Let it SNO: Massive-scale perturb-seq analysis with SCEPTRE, Nextflow, and ondisc (SNO)

Timothy Barry, Joseph Deutsch, Xihong Lin, & Eugene Katsevich

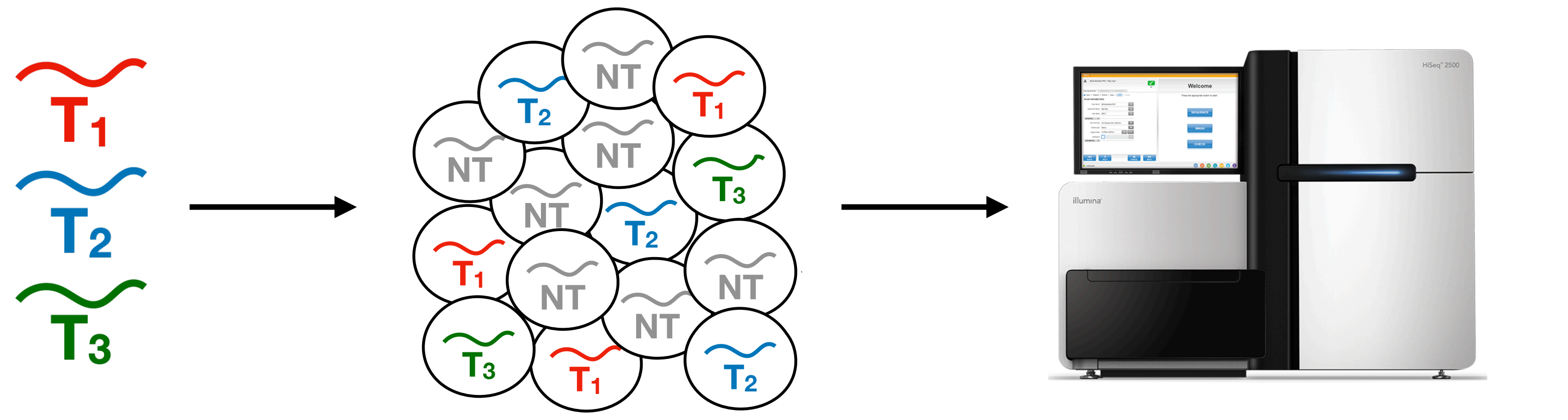
Harvard University and University of Pennsylvania

website: katsevich-lab.github.io/sceptre/

e-book: timothy-barry.github.io/sceptre-book/

Perturb-seq

Simultaneous profiling of CRISPR perturbations and whole transcriptome in single cells, with applications to drug discovery.



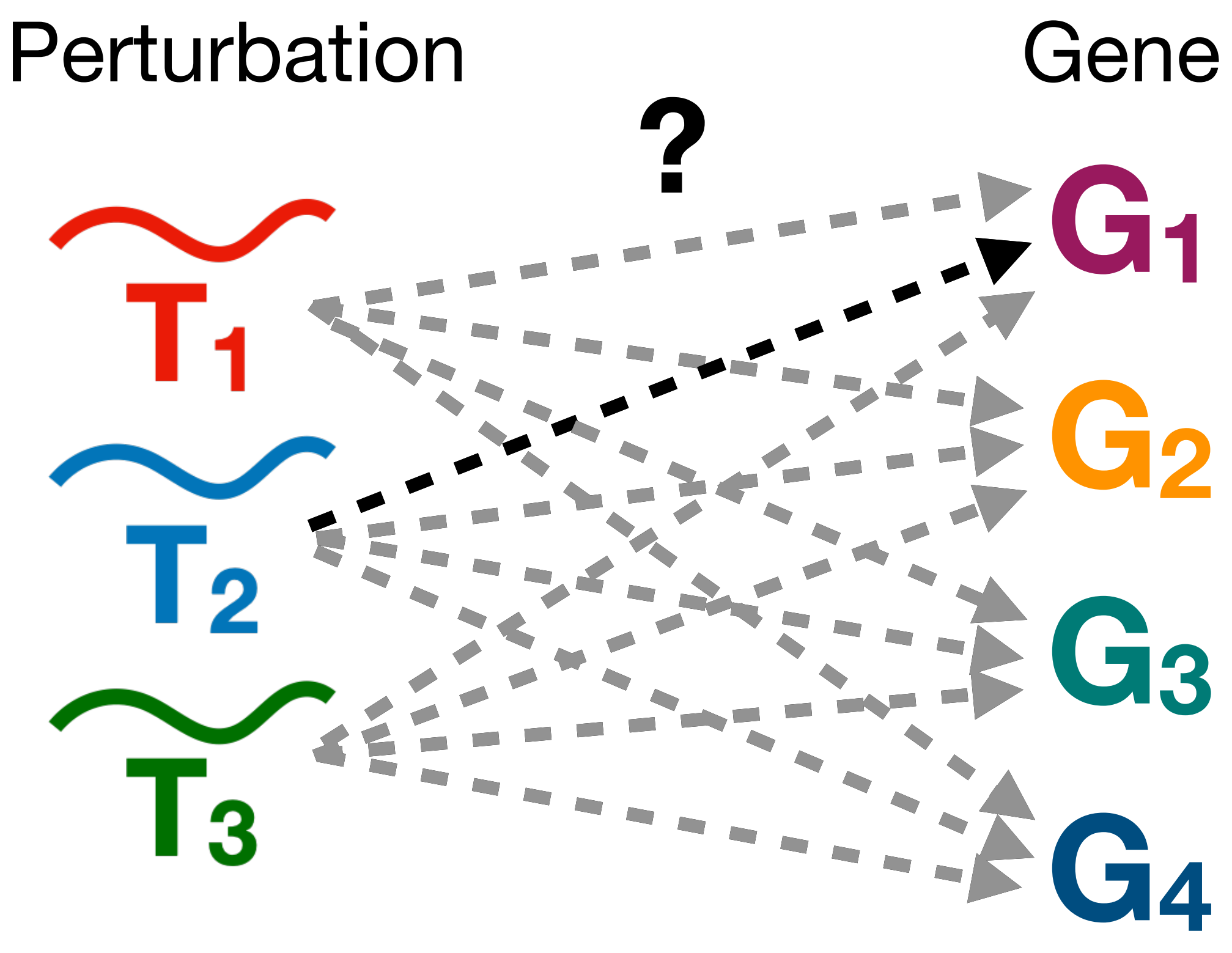
1. Design a library of CRISPR perturbations.

2. Deliver the CRISPR perturbations to cells.

3. Sequence the cells to determine the perturbation that each cell received and measure its gene expressions.

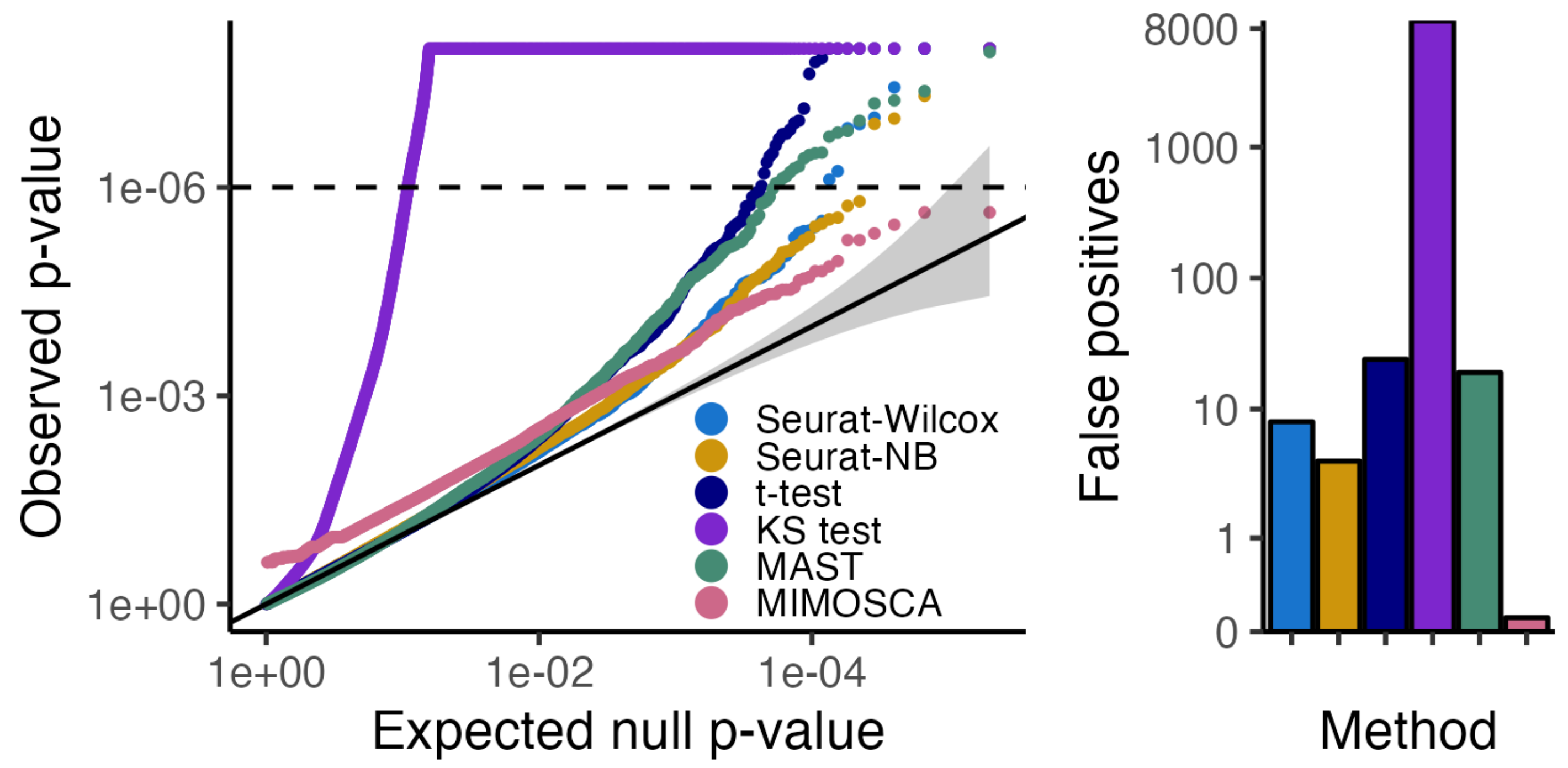
Perturb-seq data analysis

Does a perturbation affect the expression of a gene?



Statistical & computational challenges

Existing methods show miscalibration on control data and struggle to scale to large datasets.



The SNO (SCEPTRE, Nextflow, ondisc) technology stack

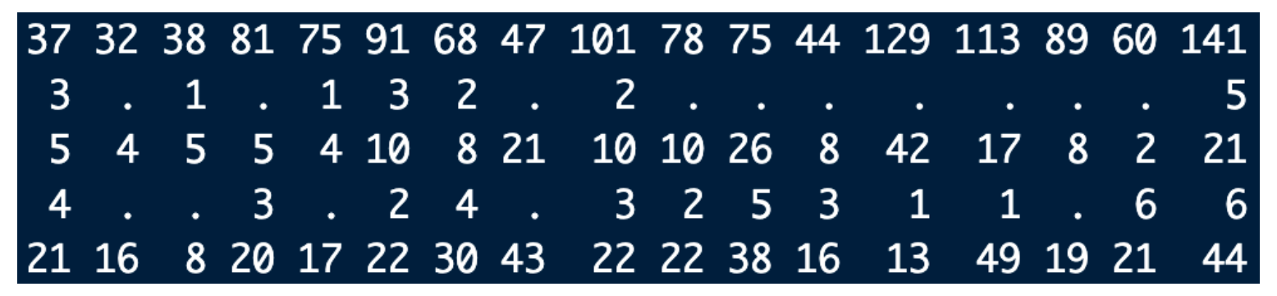
The SNO technology stack enables **statistically rigorous**, **massively scalable**, and **user-friendly** perturb-seq data analysis on laptops, clusters, and clouds. The SNO pipeline involves several steps.

1. Import data

10x GENOMICS

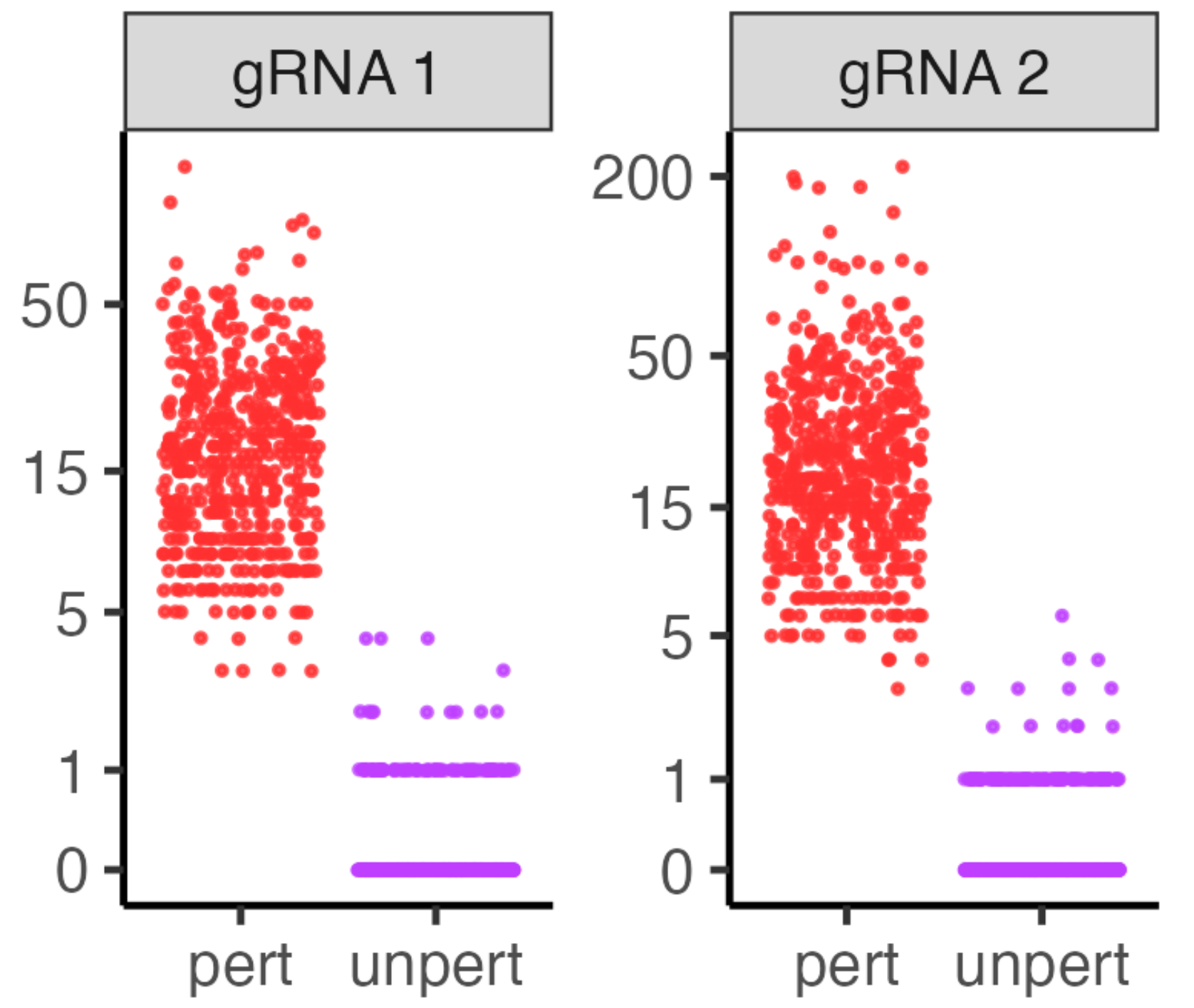
Parse BIOSCIENCES

R matrix input

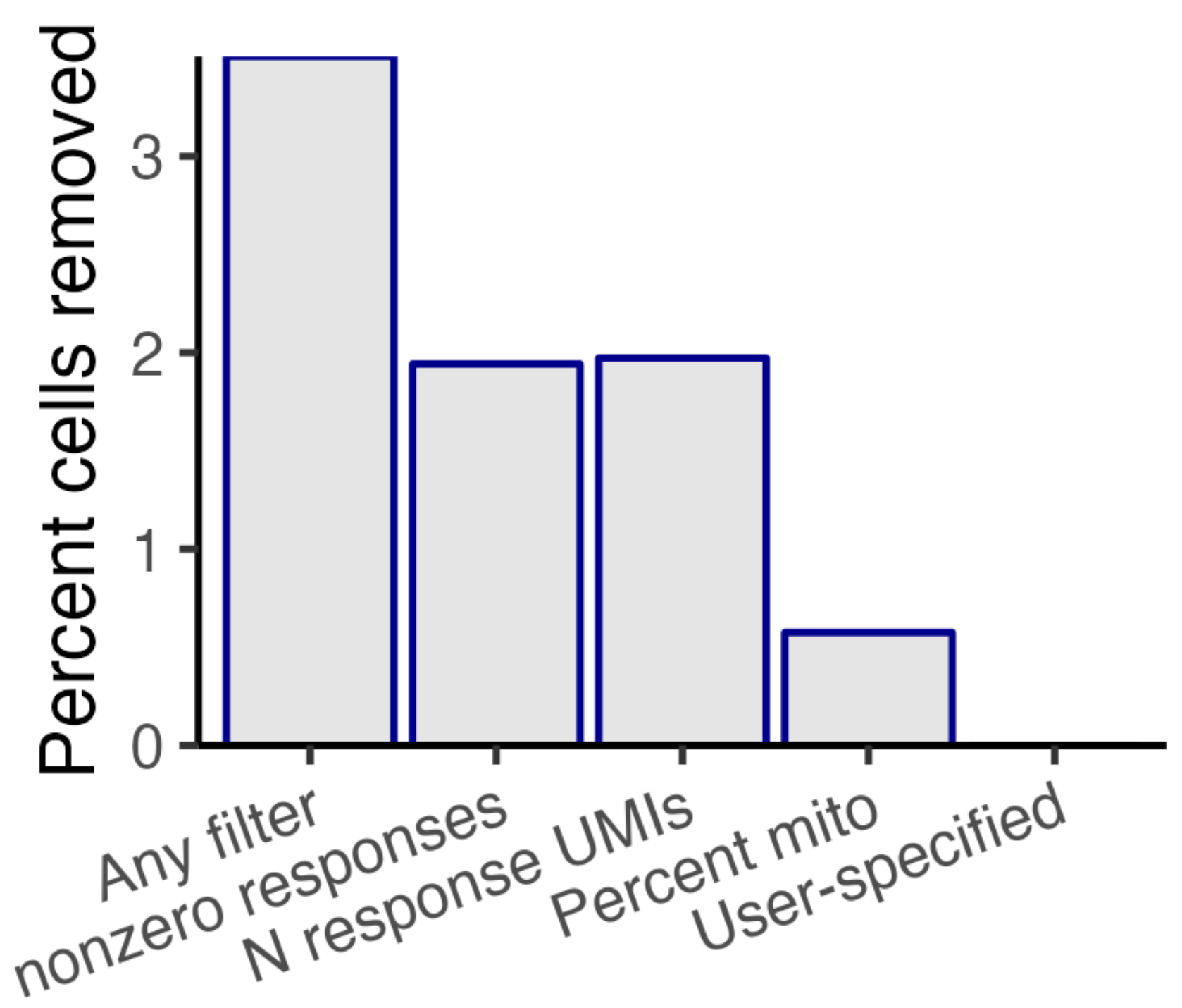


```
37 32 38 81 75 91 68 47 101 78 75 44 129 113 89 60 141
 3  .  1  .  1  3  2  .   2  .  .  .   .   .  .  .   5
 5  4  5  5  4 10  8 21  10 10 26  8  42  17  8  2  21
 4  .  .  3  .  2  4  .   3  2  5  3   1   1  .  6   6
21 16  8 20 17 22 30 43  22 22 38 16  13  49 19 21  44
```

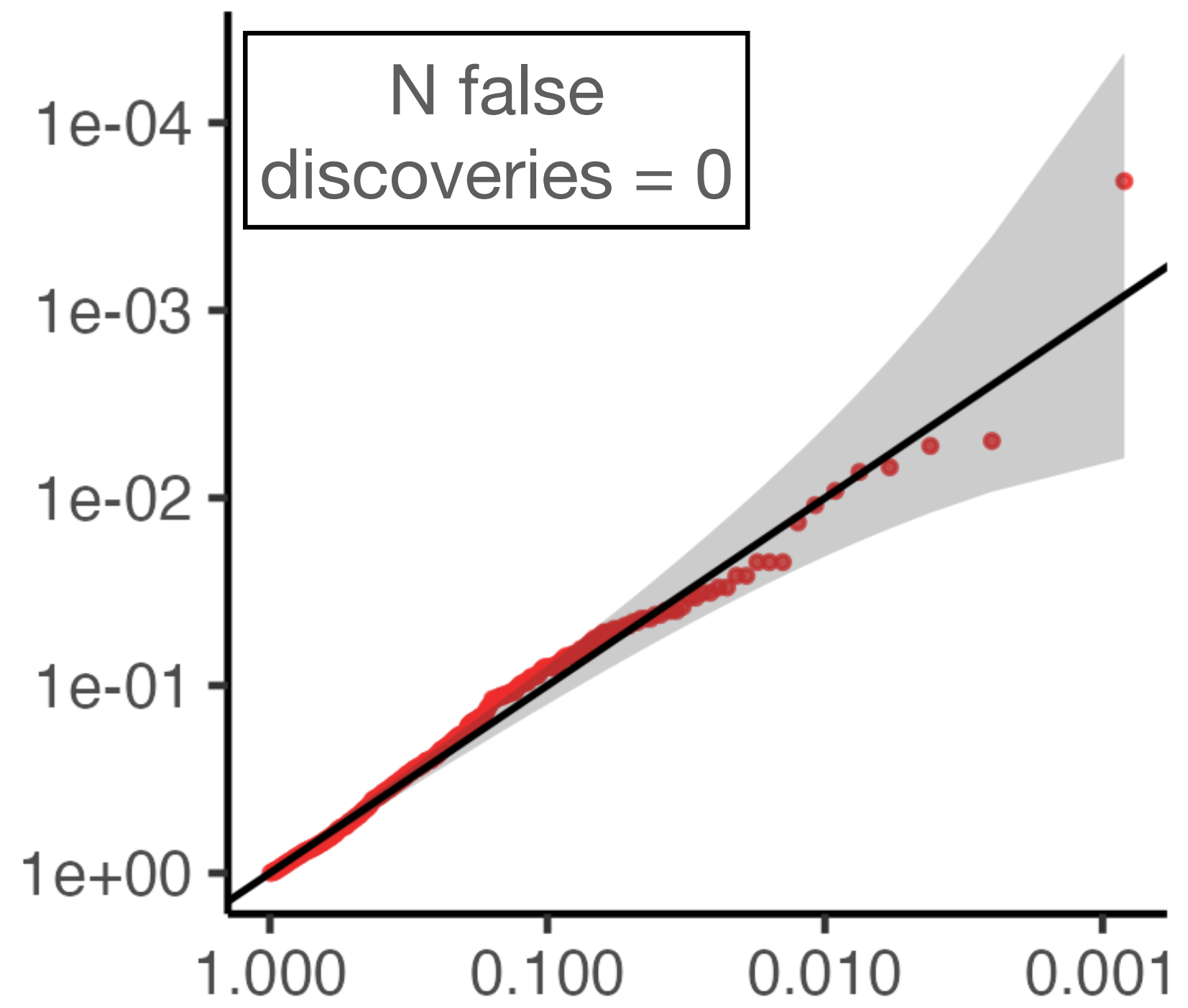
2. Assign gRNAs to cells



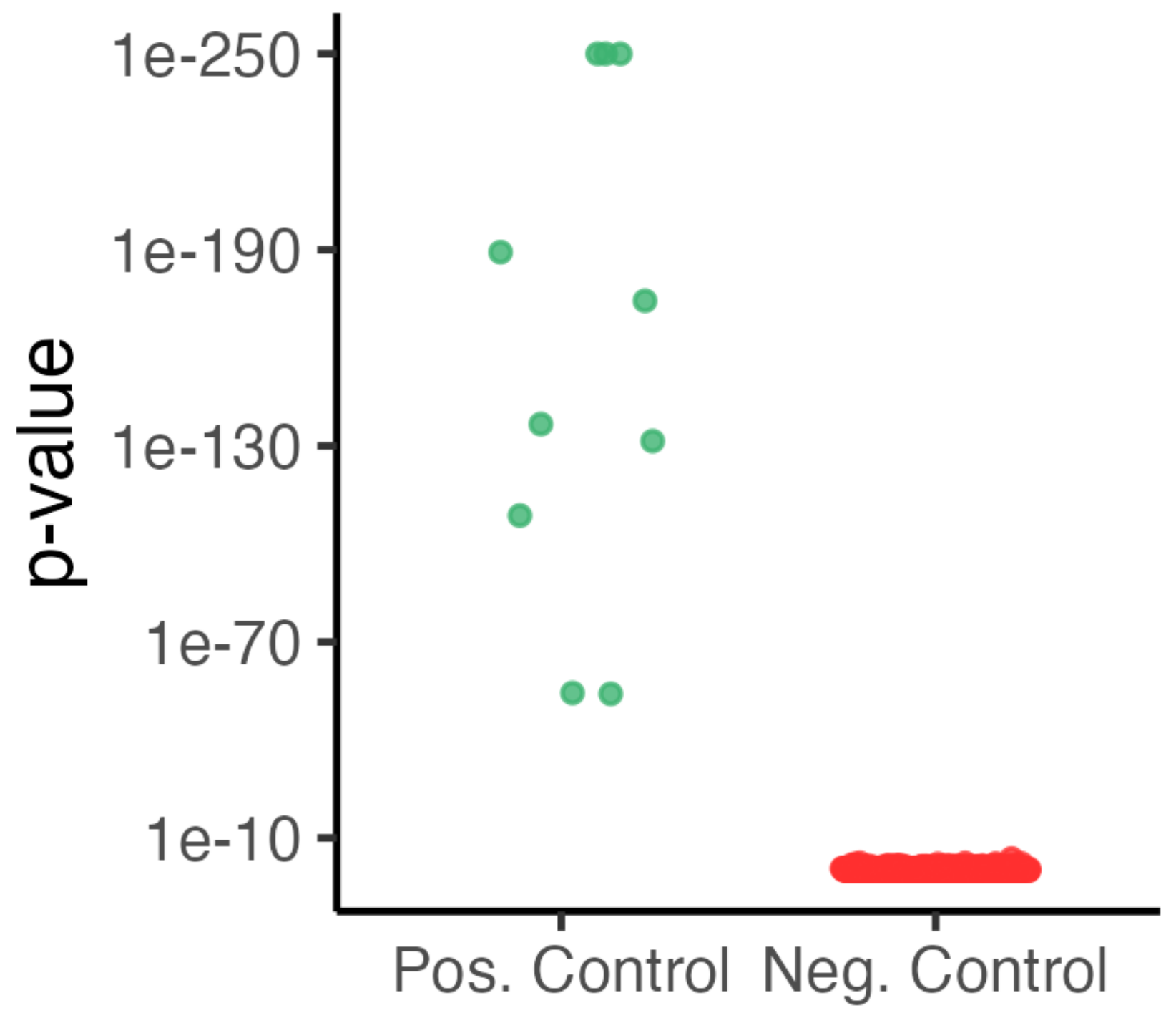
3. Run quality control



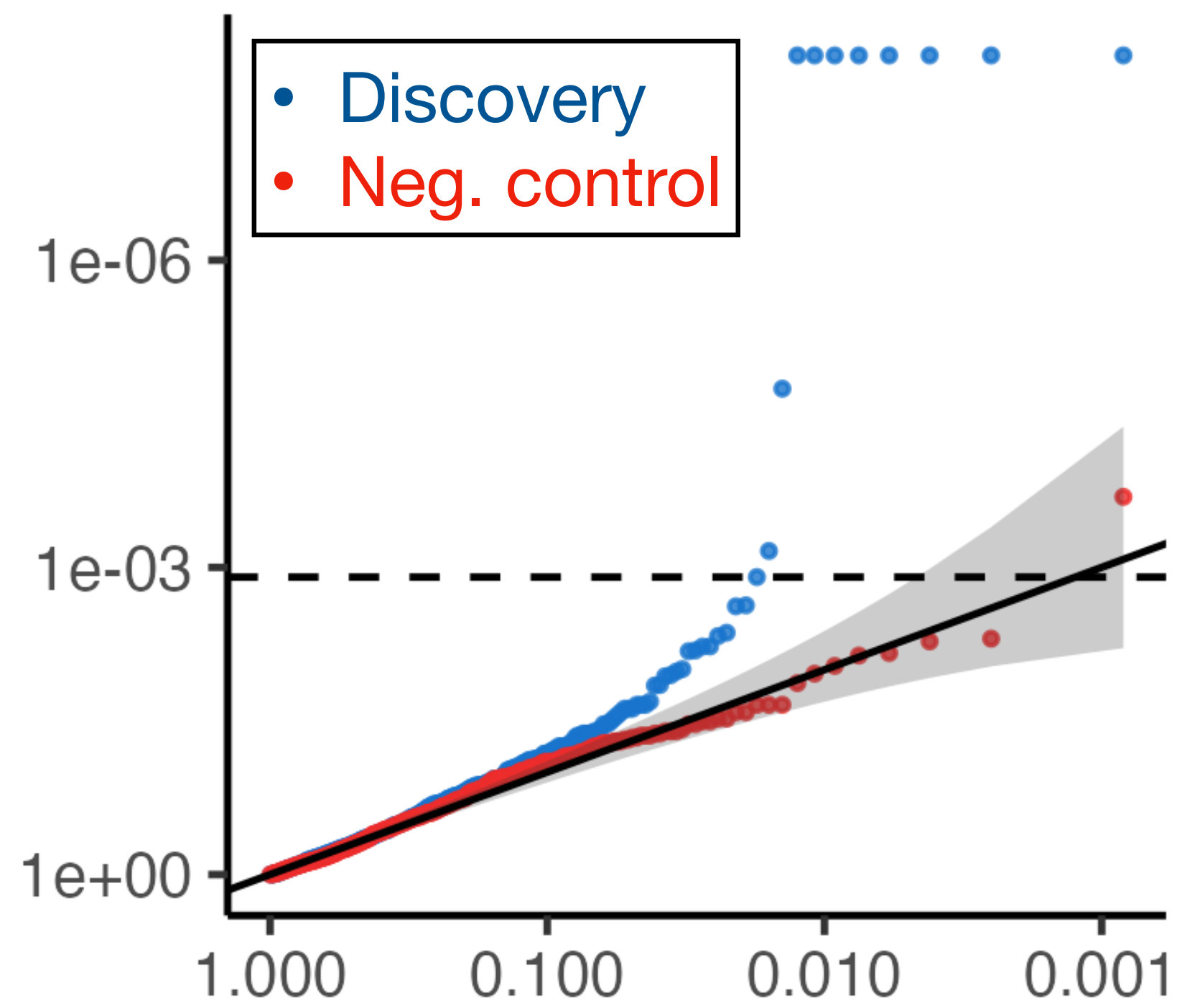
4. Run calibration check



5. Run power check

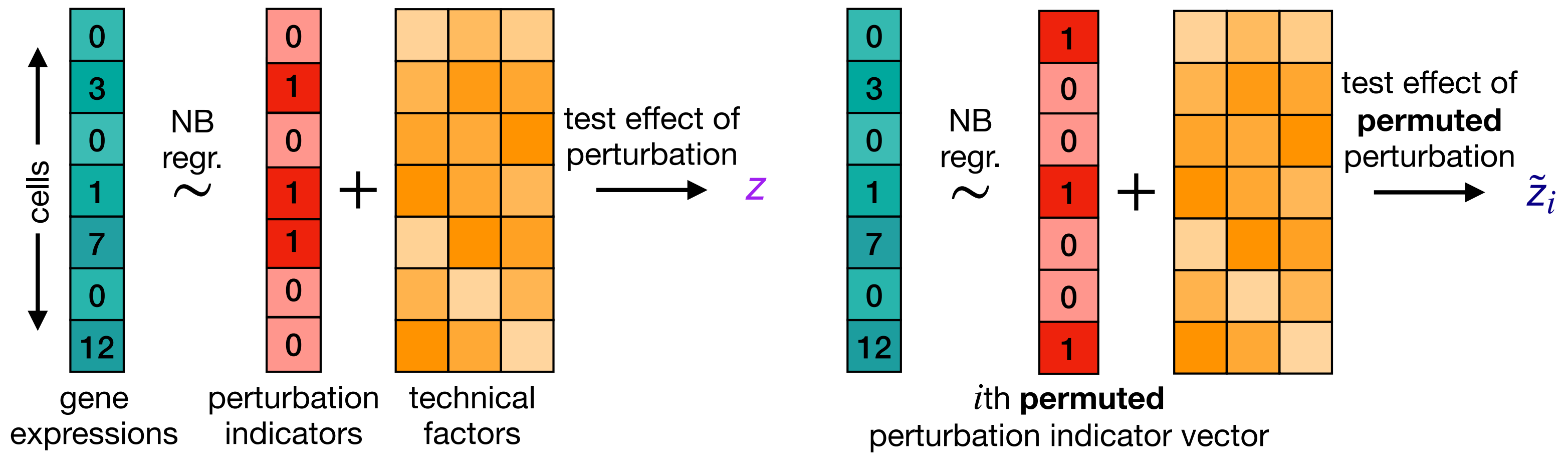


6. Run discovery analysis



New statistical & algorithmic methods

- Robust negative binomial (NB) regression by resampling NB score statistics.

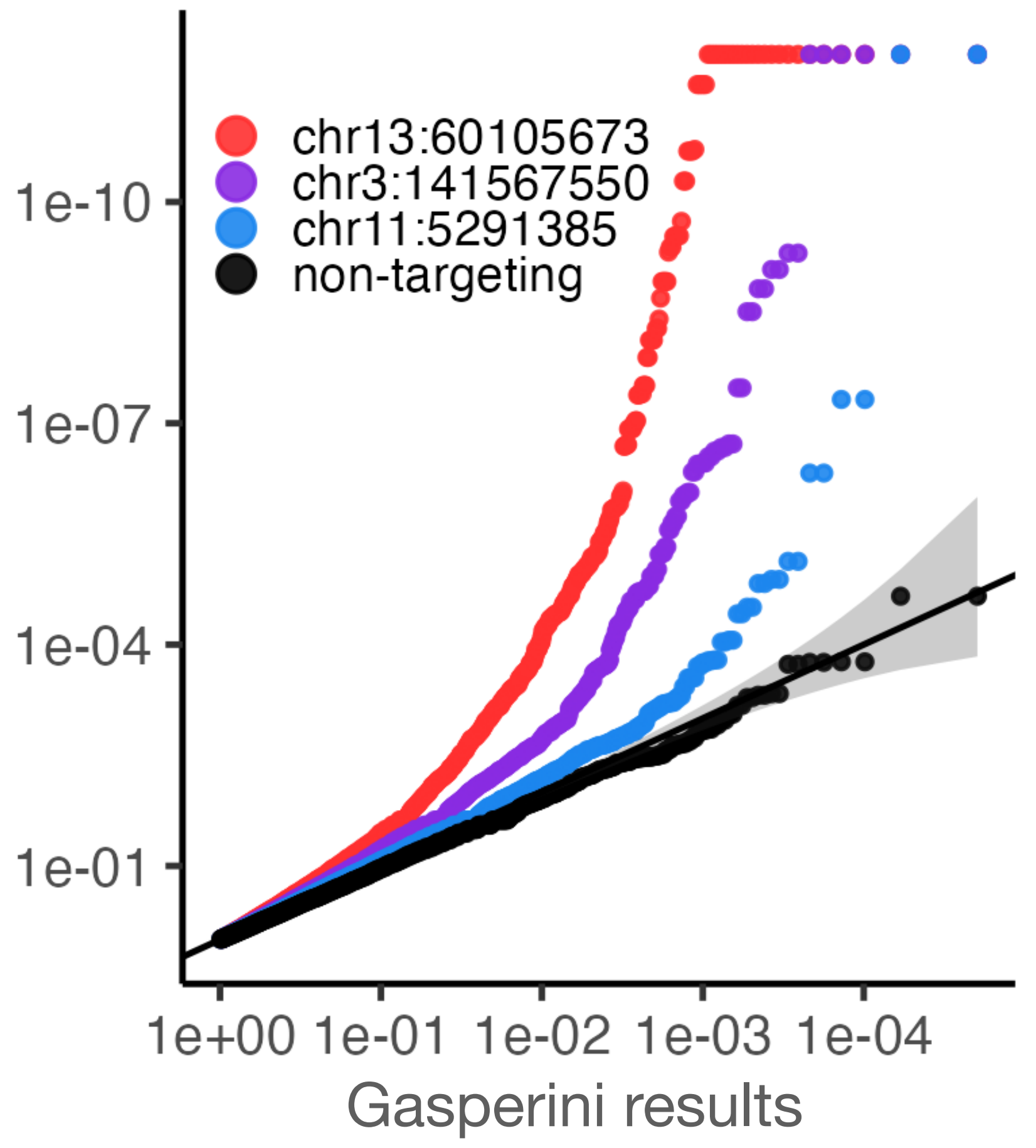


- Sparsity-exploiting algorithm for efficiently computing GLM score tests.
- Space- and time-optimal algorithm for transposing large, sparse matrices out-of-core.
- Algorithm for recycling compute across a large number of permutation tests.

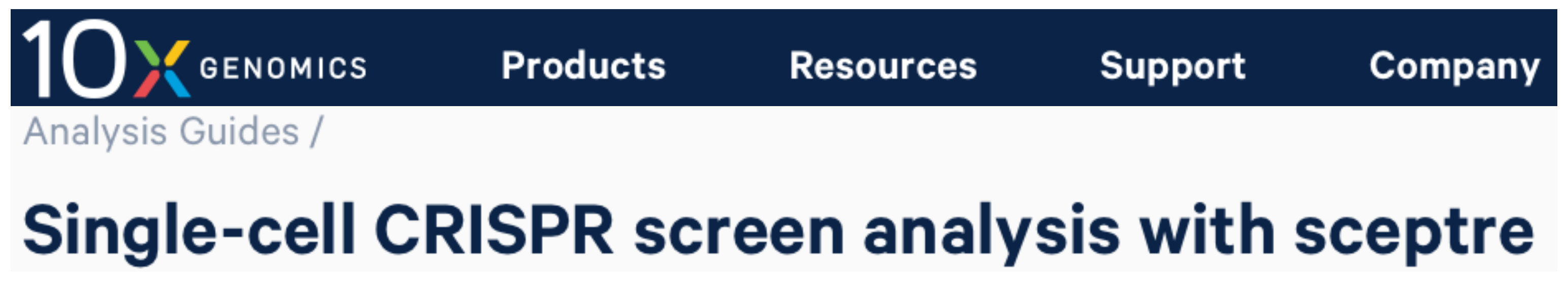
Statistical & computational performance

Trans analysis of high multiplicity-of-infection (MOI) CRISPRi screen of enhancers (Gasperini, 2019) and low-MOI CRISPRi screen of genes (Replogle 2022).

Dataset	Gasperini	Replogle
Number of cells	200K +	610K +
Number of pairs	170 million	93 million
Number of processors	152	47
Running time	8.5 hours	5.8 hours
Max memory	2.0 GB	2.0 GB



SCEPTRE is recommended by 10x Genomics!



Analysis Guides /

Single-cell CRISPR screen analysis with sceptre