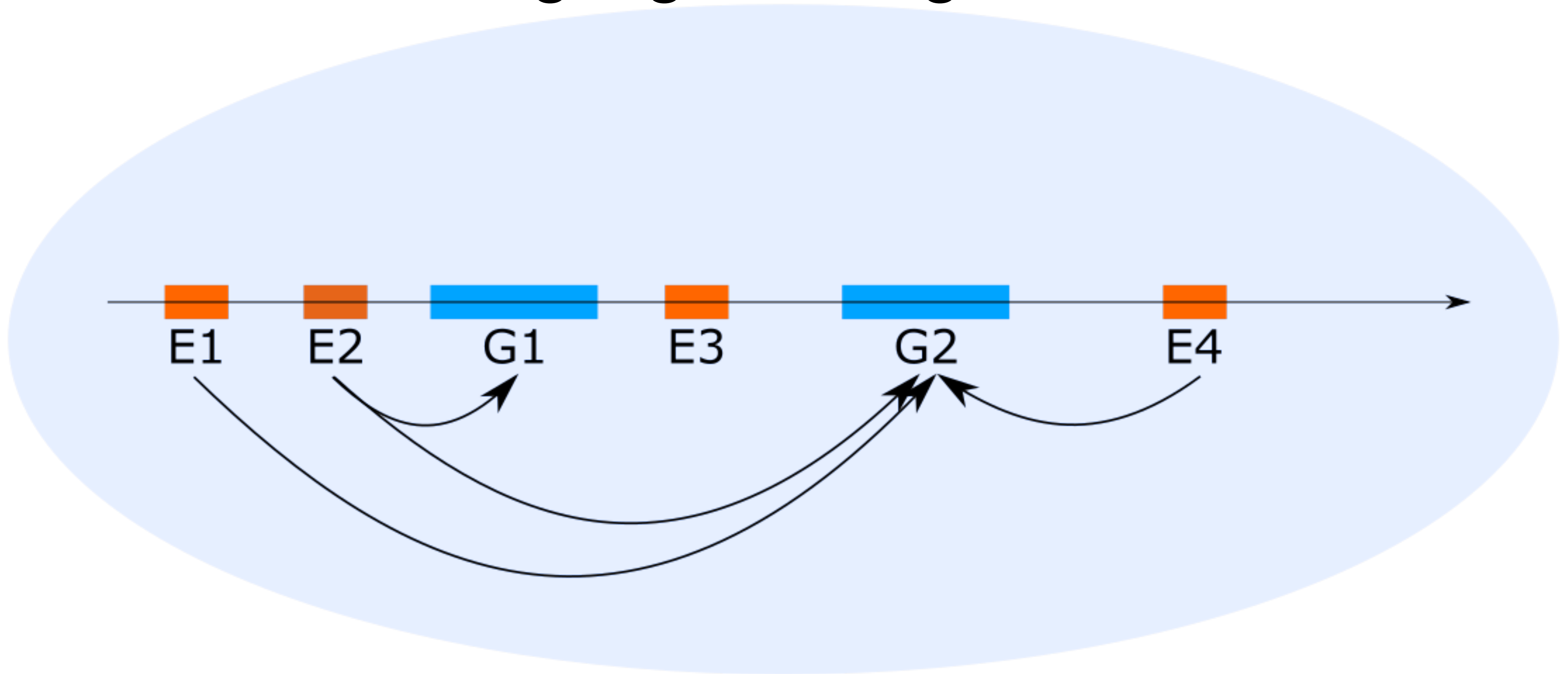
Insert CRISPR
perturbations into cells



Single-cell sequencing

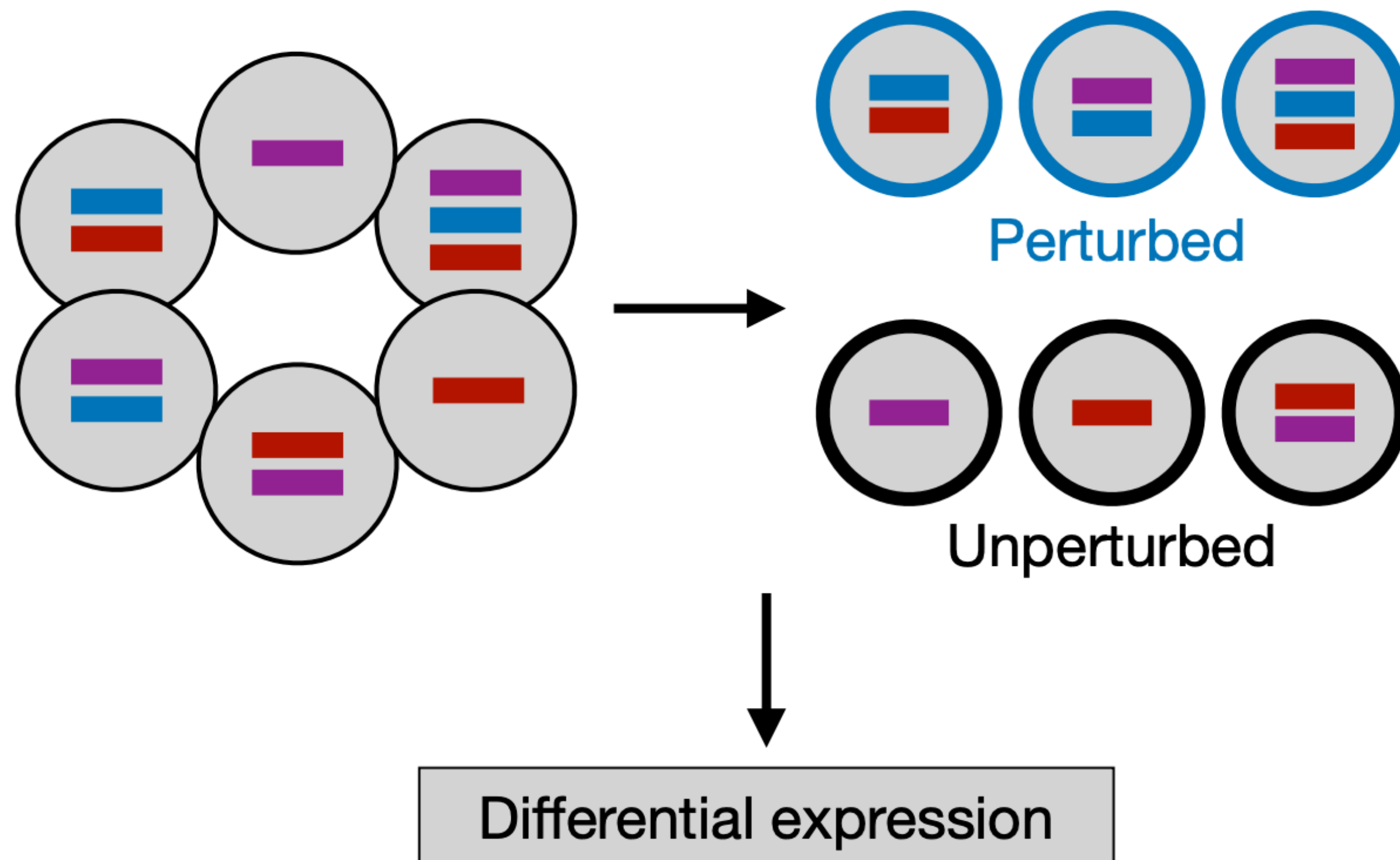


A major open challenge in genetics is mapping enhancers to target genes at genome-wide scale.



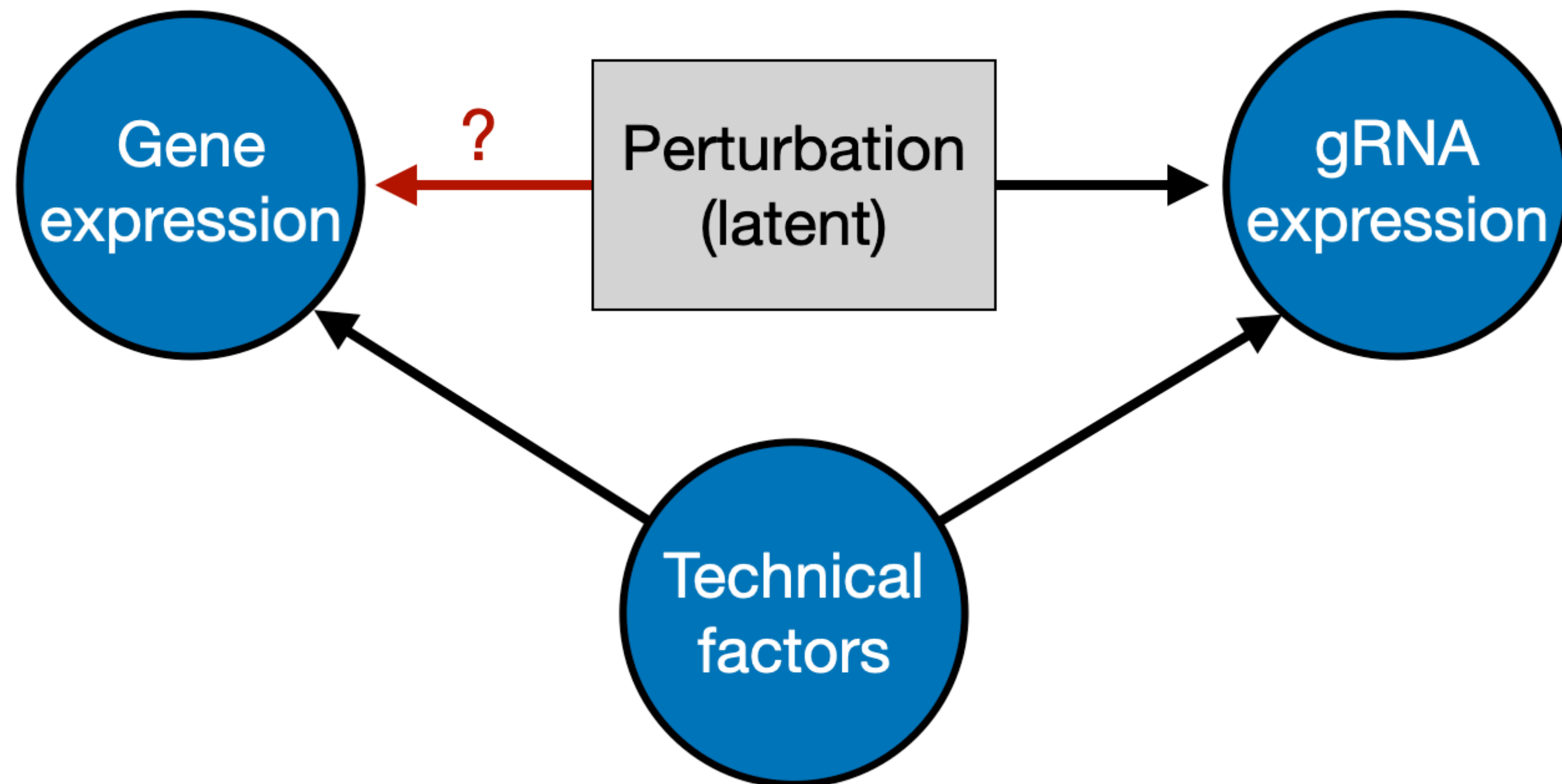
Single-cell CRISPR screens can map enhancers to their target genes.

Experimental design



1. For a given perturbation (**blue**), partition the cells into two groups: perturbed and unperturbed.
2. For a given gene, perform a differential expression analysis across these two groups of cells.

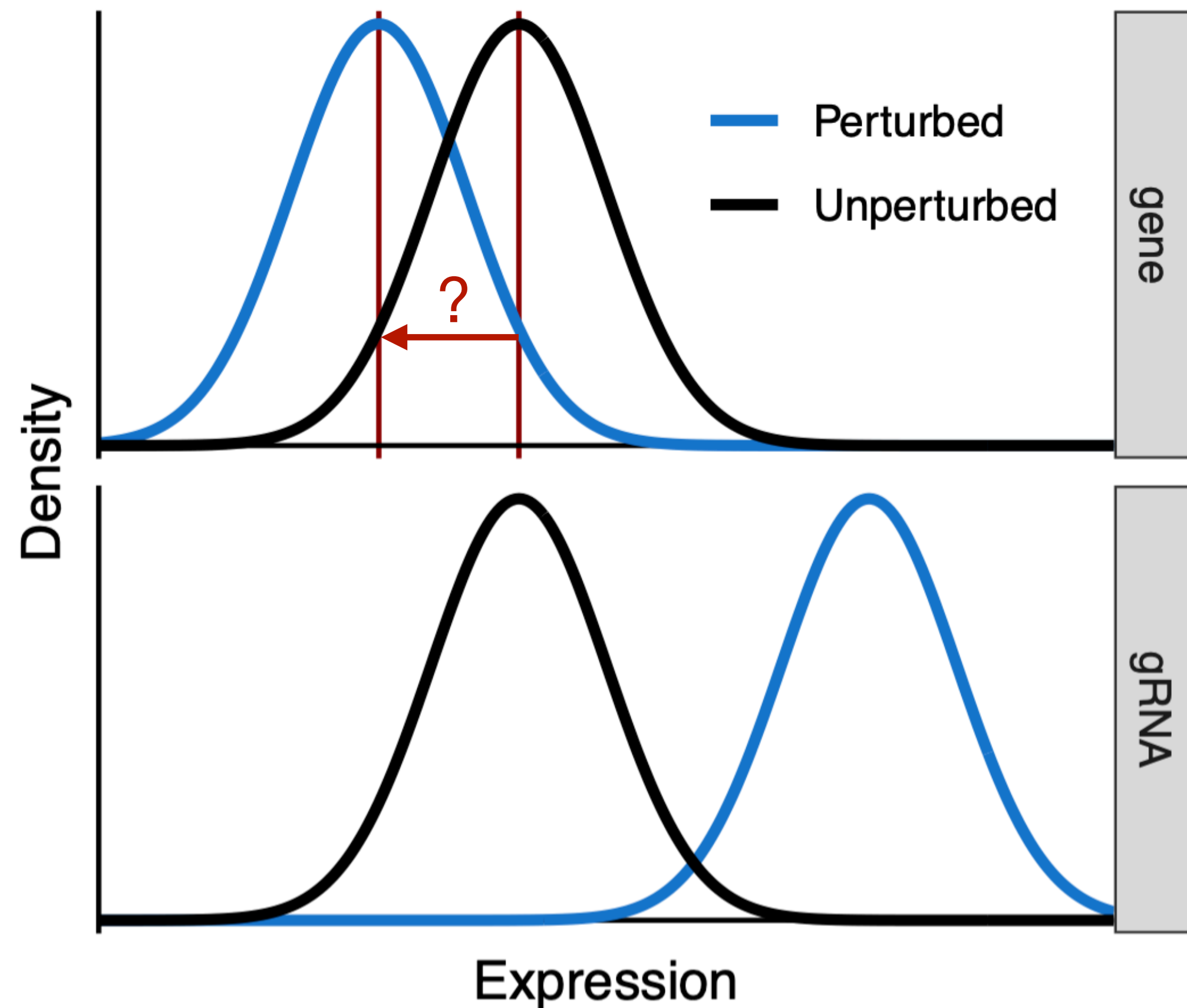
Analysis challenges



1. The "treatment variable" (perturbation presence or absence) is not directly observed; proxy molecules ("gRNAs") are observed instead.

2. "Technical factors" — experimental rather than biological sources of variation — impact the measurement of both gene and gRNA expressions and therefore act as confounders.

Analysis challenges



3. Sequenced gRNAs sometimes mapped to cells that have not received a perturbation.

Analysis challenges

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

4. The gene and gRNA data are sparse, discrete counts, rendering classical methods based on Gaussianity inapplicable.

We propose the "GLM-EIV" (GLM-based errors-in-variables) model to model the single-cell CRISPR screen data generating process.

$$\begin{aligned} \text{gene transcript count} &\sim \text{GLM}(\text{perturbation}, \text{confounders}) \\ \text{gRNA transcript count} &\sim \text{GLM}(\text{perturbation}, \text{confounders}) \\ \text{perturbation} &\sim \text{Bernoulli}(\pi) \end{aligned}$$

- The target of inference is the regression parameter linking *perturbation* to *gene transcript count* in the first GLM.
- The GLM-EIV model is an extended measurement error model.

GLM-EIV is fast and scalable.

1. EM algorithm for estimation
 - Algorithm to produce starting estimates to speed convergence
2. Analytical derivation of observed information matrix
3. Computational pipeline that scales to clusters and clouds



nextflow

The "thresholding method" is a competing method that ignores measurement error.

- For cell i and threshold $c \in \mathbb{N}$, let the imputed perturbation be

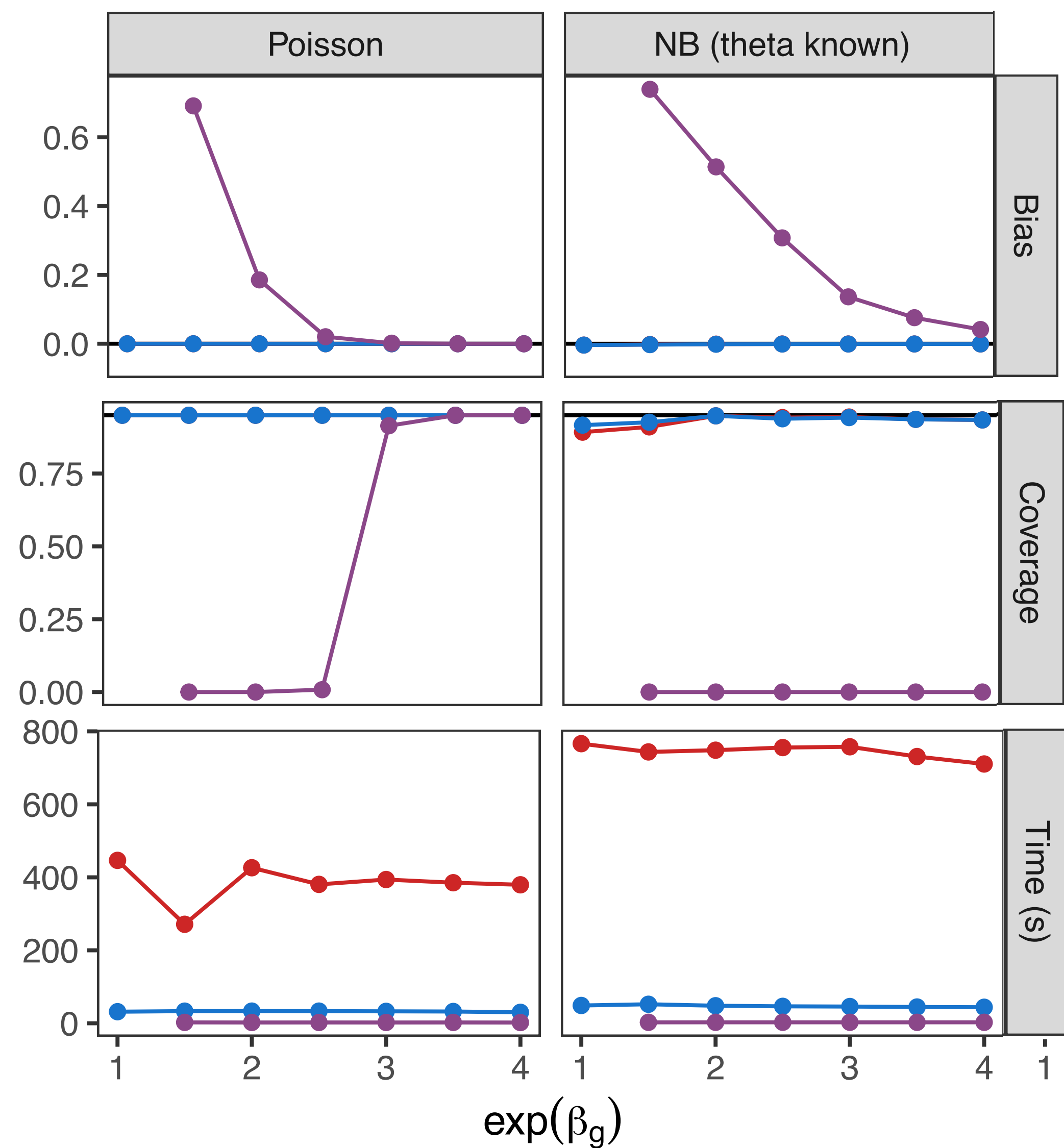
$$\widehat{perturbation}_i = I(gRNA \text{ transcript count}_i \geq c)$$

- Fit the GLM

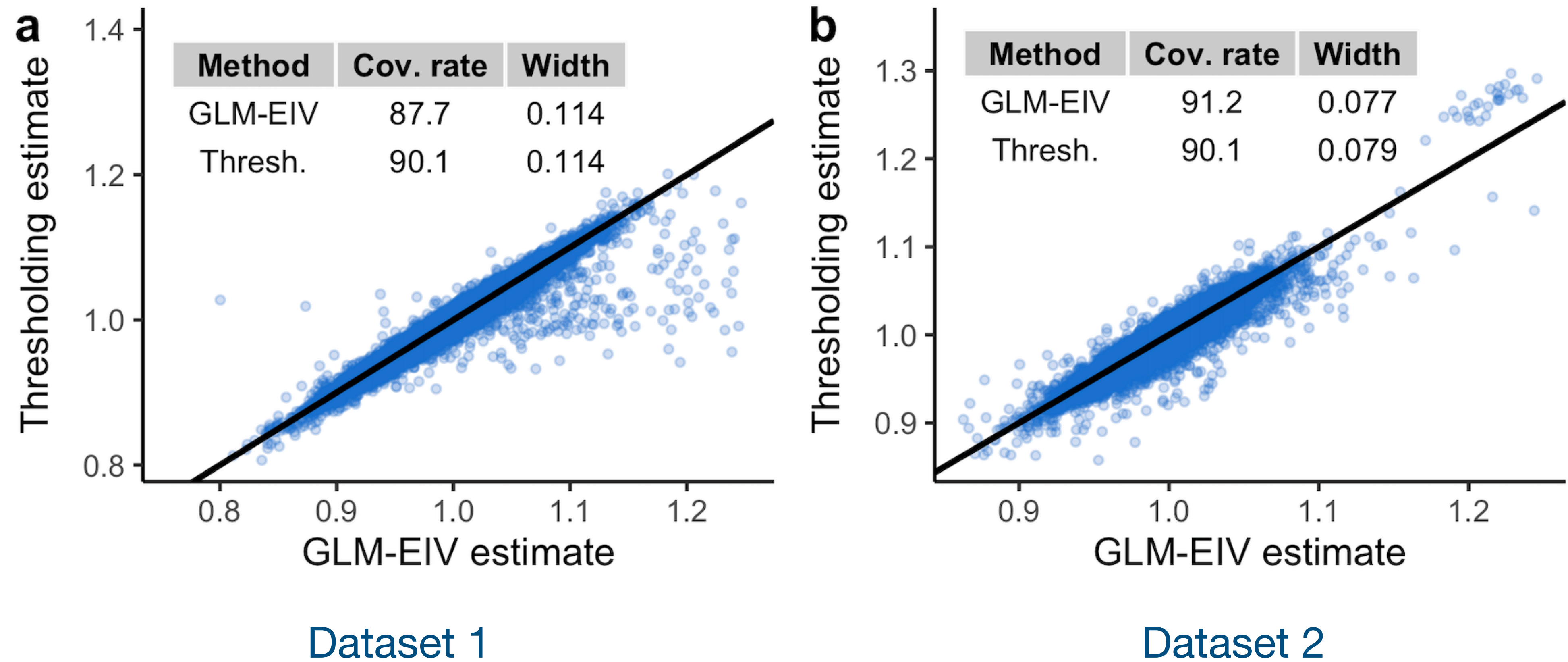
$$gene \text{ transcript count} \sim \text{GLM}(\widehat{perturbation}, confounders)$$

- We show that the thresholding method incurs strict attenuation bias (although in "easy" regions of the parameter space, the bias could be negligible).

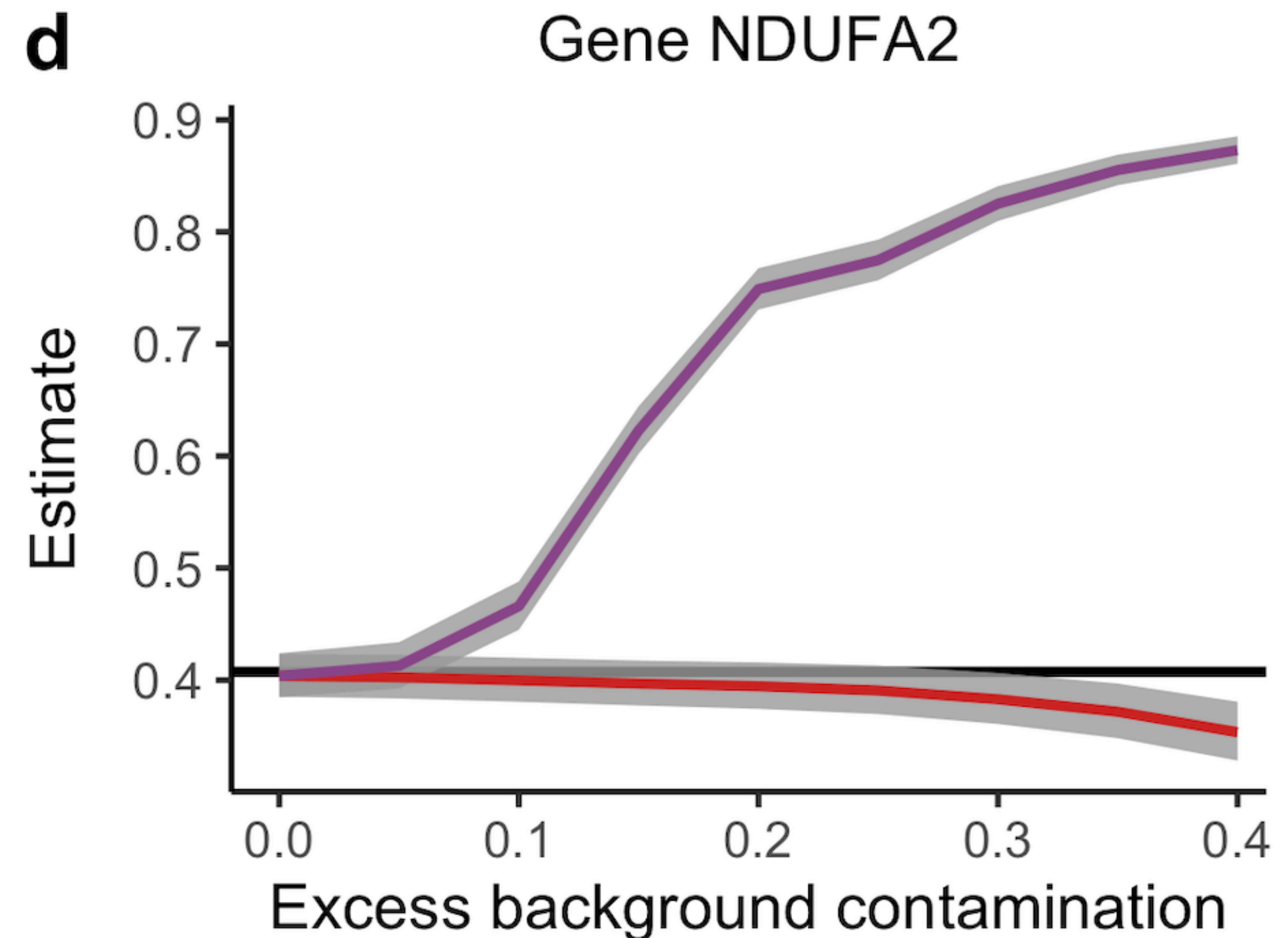
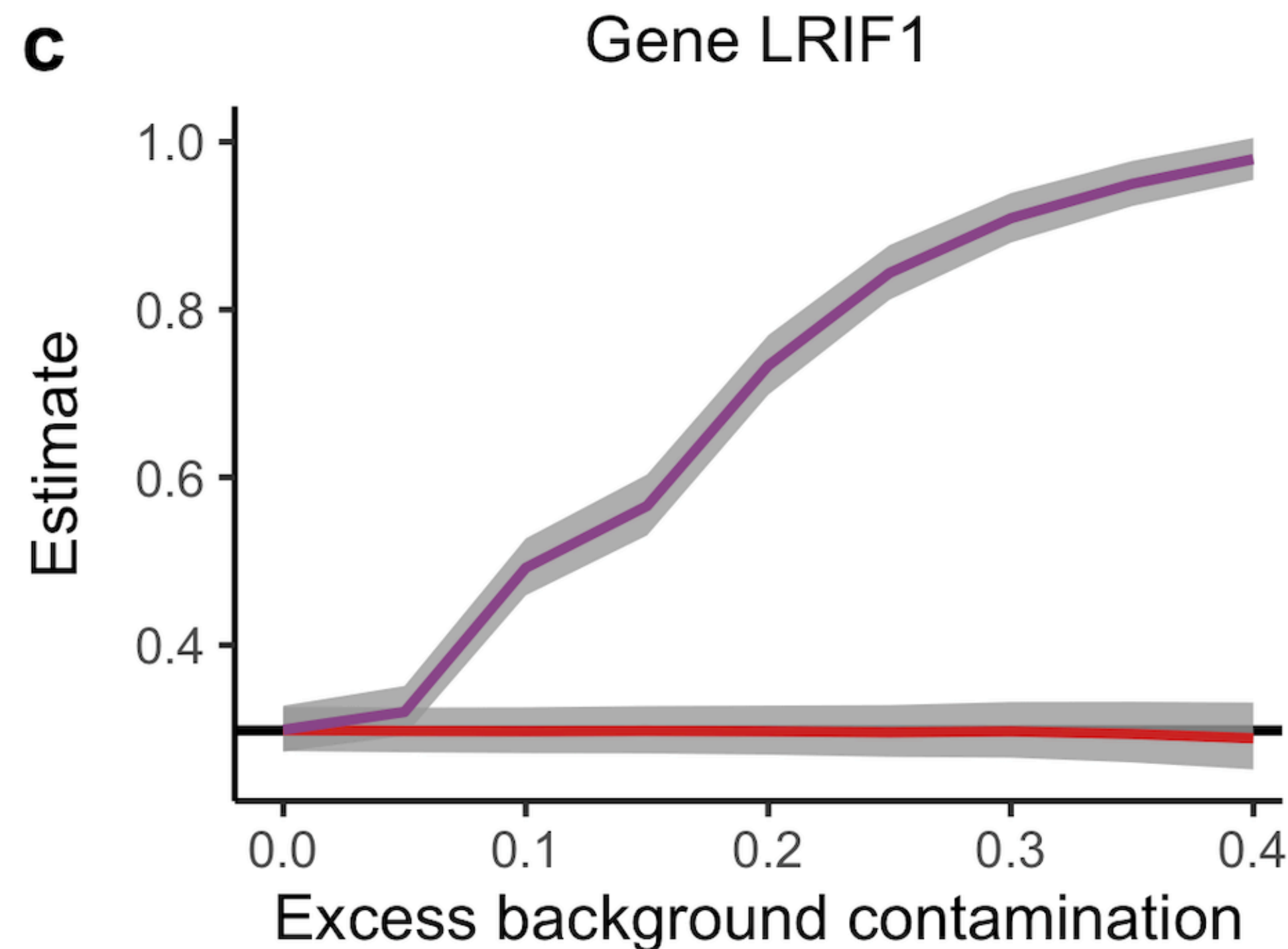
GLM-EIV is more accurate than the thresholding method on simulated data.



GLM-EIV and the thresholding method performed more similarly on real data.



When we increased the amount of background contamination in the data, GLM-EIV remained stable, while thresholding exhibited attenuation bias.



Thank you. Questions?

Manuscript code: <https://github.com/timothy-barry/glmeiv-manuscript>

ondisc package: github.com/timothy-barry/ondisc

My website: <https://timothy-barry.github.io/>