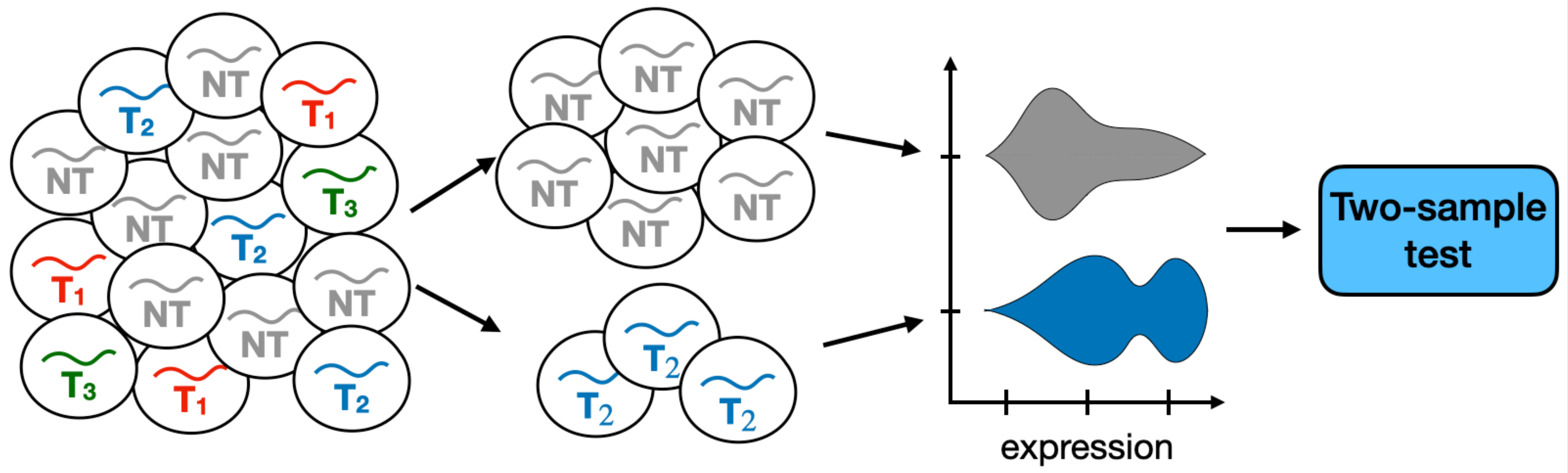
Robust inference by resampling score statistics, with application to single-cell CRISPR screens

Timothy Barry, Kaishu Mason, Kathryn Roeder, and Eugene Katsevich

Software

Single cell CRISPR screens

Single-cell CRISPR screens are an important genomics technology that could give rise to new therapeutics for human diseases.

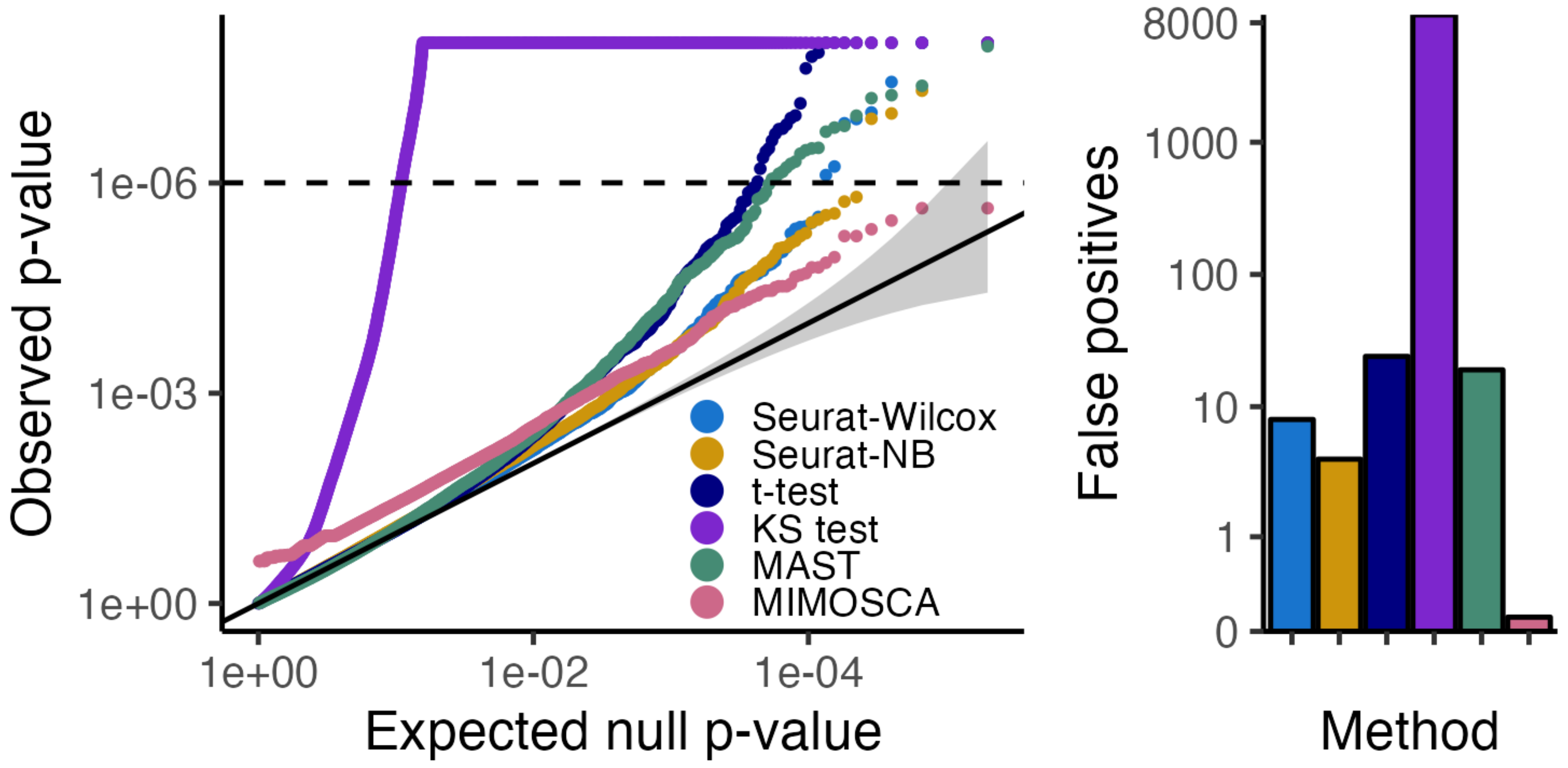


Statistical statement of the problem: We observe data $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$, where $X_i \in \{0,1\}$ is a binary treatment (i.e., the presence or absence of the CRISPR perturbation), $Y_i \in \{0,1,2,\ldots\}$ is the response (i.e., the expression of the gene), and $Z_i \in \mathbb{R}^p$ is a low-dimensional vector of "technical factors" that may or may not exert a confounding effect on X_i and Y_i. Our goal is to produce a **well-calibrated and powerful test of association** between X_i and Y_i.

We aim to apply this test of association to a large number (e.g., 100,000) of CRISPR perturbation-gene pairs, producing a discovery set that controls the false discovery rate.

Contribution 1: Large-scale benchmarking study of existing methods

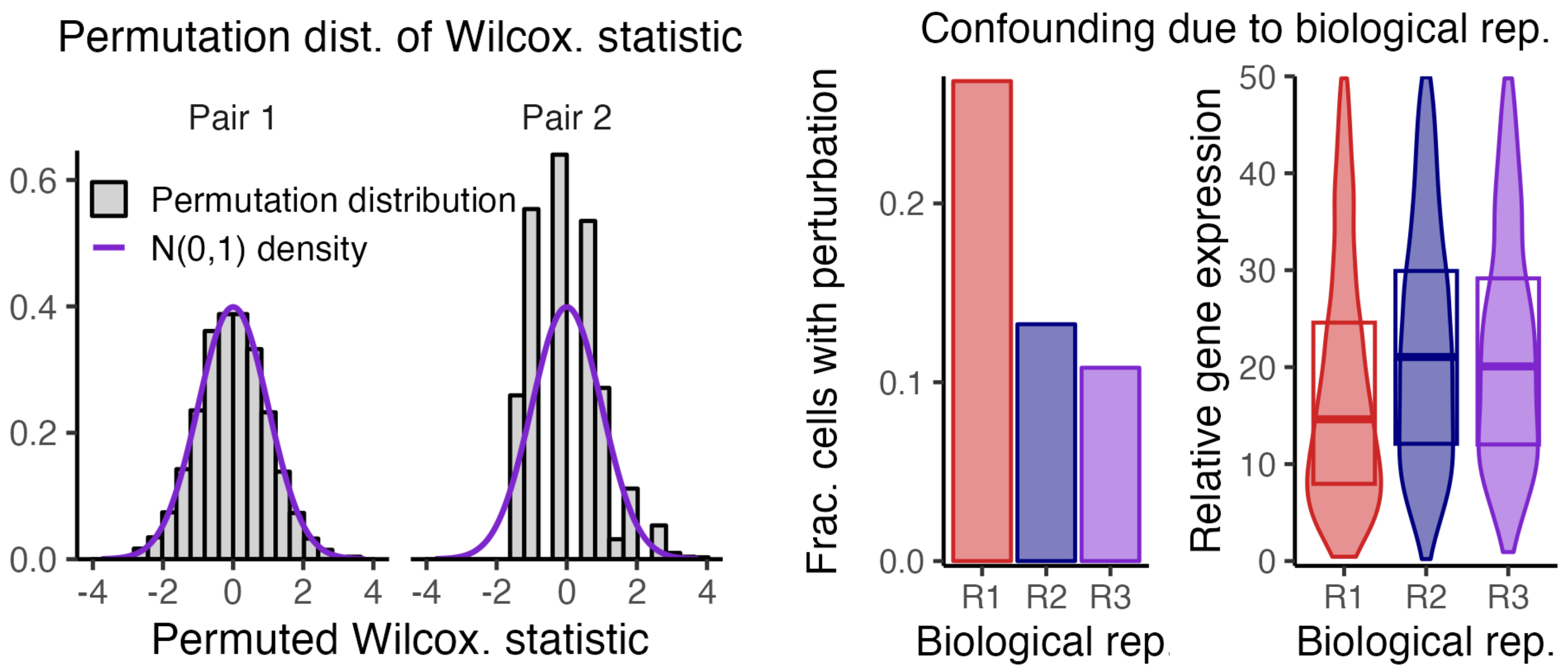
We apply **six leading methods** to analyze **negative control** CRISPR perturbation-gene pairs from **seven datasets**.



Existing methods demonstrate **miscalibration** across all datasets, suggesting that the results produced by these methods may contain **excess false positives**.

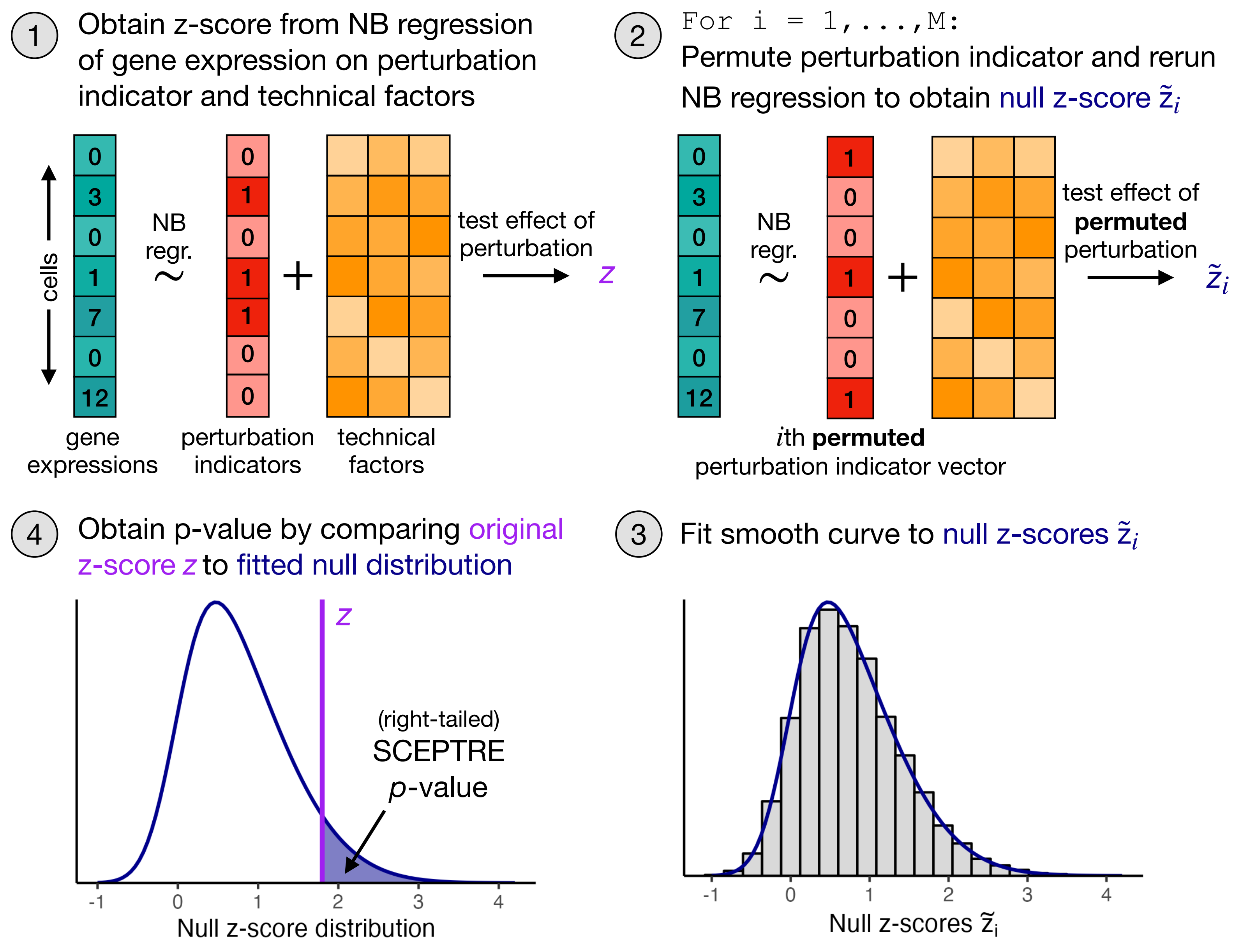
Contribution 2: Identification of core analysis challenges

We conduct an extensive empirical investigation of the data, uncovering three core analysis challenges: **sparsity**, **confounding**, and **model misspecification**. No existing method addresses all three of these challenges.



Contribution 3: A method that resolves the analysis challenges in theory and practice

SCEPTRE is a permutation test that uses a test statistic with appealing **computational** and **statistical** properties.



1. Obtain z-score from NB regression of gene expression on perturbation indicator and technical factors

2. `For i = 1,...,M:`
Permute perturbation indicator and rerun NB regression to obtain null z-score $\tilde{z}_i$

3. Fit smooth curve to null z-scores $\tilde{z}_i$

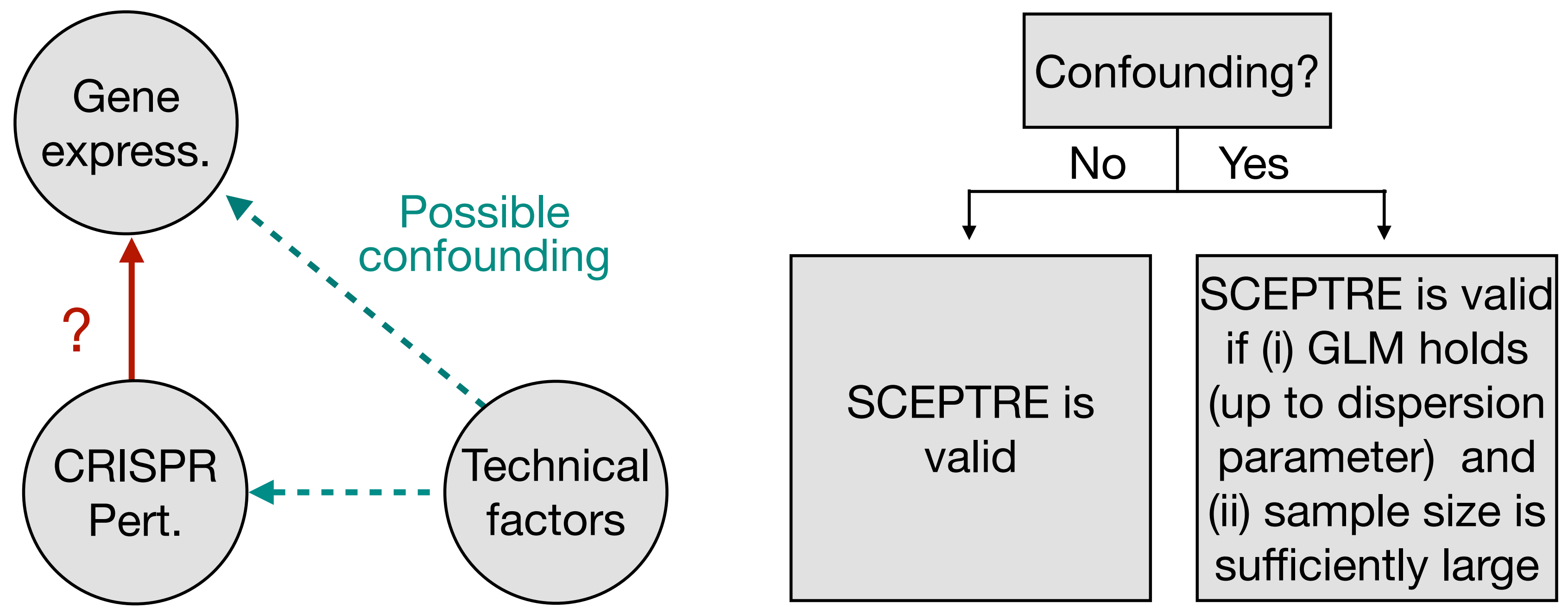
4. Obtain p-value by comparing original z-score z to fitted null distribution

cells
0
3
0
1
7
0
12
NB regr.
~
0
1
0
1
1
0
0
+
test effect of perturbation
z
gene expressions
perturbation indicators
technical factors
1
0
0
1
0
0
1
test effect of permuted perturbation
$\tilde{z}_i$
ith permuted perturbation indicator vector
z
(right-tailed) SCEPTRE p-value
-1
0
1
2
3
4
Null z-score distribution
Null z-scores $\tilde{z}_i$

SCEPTRE is nearly as fast as fitting a GLM due to several accelerations, including:

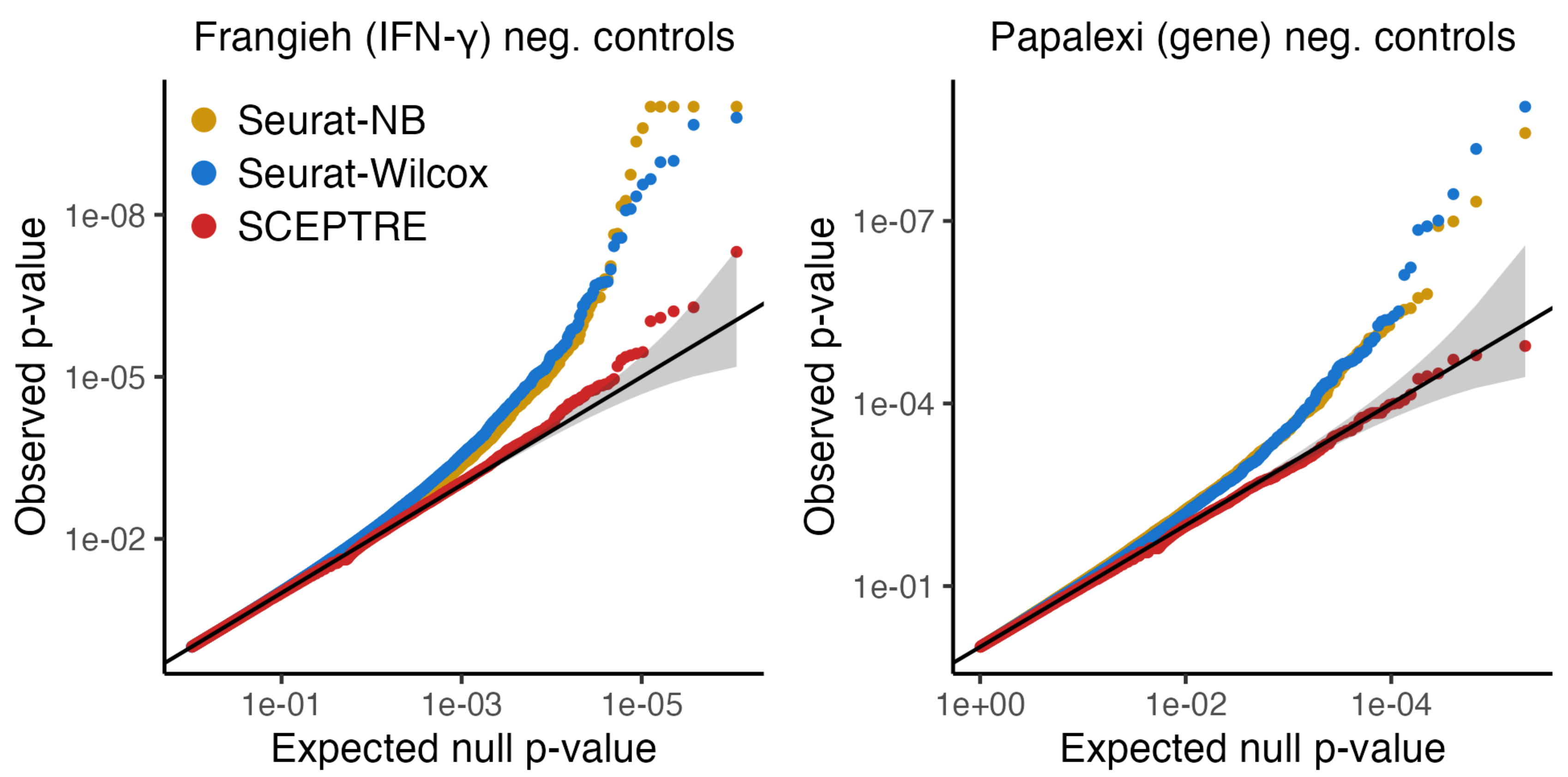
1. Use of a **score** test (rather than a **Wald** or **likelihood ratio** test) to compute the test statistics.
2. A new **algorithm** for computing **GLM score tests** (100x faster than classical algorithm for binary treatments).

Theoretically, SCEPTRE is **robust** to the calibration threats of **sparsity**, **confounding**, and **model misspecification**.


Gene express.
?
CRISPR Pert.
Possible confounding
Technical factors
Confounding?
No
Yes
SCEPTRE is valid
SCEPTRE is valid if (i) GLM holds (up to dispersion parameter) and (ii) sample size is sufficiently large

Application of SCEPTRE to real control data

SCEPTRE exhibits **better calibration** (on negative control data) and **power** (on positive control data) than competing methods.



b

Number of false positives

Dataset	SCEPTRE	Seurat-Wilcox	Seurat-NB	t-test	MAST	KS test	MIMOSCA	NT pairs
Frangieh (Co Culture)	1	13	10	89	2083	**0**	4	596344
Frangieh (Control)	**0**	7	16	69	1873	**0**	**0**	528239
Frangieh (IFN-γ)	1	15	15	67	1933	**0**	5	565502
Papalexi (Gene)	**0**	8	4	24	19	9191	**0**	100458
Papalexi (Protein)	**0**	2	**0**	1	2	**0**	**0**	36
Schraivogel	3	2	3	4	**1**	**1**	19	4357
Simulated	**0**	**0**	**0**	7	16	**0**	1	96944
Average	**0.7**	6.7	6.9	37.3	846.7	1313.1	4.1	

c

Number of true positives

Dataset	SCEPTRE	Seurat-Wilcox	Seurat-NB	t-test	MAST	KS test	MIMOSCA	PC pairs
Frangieh (Co Culture)	**103**	98	94	-	-	90	5	181
Frangieh (Control)	**77**	74	72	-	-	70	4	170
Frangieh (IFN-γ)	**94**	89	81	-	-	81	8	181
Papalexi (Gene)	**13**	12	**13**	11	11	-	0	25
Papalexi (Protein)	**2**	**2**	**2**	**2**	**2**	**2**	**2**	2
Schraivogel	22	22	21	**23**	22	19	0	25