Robust negative binomial regression by permuting score statistics

Ziang Niu

Timothy Barry

Ziang Niu

Eugene Katsevich

Xihong Lin

Differential expression aims to assess whether a gene exhibits variable activity across conditions.

Differential expression aims to assess whether a gene exhibits variable activity across conditions.

Differential expression is crucial in many genomics applications.

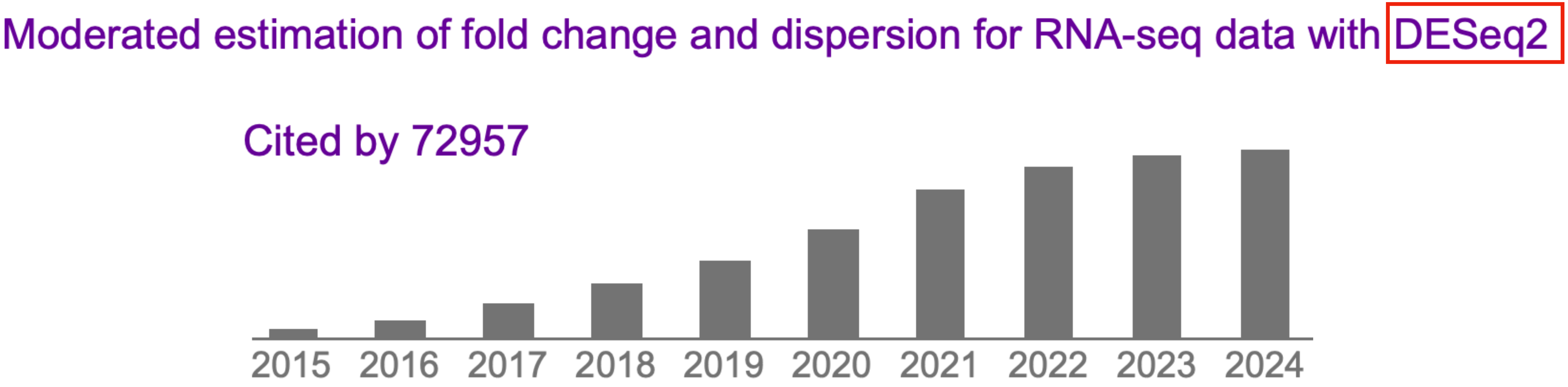
- Single-cell RNA-seq
- Bulk RNA-seq
- ChIP-seq
- Spatial transcriptomics
- CRISPR screens
- and lots more…

Negative binomial (NB) regression is the most popular method for differential expression analysis.

Popular implementations of NB regression include **DESeq2** and **MASS**, which are used by thousands of studies every year.

Negative binomial (NB) regression is the most popular method for differential expression analysis.

Popular implementations of NB regression include **DESeq2** and **MASS**, which are used by thousands of studies every year.



Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Cited by 72957

Despite its success, NB regression is a fragile method.

NB regression makes strong **parametric** and **asymptotic** assumptions.

These assumptions can break down in practice, leading to excess **false positive** and **false negative** results.

See Li et al. 2022 (*Genome Biology*)

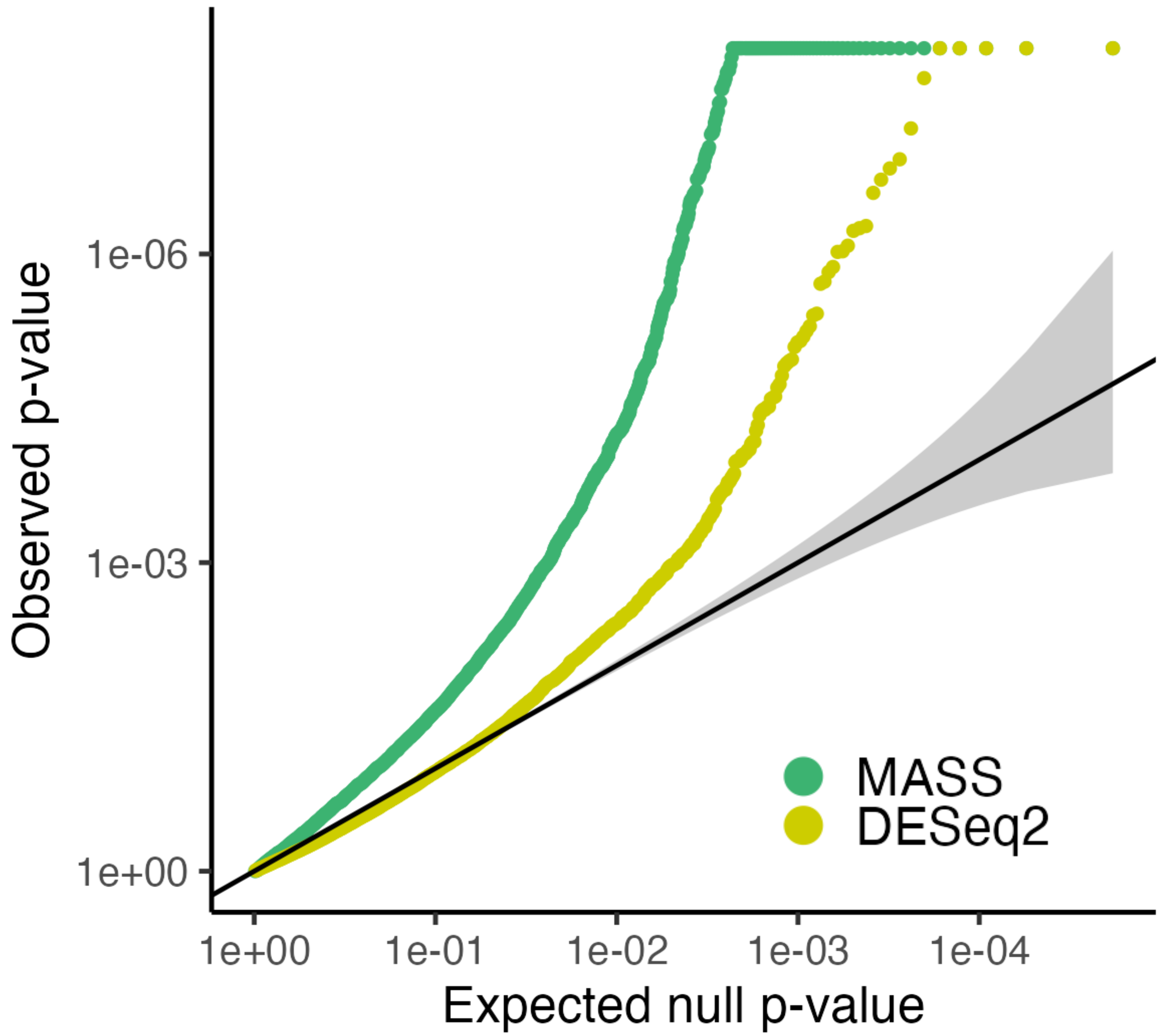
Despite its success, NB regression is a fragile method.

NB regression makes strong **parametric** and **asymptotic** assumptions.

These assumptions can break down in practice, leading to excess **false positive** and **false negative** results.

See Li et al. 2022 (*Genome Biology*)

Negative control p-values

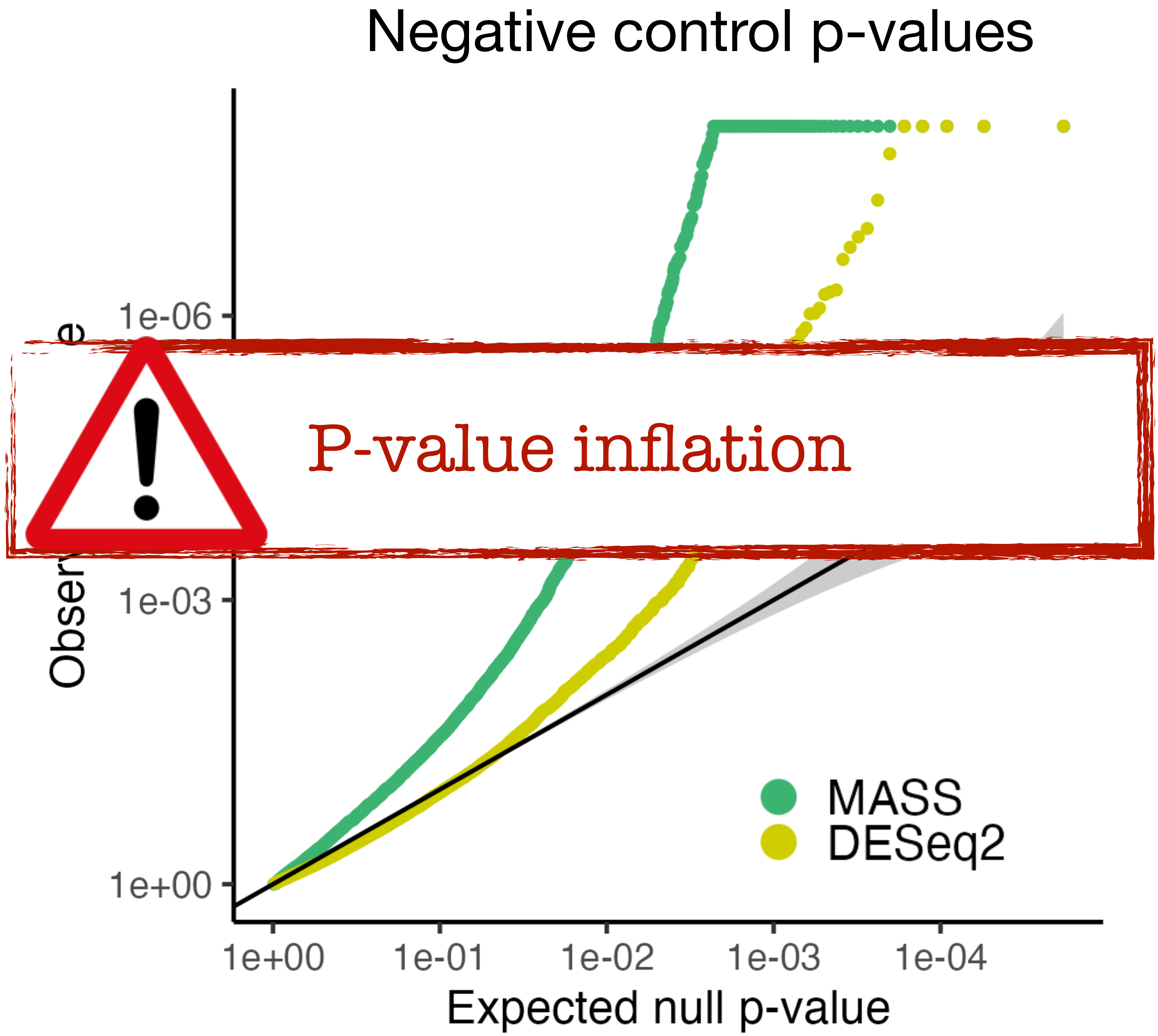


Despite its success, NB regression is a fragile method.

NB regression makes strong **parametric** and **asymptotic** assumptions.

These assumptions can break down in practice, leading to excess **false positive** and **false negative** results.

See Li et al. 2022 (*Genome Biology*)



Negative control p-values

P-value inflation

Thesis: we substantially can improve the robustness of NB regression, thereby enhancing the reliability of differential expression analysis.

We propose a framework for robust NB regression based on permuting score statistics.

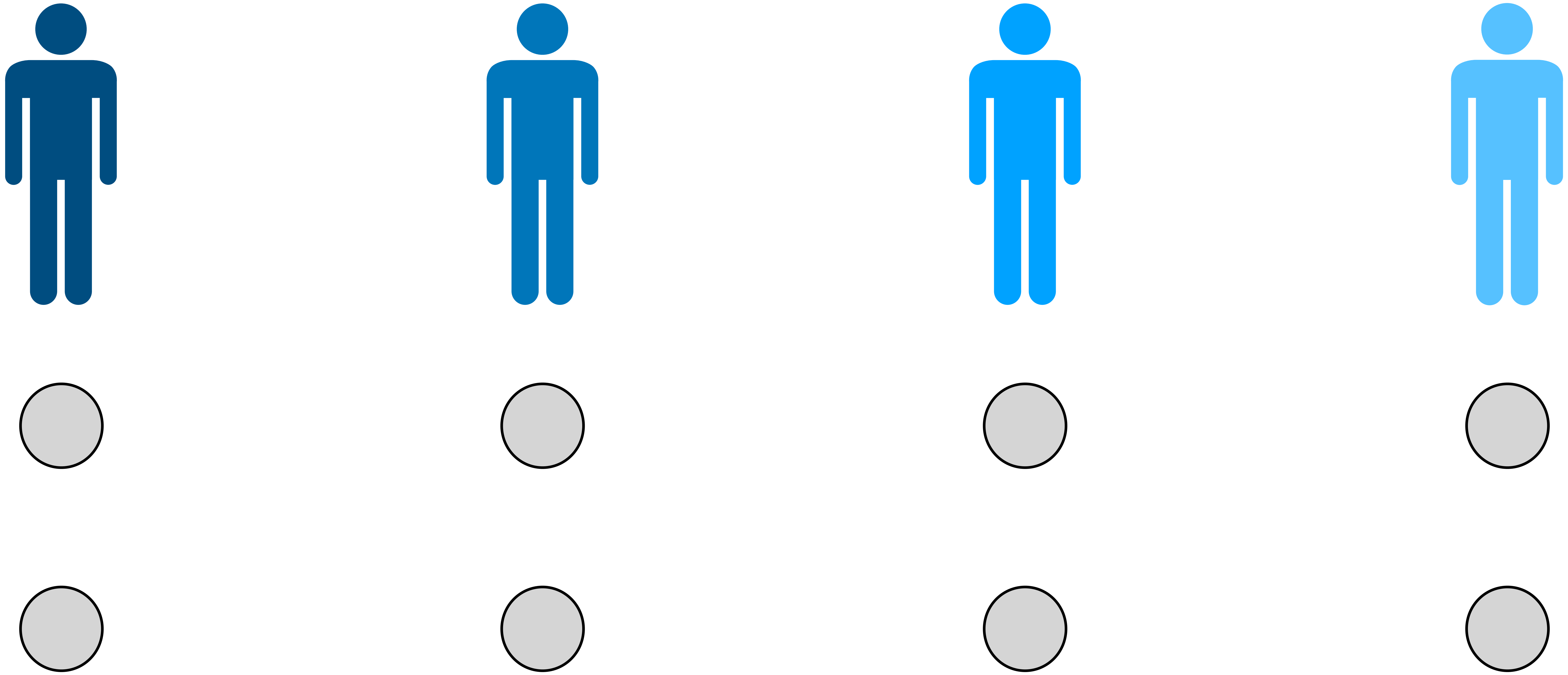
Thesis: we substantially can improve the robustness of NB regression, thereby enhancing the reliability of differential expression analysis.

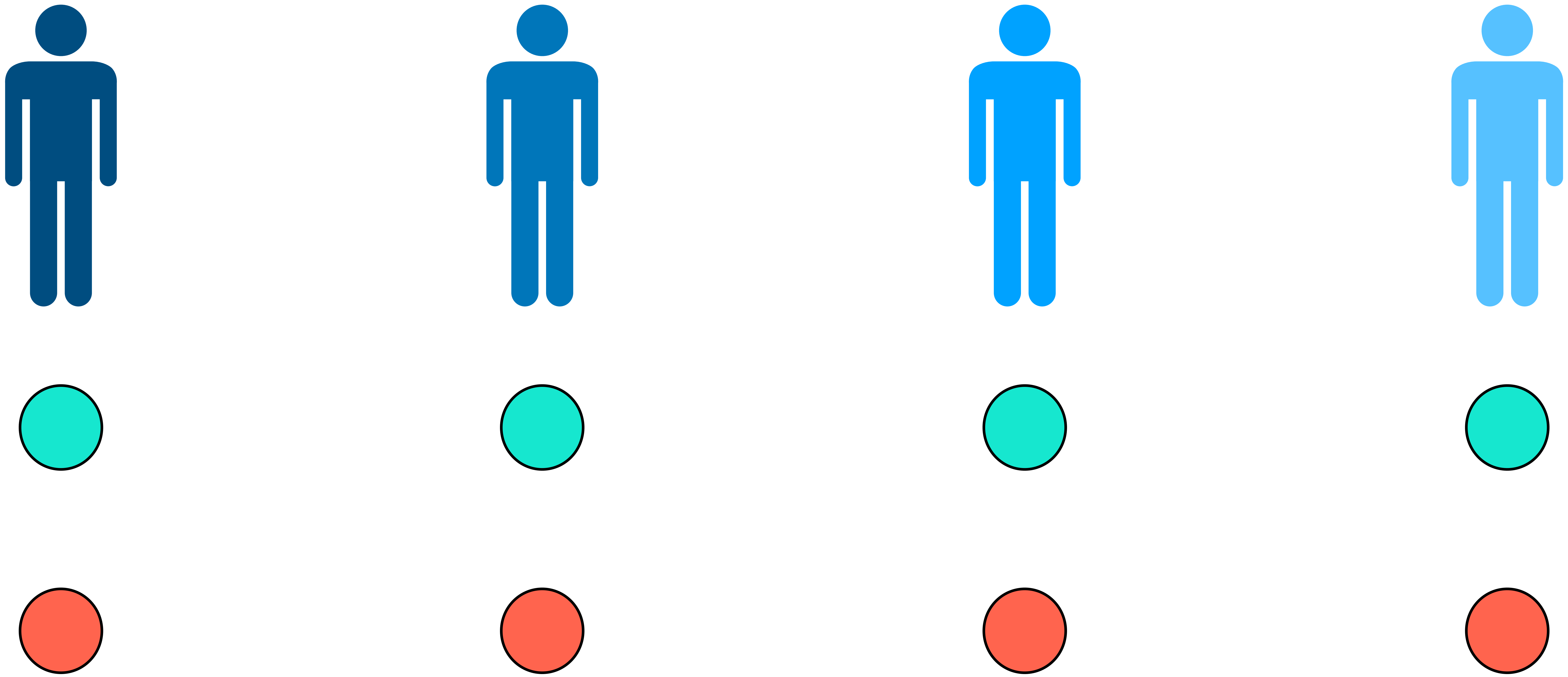
We propose a framework for robust NB regression based on permuting score statistics.

If NB regression "works"…	If NB regression fails…
Our method **matches** NB regression with respect to type-I error control, power (approximately), and compute.	Our method **outperforms** NB regression with respect to type-I error control and/or power.

Roadmap

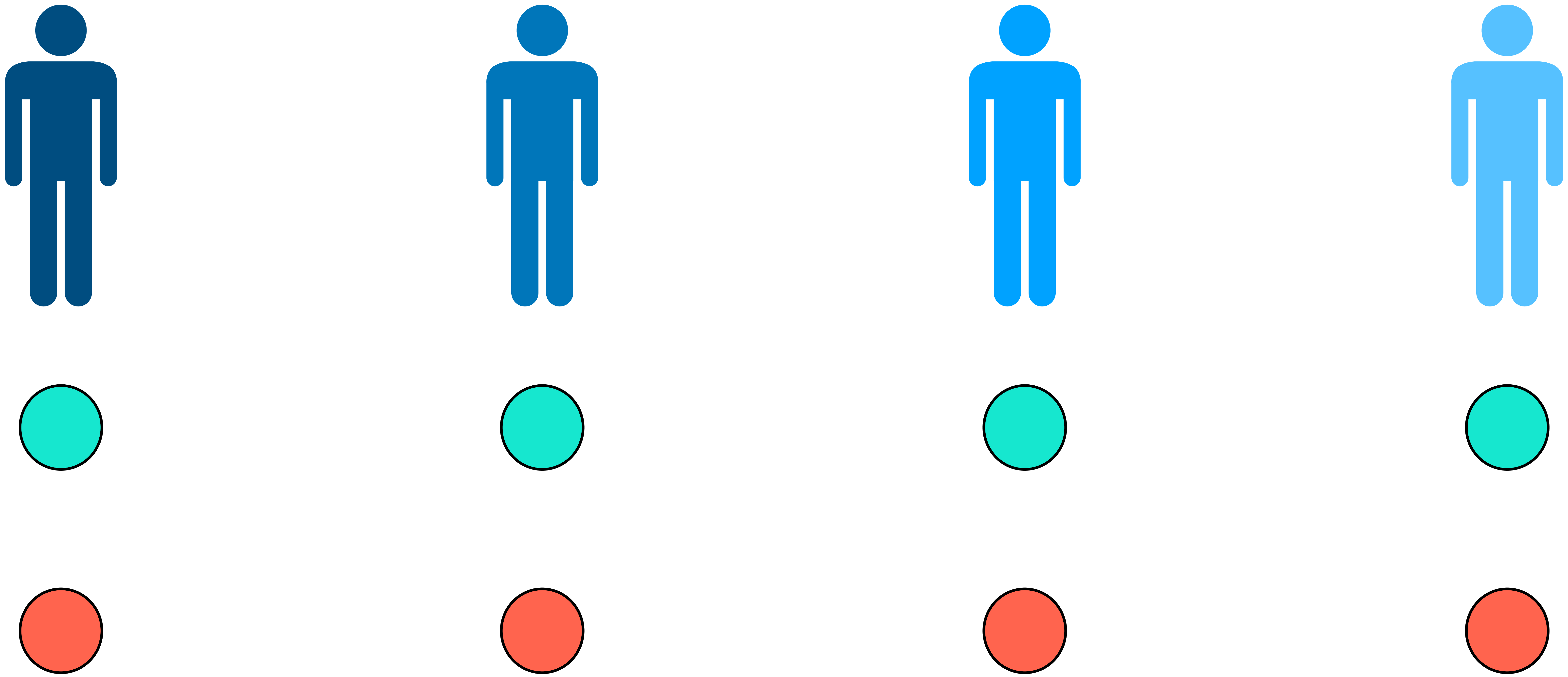
1. **Review of NB regression**
2. Permuting score statistics
3. Statistical guarantees
4. Simulations
5. Real data analysis





CRISPR perturbation

Control perturbation



CRISPR perturbation

Control perturbation

Bulk RNA sequencing

Bulk RNA sequencing

0
0
0
0
1
1
1
1

Treatment
vector

	Gene 1	Gene 2		Gene *m*
0	5	40		91
0	38	5		77
0	92	3		11
0	35	8	…	33
1	21	12		20
1	36	43		24
1	76	54		59
1	51	5		15

Treatment vector

Gene expression vectors

0
0
0
0
1
1
1
1

Treatment vector

Gene 1	Gene 2		Gene *m*
5	40		91
38	5		77
92	3		11
35	8	…	33
21	12		20
36	43		24
76	54		59
51	5		15

Gene expression vectors

Donor	Sex
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M

Covariate matrix

Statistical rendering of differential expression

Statistical rendering of differential expression

- Consider a given gene.

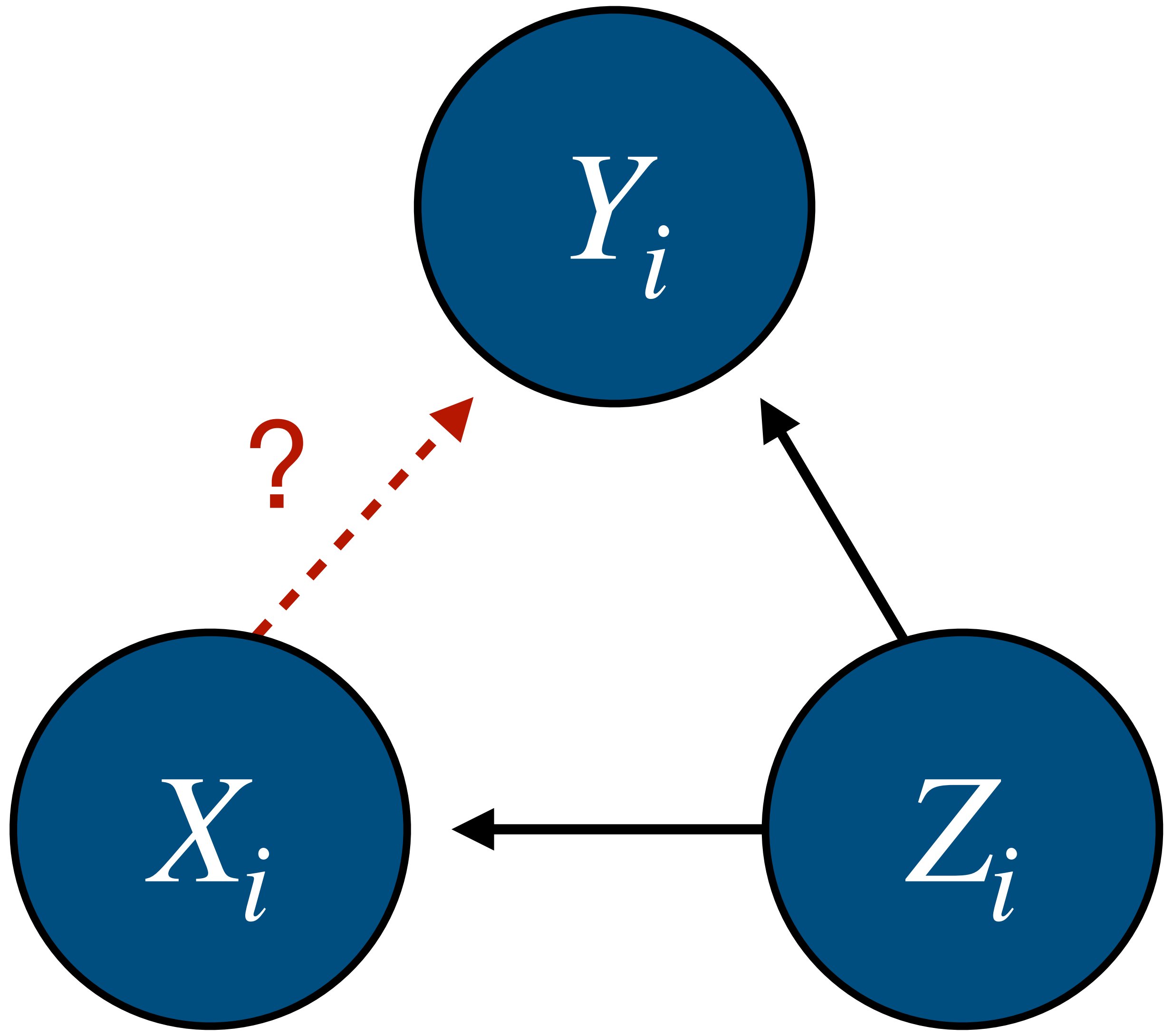
Statistical rendering of differential expression

- Consider a given gene.
- We observe *i.i.d.* tuples $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$, where
    - $Y_i \in \{0,1,2,\ldots\}$ is the gene expression
    - $X_i \in \{0,1\}$ is the treatment indicator
    - $Z_i \in \mathbb{R}^p$ is a vector of nuisance covariates.

Statistical rendering of differential expression

Statistical rendering of differential expression

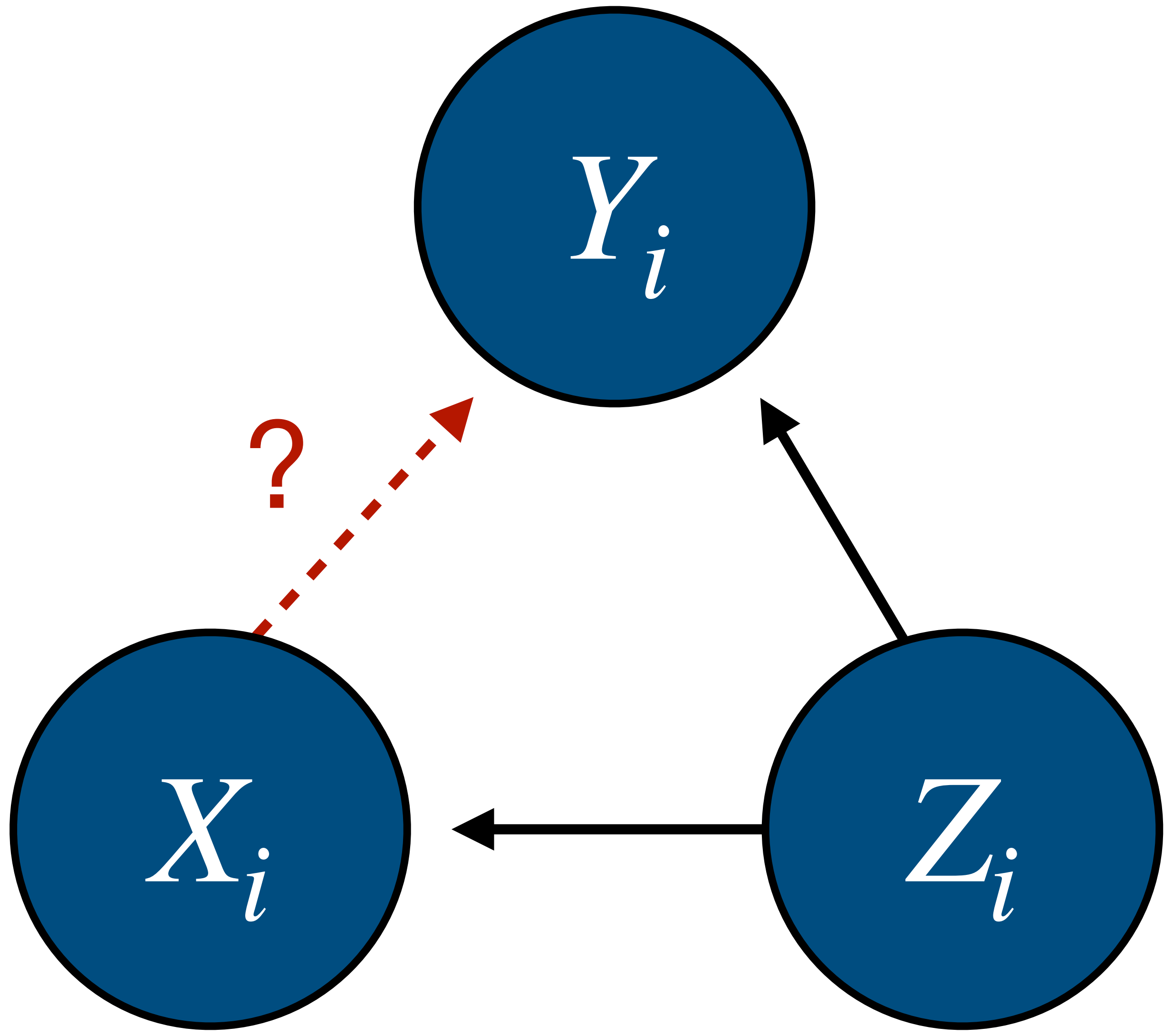
- We seek to test the *conditional independence* null hypothesis: $H_0 : X_i \perp Y_i \mid Z_i$



Statistical rendering of differential expression

- We seek to test the *conditional independence* null hypothesis: $H_0 : X_i \perp Y_i | Z_i$

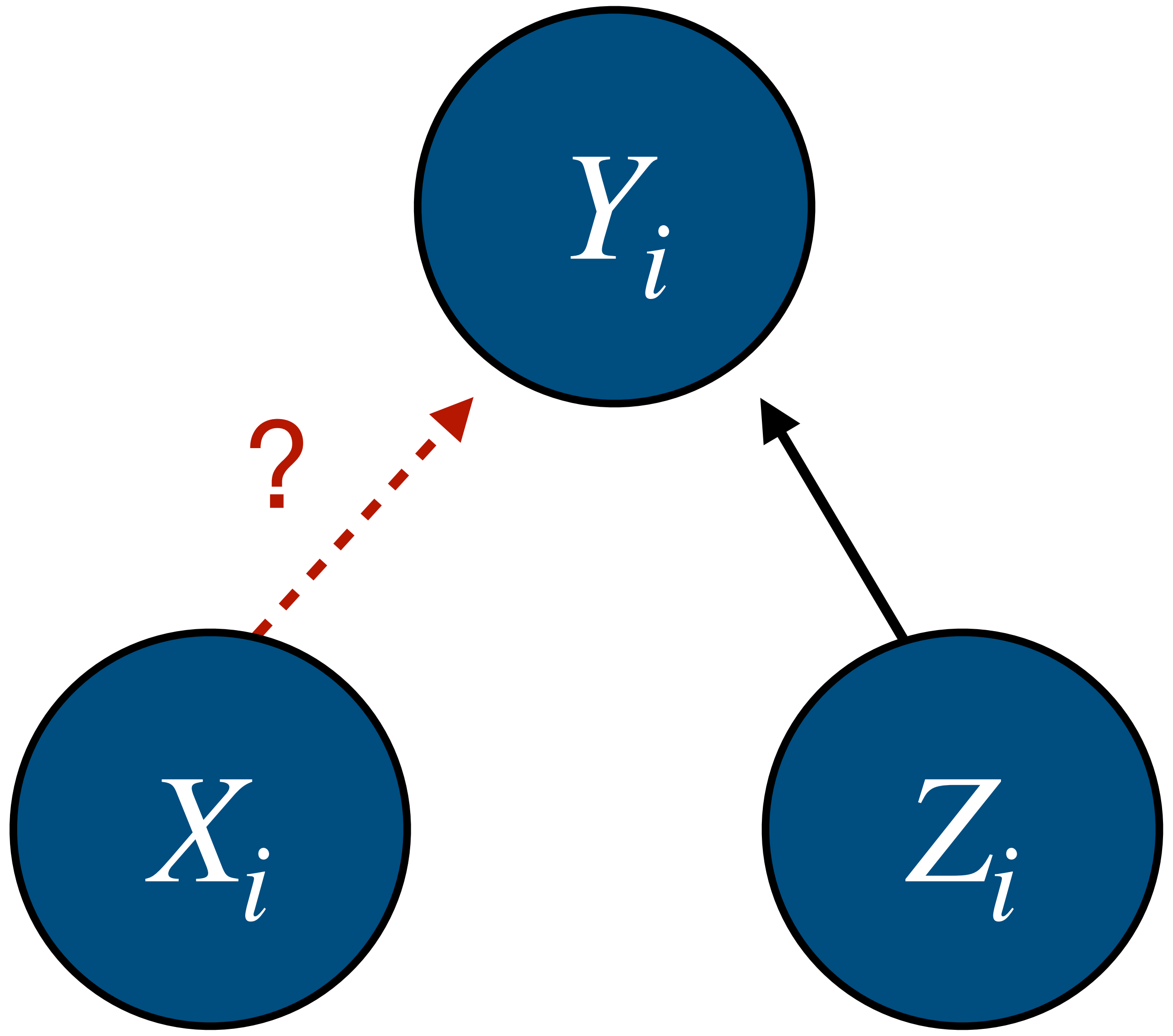
- Intuition: X_i (treatment) provides no information about Y_i (gene expression) above and beyond Z_i (nuisance covariates).



Statistical rendering of differential expression

Statistical rendering of differential expression

- In a randomized experiment, the treatment and nuisance covariates are independent: $X_i \perp Z_i$

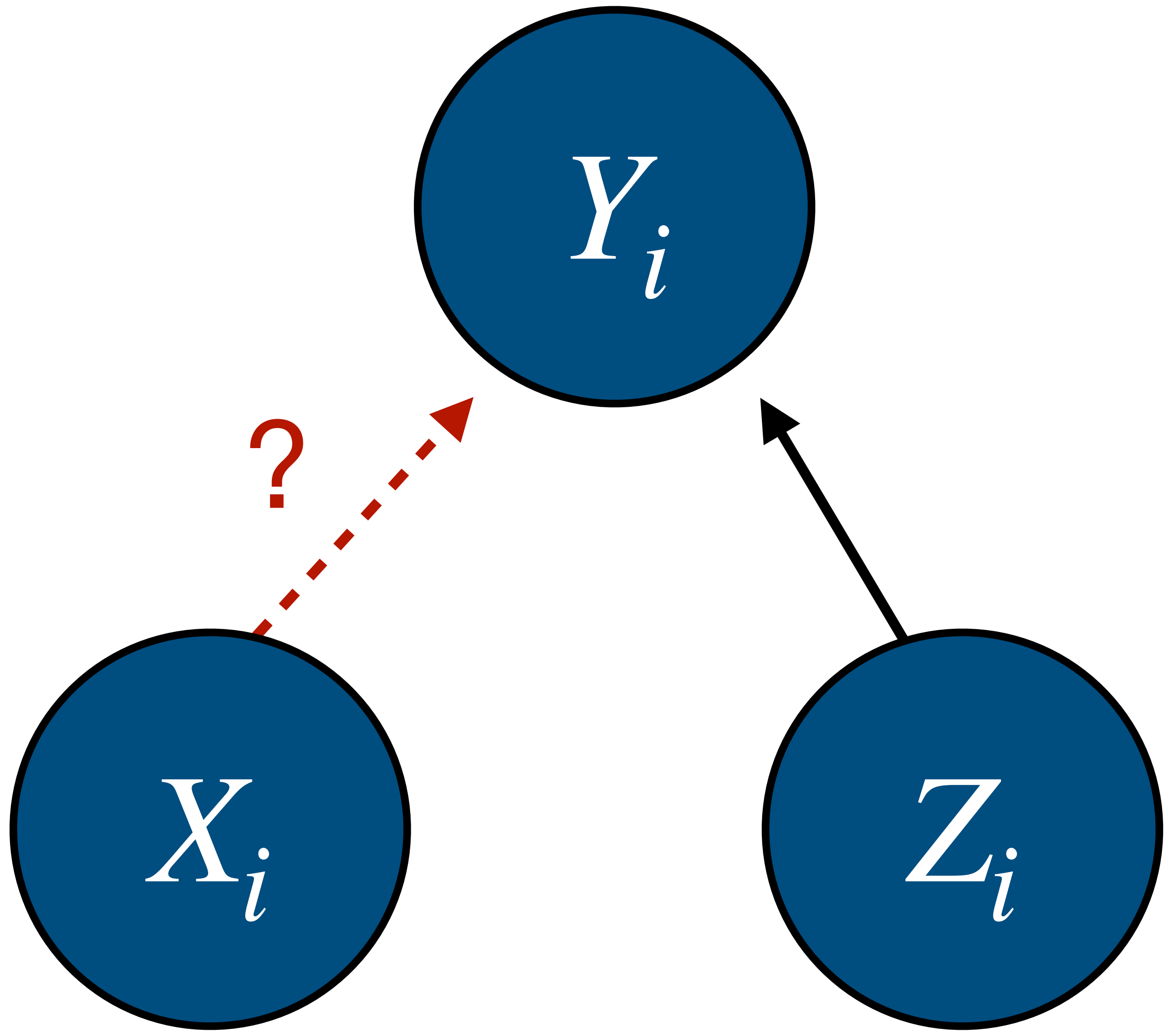


Randomized experiment

Statistical rendering of differential expression

- In a randomized experiment, the treatment and nuisance covariates are independent: $X_i \perp Z_i$

- Conditional independence and marginal independence are equivalent: $X_i \perp Y_i \iff X_i \perp Y_i \,|\, Z_i$



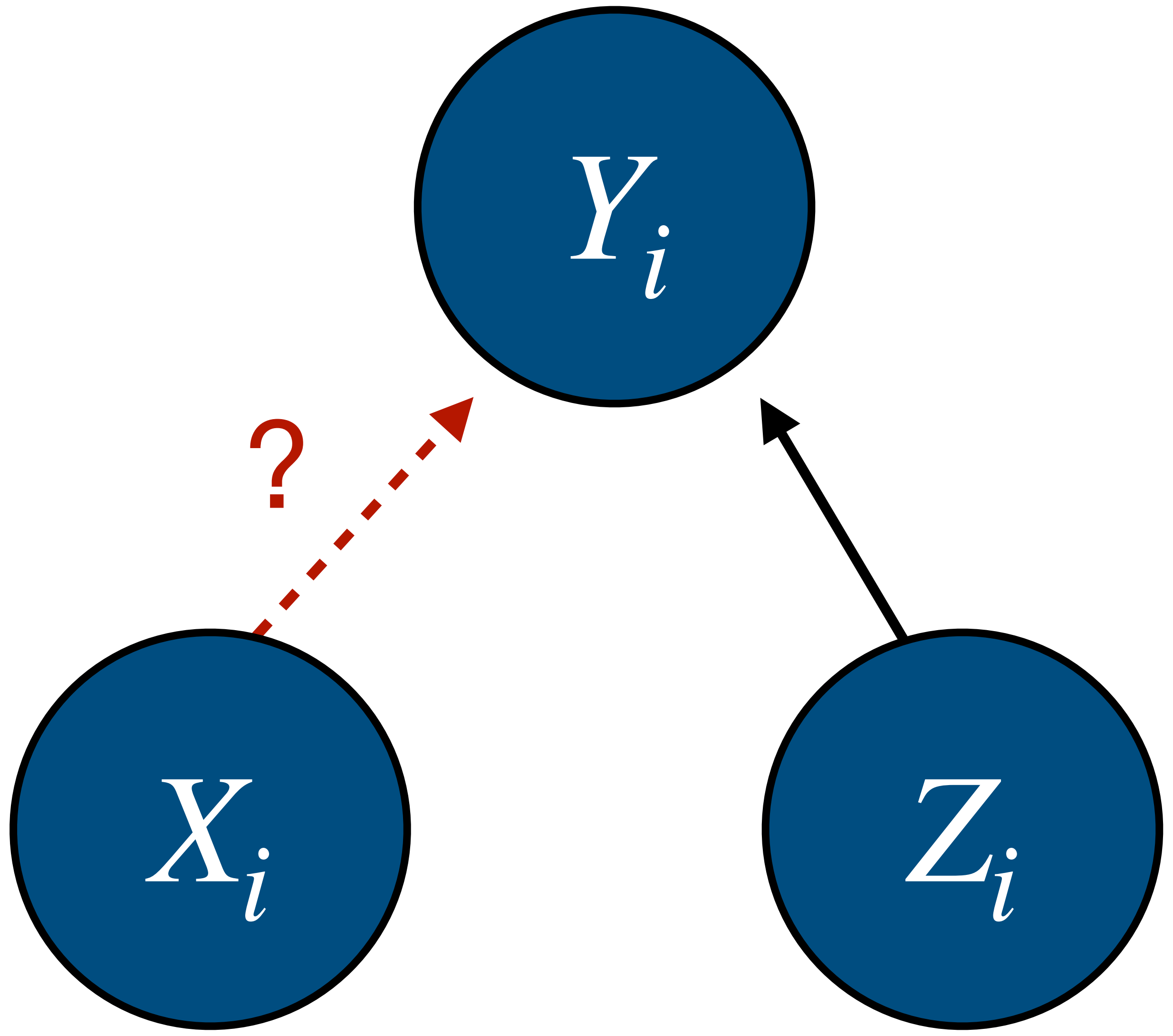
Randomized experiment

Statistical rendering of differential expression

- In a randomized experiment, the treatment and nuisance covariates are independent: $X_i \perp Z_i$

- Conditional independence and marginal independence are equivalent: $X_i \perp Y_i \iff X_i \perp Y_i \mid Z_i$

- We are interested in both observational studies and randomized experiments.



Randomized experiment

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

γ: treatment coefficient ($X_i \perp Y_i \mid Z_i \iff \gamma = 0$)

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

γ: treatment coefficient ($X_i \perp Y_i \,|\, Z_i \iff \gamma = 0$)

NB: negative binomial distribution

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \mathrm{NB}_{\phi}(\mu_i) \end{cases}$$

γ: treatment coefficient ($X_i \perp Y_i \mid Z_i \iff \gamma = 0$)

NB: negative binomial distribution

ϕ: dispersion parameter (setting $\phi = 0$ yields a Poisson distribution)

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

γ: treatment coefficient ($X_i \perp Y_i \,|\, Z_i \iff \gamma = 0$)

NB: negative binomial distribution

ϕ: dispersion parameter (setting $\phi = 0$ yields a Poisson distribution)

The dispersion parameter ϕ typically is estimated from data.

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

For each gene, fit an NB GLM and test $\gamma = 0$ via a Wald test.

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

For each gene, fit an NB GLM and test $\gamma = 0$ via a Wald test.

Input the resulting p-values $p_1, \ldots, p_m$ to the BH procedure.

Negative binomial (NB) regression is the standard approach to testing differential expression.

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \end{cases}$$

For each gene, fit an NB GLM and test $\gamma = 0$ via a Wald test.

Input the resulting p-values $p_1, \ldots, p_m$ to the BH procedure.

Obtain a discovery set that controls the false discovery rate (FDR).

NB regression makes strong **parametric** and **asymptotic** assumptions, which can fail to hold in practice.

NB regression makes strong **parametric** and **asymptotic** assumptions, which can fail to hold in practice.

	Parametric assumptions
What is assumed?	NB model correctly specified
When are these assumptions violated?	Missing interaction term, wrong link function, presence of zero inflation, etc.

NB regression makes strong **parametric** and **asymptotic** assumptions, which can fail to hold in practice.

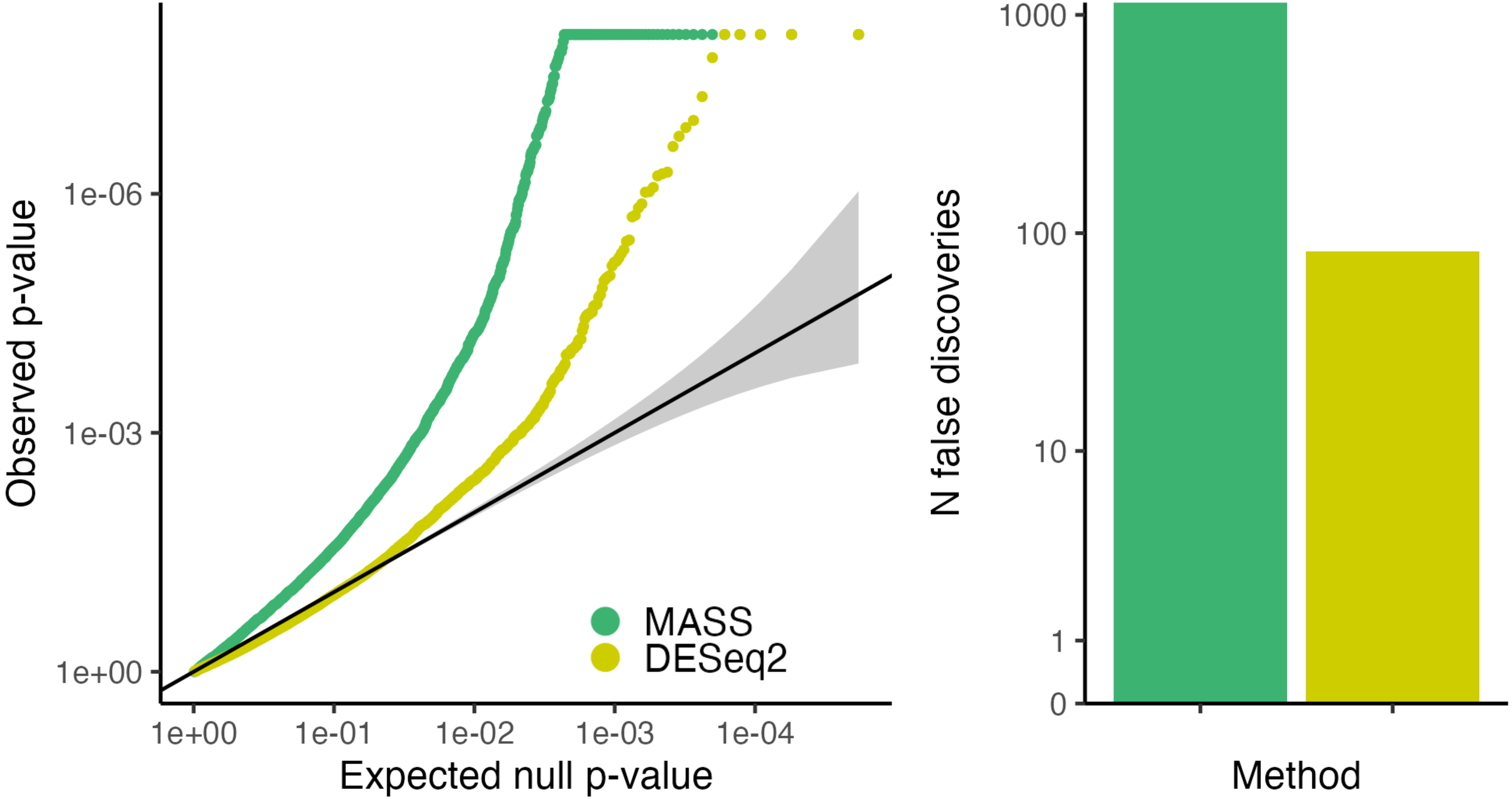
	Parametric assumptions	Asymptotic assumptions
What is assumed?	NB model correctly specified	Amount of “information” contained within sample sufficiently large
When are these assumptions violated?	Missing interaction term, wrong link function, presence of zero inflation, etc.	Small sample size, small counts

NB regression makes strong **parametric** and **asymptotic** assumptions, which can fail to hold in practice.

	Parametric assumptions	Asymptotic assumptions
What is assumed?	NB model correctly specified	Amount of “information” contained within sample sufficiently large
When are these assumptions violated?	Missing interaction term, wrong link function, presence of zero inflation, etc.	Small sample size, small counts

*Asymptotic breakdown can have two consequences: (1) lead to a poor estimate $\hat{\phi}$ of the dispersion parameter ϕ; (2) poor normal approximation.

NB regression (implemented as **MASS** and **DESeq2**) can fail to control type-I error on real data.



Roadmap

1. Review of NB regression
2. **Permuting score statistics**
3. Statistical guarantees
4. Simulations
5. Real data analysis

Consider i.i.d. data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ generated from the NB GLM:

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_\phi(\mu_i)\,. \end{cases}$$

Consider i.i.d. data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ generated from the NB GLM:

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_\phi(\mu_i)\,. \end{cases}$$

Let $\bar{\phi}$ be the dispersion parameter that we plug into the model.

Consider i.i.d. data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ generated from the NB GLM:

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_\phi(\mu_i)\,. \end{cases}$$

Let $\bar{\phi}$ be the dispersion parameter that we plug into the model.

Let $\hat{\beta}_n$ be the estimate for β that results from regressing Y onto Z, i.e. $\hat{\beta}_n$ solves the score equation

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_i(Y_i - \mu_i)}{1 + \bar{\phi}\mu_i} = 0,\ \mu_i \equiv \exp(\beta^\top Z_i)\,.$$

Consider i.i.d. data $\{(X_i, Y_i, Z_i)\}_{i=1}^{n}$ generated from the NB GLM:

$$\begin{cases} \log(\mu_i) = \gamma X_i + \beta^T Z_i \\ Y_i \sim \text{NB}_{\phi}(\mu_i) \, . \end{cases}$$

Let $\bar{\phi}$ be the dispersion parameter that we plug into the model.

Let $\hat{\beta}_n$ be the estimate for β that results from regressing Y onto Z, i.e. $\hat{\beta}_n$ solves the score equation

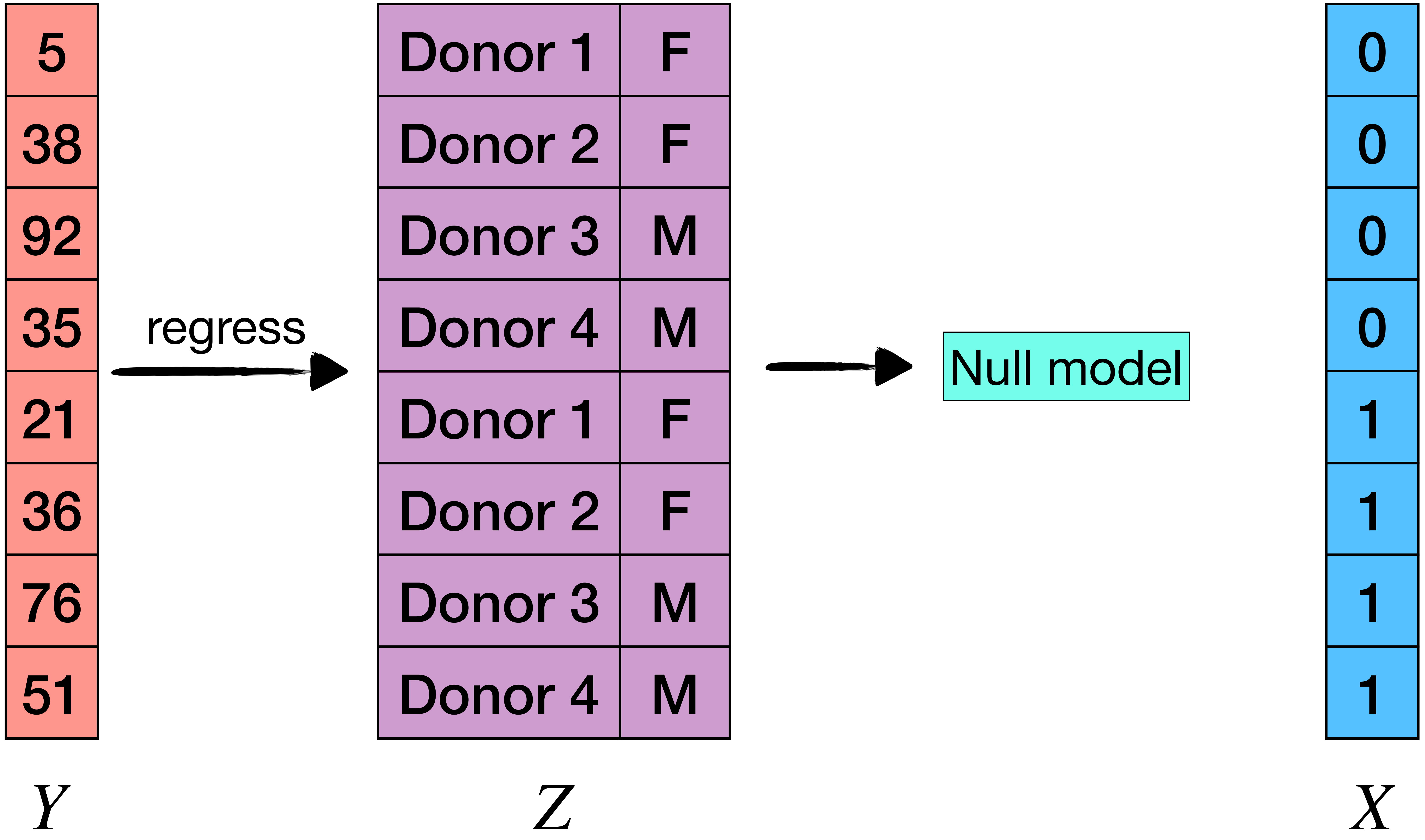
$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i(Y_i - \mu_i)}{1+\bar{\phi}\mu_i} = 0, \; \mu_i \equiv \exp(\beta^\top Z_i) \, .$$

Let $T_n(X, Y, Z)$ be the score test statistic for testing the null hypothesis $\gamma = 0$.

We propose a permutation test based on an NB GLM score test statistic.

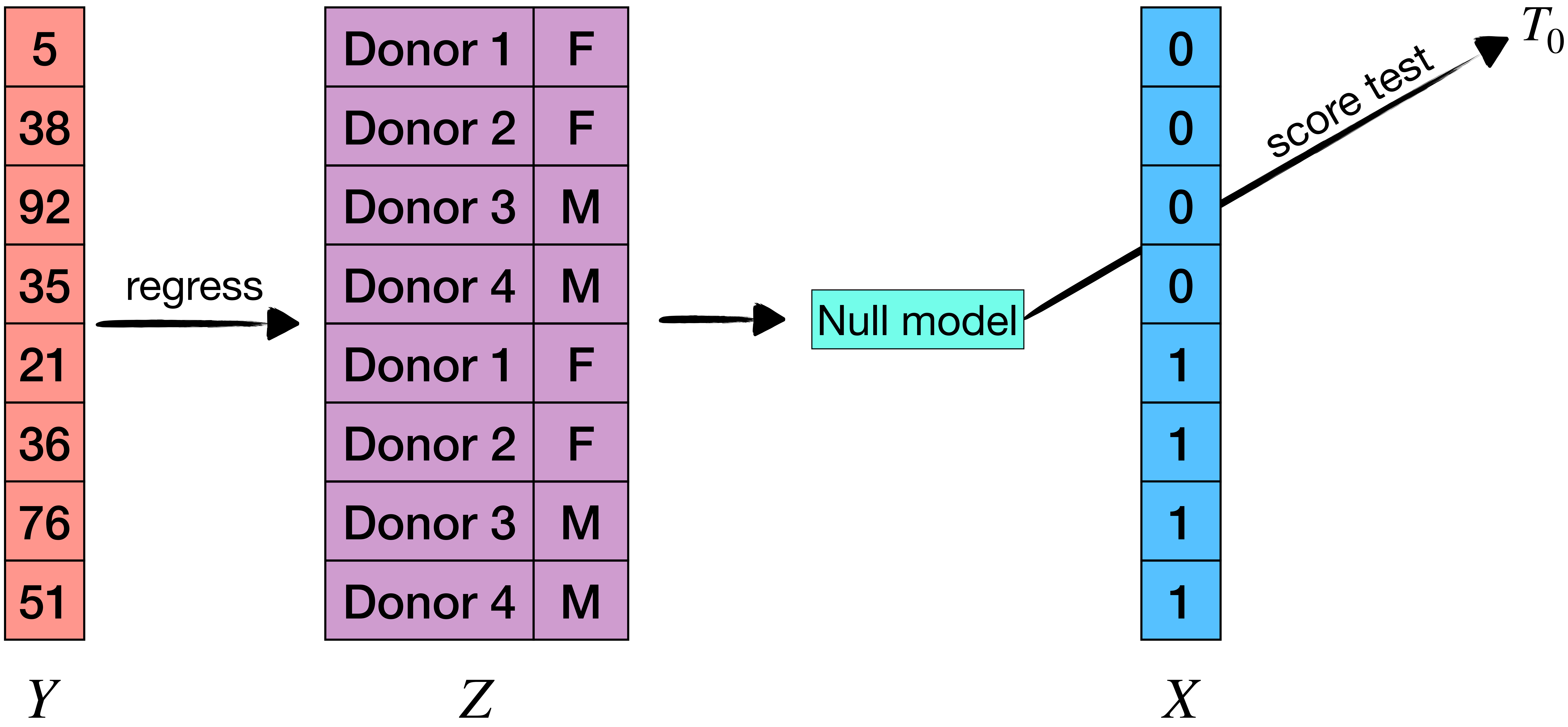
Y	Z		X
5	Donor 1	F	0
38	Donor 2	F	0
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	1
76	Donor 3	M	1
51	Donor 4	M	1

We propose a permutation test based on an NB GLM score test statistic.



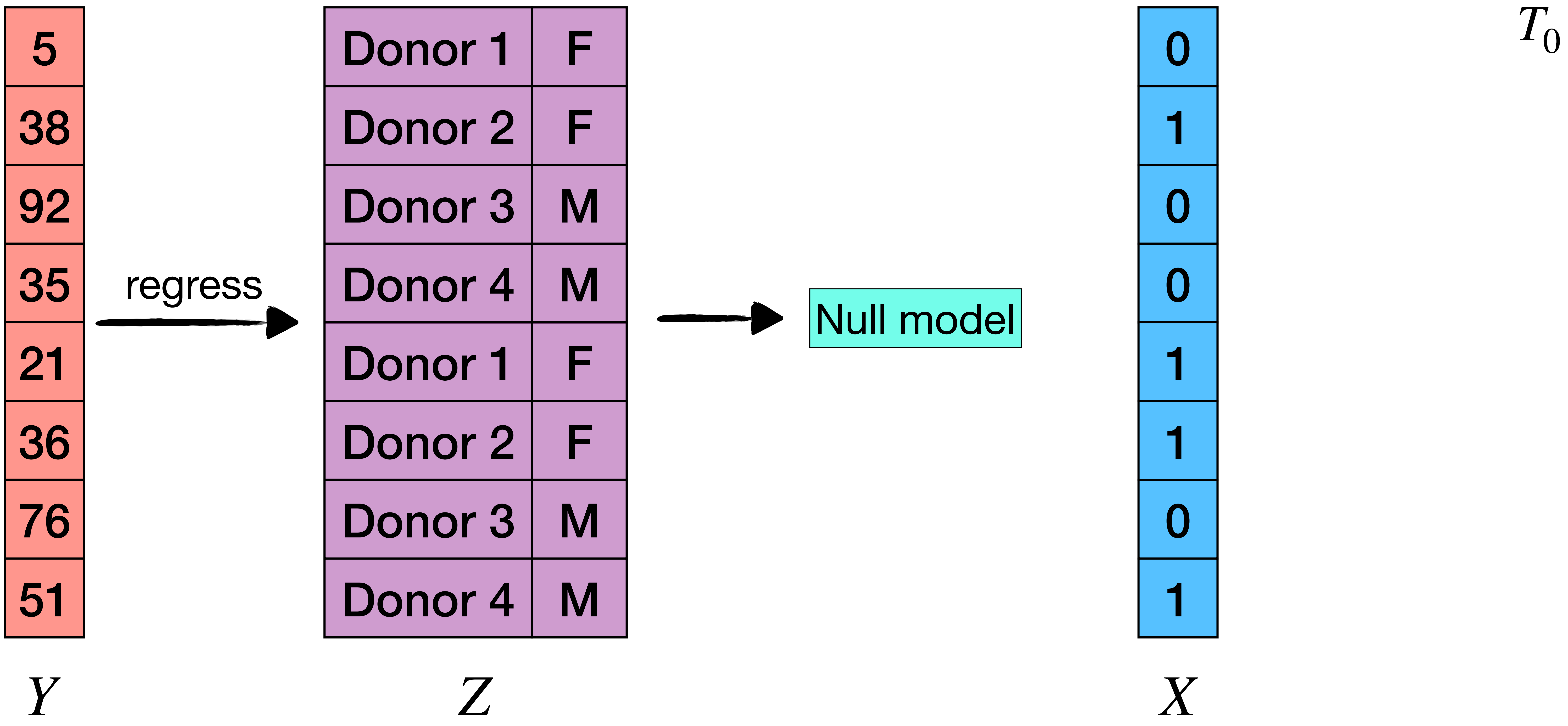
Y	Z		X
5	Donor 1	F	0
38	Donor 2	F	0
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	1
76	Donor 3	M	1
51	Donor 4	M	1

We propose a permutation test based on an NB GLM score test statistic.



5	Donor 1	F	0
38	Donor 2	F	0
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	1
76	Donor 3	M	1
51	Donor 4	M	1
Y	Z		X

We propose a permutation test based on an NB GLM score test statistic.



Y
5
38
92
35
21
36
76
51

regress

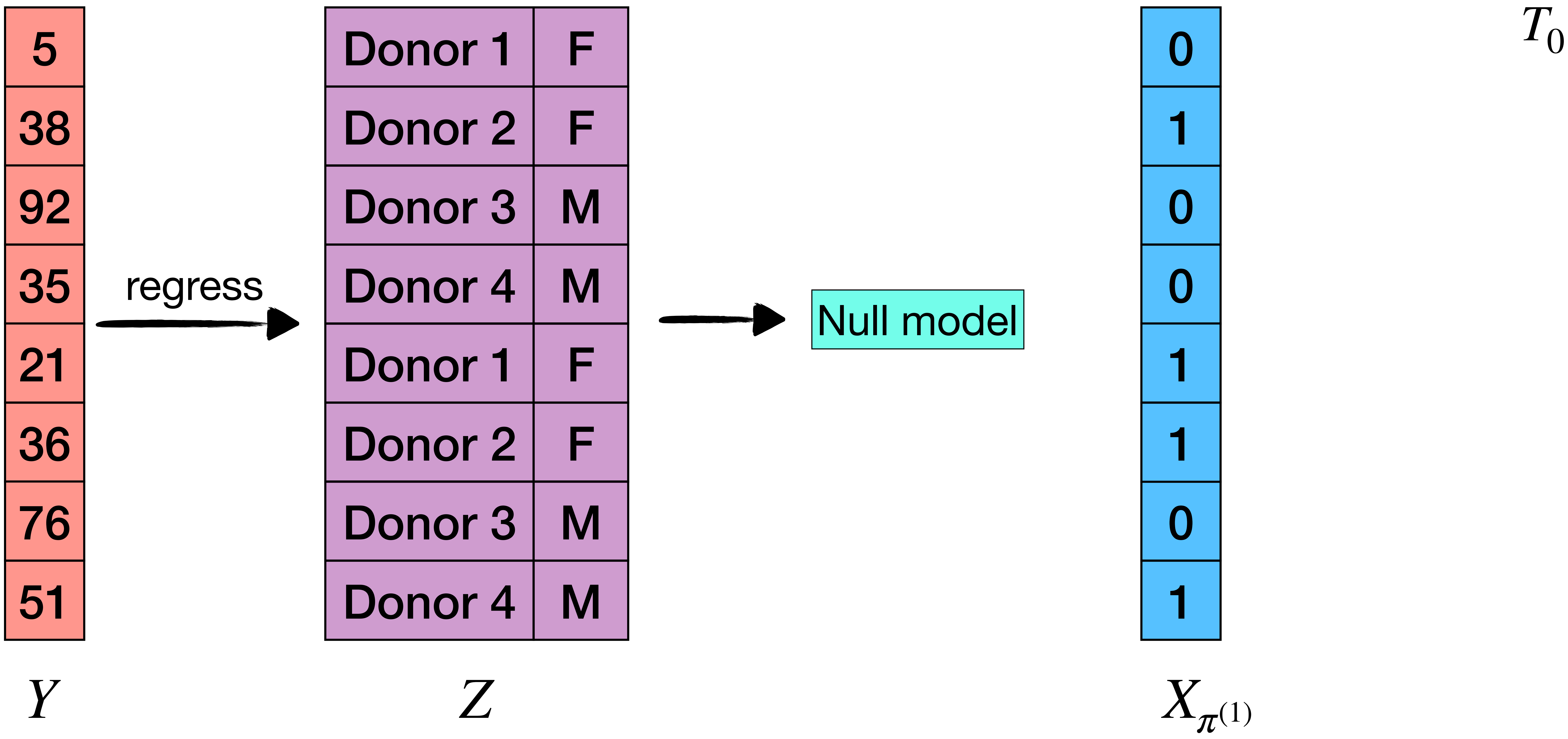
Z	
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M

Null model

X
0
1
0
0
1
1
0
1

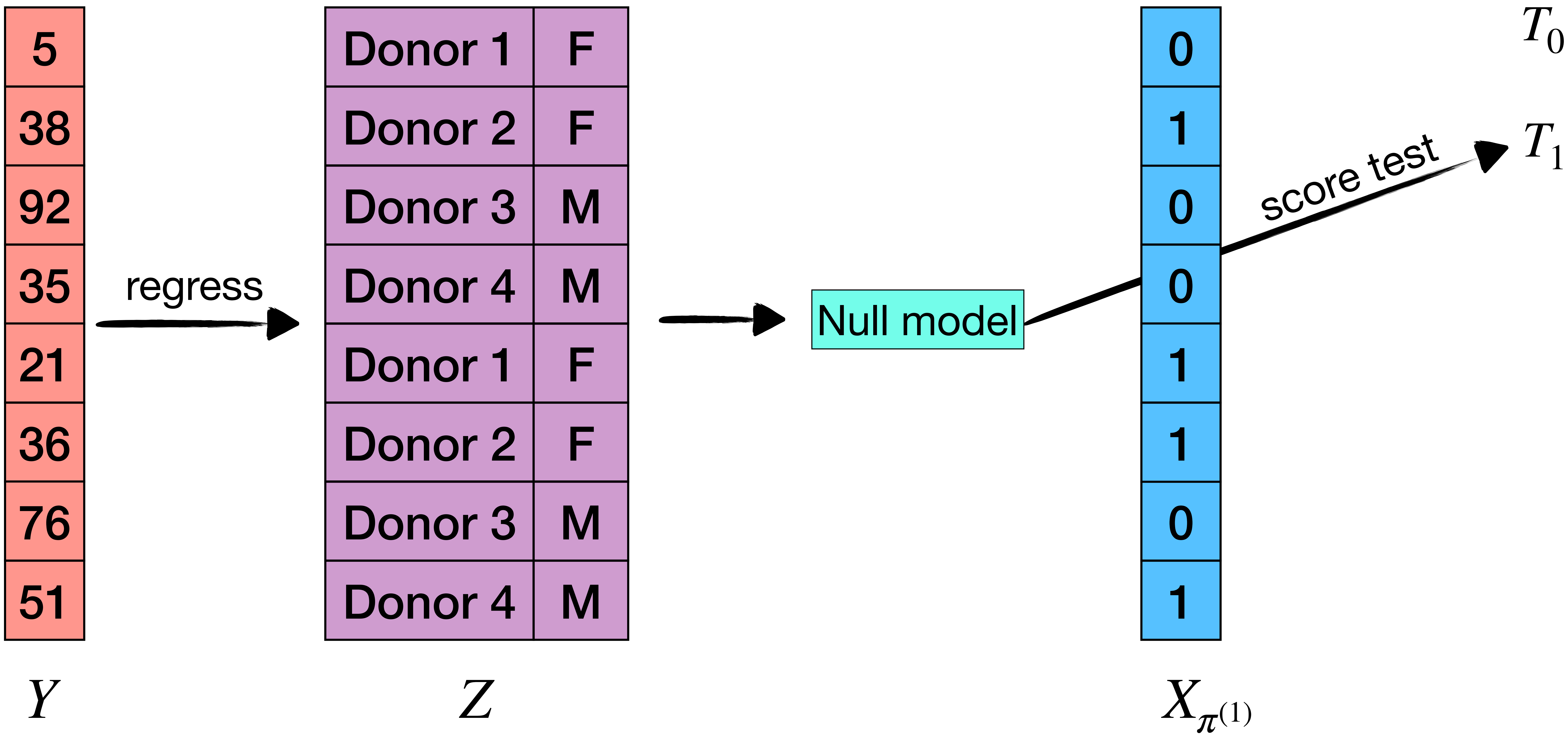
T_0

We propose a permutation test based on an NB GLM score test statistic.



Y	Z		$X_{\pi^{(1)}}$
5	Donor 1	F	0
38	Donor 2	F	1
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	1
76	Donor 3	M	0
51	Donor 4	M	1

We propose a permutation test based on an NB GLM score test statistic.

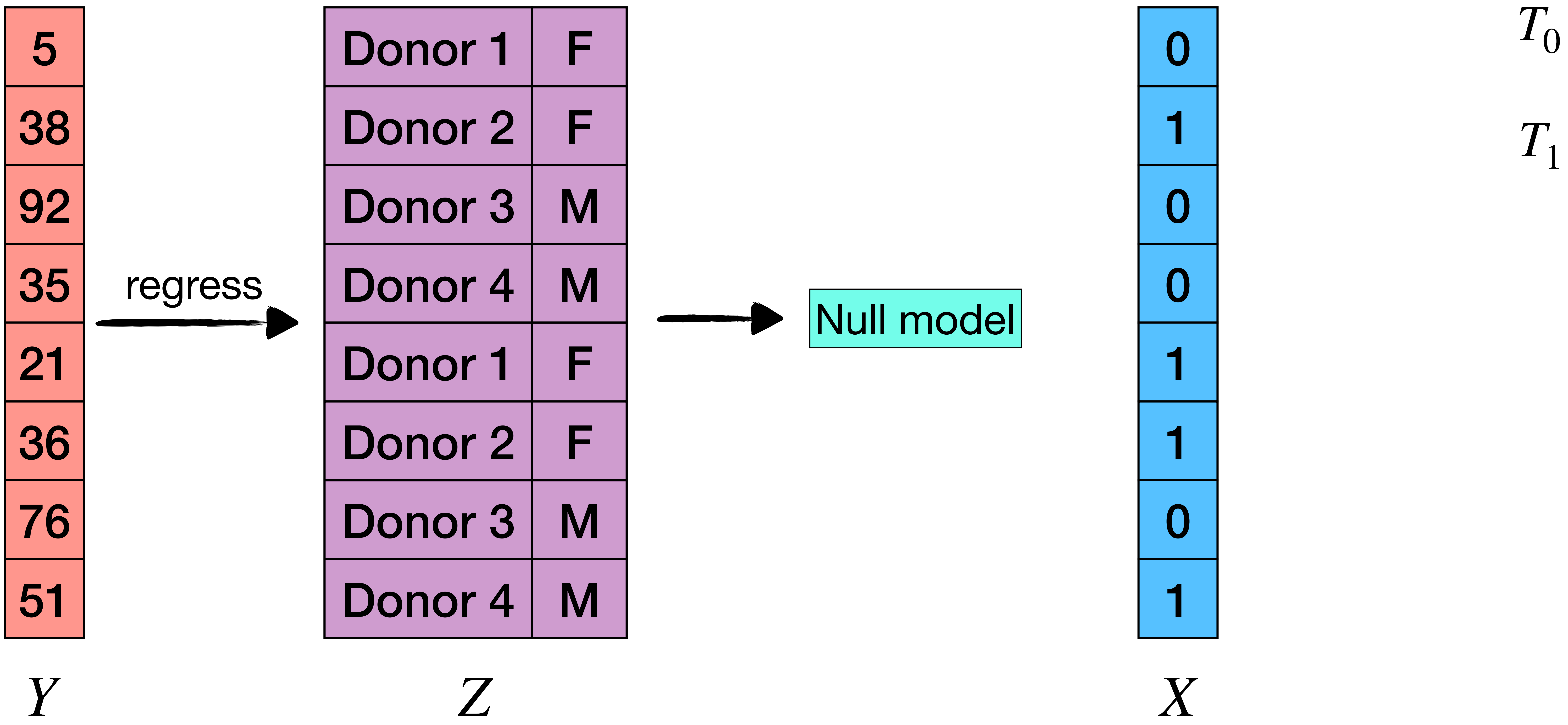


Y
5
38
92
35
21
36
76
51

Z	
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M

$X_{\pi^{(1)}}$
0
1
0
0
1
1
0
1

We propose a permutation test based on an NB GLM score test statistic.



5	Donor 1	F	0
38	Donor 2	F	1
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	1
76	Donor 3	M	0
51	Donor 4	M	1
Y	Z		X

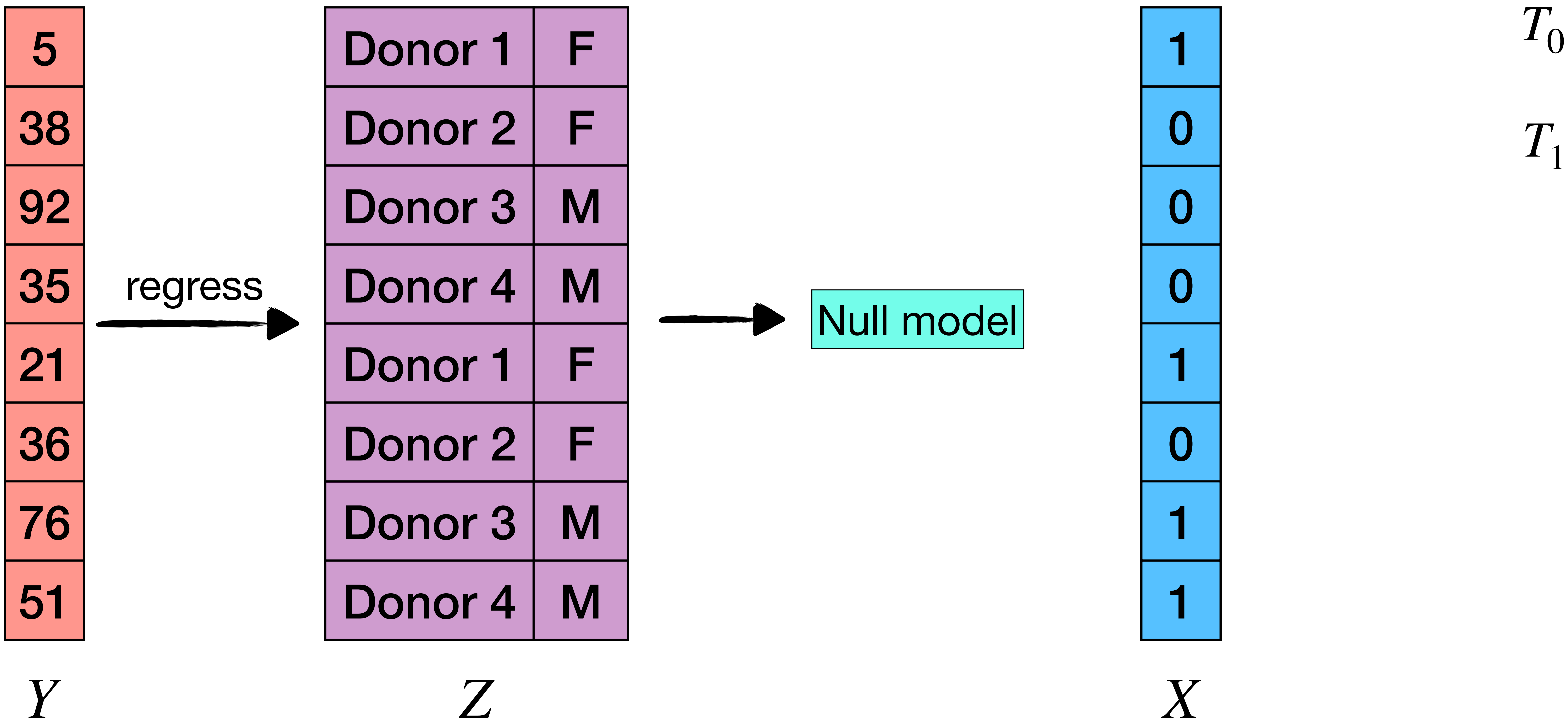
regress

Null model

T_0

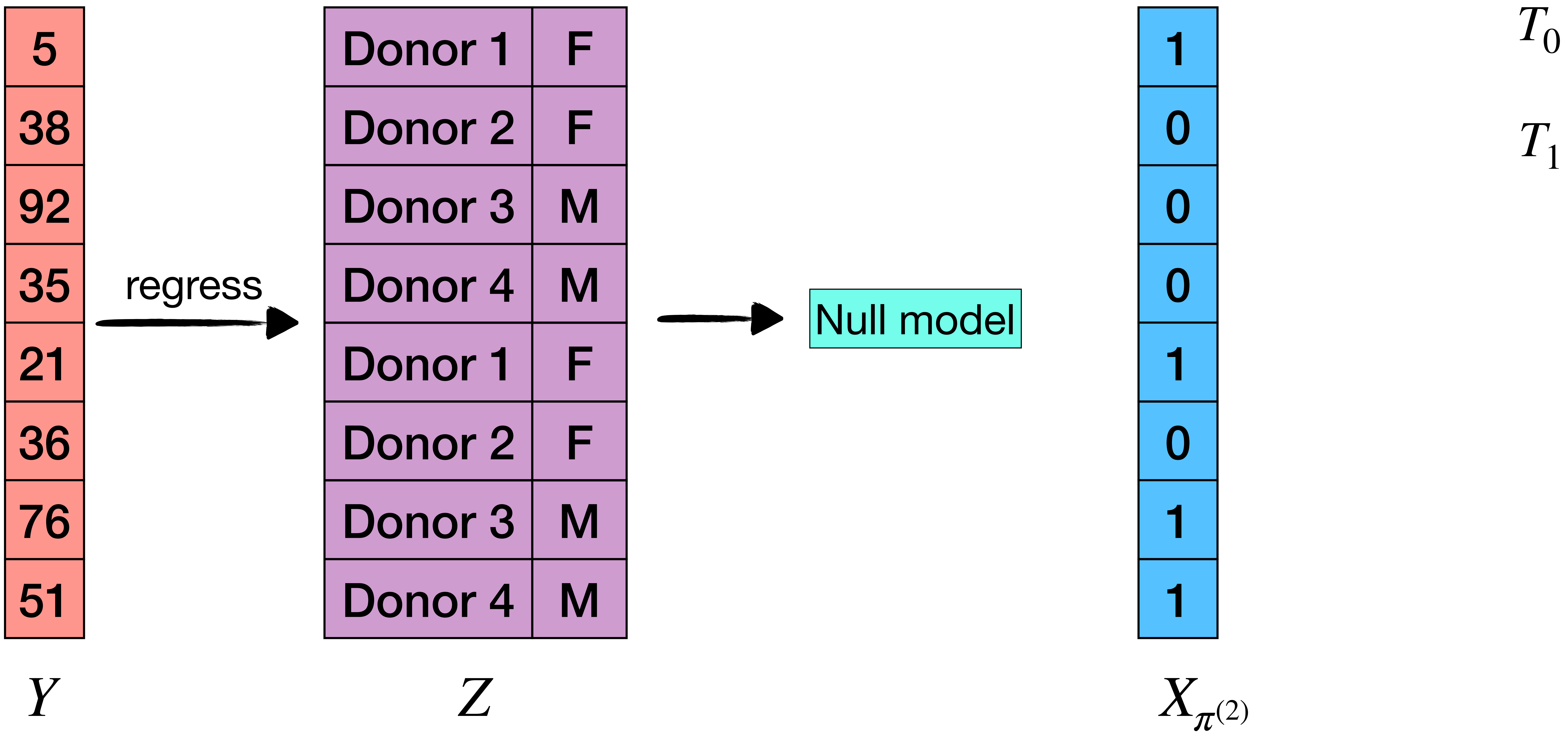
T_1

We propose a permutation test based on an NB GLM score test statistic.



Y	Z		X
5	Donor 1	F	1
38	Donor 2	F	0
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	0
76	Donor 3	M	1
51	Donor 4	M	1

We propose a permutation test based on an NB GLM score test statistic.

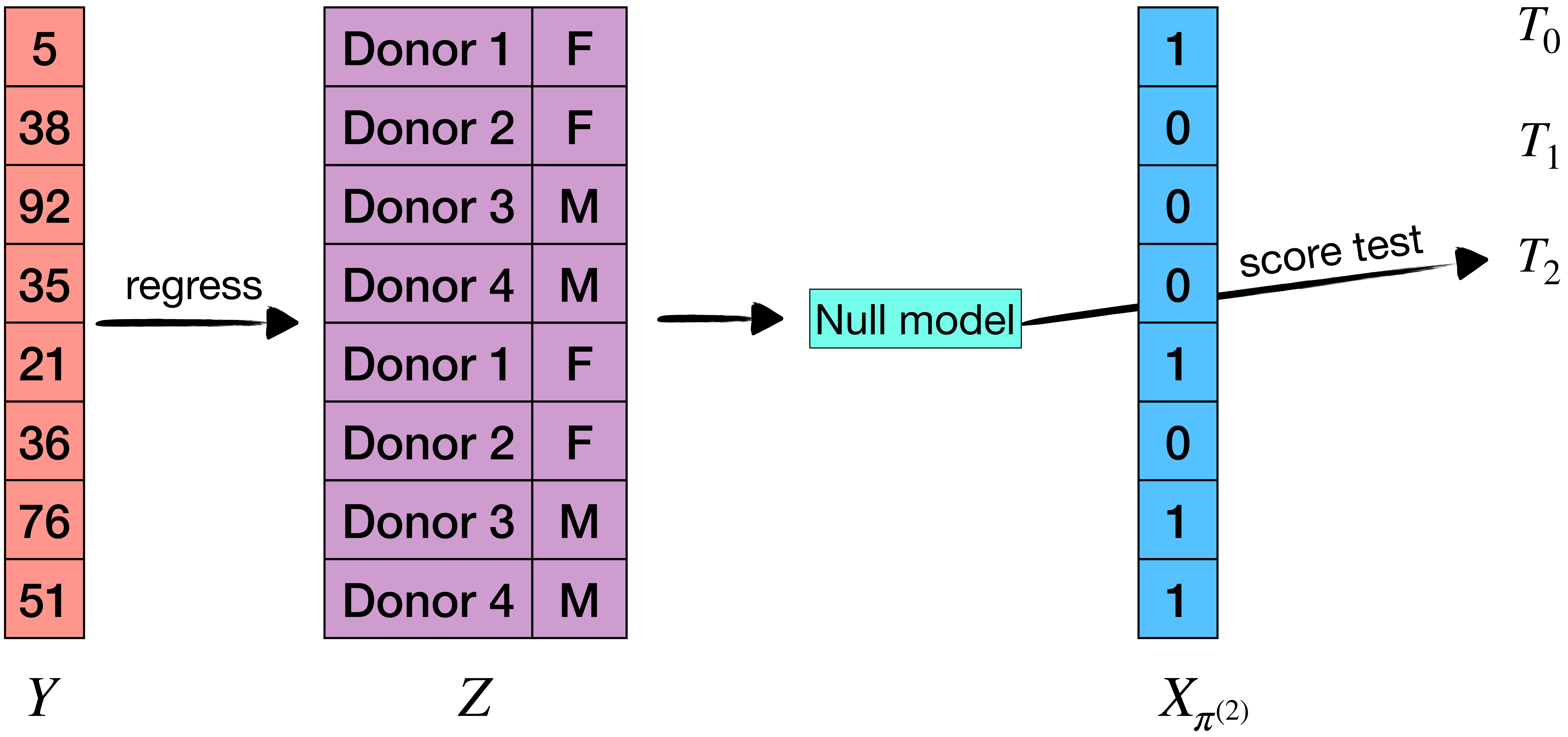


Y
5
38
92
35
21
36
76
51

Z	
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M

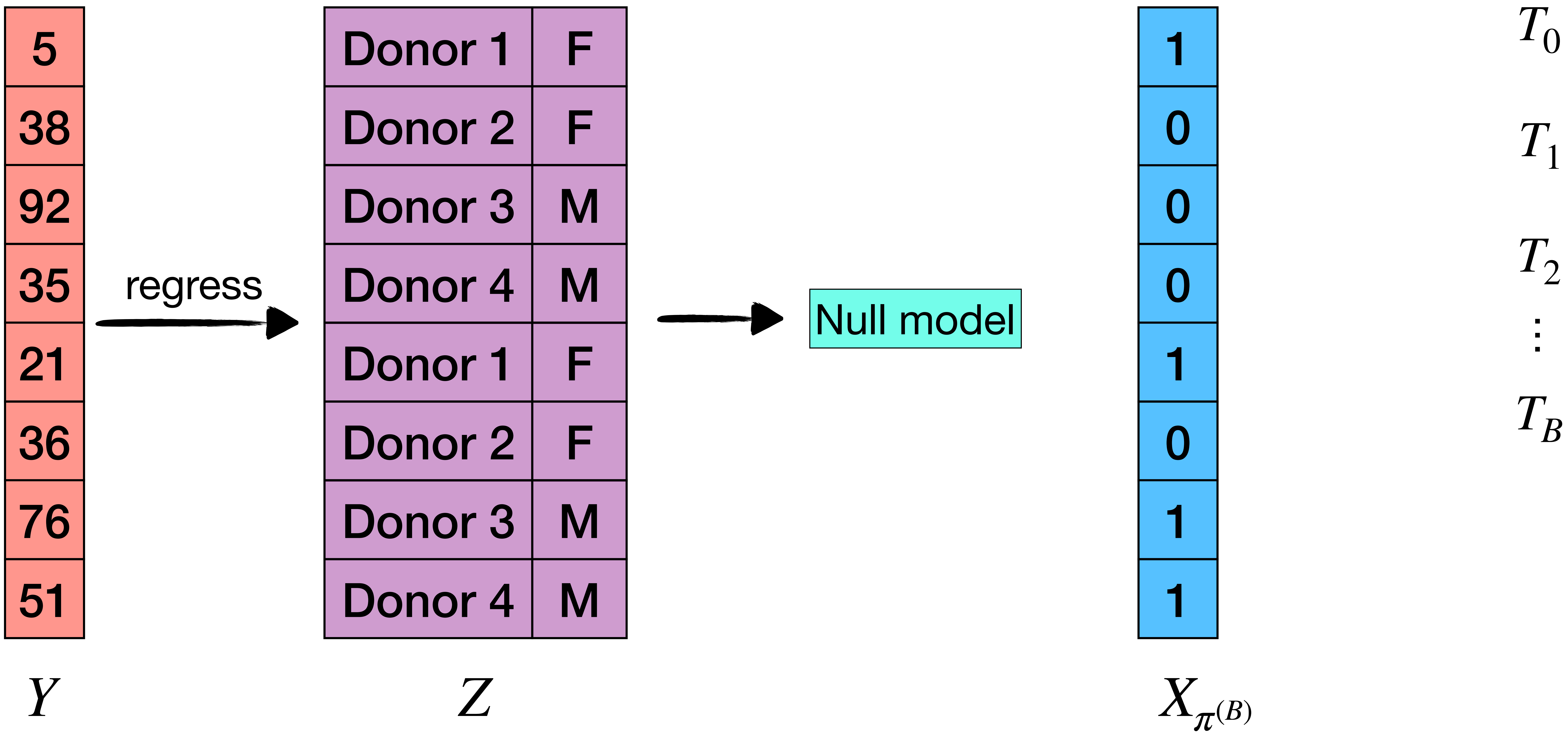
$X_{\pi^{(2)}}$
1
0
0
0
1
0
1
1

We propose a permutation test based on an NB GLM score test statistic.



Y	Z		$X_{\pi^{(2)}}$
5	Donor 1	F	1
38	Donor 2	F	0
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	0
76	Donor 3	M	1
51	Donor 4	M	1

We propose a permutation test based on an NB GLM score test statistic.



5
38
92
35
21
36
76
51

Y

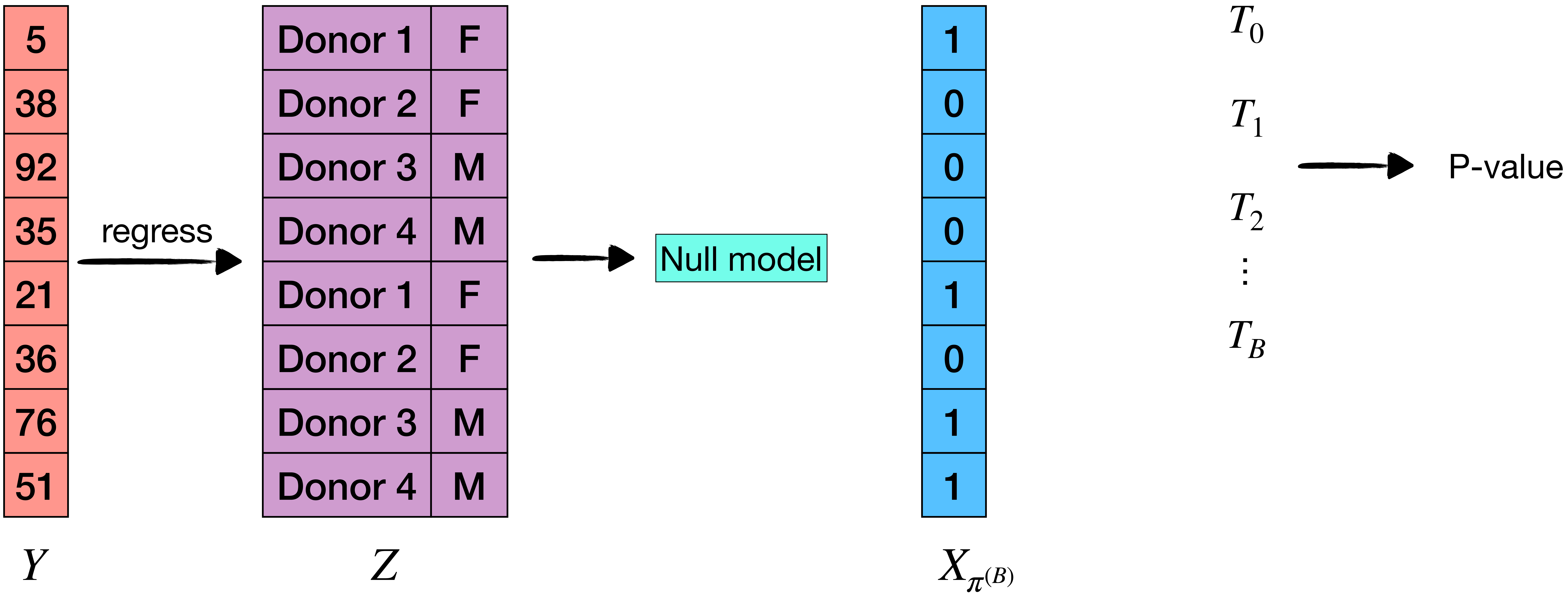
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M

Z

1
0
0
0
1
0
1
1

$X_{\pi^{(B)}}$

We propose a permutation test based on an NB GLM score test statistic.



Y
5
38
92
35
21
36
76
51

Z	
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M
Donor 1	F
Donor 2	F
Donor 3	M
Donor 4	M

$X_{\pi^{(B)}}$
1
0
0
0
1
0
1
1

We propose a permutation test based on an NB GLM score test statistic.

Y	Z		$X_{\pi^{(B)}}$
5	Donor 1	F	1
38	Donor 2	F	0
92	Donor 3	M	0
35	Donor 4	M	0
21	Donor 1	F	1
36	Donor 2	F	0
76	Donor 3	M	1
51	Donor 4	M	1

regress
Null model
T_0
T_1
T_2
$\vdots$
T_B
P-value

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T_b \geq T_0)}{B+1}$$

The proposed method is fast for four reasons.

1. Efficient algorithm for computing GLM score tests.

2. Adaptive permutation testing via anytime-valid inference.

3. C++ implementation.

Roadmap

1. Review of NB regression
2. Permuting score statistics
3. **Statistical guarantees**
4. Simulations
5. Real data analysis

Suppose the NB GLM is correctly specified and the regularity conditions hold.

Suppose the NB GLM is correctly specified and the regularity conditions hold.

Proposition 1 (sampling distribution)

$$T_n(X, Y, Z) \xrightarrow{d} N(0, \sigma_s^2)$$

Suppose the NB GLM is correctly specified and the regularity conditions hold.

Proposition 1 (sampling distribution)

$$T_n(X, Y, Z) \xrightarrow{d} N(0, \sigma_s^2)$$

Proposition 2 (permutation distribution)

Let X_π denote a randomly permuted version of X. Then

$$T_n(X_\pi, Y, Z) \,|\, (X, Y, Z) \xrightarrow{d} N(0, \sigma_p^2)$$

Suppose the NB GLM is correctly specified and the regularity conditions hold.

Proposition 1 (sampling distribution)

$$T_n(X, Y, Z) \xrightarrow{d} N(0, \sigma_s^2)$$

Proposition 2 (permutation distribution)

Let X_π denote a randomly permuted version of X. Then

$$T_n(X_\pi, Y, Z) \,|\, (X, Y, Z) \xrightarrow{d} N(0, \sigma_p^2)$$

σ_s^2 and σ_p^2 are constants that depend on the model parameters.

If the dispersion correct (i.e., $\bar{\phi} = \phi$), then $\sigma_s^2 = \sigma_p^2 = 1$.

If the dispersion is incorrect (i.e., $\bar{\phi} \neq \phi$), then $\sigma_s^2 \approx \sigma_p^2$.

Suppose the NB GLM is correctly specified and the regularity conditions hold.

Proposition 1 (sam tribution)

$$T_{}(X \quad) \xrightarrow{} N(0,\sigma_s^2)$$

M estimation theory

Proposition 2 (permu on)

Let X_π denote a randomly ion of X. Then

$$T_n(X \quad , Z) \xrightarrow{d} N(0,\sigma_p^2)$$

Central limit theorem for ranks

(See Diciccio and Romano 2017)

σ_s^2 and σ_p^2 are co ants that depend on the model parameters.

If the dispersion correct (i.e., $\bar{\phi} = \phi$), then σ_s^2=σ_p^2 $= 1$.

If the dispersion is incorrect (i.e., $\bar{\phi} \neq \phi$), then σ_s^2≈σ_p^2.

Theorem: confounder adjustment via marginal permutations (CAMP)

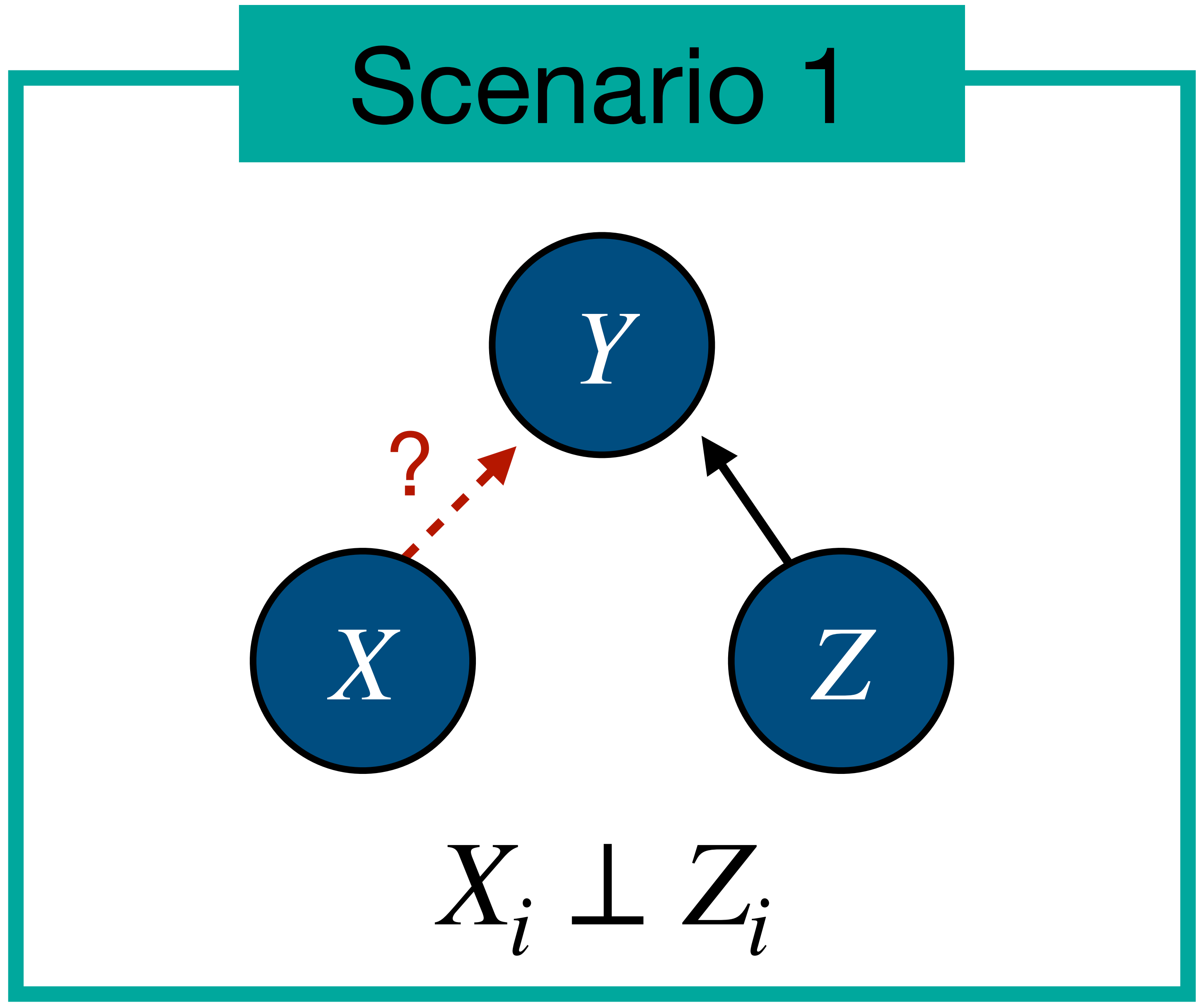
- Consider the conditional independence null hypothesis, $Y_i \perp X_i \mid Z_i$.

- **CAMP** (informal): We have two separate chances to control type-I error.

Theorem: confounder adjustment via marginal permutations (CAMP)

- Consider the conditional independence null hypothesis, $Y_i \perp X_i \mid Z_i$.

- **CAMP** (informal): We have two separate chances to control type-I error.

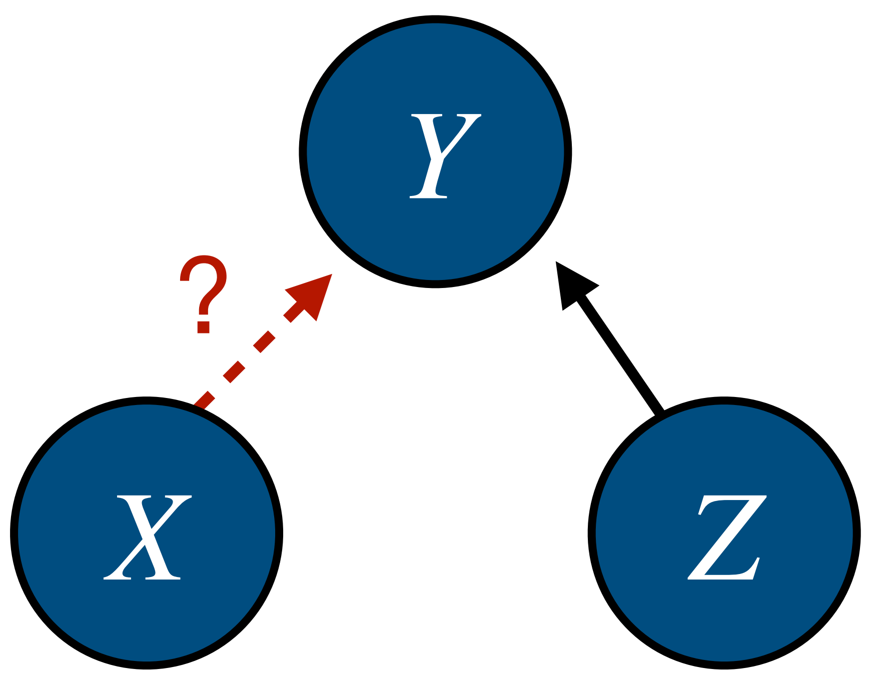


Theorem: confounder adjustment via marginal permutations (CAMP)

- Consider the conditional independence null hypothesis, $Y_i \perp X_i \mid Z_i$.

- **CAMP** (informal): We have two separate chances to control type-I error.

Scenario 1



$X_i \perp Z_i$

Scenario 2

✓ GLM specified correctly

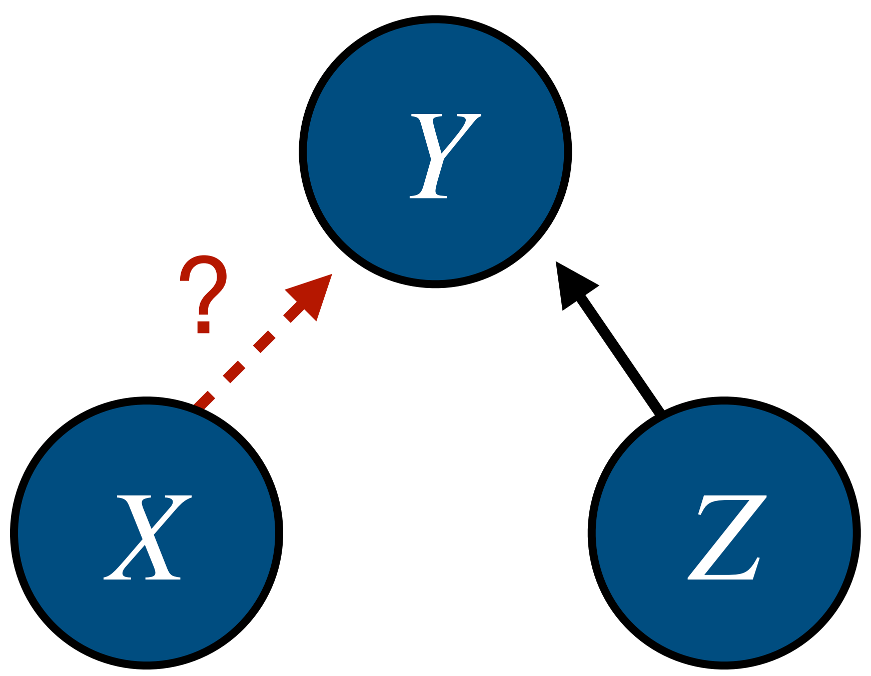
✓ Sample size sufficiently large

Theorem: confounder adjustment via marginal permutations (CAMP)

- Consider the conditional independence null hypothesis, $Y_i \perp X_i \mid Z_i$.

- **CAMP** (informal): We have two separate chances to control type-I error.

Scenario 1



$X_i \perp Z_i$

Scenario 2

✓ GLM specified correctly

✓ Sample size sufficiently large

- CAMP is related to — but distinct from — double robustness.

CAMP: Scenario 1

CAMP: Scenario 1

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B + 1}$$

CAMP: Scenario 1

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B+1}$$

Let $\phi_n(X, Y, Z) = \mathbf{1}(p_{\text{perm}} \leq \alpha)$ be the level-α test, based on the permuted score statistic. Let $\mathscr{K}$ be set of distributions for which $X_i \perp Z_i$. Let $\mathscr{N}$ be the set of distributions for which $X_i \perp Y_i \,|\, Z_i$. Then

$$\sup_{\mathscr{L} \in \mathscr{K} \cap \mathscr{N}} \mathbb{E}_{\mathscr{L}}[\phi_n(X, Y, Z)] - \alpha = 0.$$

CAMP: Scenario 1

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B+1}$$

Let $\phi_n(X, Y, Z) = \mathbf{1}(p_{\text{perm}} \leq \alpha)$ be the level-α test, based on the permuted score statistic. Let $\mathscr{K}$ be set of distributions for which $X_i \perp Z_i$. Let $\mathscr{N}$ be the set of distributions for which $X_i \perp Y_i \,|\, Z_i$. Then

$$\sup_{\mathscr{L} \in \mathscr{K} \cap \mathscr{N}} \mathbb{E}_{\mathscr{L}}[\phi_n(X, Y, Z)] - \alpha = 0.$$

Intuition: If $X_i \perp Z_i$, then type-I error is controlled under *arbitrary model misspecification* and *in finite samples.*

CAMP: Scenario 2

CAMP: Scenario 2

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B + 1}$$

CAMP: Scenario 2

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B + 1}$$

Let $\phi_n(X, Y, Z) = \mathbf{1}(p_{\text{perm}} \leq \alpha)$. Let $\mathcal{N}$ be the set of distributions for which $X_i \perp Y_i \mid Z_i$. Then

CAMP: Scenario 2

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B+1}$$

Let $\phi_n(X, Y, Z) = \mathbf{1}(p_{\text{perm}} \leq \alpha)$. Let $\mathcal{N}$ be the set of distributions for which $X_i \perp Y_i \mid Z_i$. Then

$$\sup_{\mathcal{L} \in \mathcal{N}} \limsup_{n \to \infty} \mathbb{E}_{\mathcal{L}}[\phi_n(X, Y, Z)] - \alpha \leq \frac{1}{4}\Phi^{-1}(1-\alpha)(1 - \sigma_p/\sigma_s)$$

CAMP: Scenario 2

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B + 1}$$

Let $\phi_n(X, Y, Z) = \mathbf{1}(p_{\text{perm}} \leq \alpha)$. Let $\mathcal{N}$ be the set of distributions for which $X_i \perp Y_i \,|\, Z_i$. Then

If $\bar{\phi} \neq \phi$, then $\sigma_p/\sigma_s \approx 1$, so

$$\sup_{\mathcal{L} \in \mathcal{N}} \limsup_{n \to \infty} \mathbb{E}_{\mathcal{L}}[\phi_n(X, Y, Z)] - \alpha \leq \frac{1}{4}\Phi^{-1}(1 - \alpha)(1 - \sigma_p/\sigma_s) \quad \approx 0$$

CAMP: Scenario 2

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(T(X_{\pi^{(b)}}, Y, Z) \geq T_0(X, Y, Z))}{B+1}$$

Let $\phi_n(X, Y, Z) = \mathbf{1}(p_{\text{perm}} \leq \alpha)$. Let $\mathscr{N}$ be the set of distributions for which $X_i \perp Y_i \,|\, Z_i$. Then

If $\bar{\phi} \neq \phi$, then $\sigma_p/\sigma_s \approx 1$, so

$$\sup_{\mathscr{L} \in \mathscr{N}} \limsup_{n \to \infty} \mathbb{E}_{\mathscr{L}}[\phi_n(X, Y, Z)] - \alpha \leq \frac{1}{4}\Phi^{-1}(1-\alpha)(1 - \sigma_p/\sigma_s) \quad \approx 0$$

Intuition: If the NB GLM is correctly specified up to its dispersion parameter, then type-I error is controlled in large samples.

Roadmap

1. Review of NB regression

2. Permuting score statistics

3. Statistical guarantees

4. **Simulations**

5. Real data analysis

We evaluated four methods in our simulation studies.

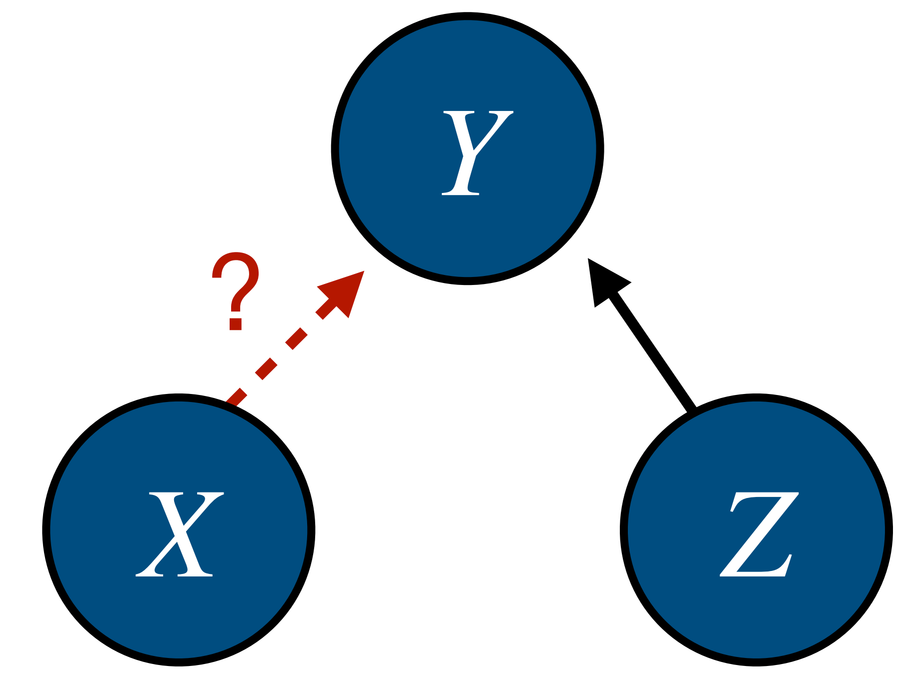
1. Standard NB regression
   - Implemented as MASS
2. Permuting score statistics (ours)
   - Implemented as “robust MASS”
3. Finite-sample Mann-Whitney (MW) test
4. Permuting NB GLM residuals

Finite-sample MW test

- Nonparametric, finite-sample valid test of independence between X_i and Y_i.

- No adjustment for covariates.

- Valid test of $X_i \perp Y_i \mid Z_i$ when $X_i \perp Z_i$.

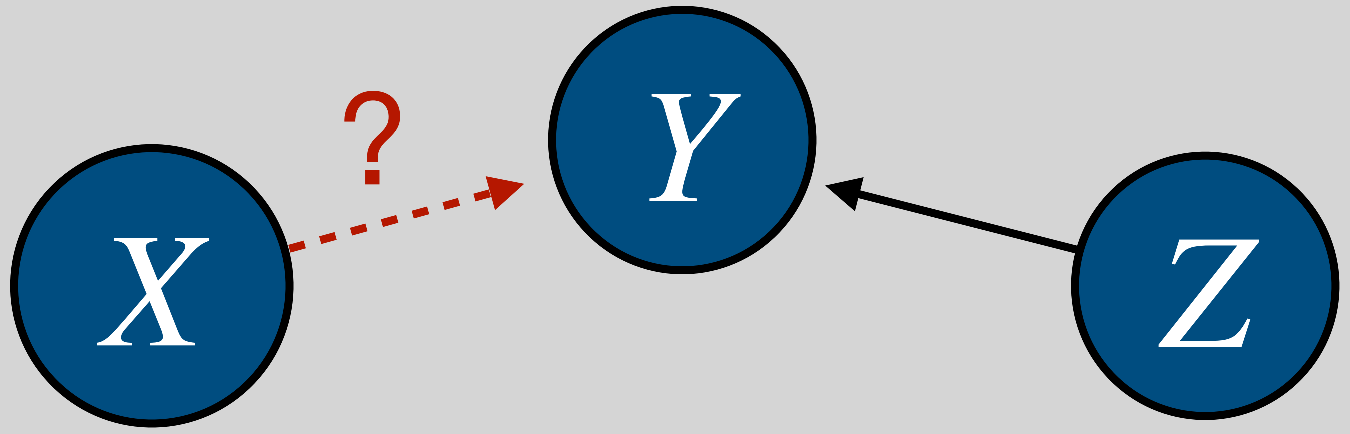


$X_i \perp Z_i$

Permuting NB GLM residuals

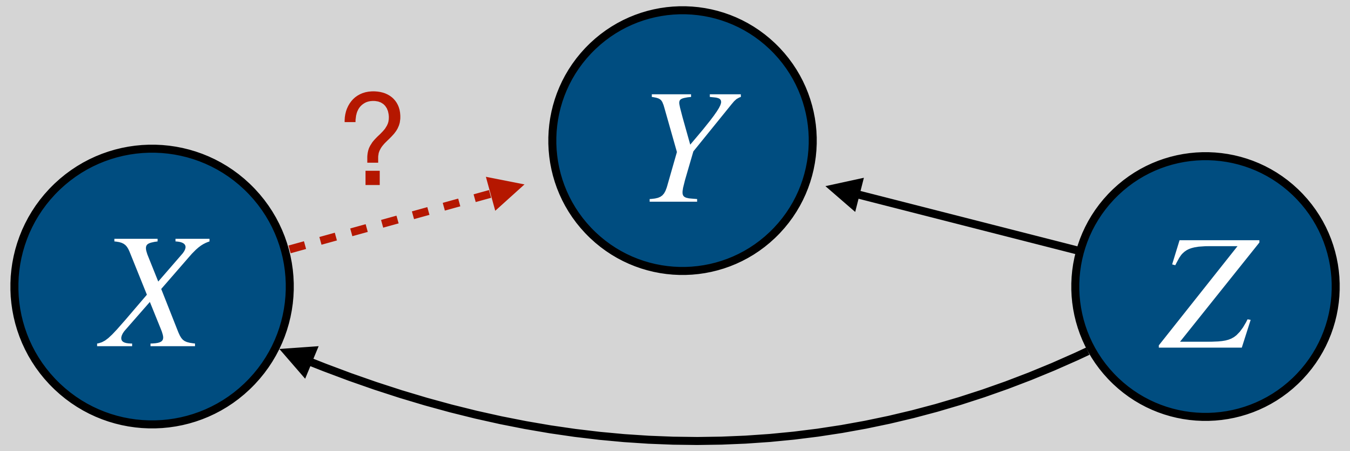
1. Fit null model (i.e., regress Y onto Z via NB GLM).
2. Extract residuals from fitted GLM.
3. Test for association between residuals and X via permutation test.

Simulation 1



Type-I error
False discovery rate
0.10
0.08
0.06
0.04
0.02
0.00
0.050
0.075
0.100
0.125
Effect size
Power
N true discoveries
40
30
20
10
0
0.050
0.075
0.100
0.125
Effect size
Compute
Running time (s)
10
8
6
4
2
0
Method
Method
🟩 MASS
🟪 Permuting residuals
🟥 MW test (finite–sample)
🟦 Robust MASS (ours)

Simulation 2



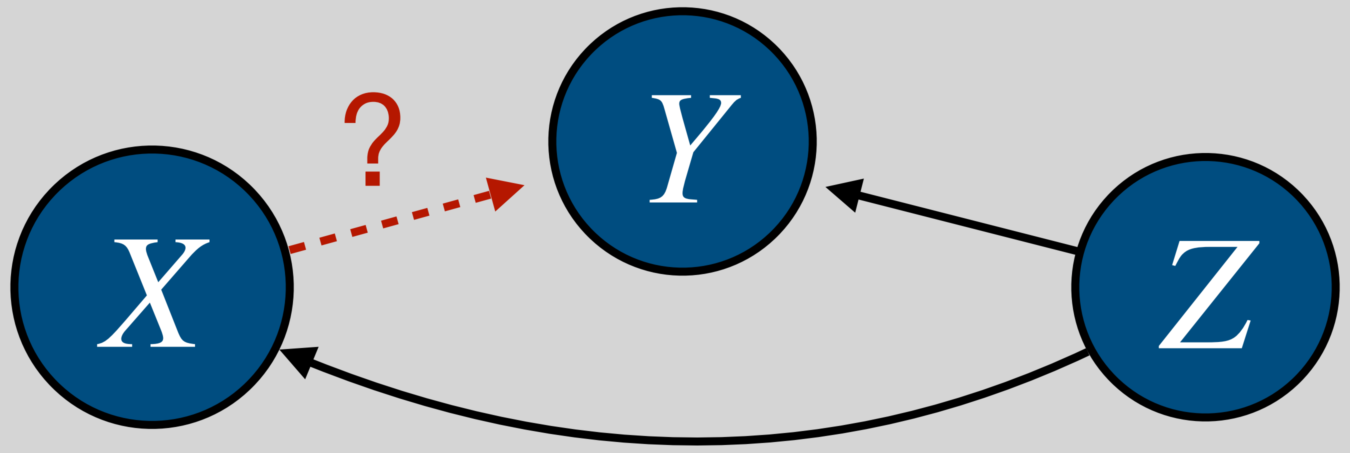
Type-I error

Power

Compute

False discovery rate
0.0
0.1
0.2
0.3
0.4
X, Z dependence
0.0
0.2
0.4
0.6
0.8
N true discoveries
0
5
10
15
20
25
Running time (s)
0
2
4
6
8
10
Method
Method
MASS
Permuting residuals
MW test (finite−sample)
Robust MASS (ours)

Simulation 3



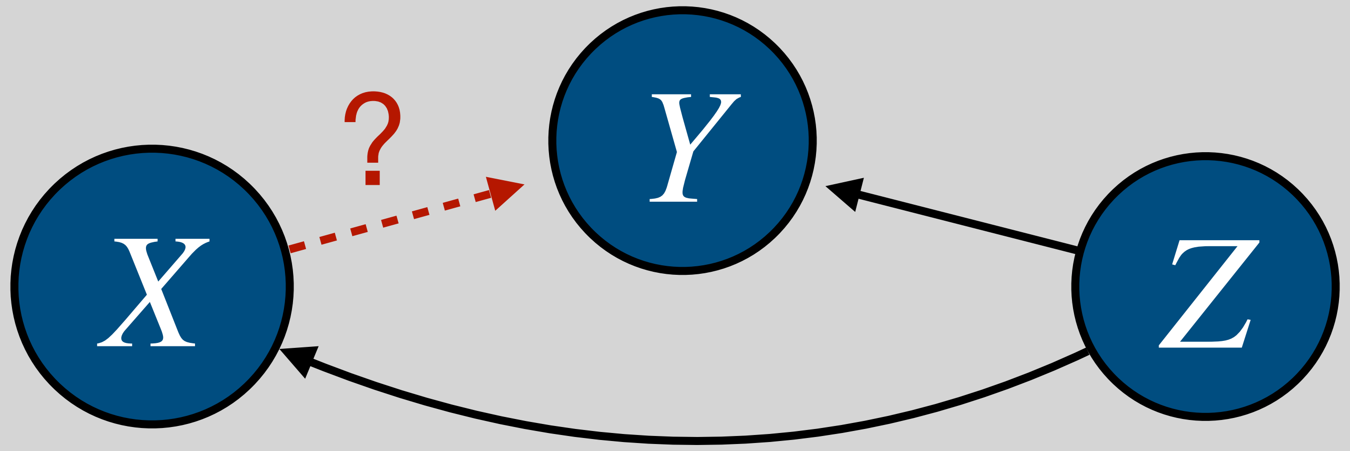
Type-I error

Power

Compute

False discovery rate
0.4
0.3
0.2
0.1
0.0
250
500
750
1000
Sample size (n)
N true discoveries
50
40
30
20
10
0
250
500
750
1000
Sample size (n)
Running time (s)
16
12
8
4
0
Method
Method
MASS
Permuting residuals
MW test (finite−sample)
Robust MASS (ours)

Simulation 4

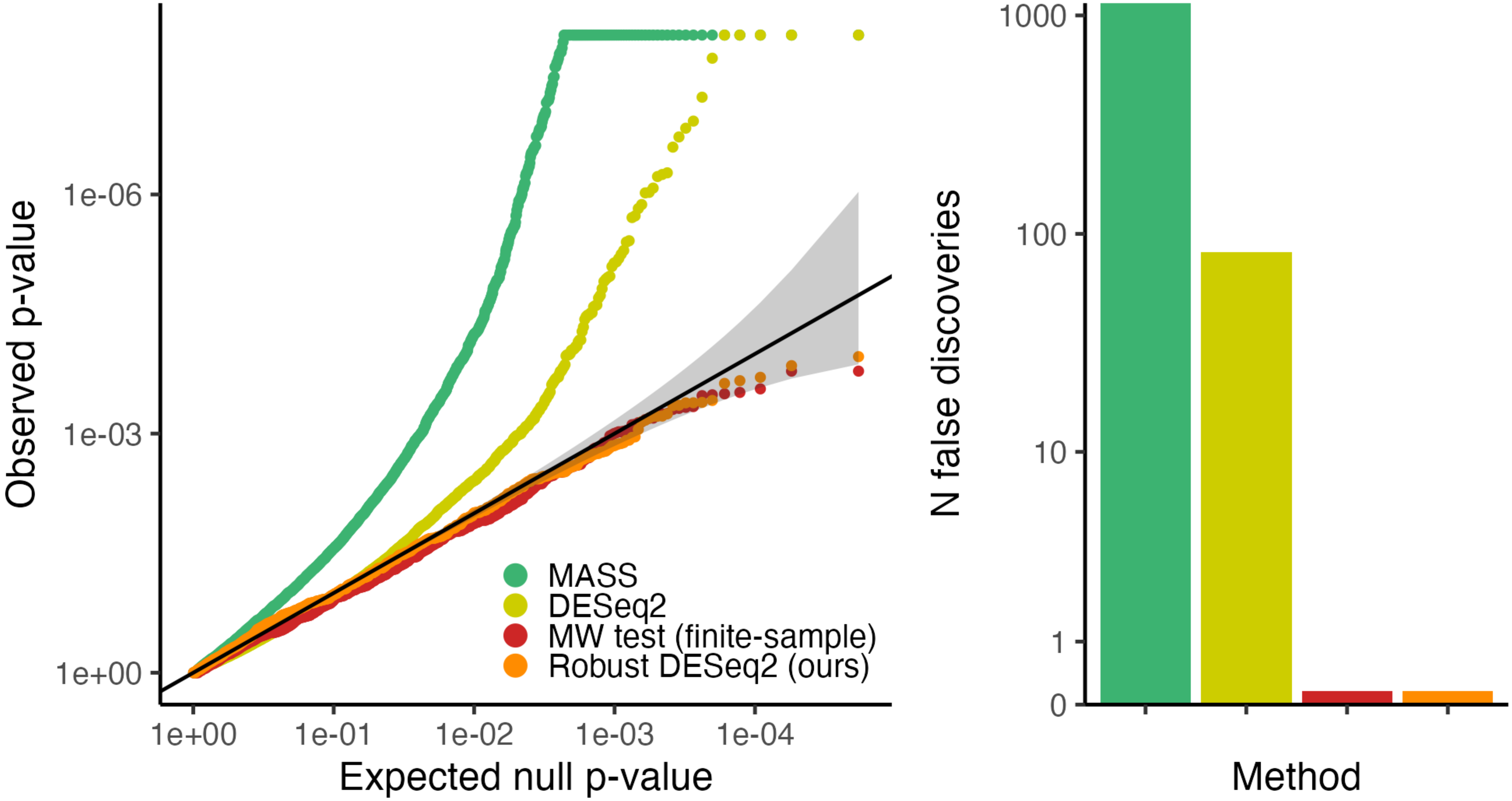


Type-I error
False discovery rate
0.3
0.2
0.1
0.0
0.00
0.04
0.08
0.12
0.16
Proportion zero
Power
N true discoveries
40
30
20
10
0
0.00
0.04
0.08
0.12
0.16
Proportion zero
Compute
Running time (s)
7.5
5.0
2.5
0.0
Method
Method
MASS
Permuting residuals
MW test (finite–sample)
Robust MASS (ours)

Roadmap

1. Review of NB regression
2. Permuting score statistics
3. Statistical guarantees
4. Simulations
5. **Real data analysis**

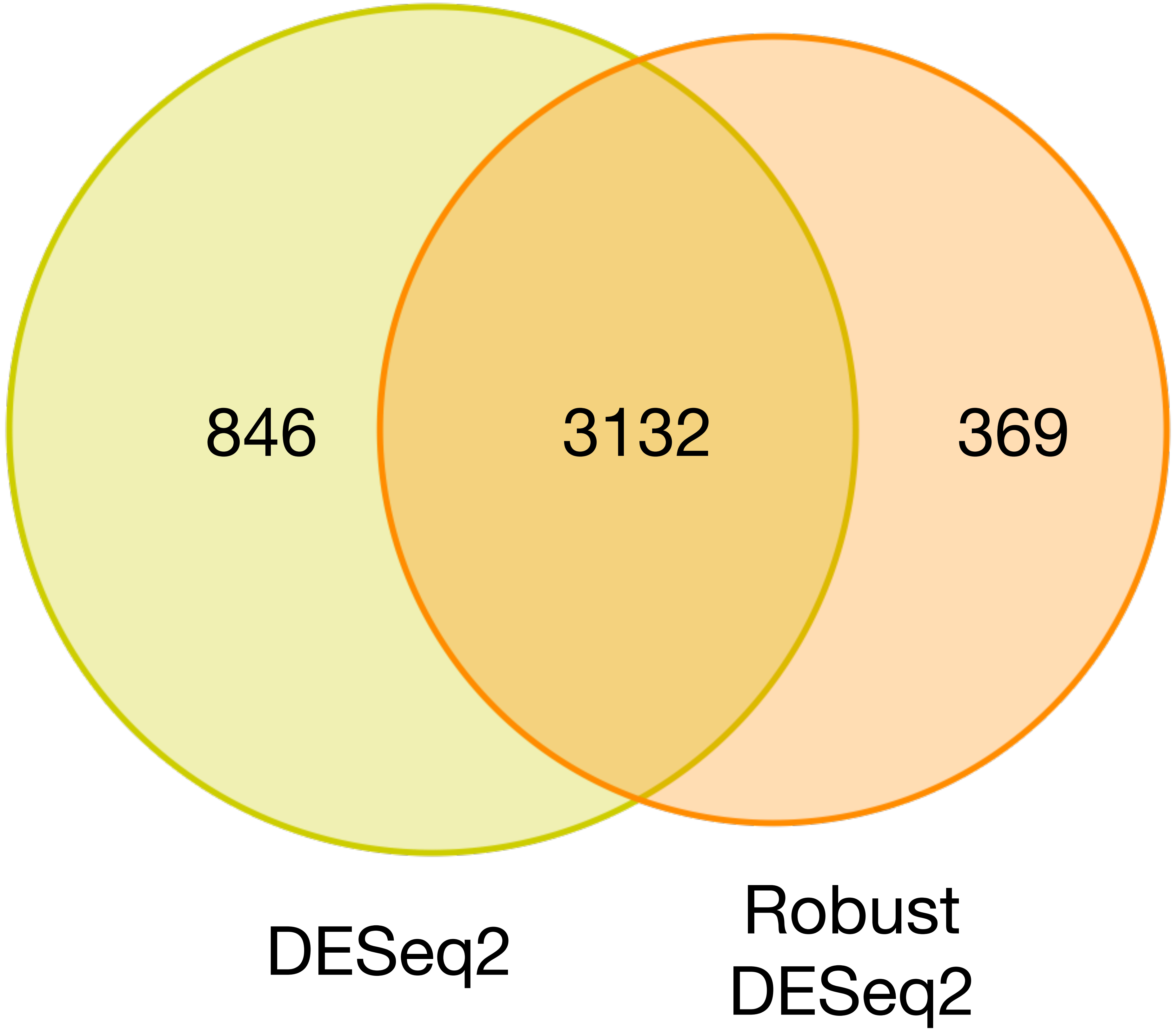
Negative control analysis



Discovery analysis

Method	N rejections*	Positive control gene rejected	Running time
MASS	8002	✓	345.2 s
DESeq2	3978	✓	14.4 s
MW (finite)	292	✓	14.5 s
Robust DESeq2	3501	✓	48.7 s

*out of 27,304 genes



Conclusion

NB regression is a popular method for differential expression testing.

Violating the assumptions of NB regression leads to inaccurate results.

Permuting score statistics improves robustness of the NB model.