Technologies for single-cell CRISPR screen data analysis

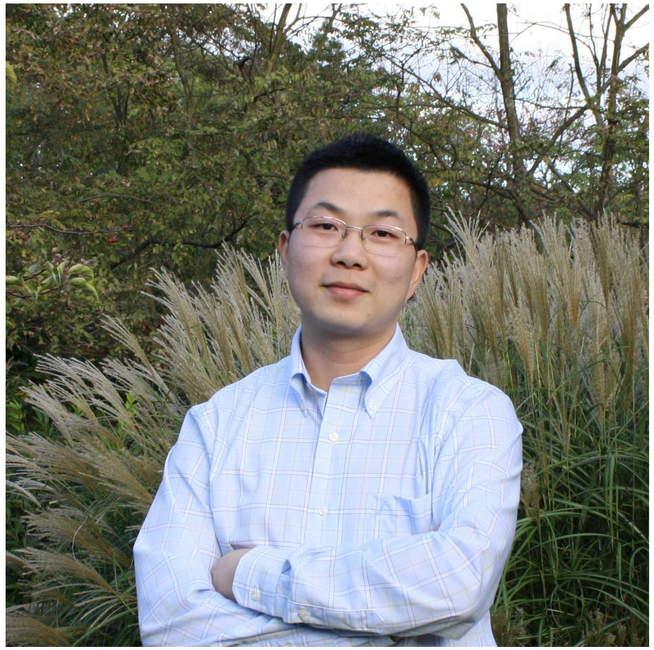
Tim Barry

Advisors

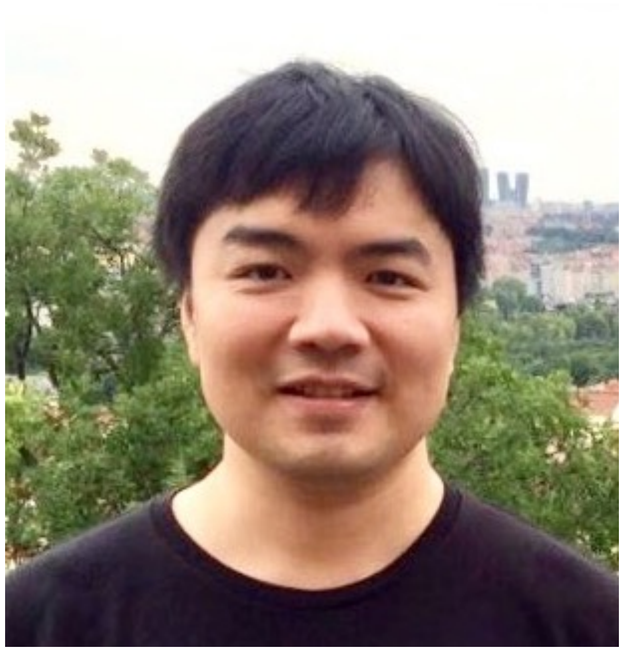
Kathryn Roeder

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

Committee members



Jing Lei



Jian Ma (Computational Biology Dept., CMU)

Aaditya Ramdas

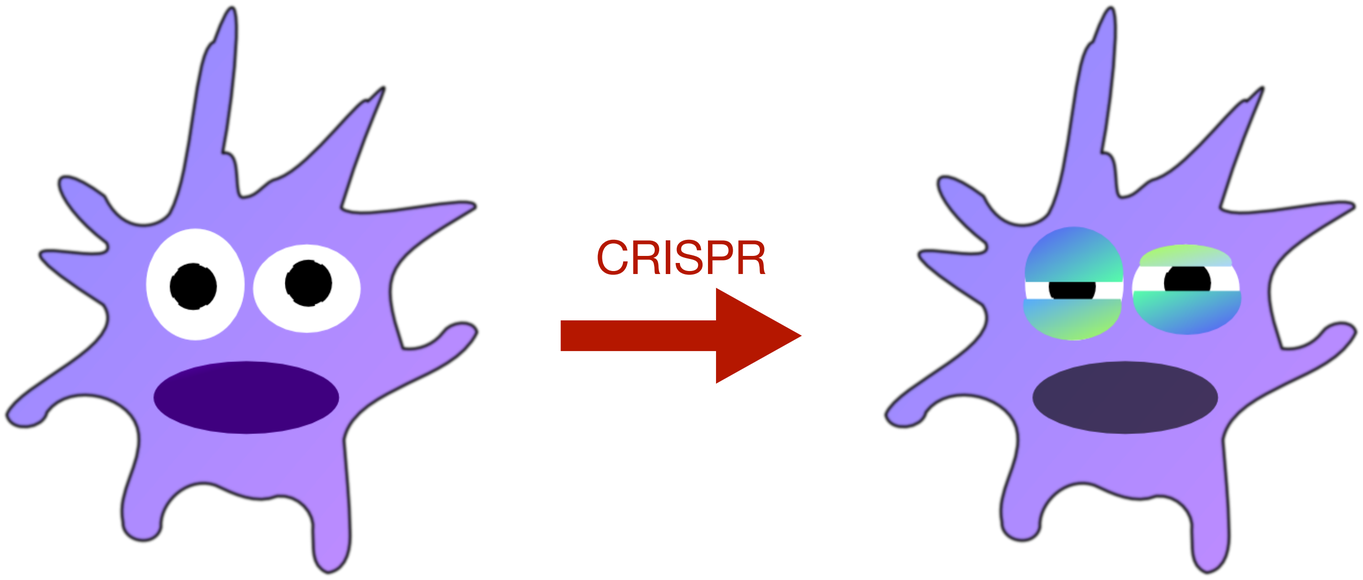
CRISPR is a genome engineering technology that can be used to modify living organisms in incredible ways.

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

CRISPR

CRISPR accelerates biological discovery.

1. Identify a region of the genome (e.g., a gene) with unknown function.
2. Perturb this region of the genome with CRISPR.
3. See what happens!



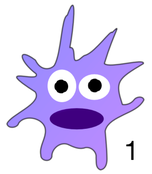
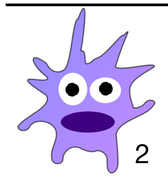
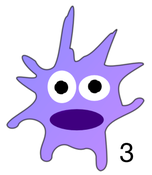
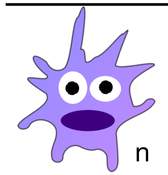
Normal cell

CRISPR-edited cell

What might be the function of the gene *targeted* by CRISPR?

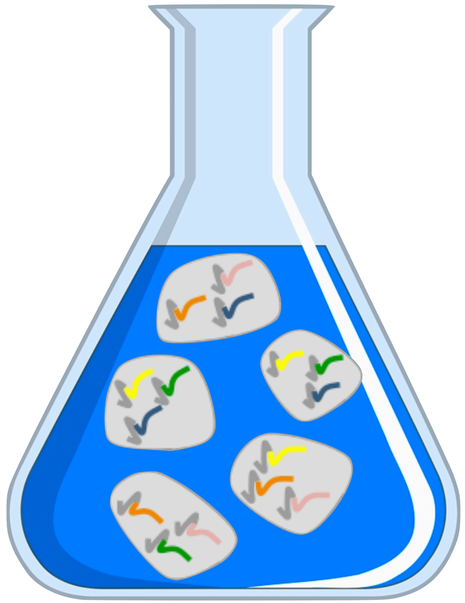
Single-cell RNA sequencing is a technology for measuring gene expressions in individual cells.

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.

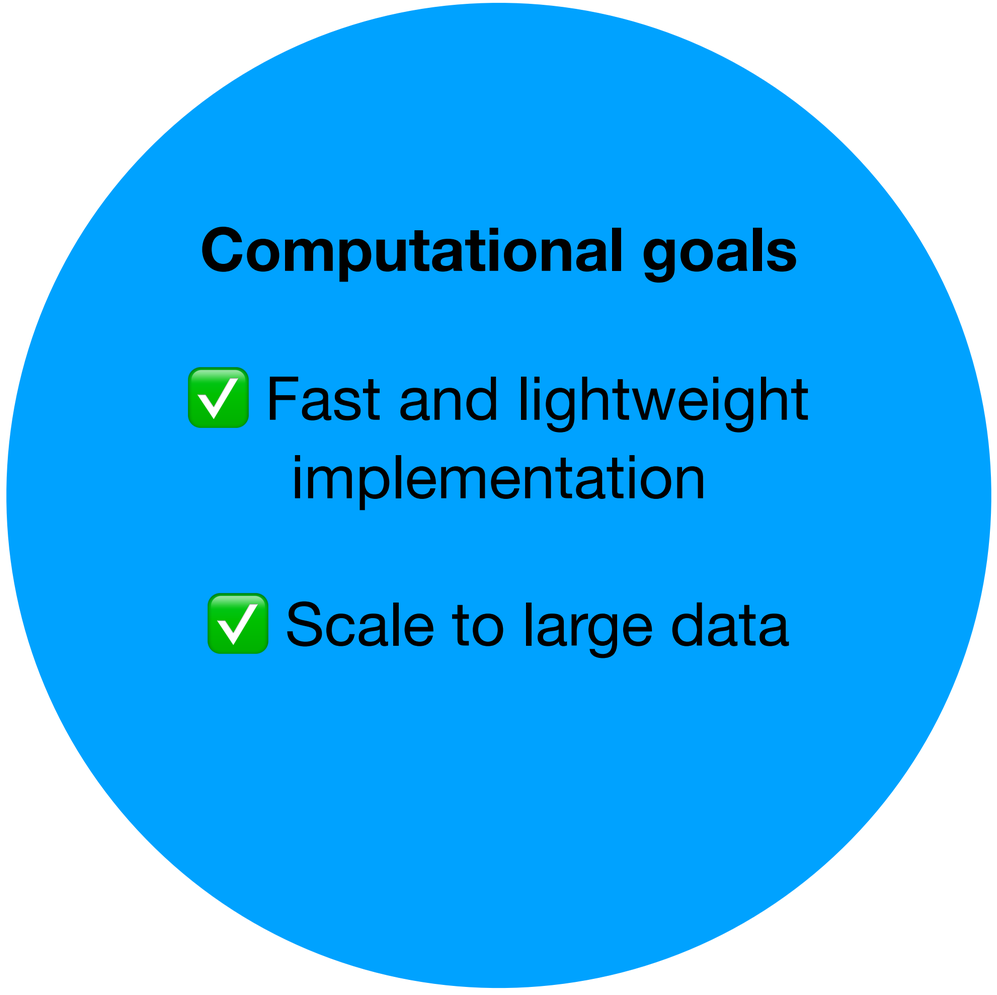
Insert CRISPR perturbations into cells



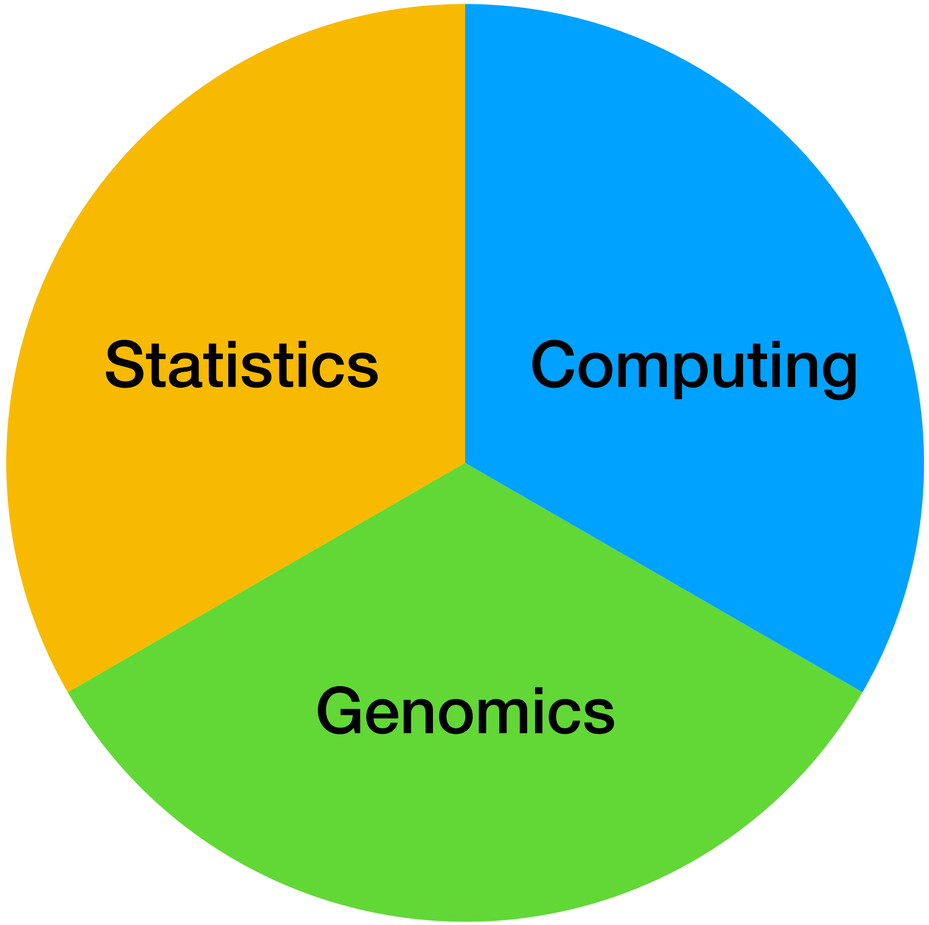
Single-cell sequencing

- Single cell CRIPSR screens could transform biology and medicine. However, they pose major challenges.

The broad objective of this thesis is to develop statistically rigorous and computationally efficient tools for single-cell CRISPR screen analysis.



Three (completed or mostly completed) projects.



1. **Statistics**: "Exponential family measurement error models for single-cell CRISPR screens," by **T Barry**, E Katsevich, K Roeder. *Under review at Biostatistics*.

2. **Genomics**: "SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis," by **T Barry**, X Wang, J Morris, K Roeder, E Katsevich. *Genome Biology*, 2021.

3. **Computing**: "Ondisc: efficient algorithms and functional data structures for out-of-core and distributed single-cell analysis," by **T Barry**, S Dai, E Katsevich, K Roeder. *In preparation*.

Roadmap

1. Single-cell CRISPR screen overview

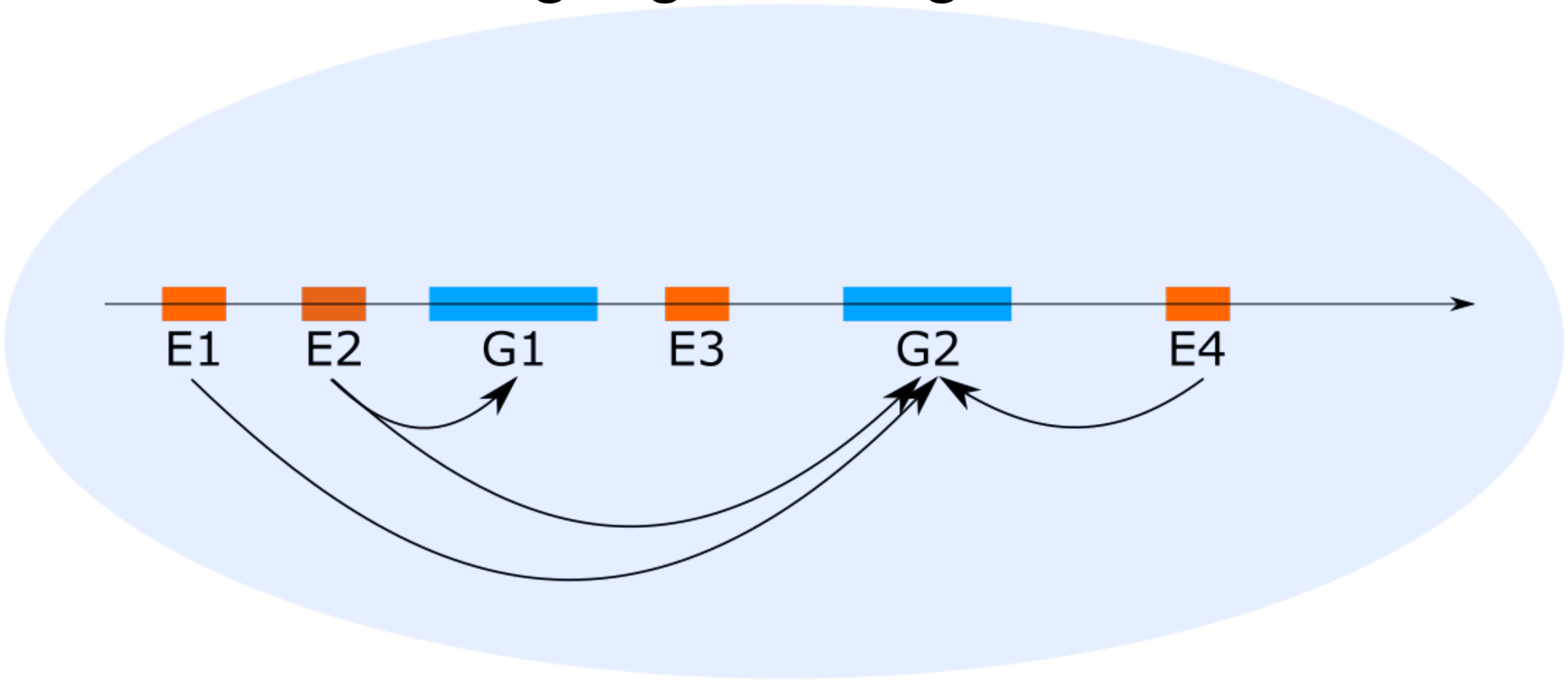
2. Exponential family measurement error models...

3. SCEPTRE

4. ondisc

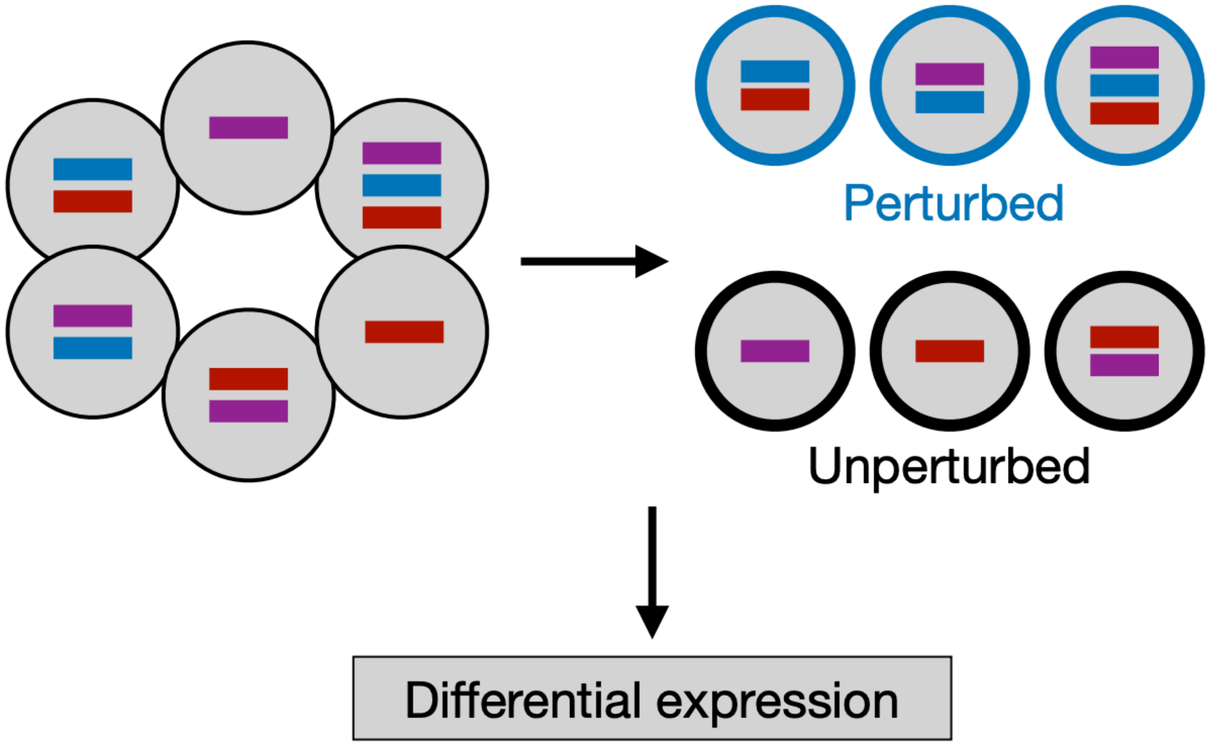
5. Future work

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.

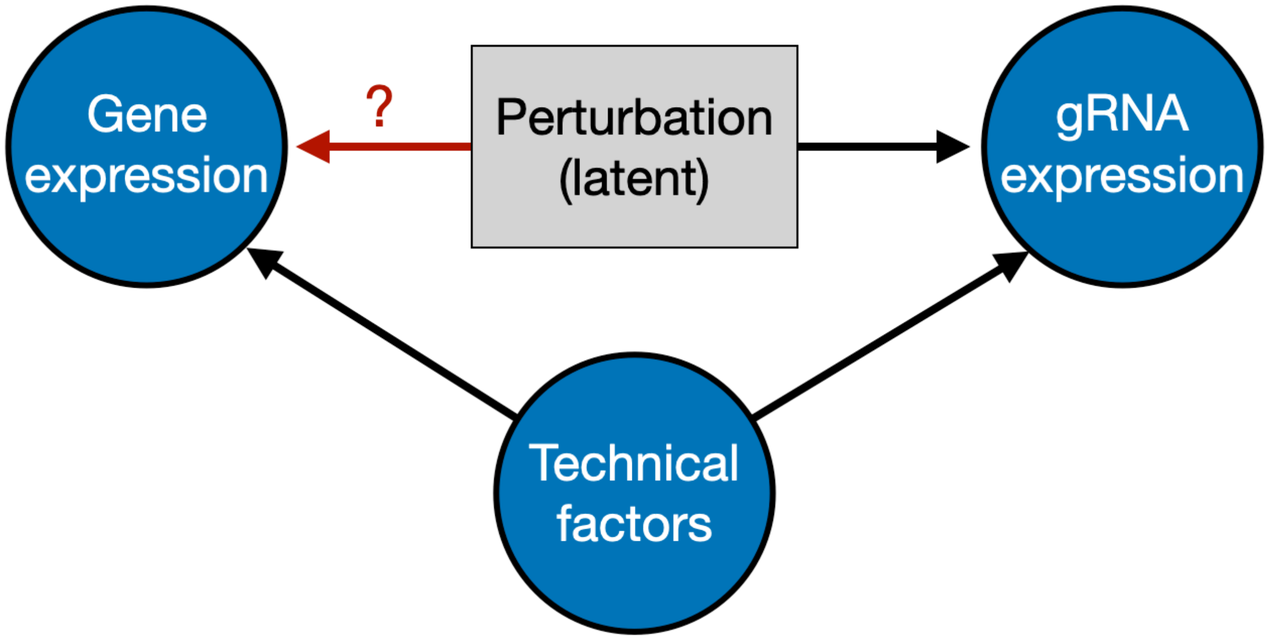
Experimental design



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.

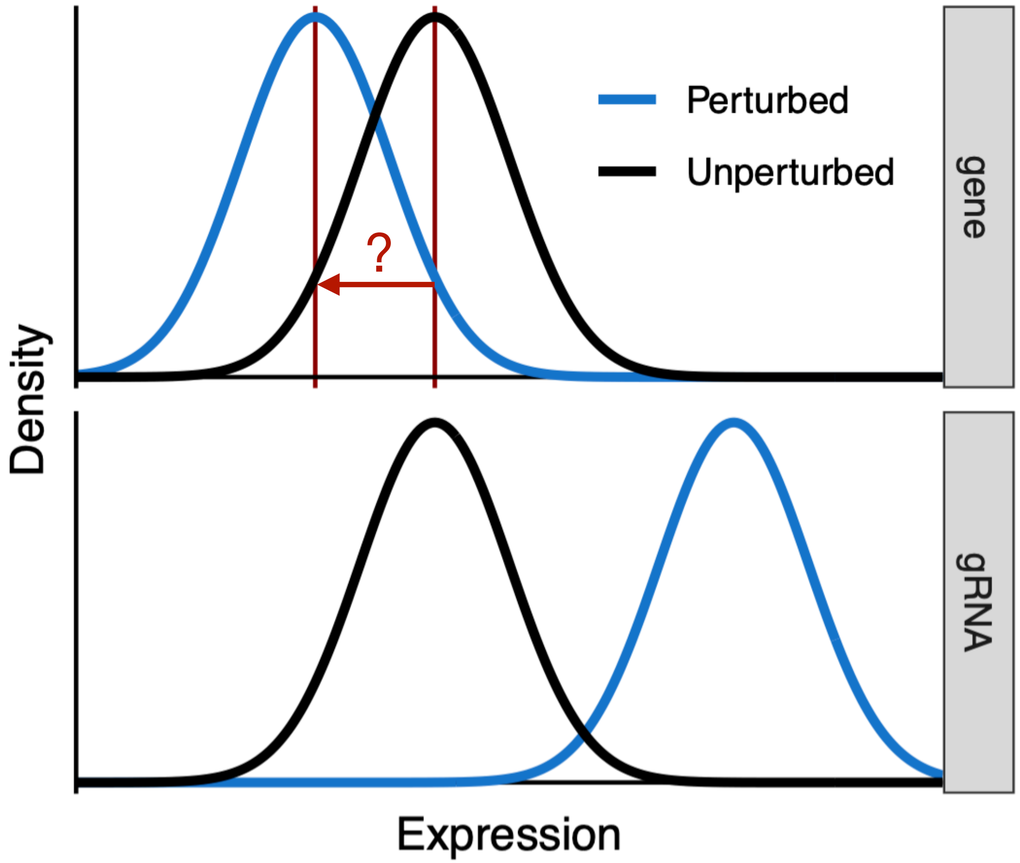
Analysis challenges



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.

Roadmap

1. Single-cell CRISPR screen overview

2. Exponential family measurement error models...

2. SCEPTRE

3. ondisc

4. Future work

Exponential measurement family measurement errors models for single-cell CRISPR screens (2022)

Problem:

The "treatment" variable (perturbation presence or absence) is measured with error, complicating differential expression analysis.

Objectives:

1. Derive a model for the entire data-generating mechanism, including the measurement process.

2. Obtain methods for estimation and inference in this model.

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

$$\textit{gene transcript count} \sim \text{GLM}(\textit{perturbation}, \textit{confounders})$$
$$\textit{gRNA transcript count} \sim \text{GLM}(\textit{perturbation}, \textit{confounders})$$
$$\textit{perturbation} \sim \text{Bernoulli}(\pi)$$

- 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
   - Excellent starting estimates to speed convergence
2. Analytical derivation of observed information matrix
3. Computational pipeline that scales to clusters and clouds

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

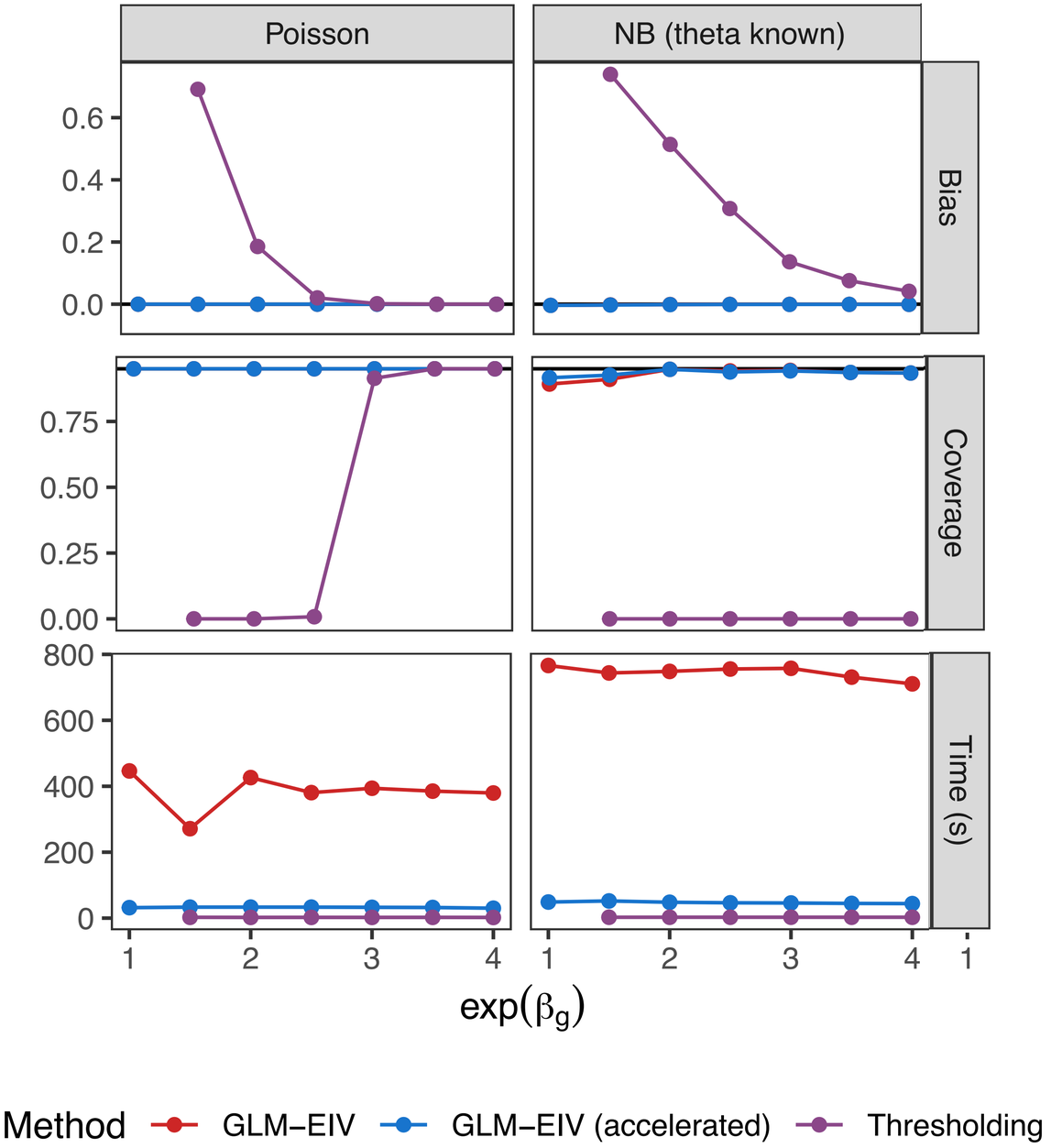
$$\widehat{\textit{perturbation}_i} = I(\textit{gRNA transcript count}_i \geq c)$$

- Fit the GLM

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

- 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.



Roadmap

1. Single-cell CRISPR screen overview

2. Exponential family measurement error models...

3. SCEPTRE

4. ondisc

5. Future work

SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis (2021)

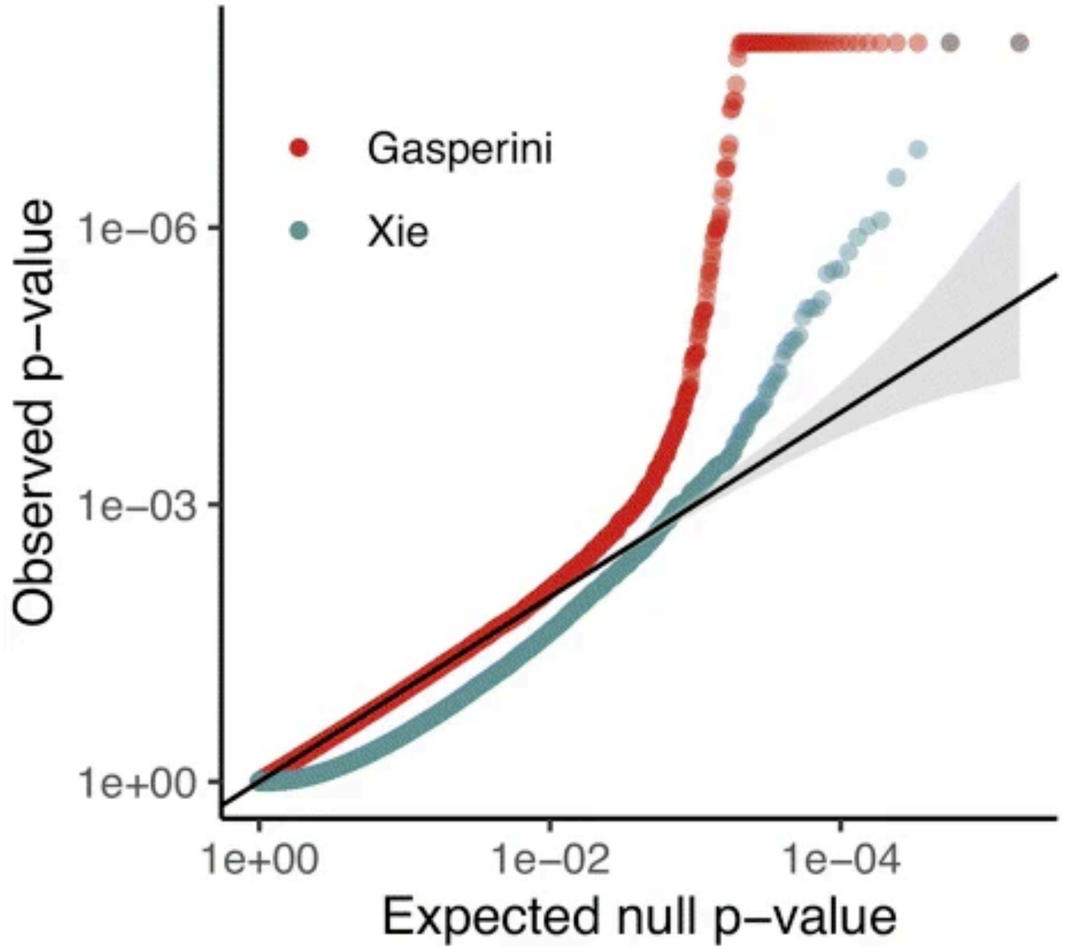
Problem:

Calibrating *tests* of association between a CRISPR perturbation and the expression of a gene is challenging.

Objective:

Develop a well-calibrated and powerful test of association in single-cell CRISPR screen analysis; implement this test in an efficient and scalable software package.

Existing hypothesis testing methods demonstrate severe *p*-value inflation on negative control data.



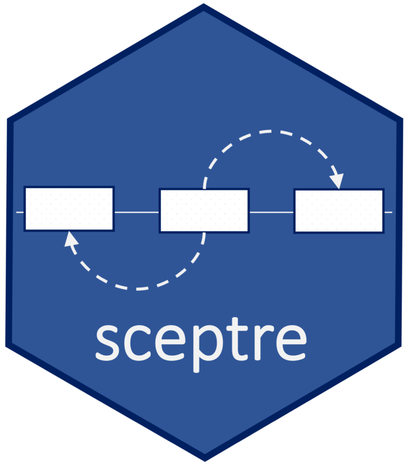
Gasperini: negative binomial regression

Xie: nonparametric, chi-squared-like test of independence

Conditional randomization simultaneously enables confounder adjustment and protects against expression model misspecification.

	Adjusts for technical factors	Is robust to expression model misspecification
Parametric method	✅	❌
Nonparametric method	❌	✅
Conditional randomization test	✅	✅

We develop SCEPTRE, a custom implementation of the conditional randomization test for single-cell CRISPR screen data.



SCEPTRE

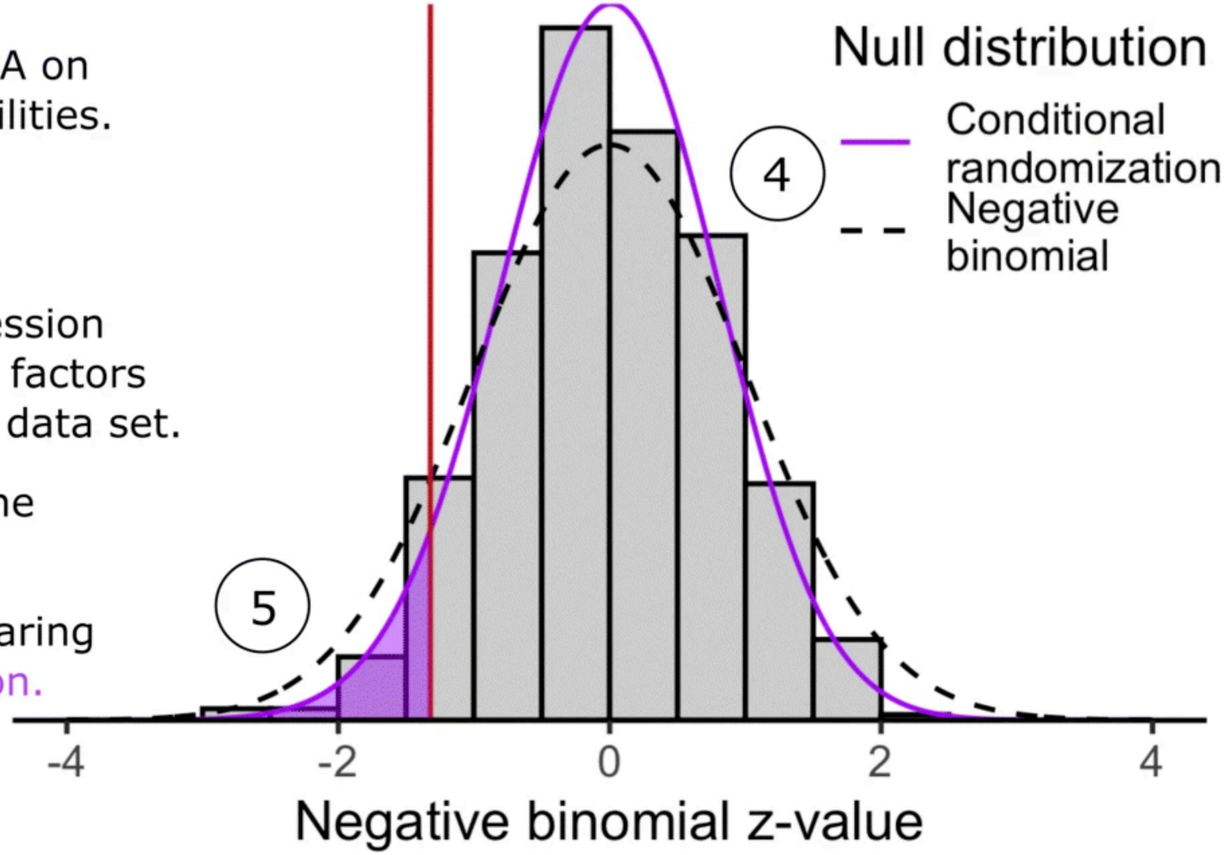
Step 1: Fit logistic regression of gRNA on technical factors to get fitted probabilities.

Step 2: Repeatedly resample gRNAs for each cell based on probabilities.

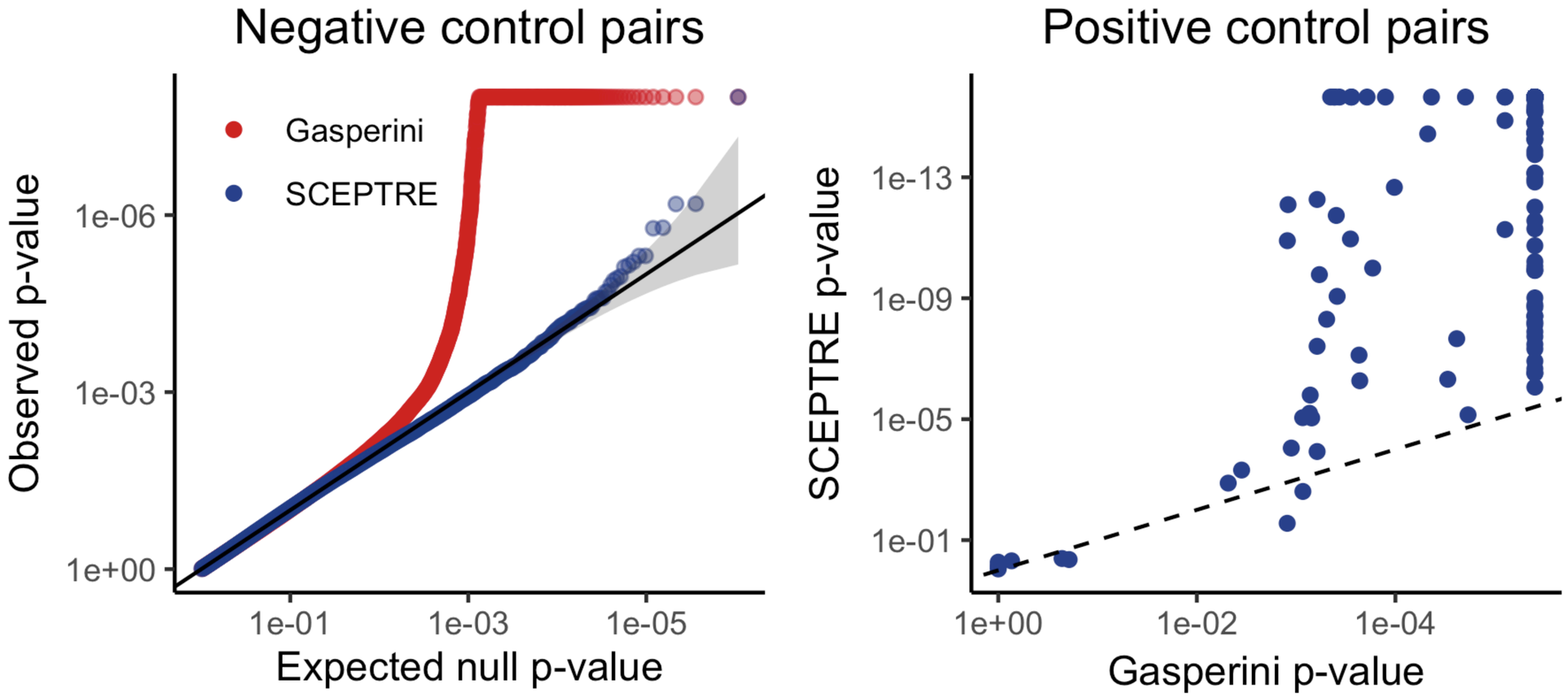
Step 3: Fit a negative binomial regression of expression on gRNA and technical factors for original data and each reshuffled data set.

Step 4: Fit a skew-t distribution to the set of resampled z-values.

Step 5: Compute a p-value by comparing original z-value to the null distribution.



SCEPTRE demonstrates superior calibration and power on negative and positive control data, respectively.



Roadmap

1. Single-cell CRISPR screen overview

2. Exponential family measurement error models...

3. SCEPTRE

4. ondisc

5. Future work

`ondisc`: efficient algorithms and functional data structures for out-of-core and distributed single-cell analysis (in preparation)

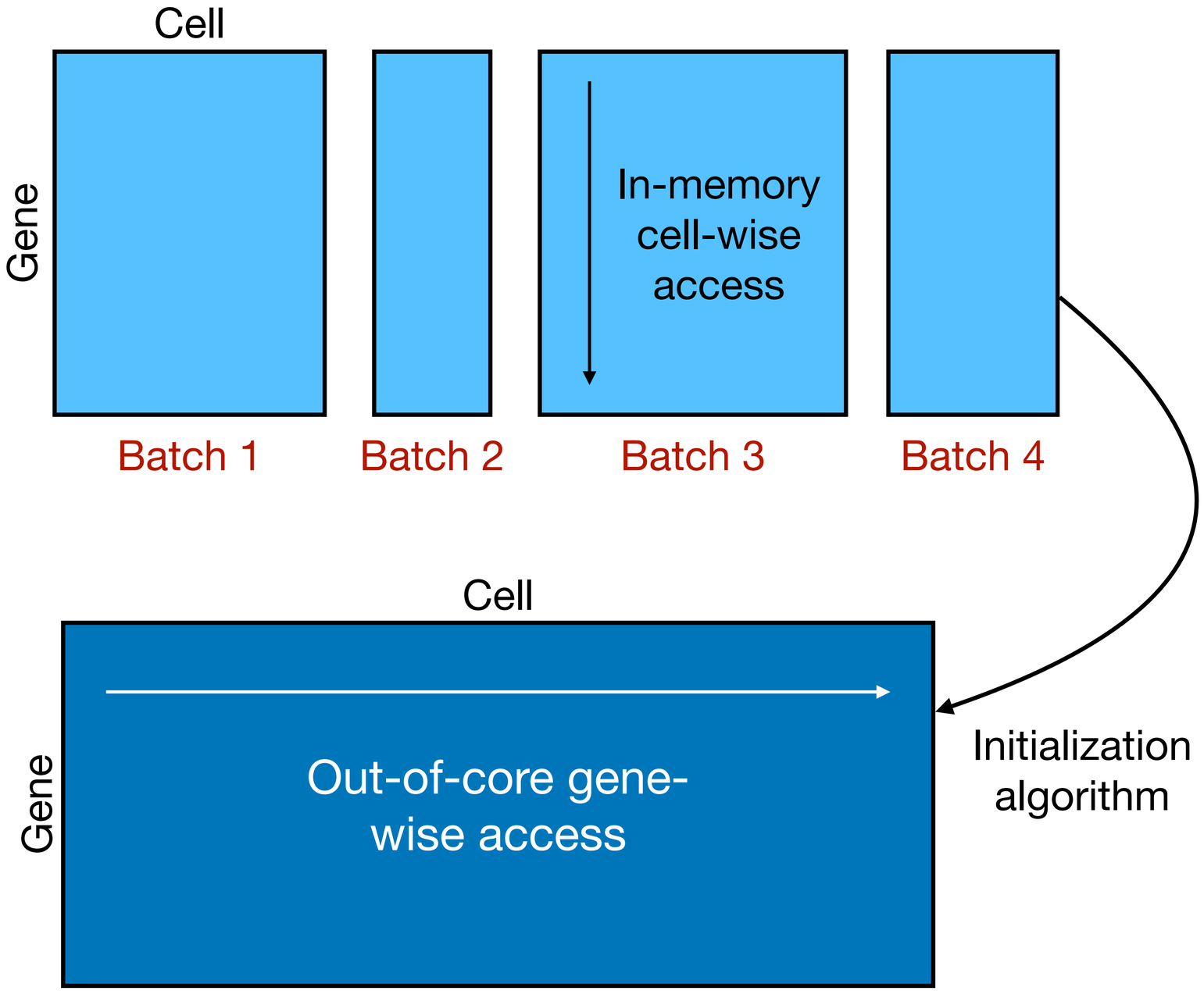
Problem:

Large single-cell datasets (including single-cell CRISPR screen datasets) pose serious computational challenges.

Objective:

1. Build a platform that powers (i) **out-of-core** and (ii) **distributed** computing on single-cell data.

2. Implement this framework in an R/C++ package.



Large single-cell data pose several challenges:

1. May not fit in memory
2. Sparse matrix format
3. Access to cells, not genes

The `ondisc` initialization algorithm converts a large, cell-accessible sparse matrix into gene-accessible form out-of-core.

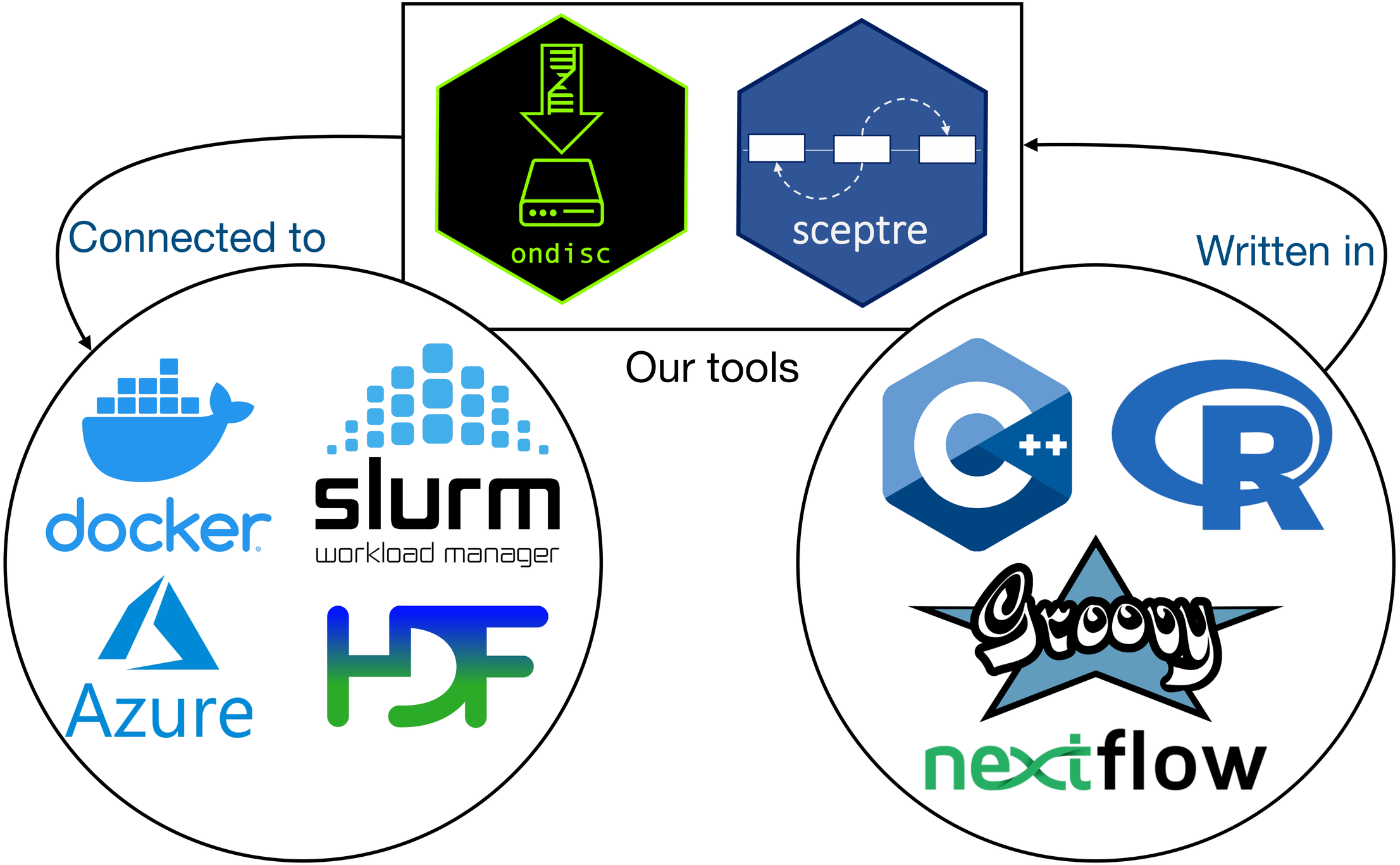
The `ondisc` initialization algorithm is equivalent to transposing a large, sparse matrix out-of-core.

- We introduce a generalized radix sort algorithm for this purpose. We prove that the algorithm is asymptotically optimal in time and space:
  - Running time: O(*r*)
  - Disk space: O(*r*)
  - Memory space: O(1),

  where *r* is the number of nonzero elements in the sparse matrix.

`ondisc` implements (or facilitates the implementation of) out-of-core and/or distributed versions of common single-cell analysis tasks.

- Gene-gene coexpression analysis
- PCA on a normalized gene expression matrix
- Differential expression and single-cell CRISPR screen analysis


Connected to
ondisc
sceptre
Written in
Our tools
docker
slurm
workload manager
Azure
C++
R
Groovy
nextflow
Clouds, clusters, and databases
Languages

Roadmap

1. Single-cell CRISPR screen overview

2. Exponential family measurement error models...

3. SCEPTRE

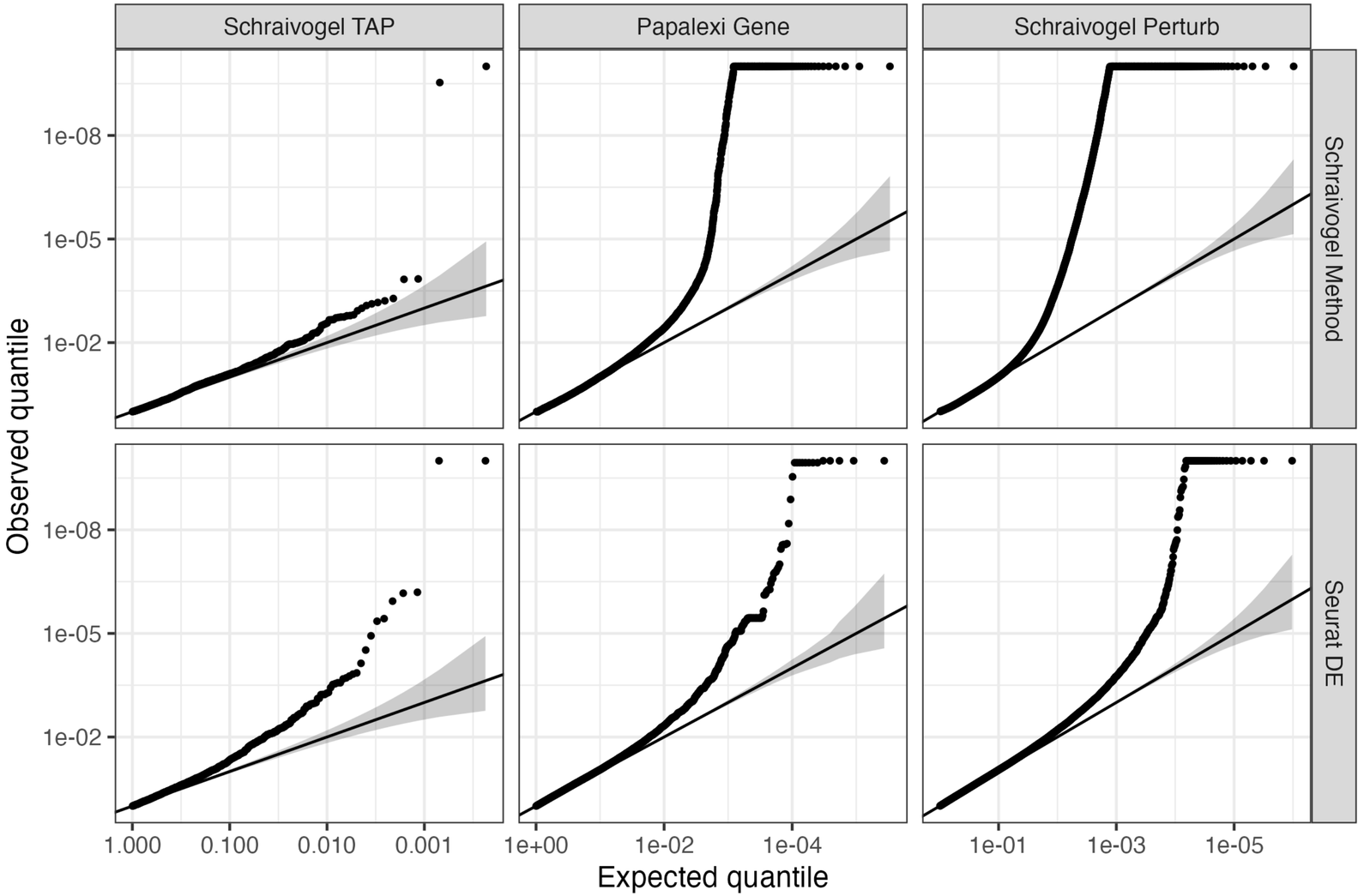
4. ondisc

5. Future work

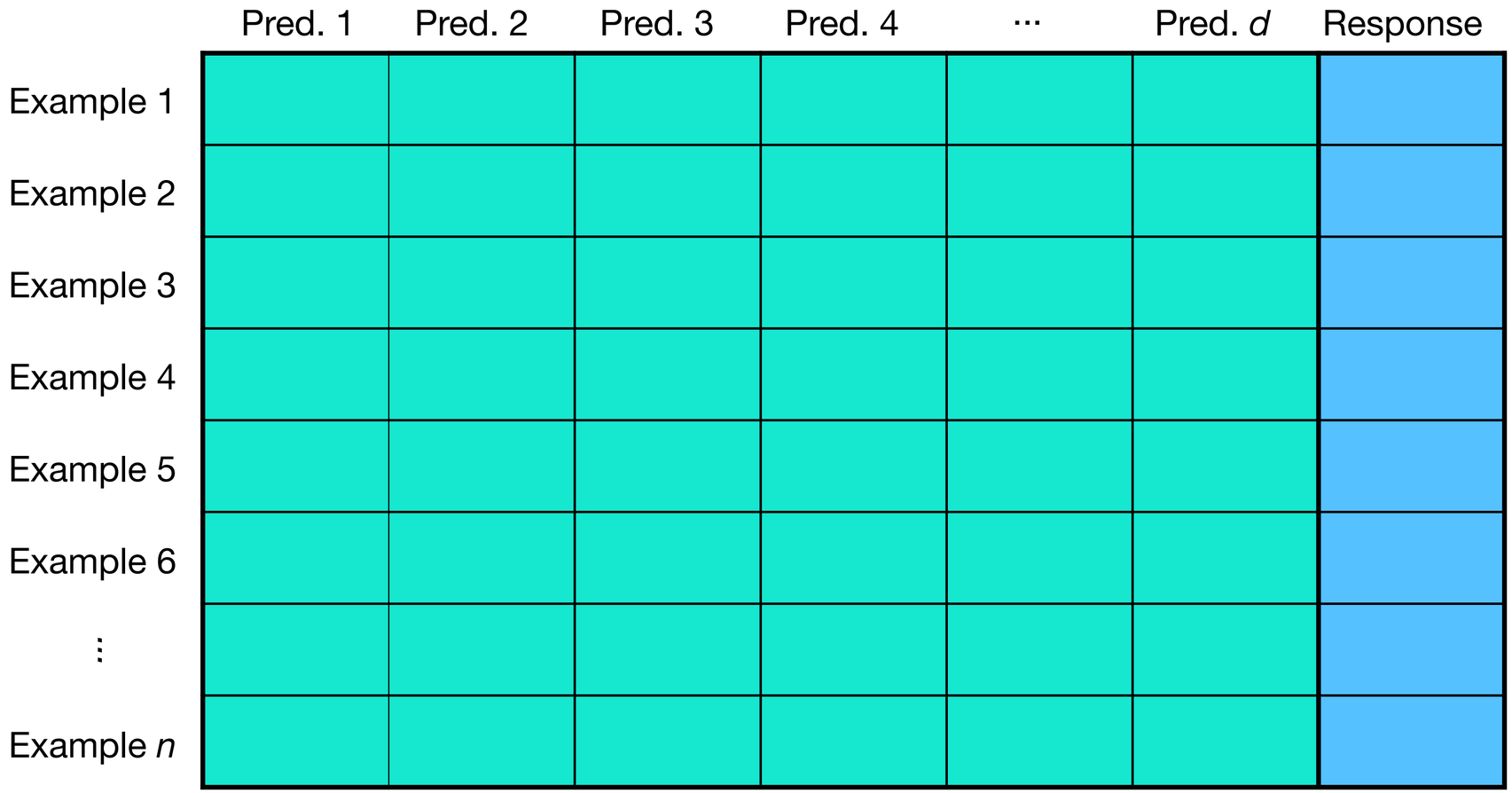
SCEPTRE2 will extend SCEPTRE to apply to a fundamentally different class of single-cell CRISPR screen datasets.

- Generalize statistical methodology.
- Improve computational performance.
- Benchmark on ~10 representative datasets.

Preliminary results indicate that existing methods are computationally inefficient and severely miscalibrated on negative control data.

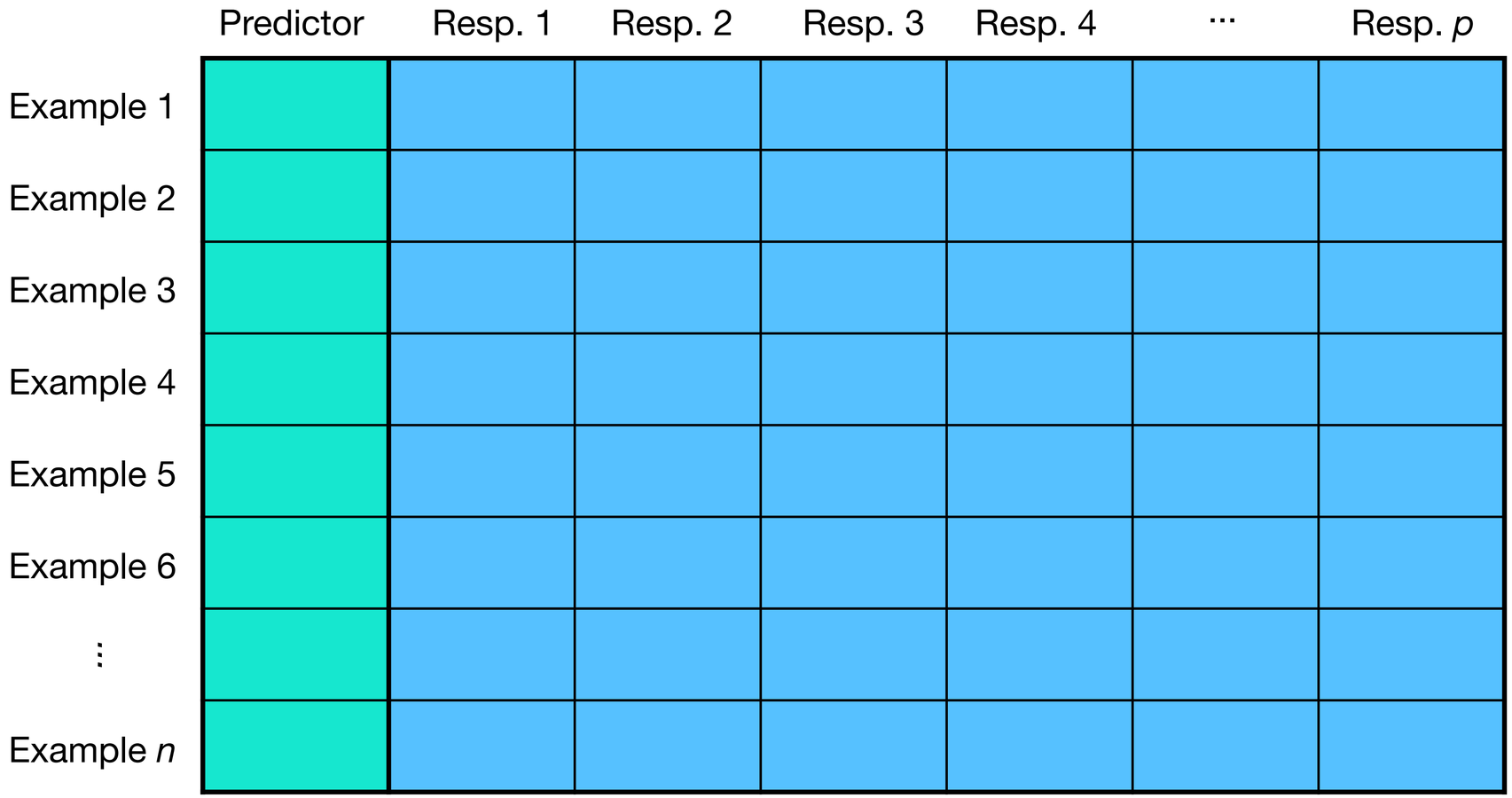


High throughput data are data on which thousands (or more) of variables are measured.



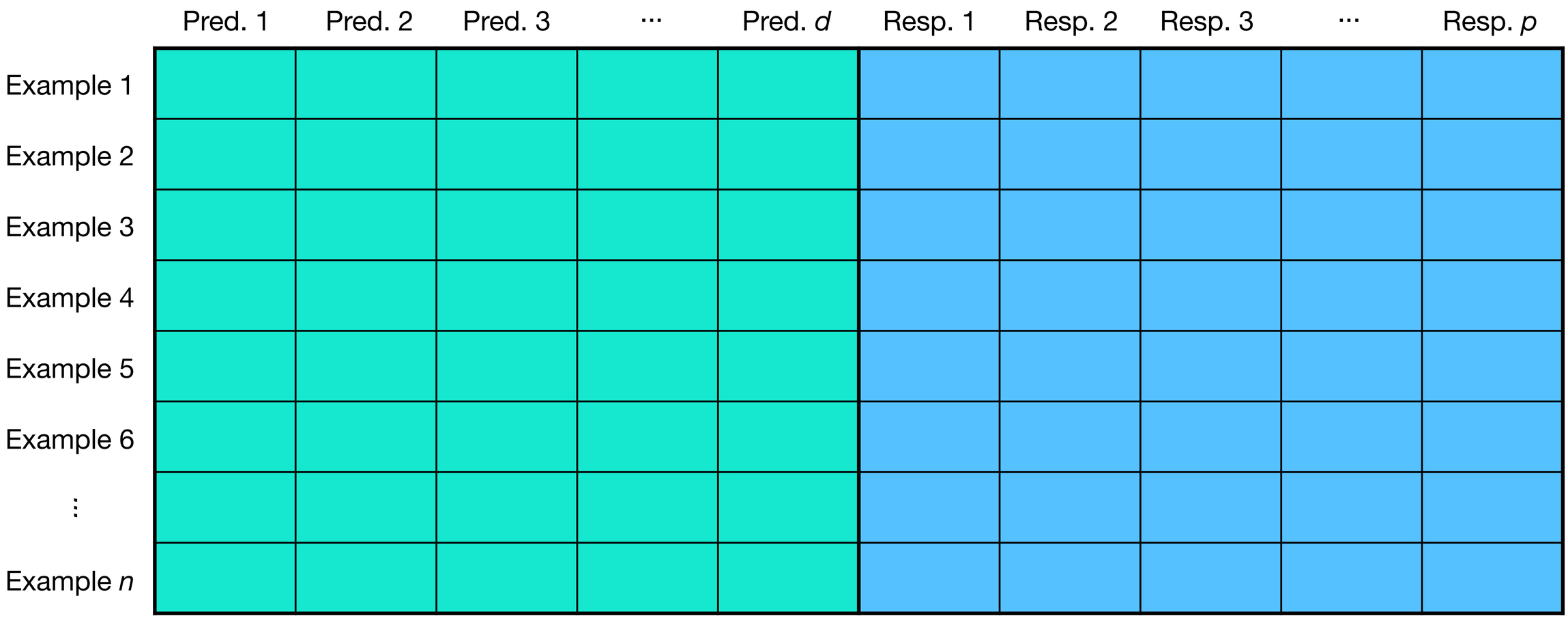
Many predictors and one response (e.g., genetic association study)

High throughput data are data on which thousands (or more) of variables are measured.



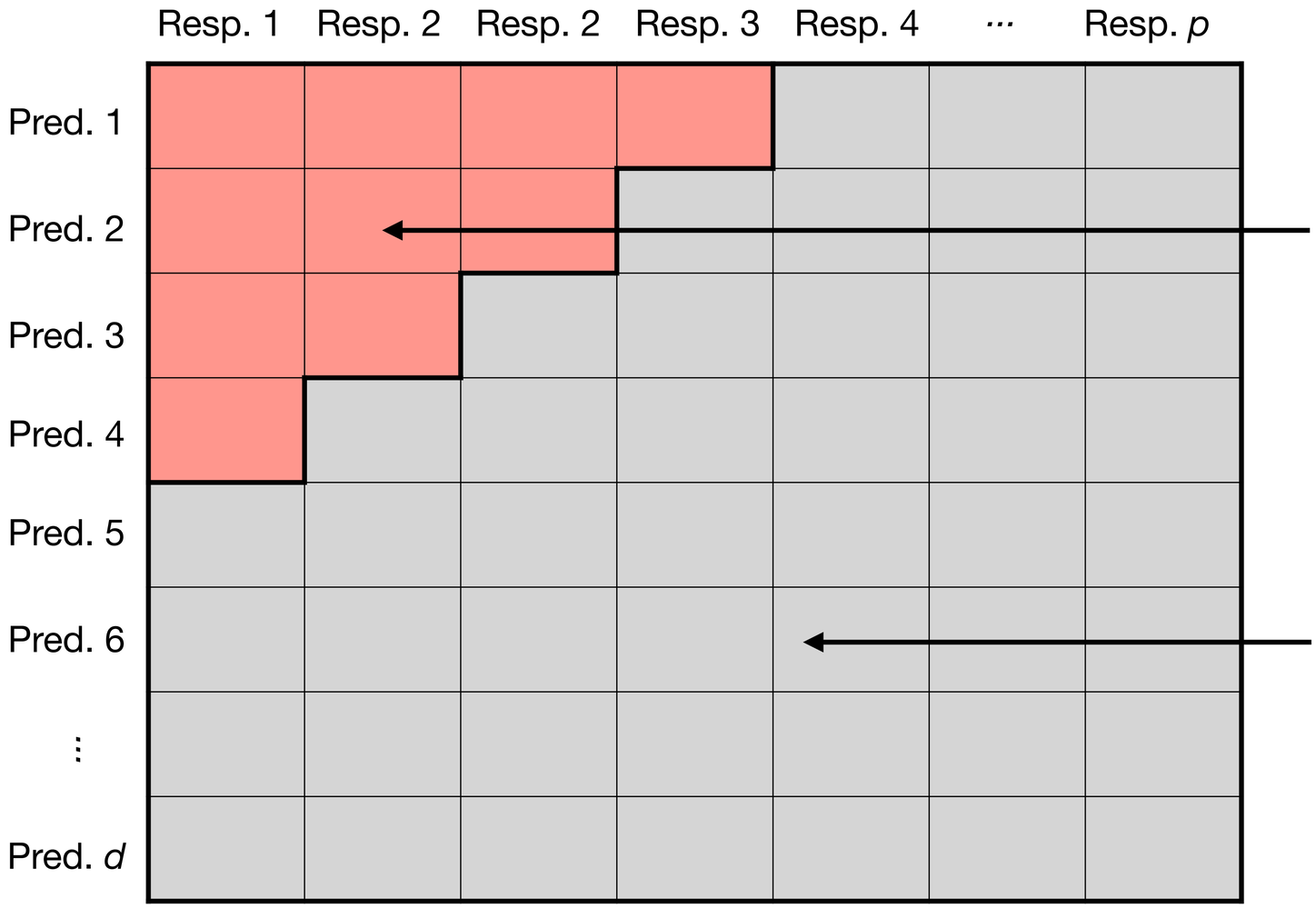
Many responses and one predictor (e.g., differential gene expression study)

"Multi-axis high throughput data" are data on which many predictors *and* many responses are measured.



E.g., single-cell CRISPR screen, genetic variant-gene expression association study, etc.

The goal is to test for association between some pre-selected subset of predictors and responses.



Discovery set

"Leftover set" (typically contains known negative and positive control pairs)

MATHS (multi-axis testing of hypotheses by splitting) is a selective inference framework for boosting the calibration and power of hypothesis testing procedures on "multi-axis high-throughput" data.

- **Idea:** 1) Audition competing hypothesis testing procedures $\{\phi_1, \ldots, \phi_B\}$ on the "leftover" pairs, selecting the procedure ϕ^* that maximizes power while controlling type-I error. 2) Apply ϕ^* to the discovery set.

- **Applications**: (i) calibrating ML-based tests of association, (ii) determining the "correct" set of confounders to control for, (iii) checking parametric model specification, etc.

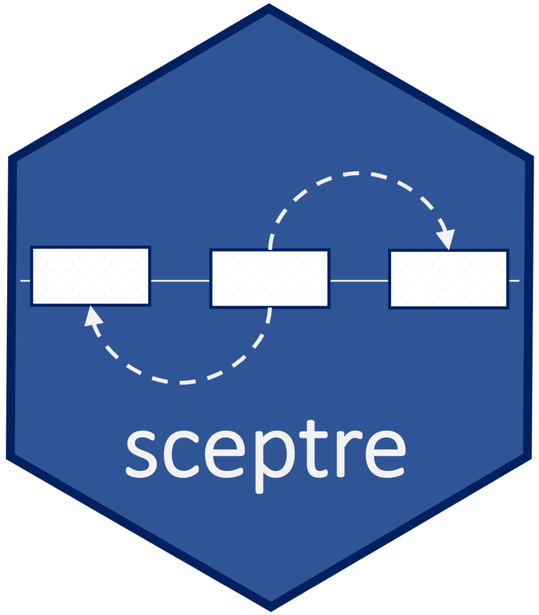
- **Progress**: Preliminary concentration inequality results for type-I error of ϕ^*.

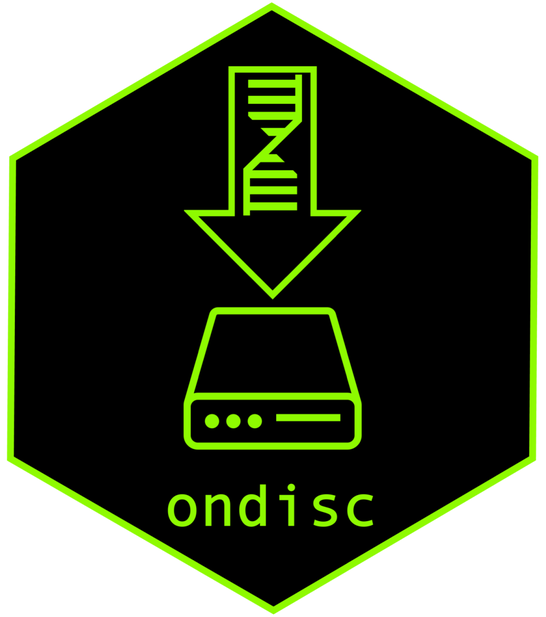
- **To do**: More theoretical results, implement in software, apply to data.

Timeline

- **October 2022**: Preprint SCEPTRE version 2 manuscript.
- **December 2022:** Preprint `ondisc` manuscript.
- **Spring 2023**: Make progress on MATHS project; write thesis.
- **May 2023:** Graduate.

Thank you
Try our software:





github.com/katsevich-lab/sceptre

github.com/timothy-barry/ondisc