Computationally efficient and statistically accurate conditional independence testing with spaCRT

International Seminar on Selective Inference
November 4, 2024

Ziang Niu
University of Pennsylvania
https://ziangniu6.github.io

Co-authors

Jyotishka Ray Choudhury
(Georgia Tech)

Eugene Katsevich
(UPenn)

A model for genetic origins of human diseases

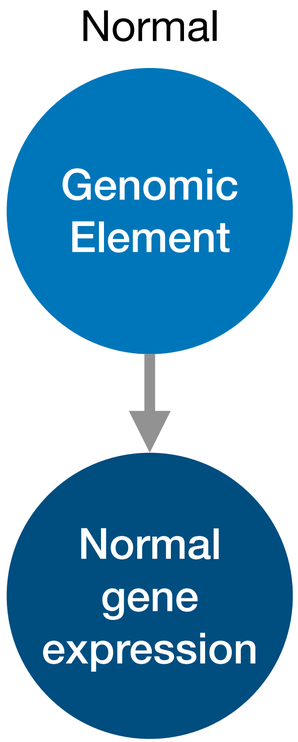
A model for genetic origins of human diseases

Normal

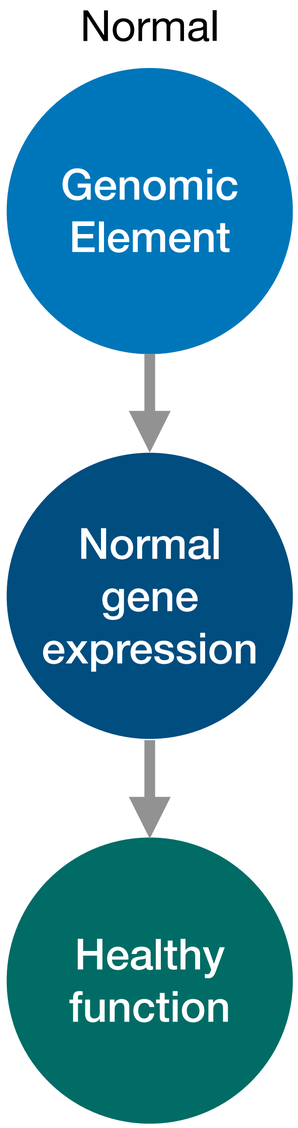
A model for genetic origins of human diseases

Normal

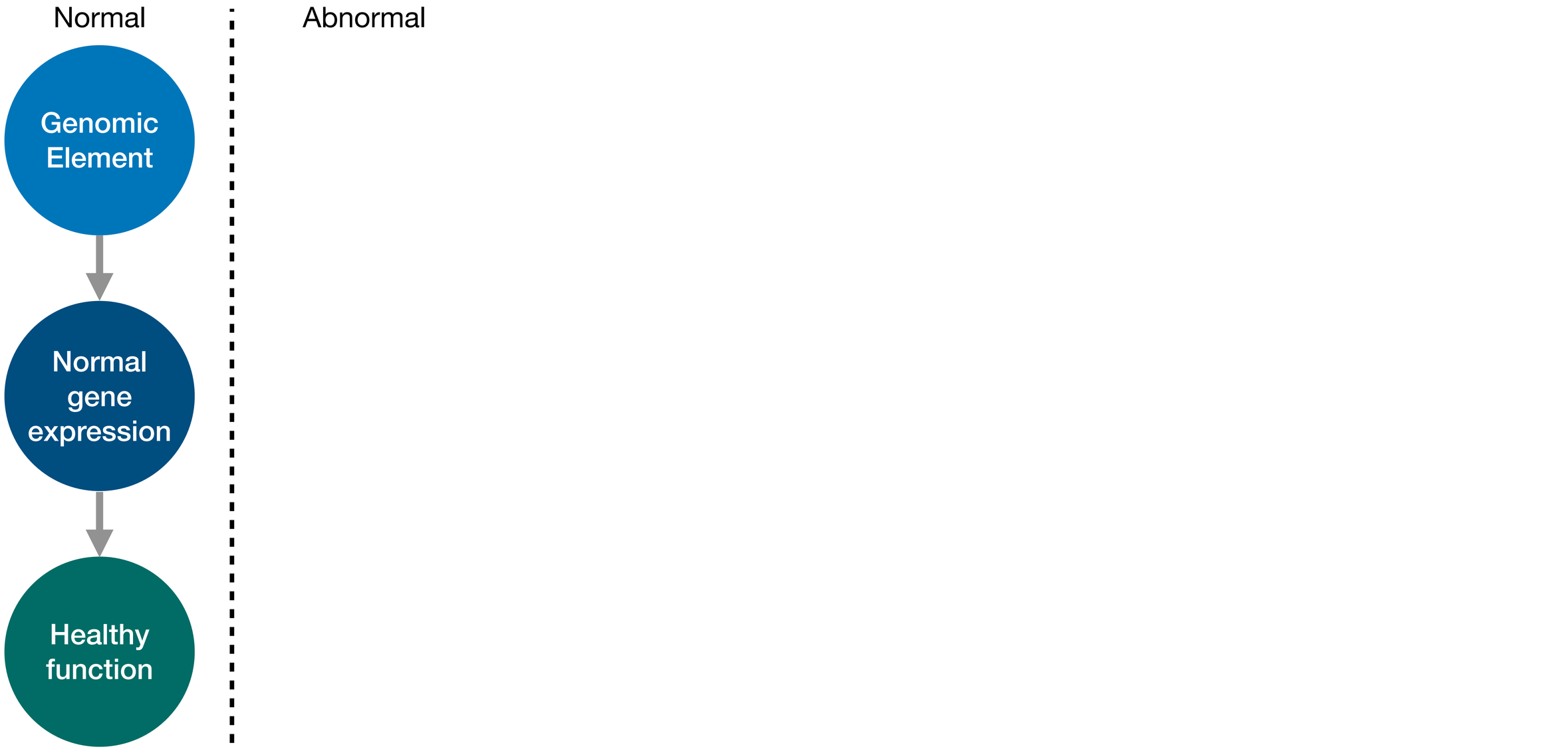
A model for genetic origins of human diseases



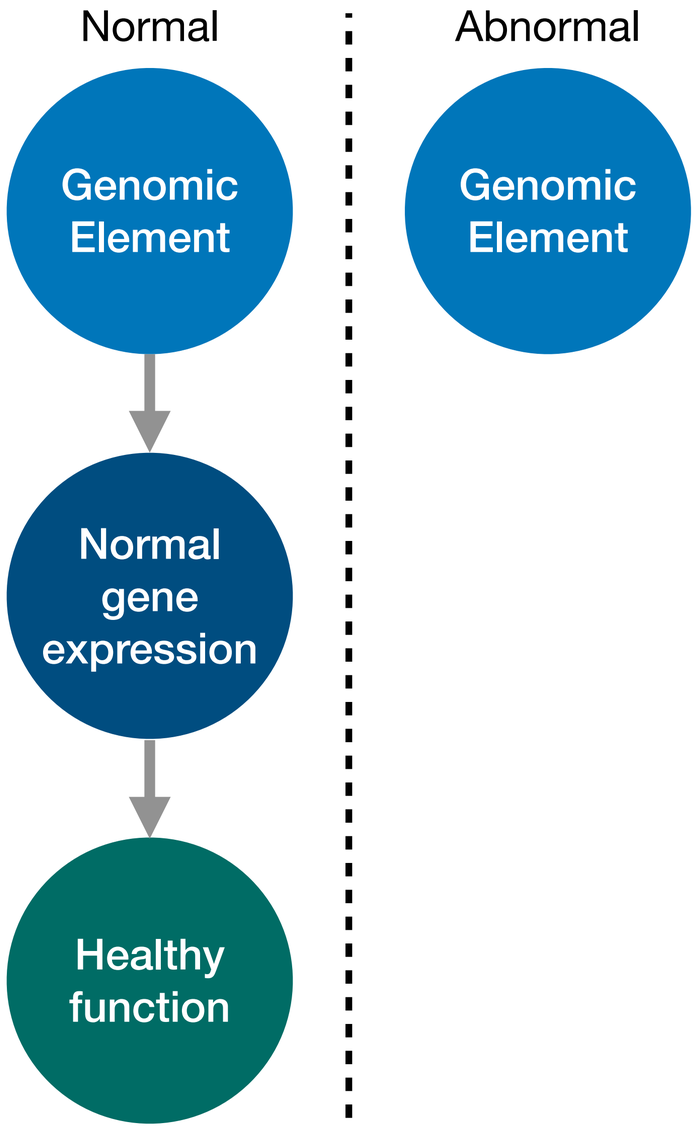
A model for genetic origins of human diseases



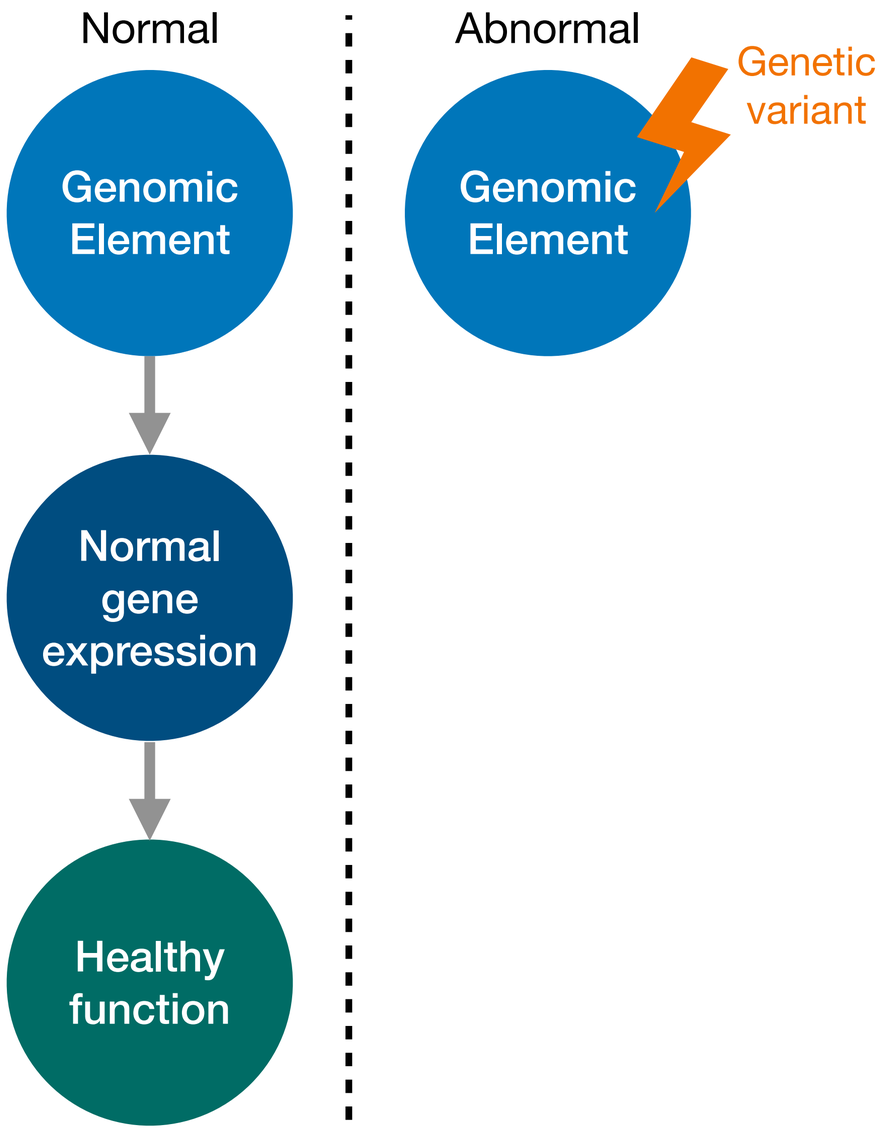
A model for genetic origins of human diseases



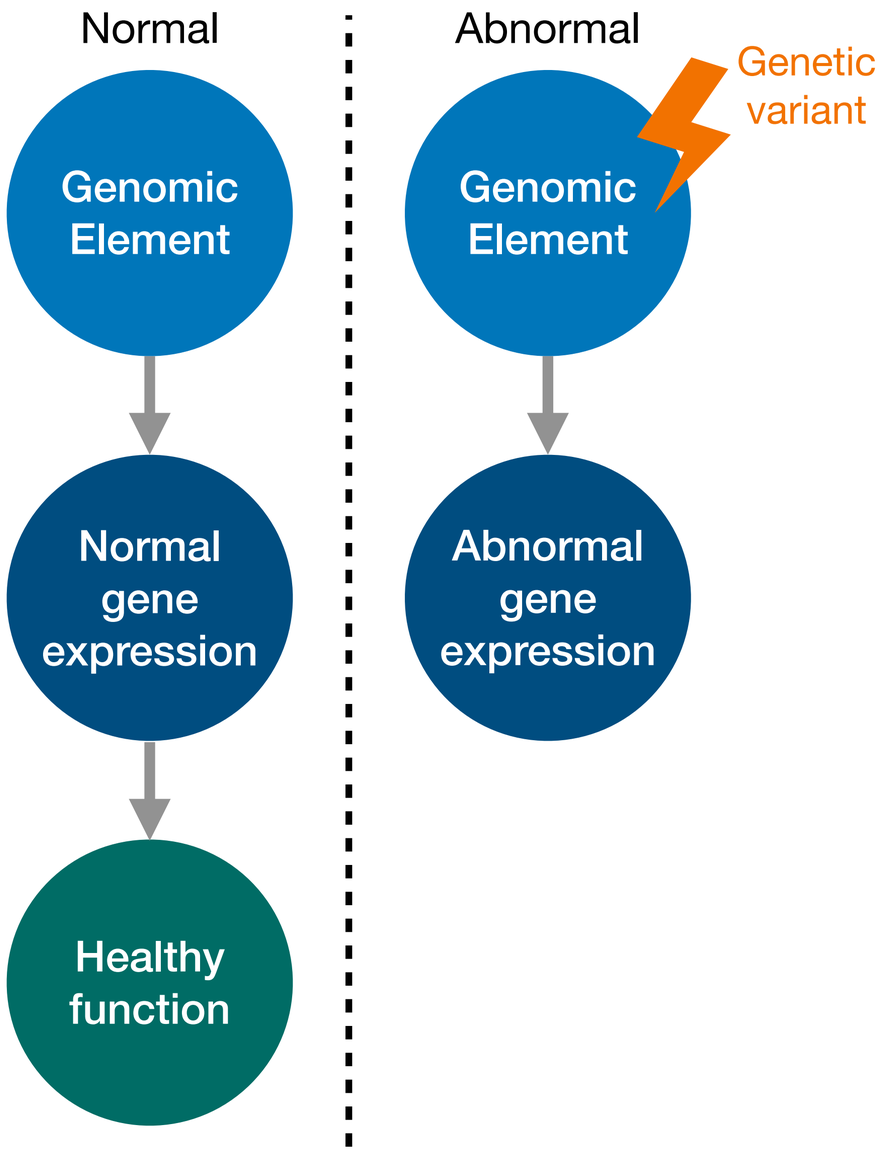
A model for genetic origins of human diseases



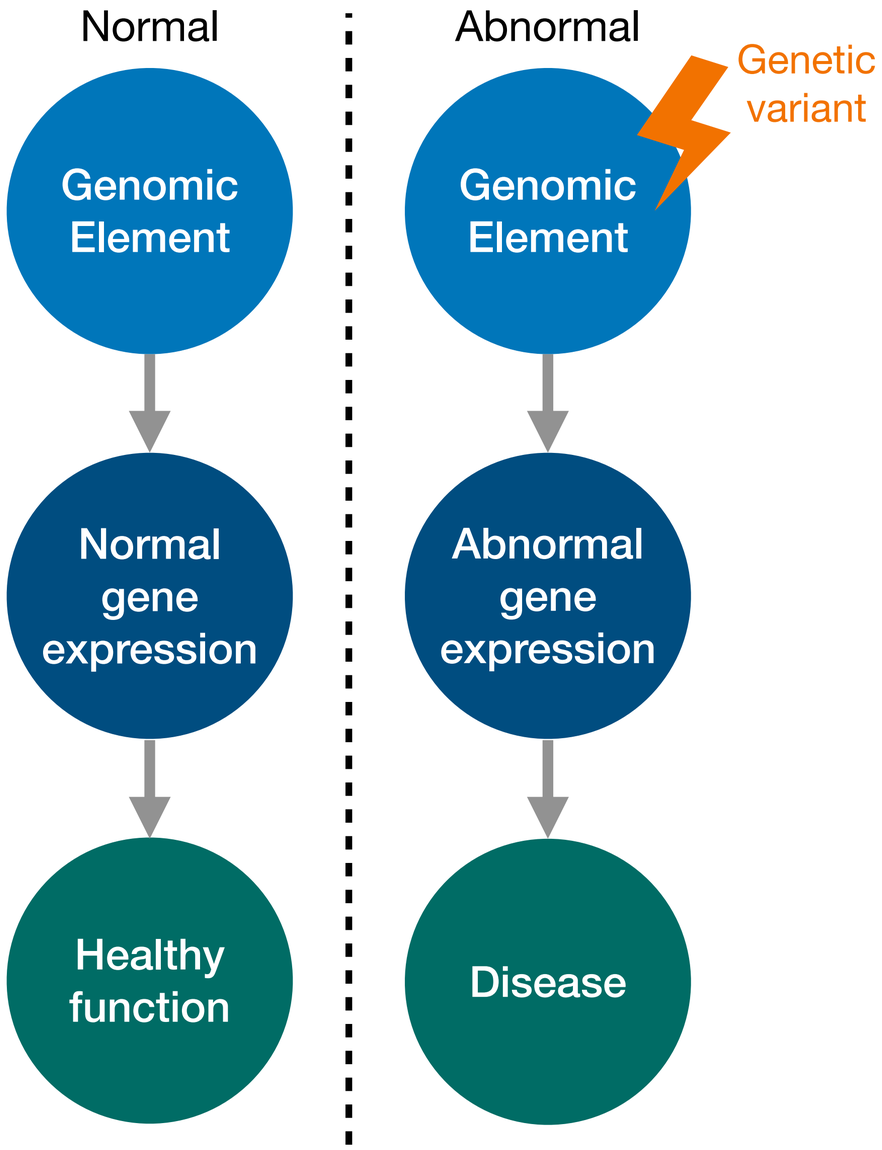
A model for genetic origins of human diseases



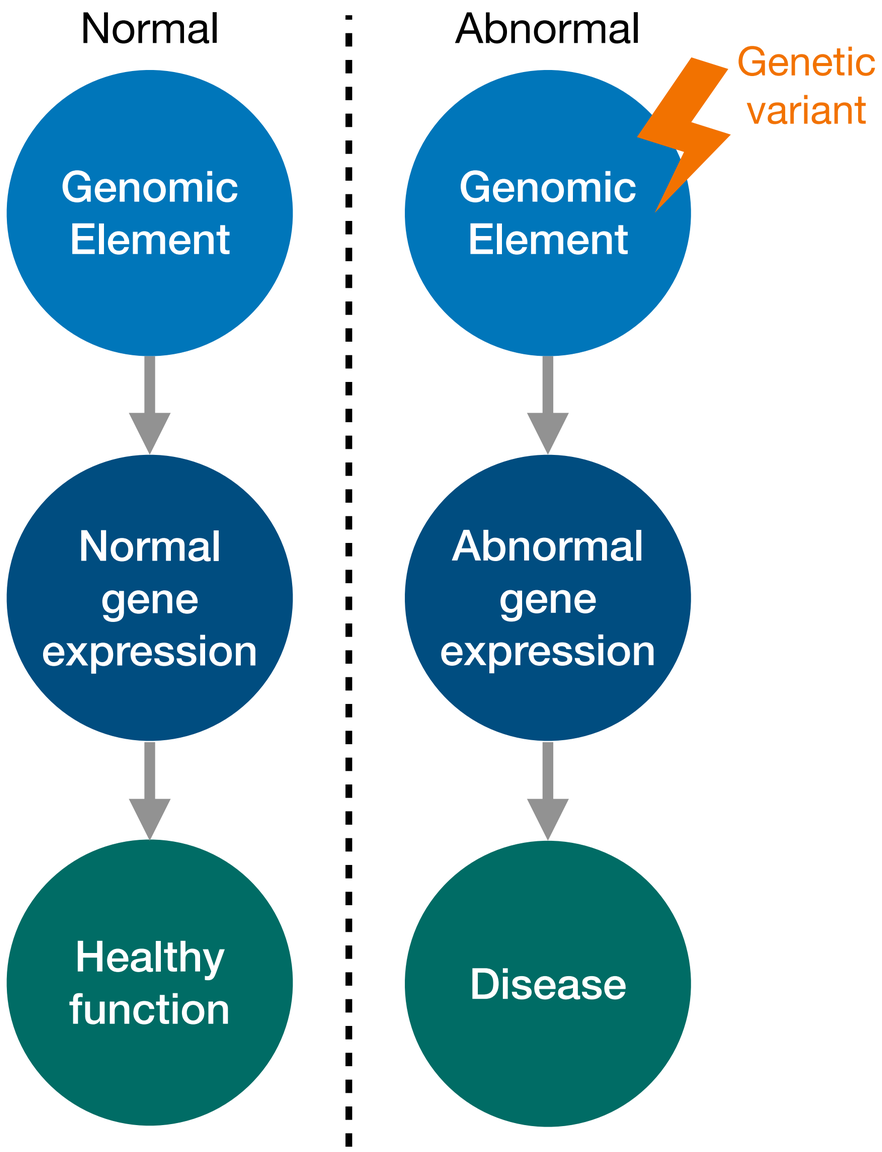
A model for genetic origins of human diseases

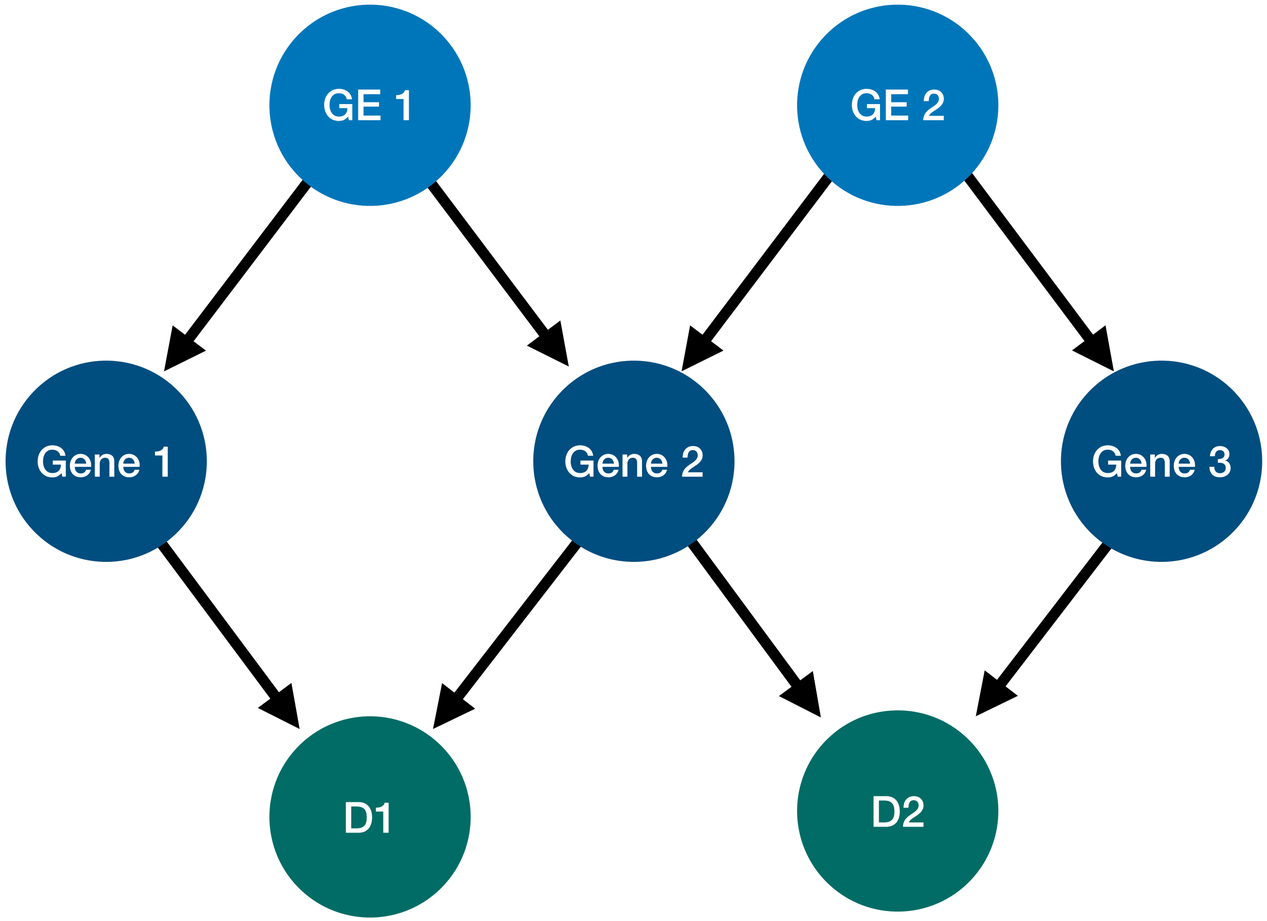


A model for genetic origins of human diseases

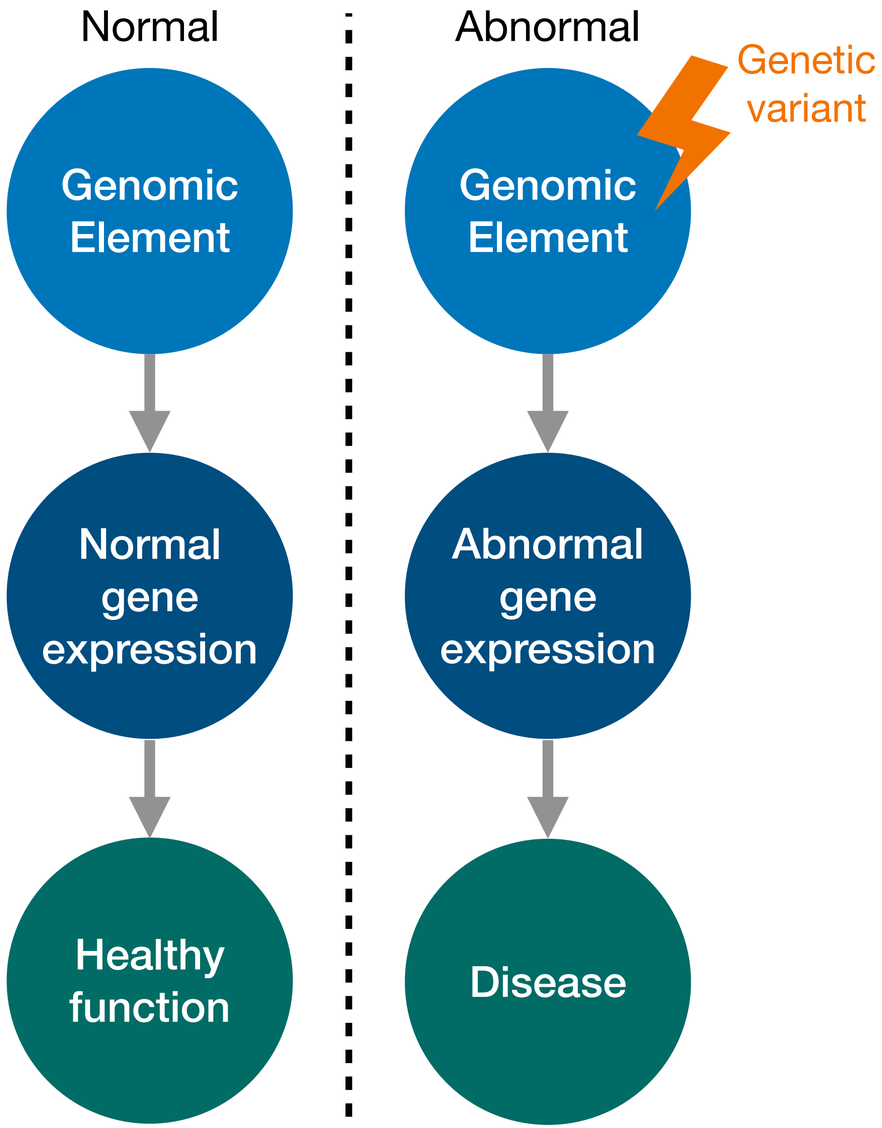


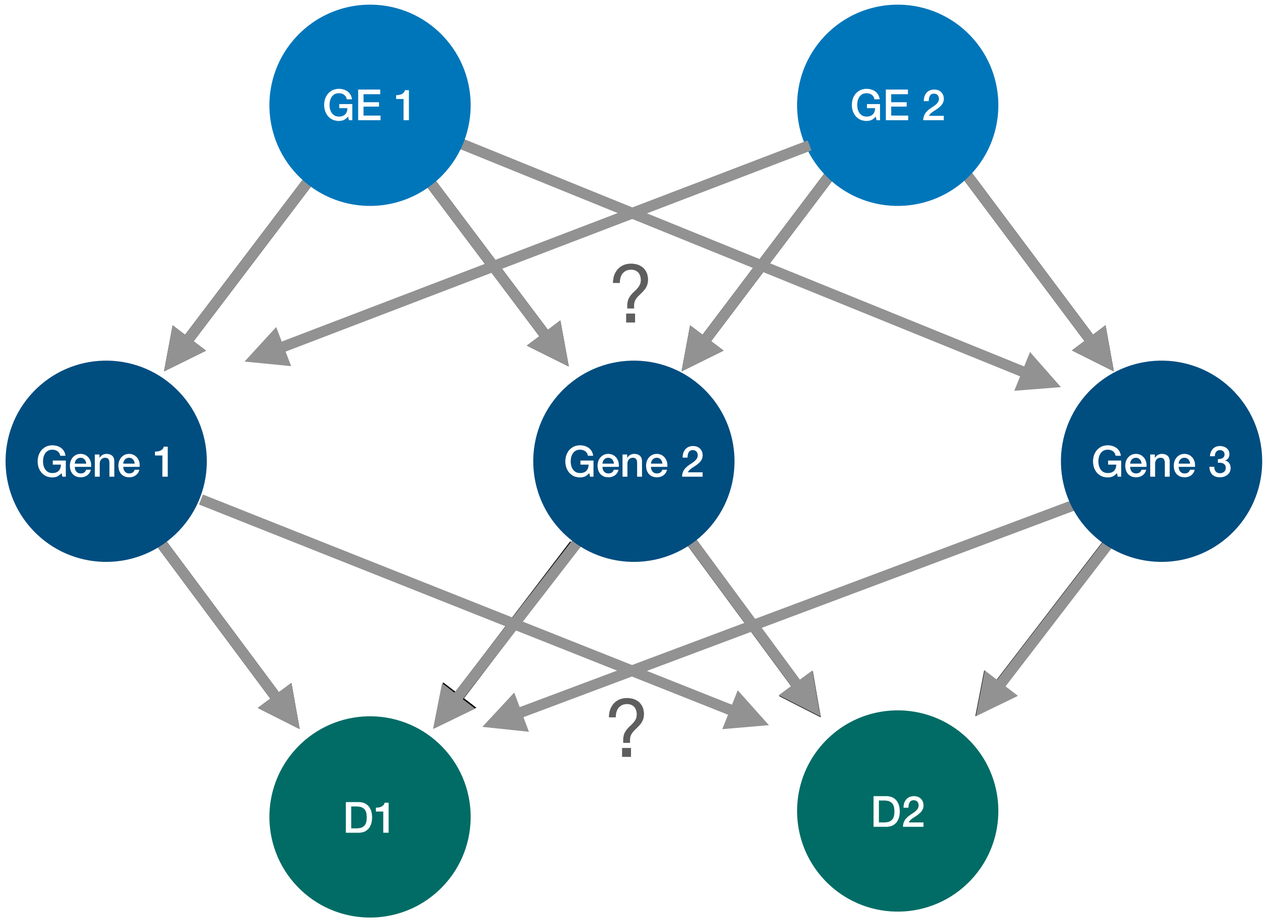
A model for genetic origins of human diseases



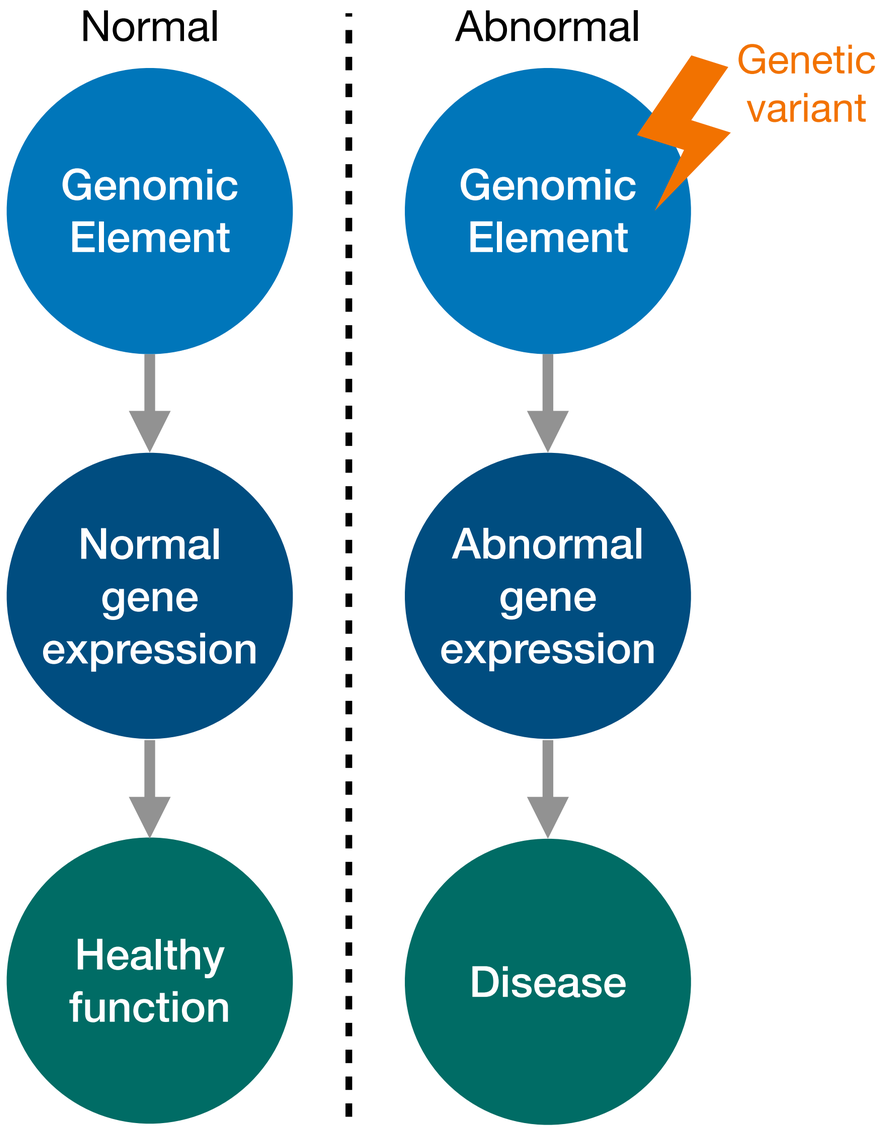


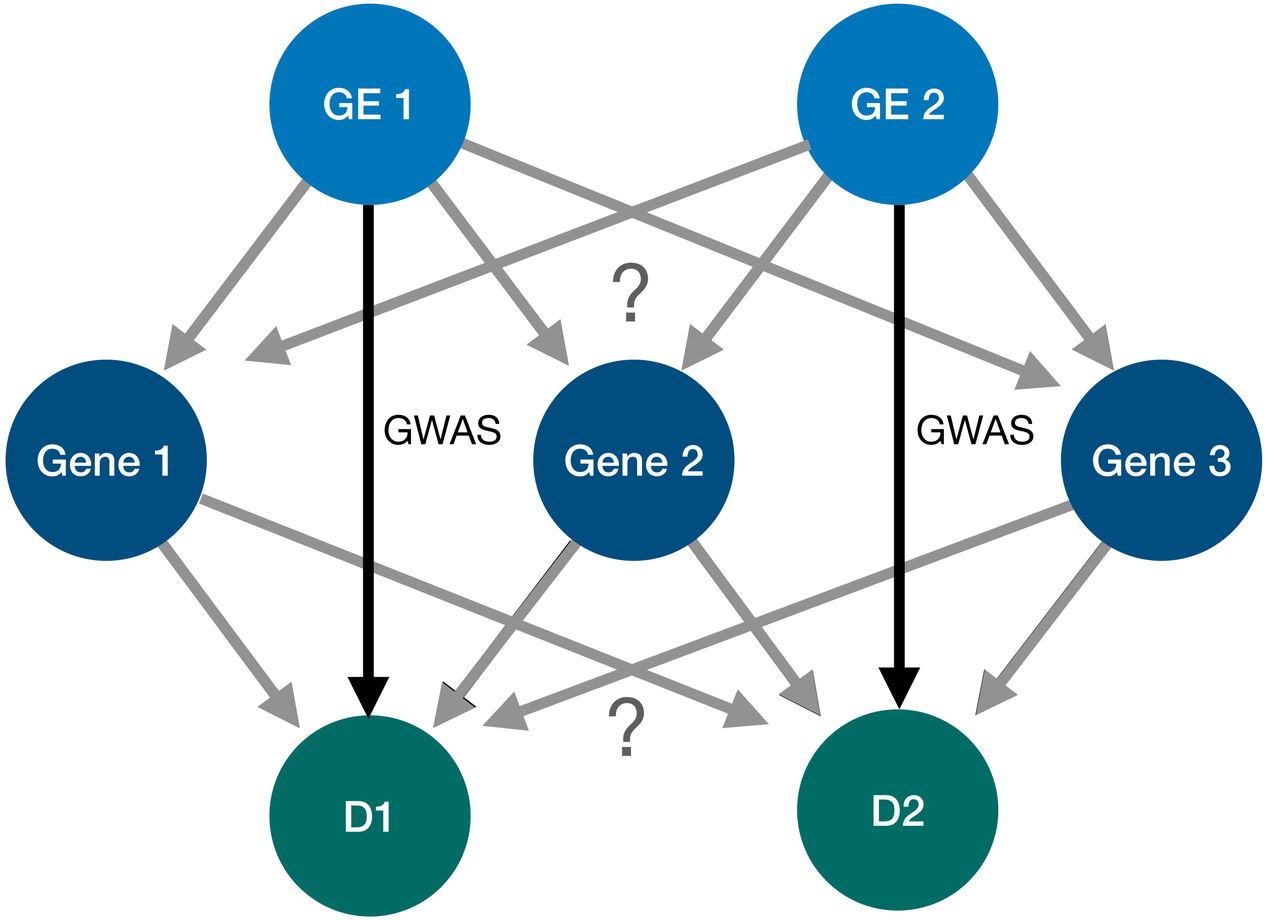
Uncovering genetic origins of human diseases



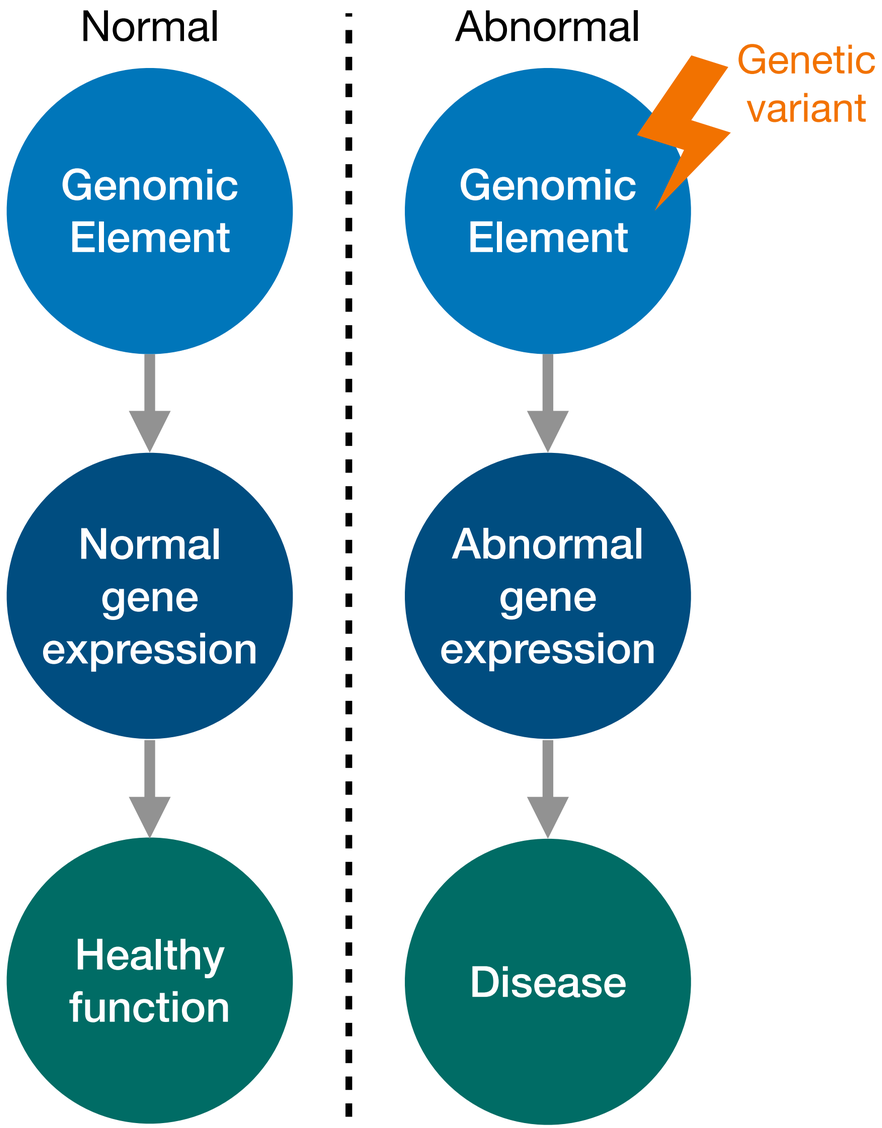


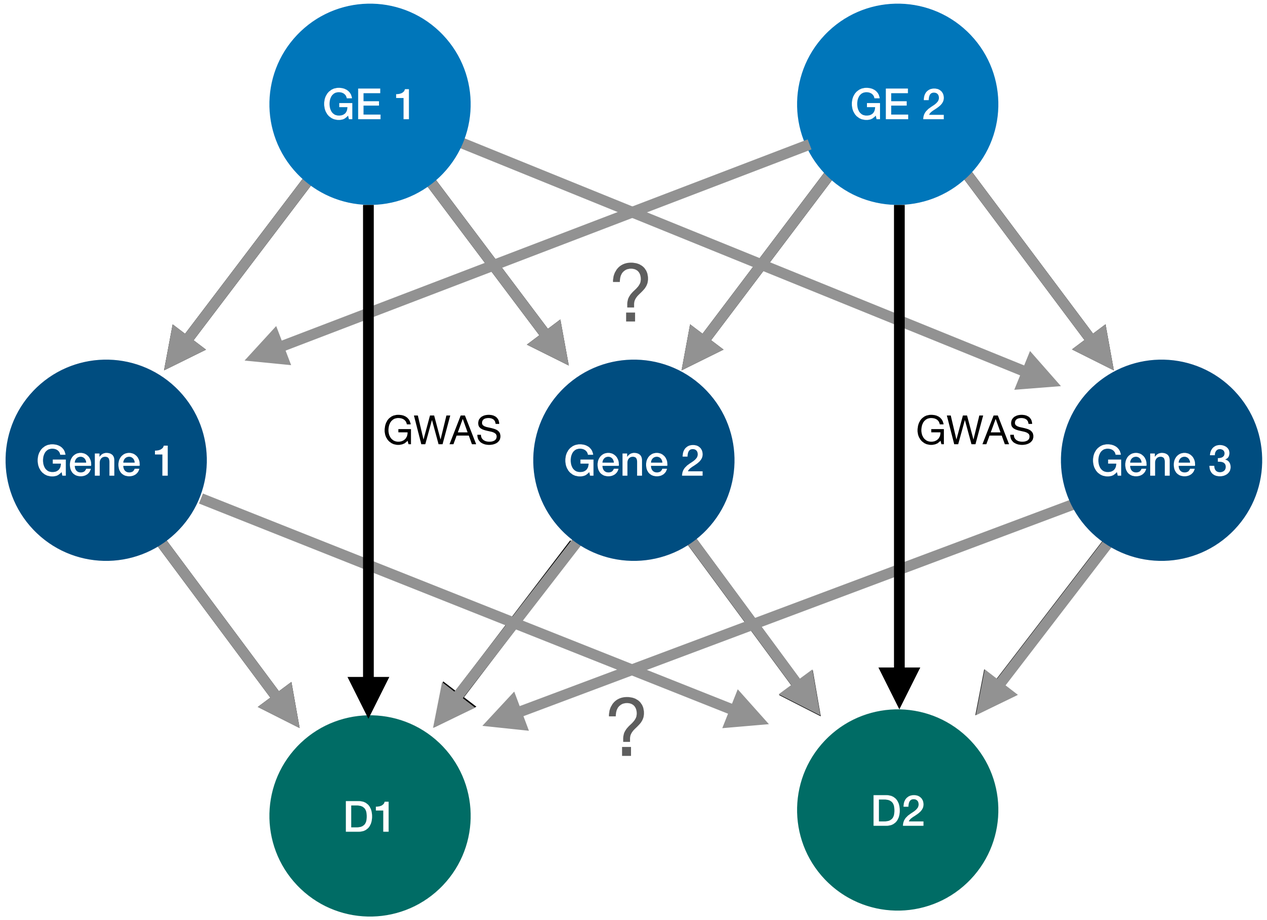
Uncovering genetic origins of human diseases





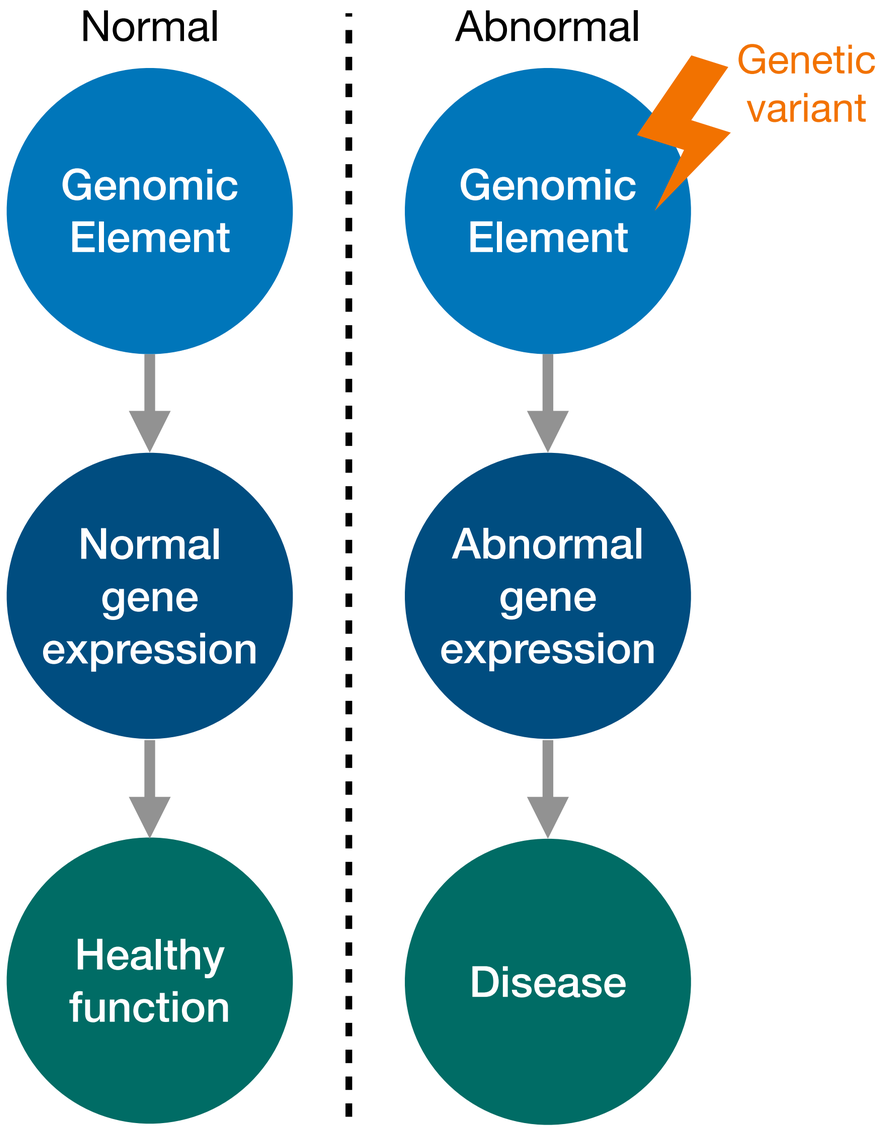
Uncovering genetic origins of human diseases

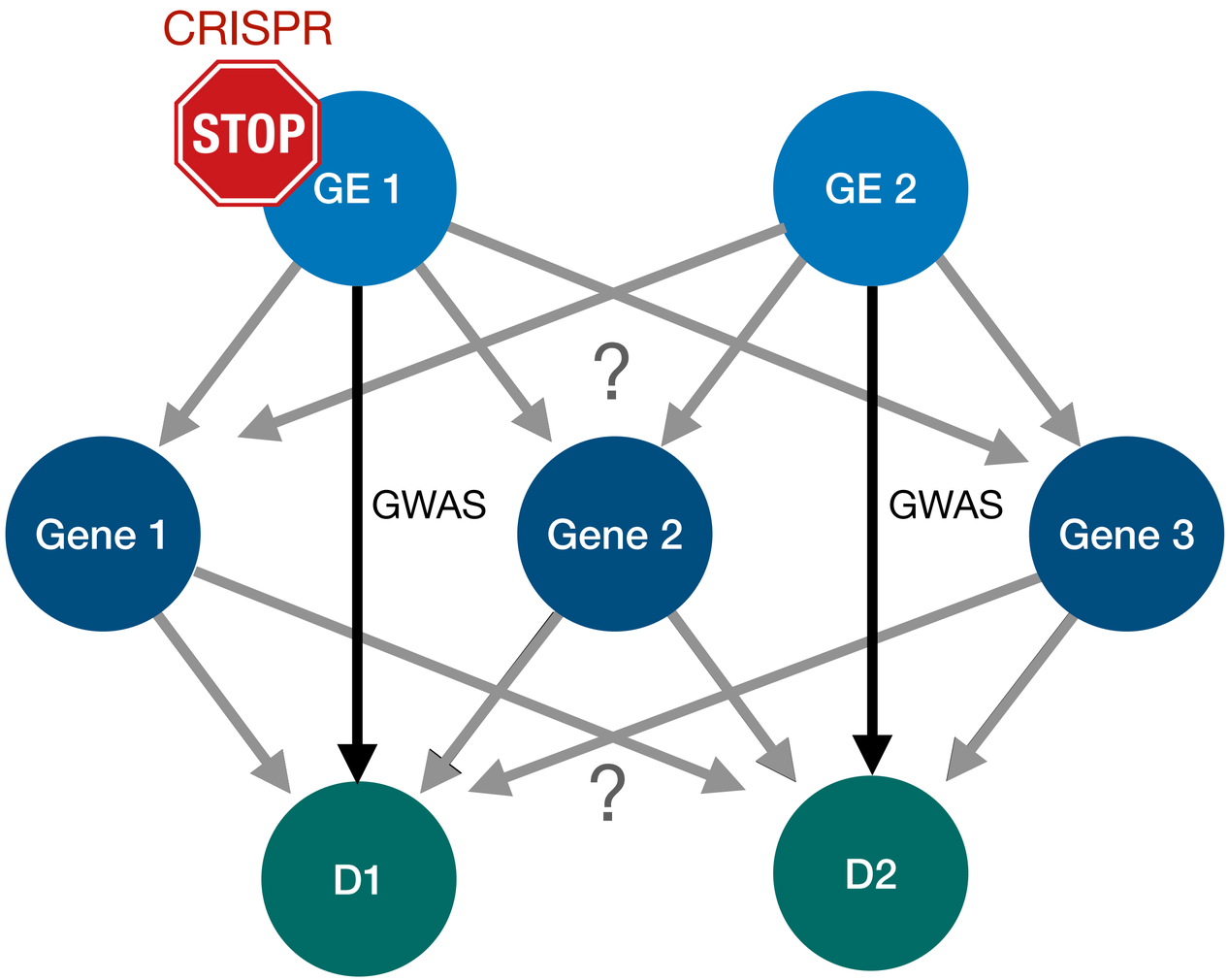




Single-cell CRISPR screens

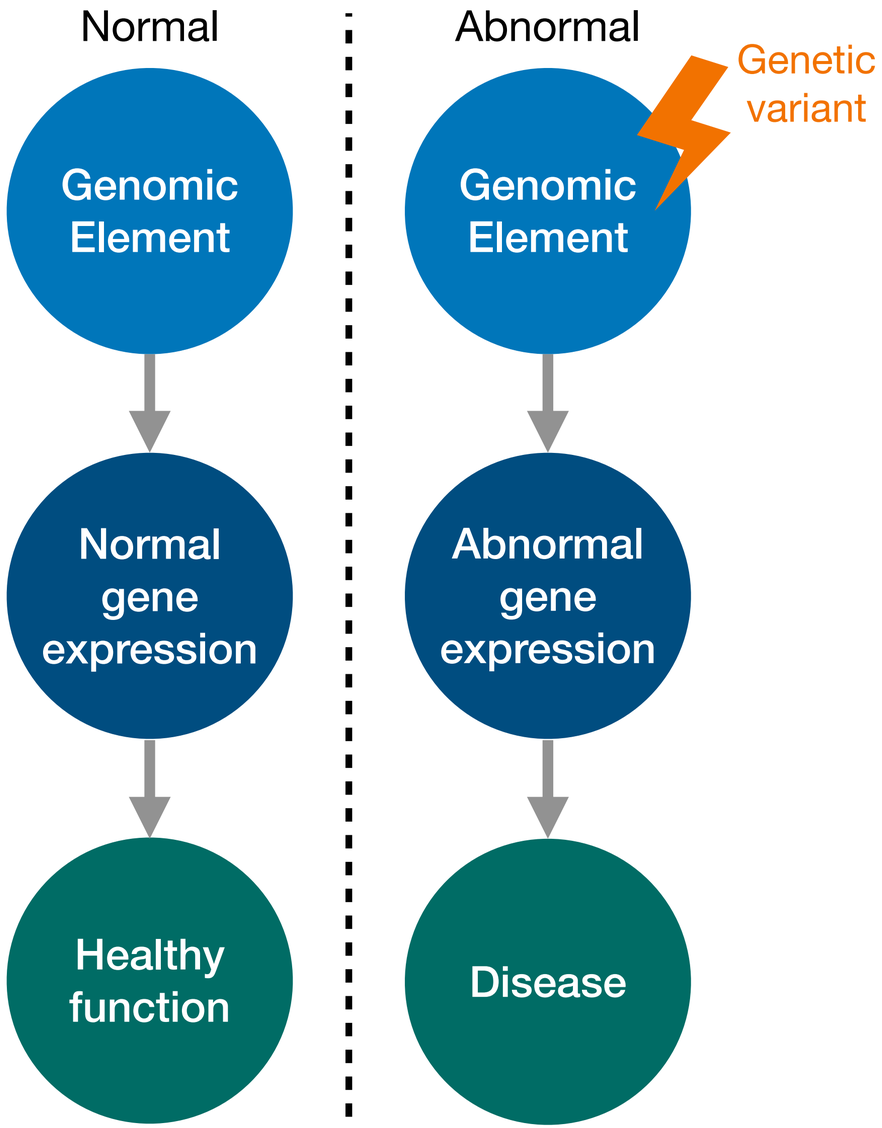
Uncovering genetic origins of human diseases

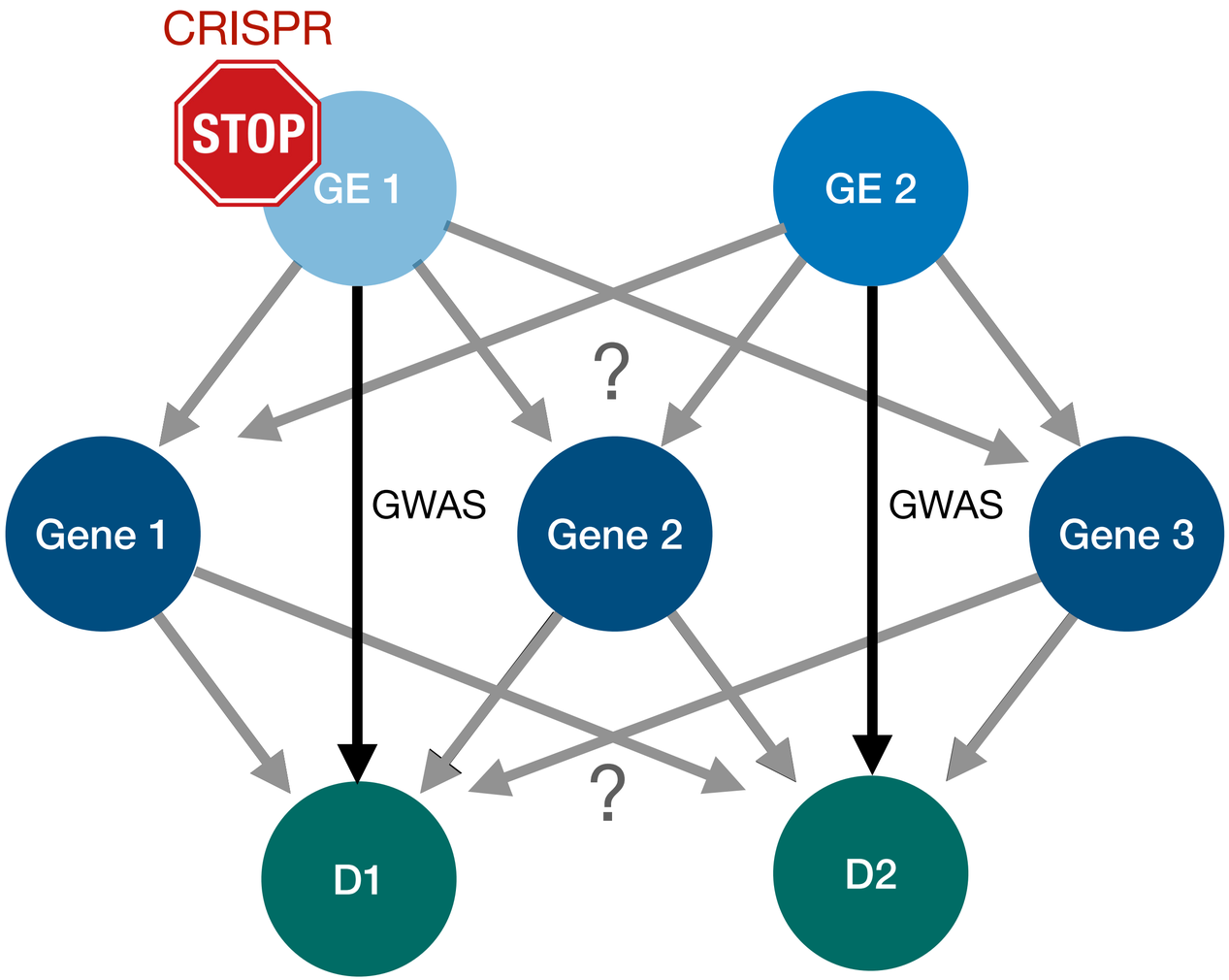




Single-cell CRISPR screens

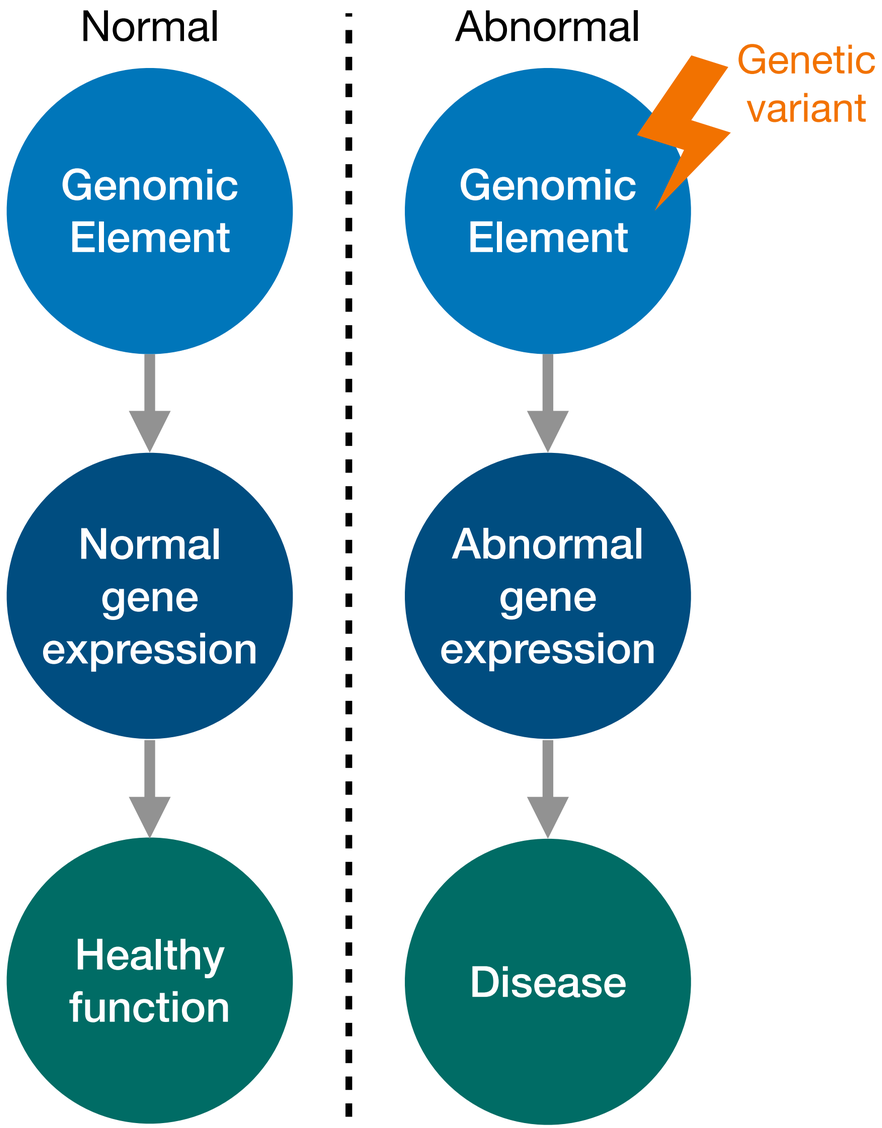
Uncovering genetic origins of human diseases

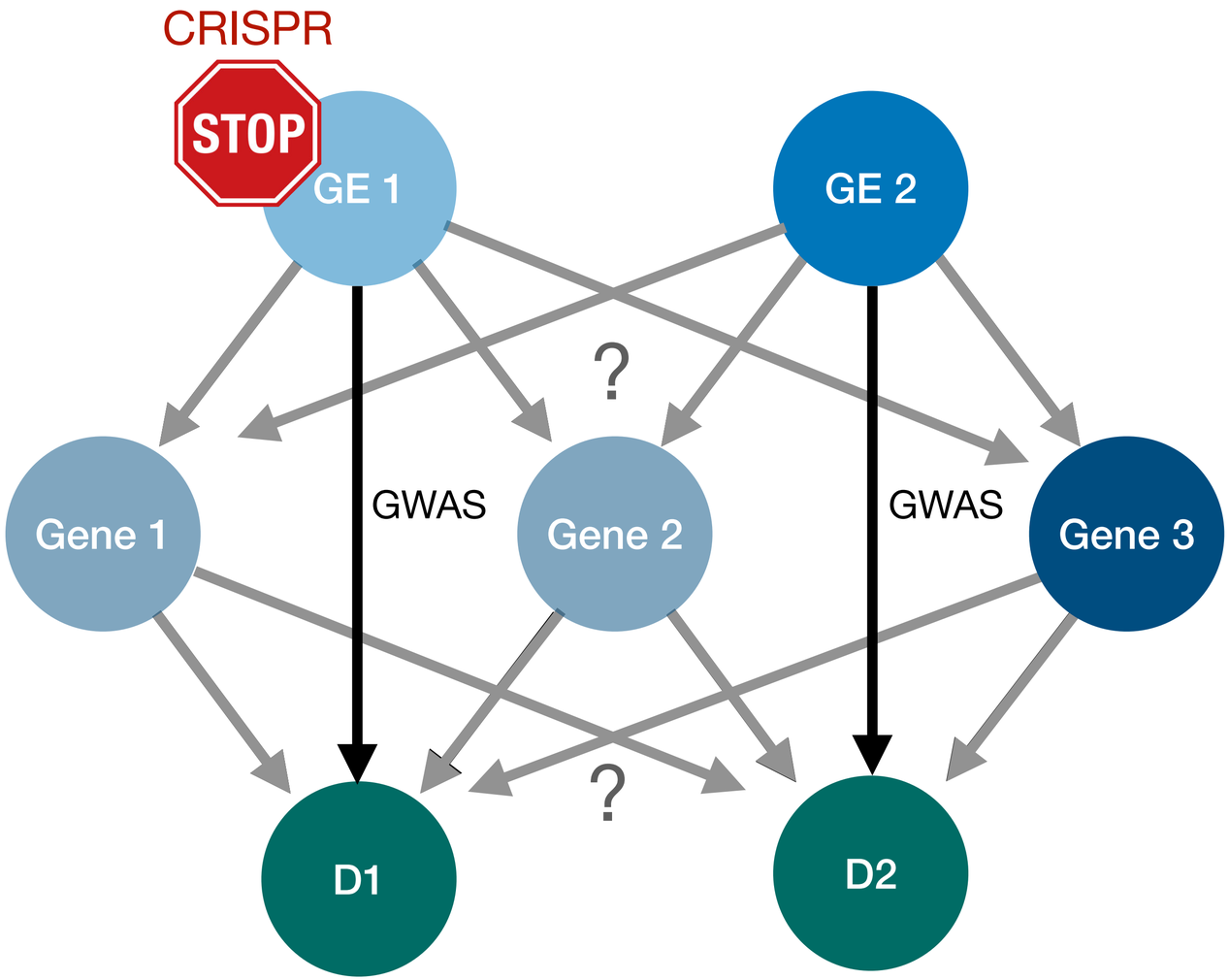




Single-cell CRISPR screens

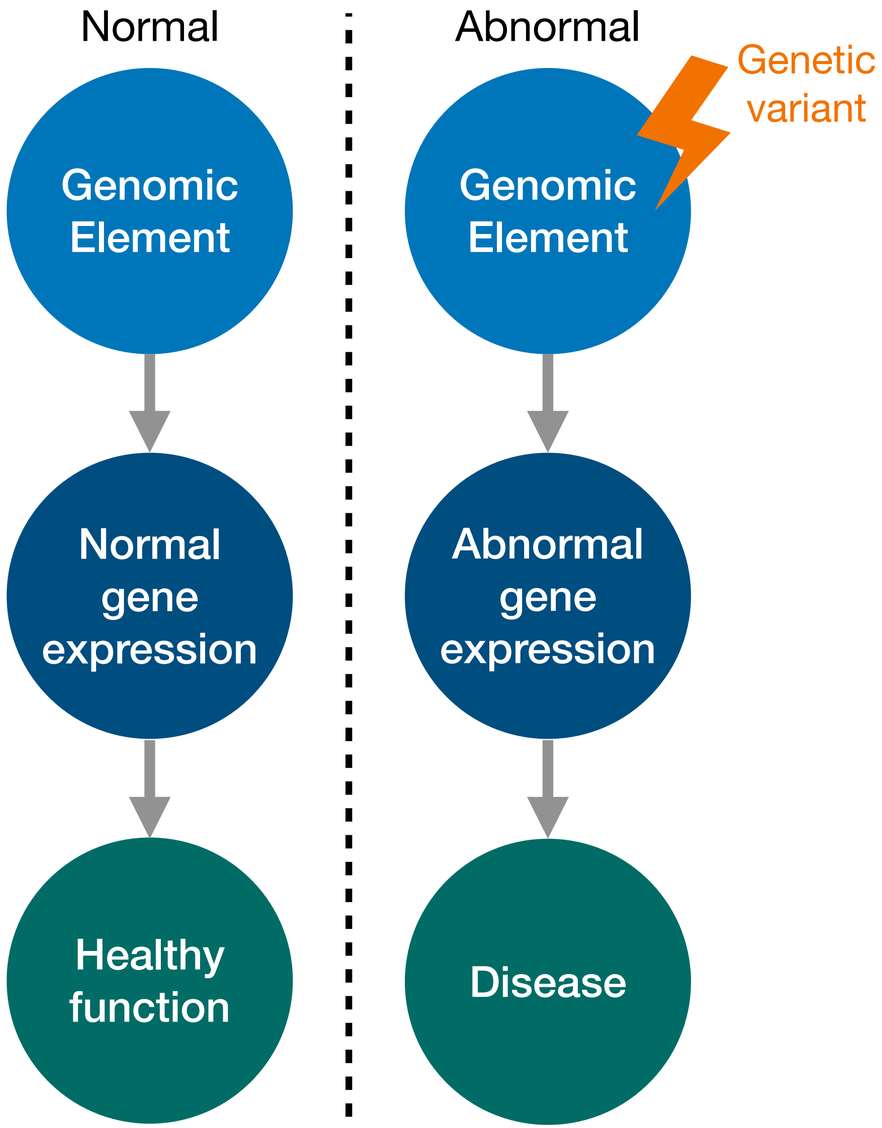
Uncovering genetic origins of human diseases

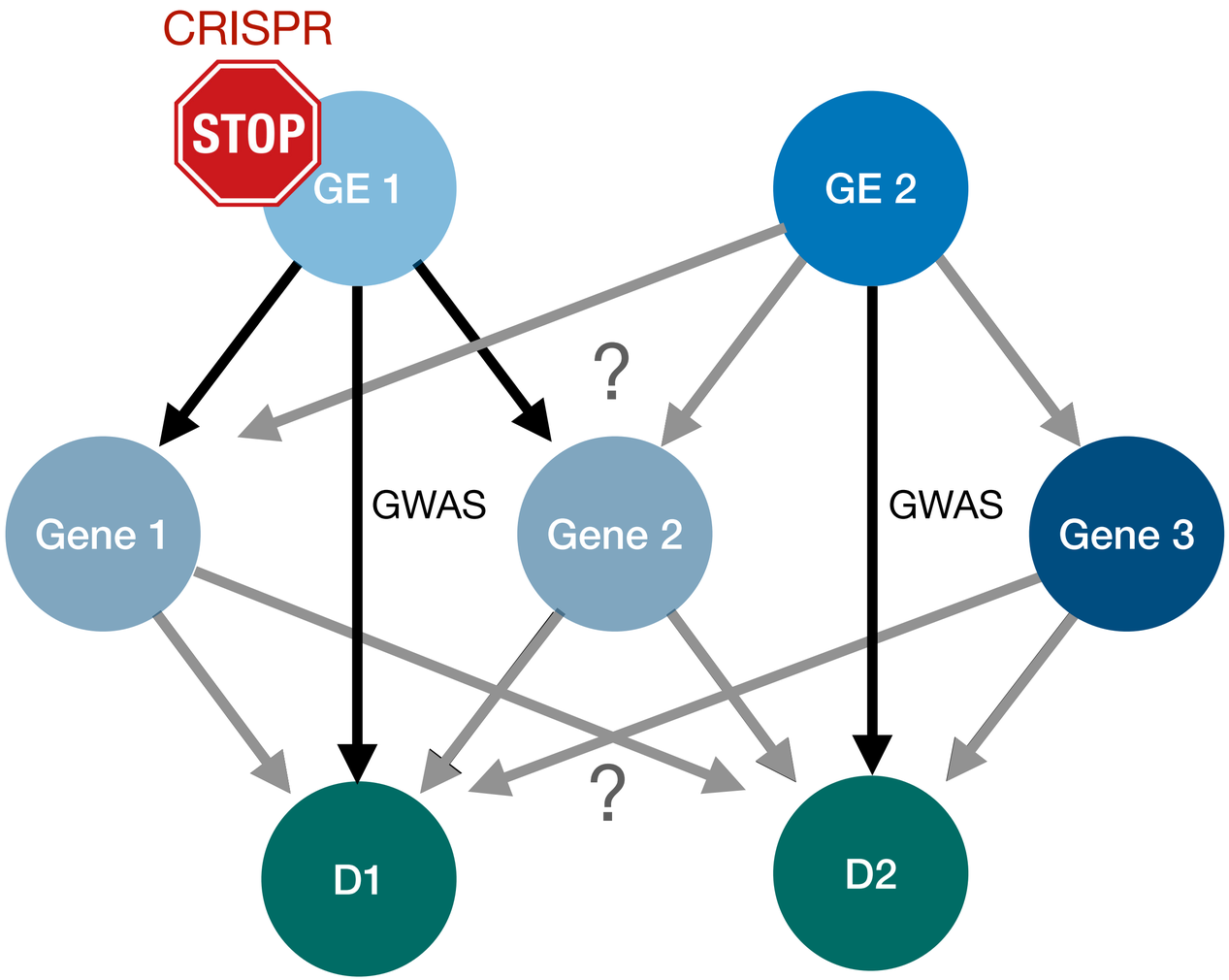




Single-cell CRISPR screens

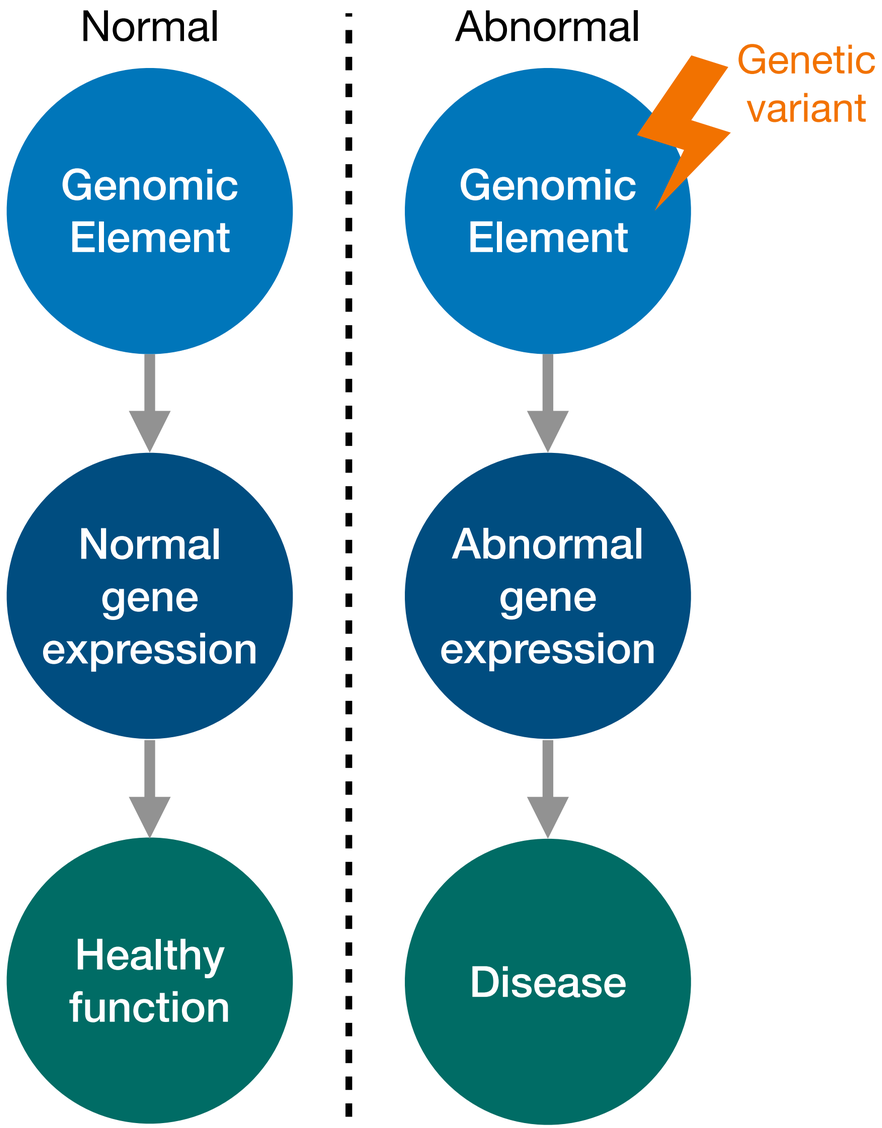
Uncovering genetic origins of human diseases

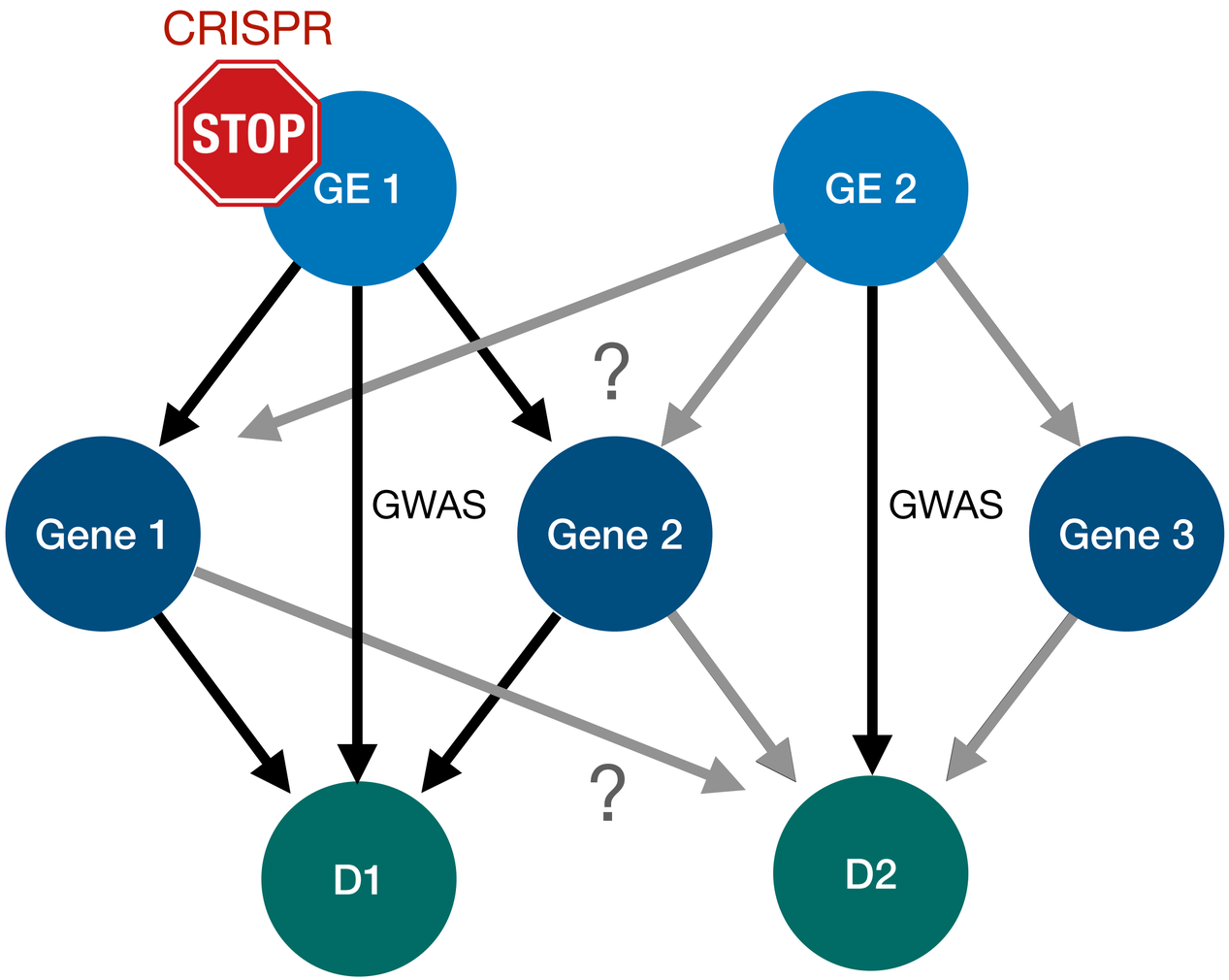




Single-cell CRISPR screens

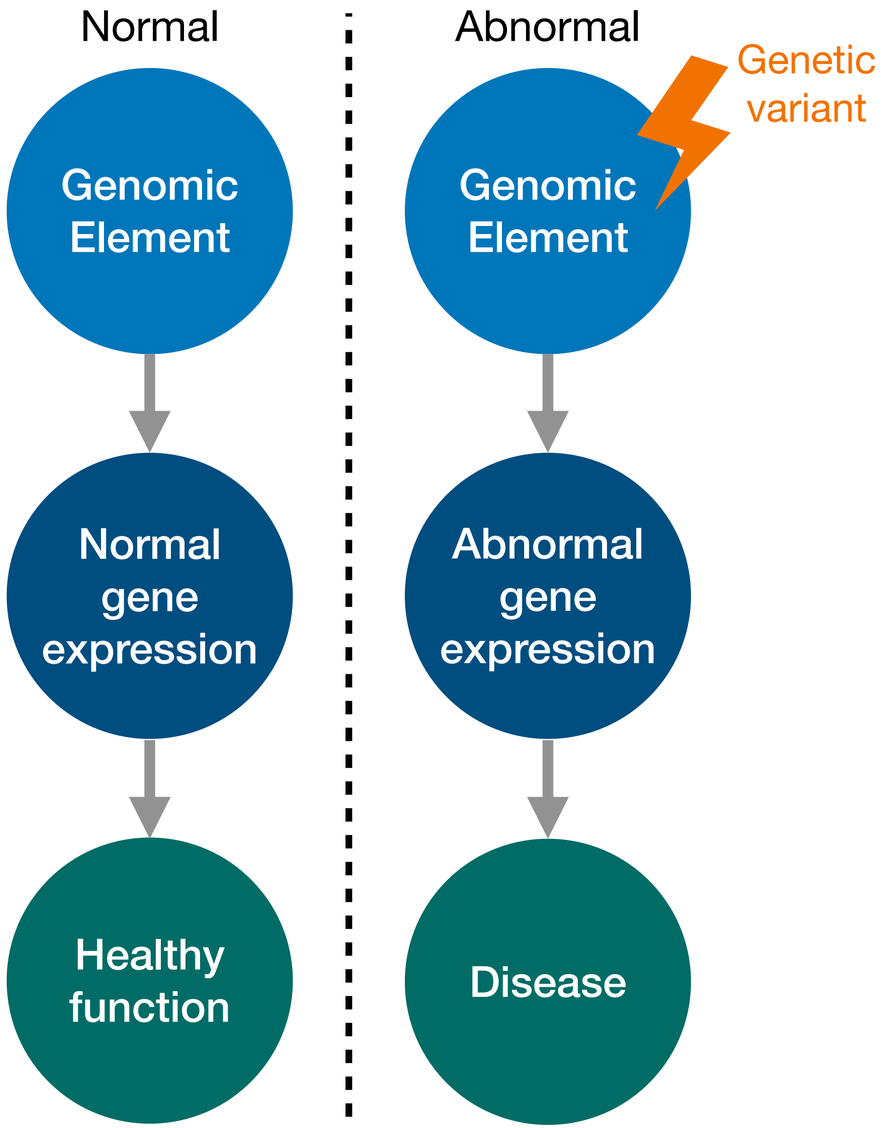
Uncovering genetic origins of human diseases

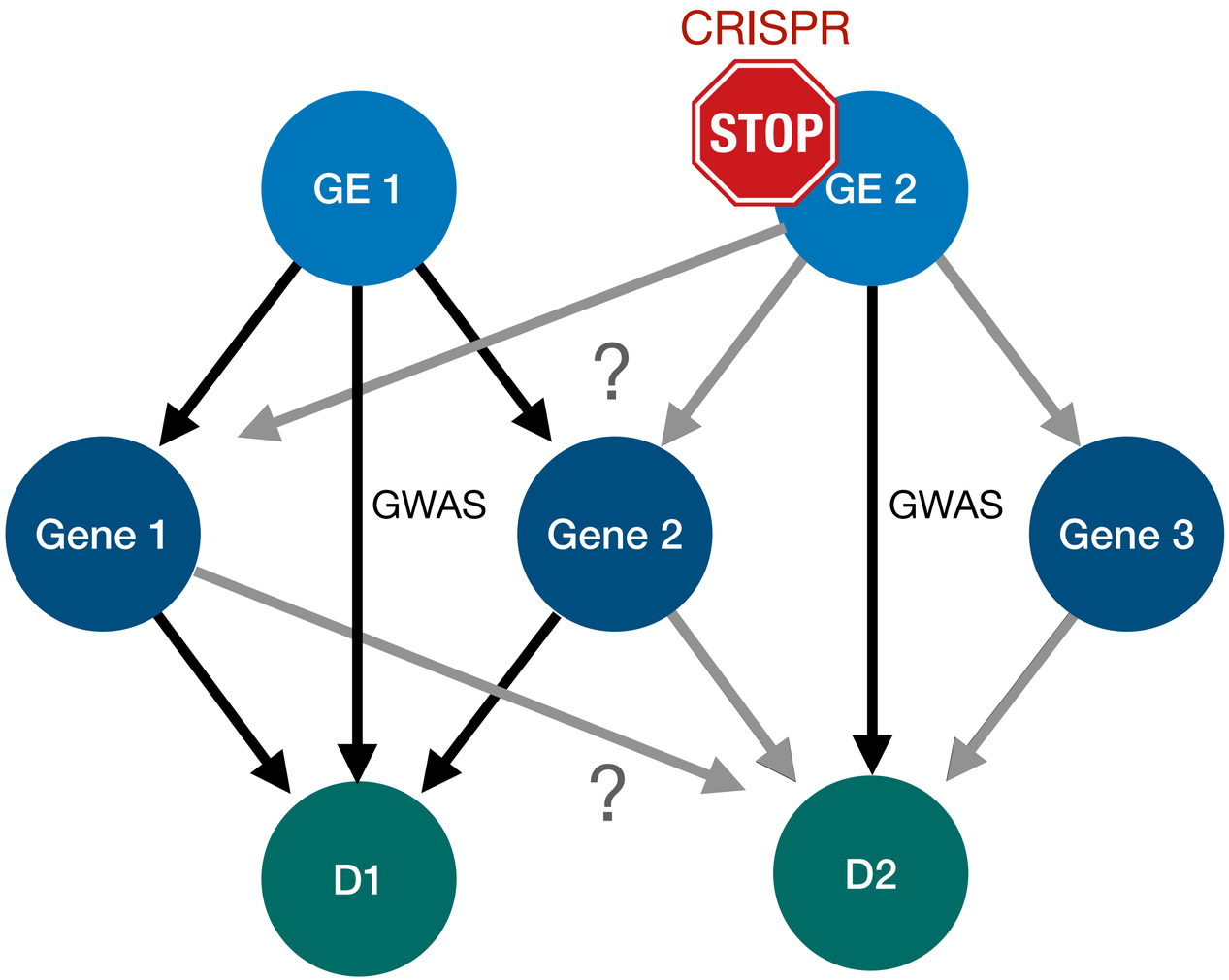




Single-cell CRISPR screens

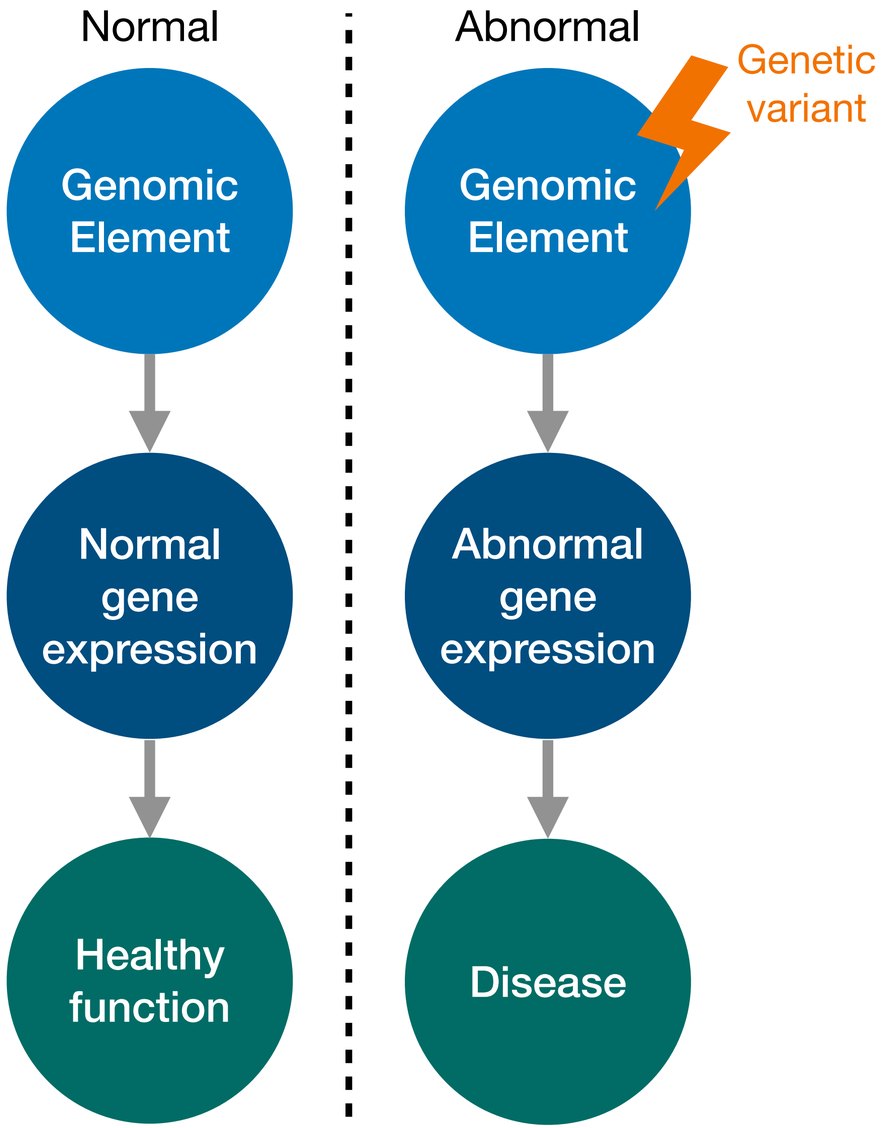
Uncovering genetic origins of human diseases

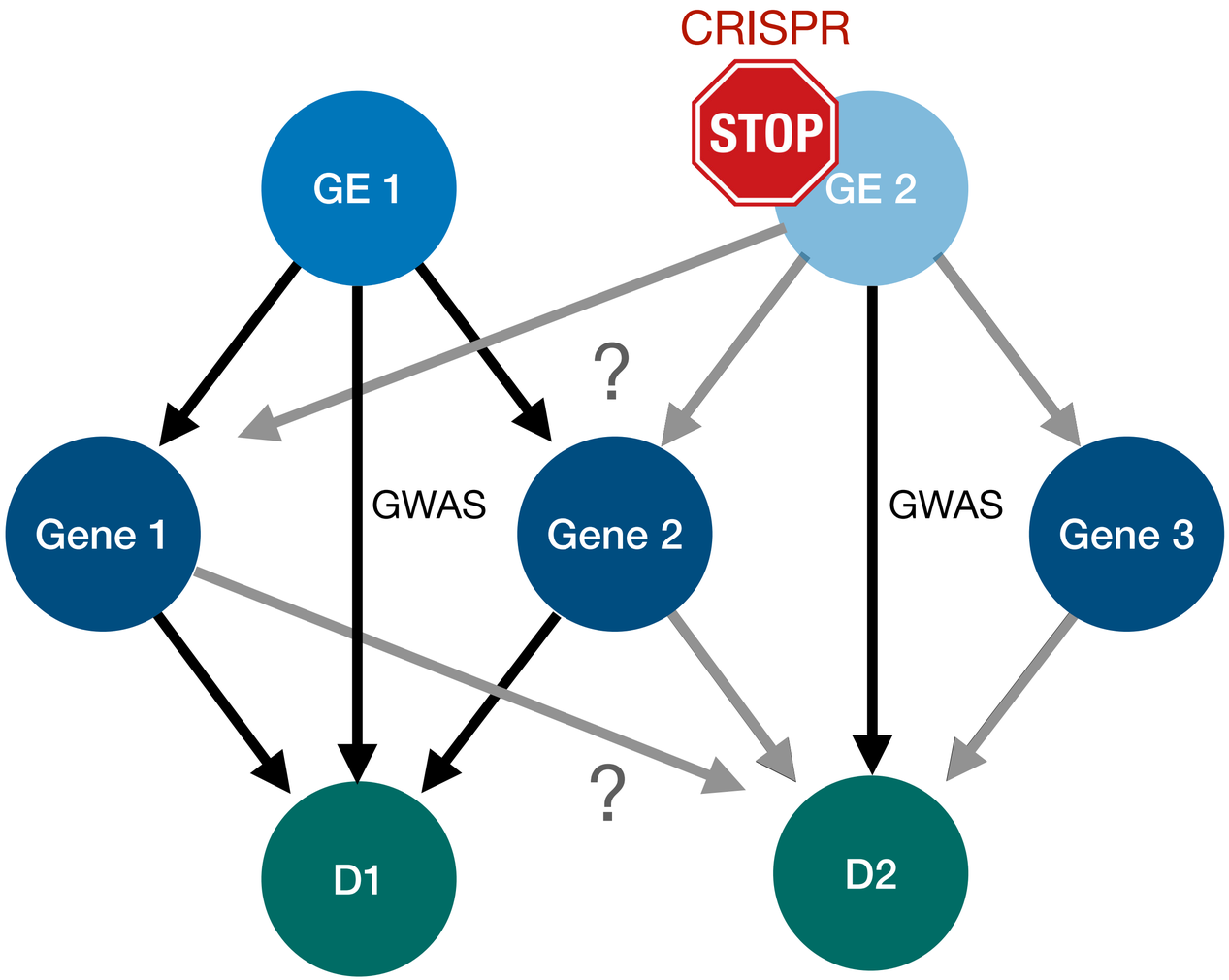




Single-cell CRISPR screens

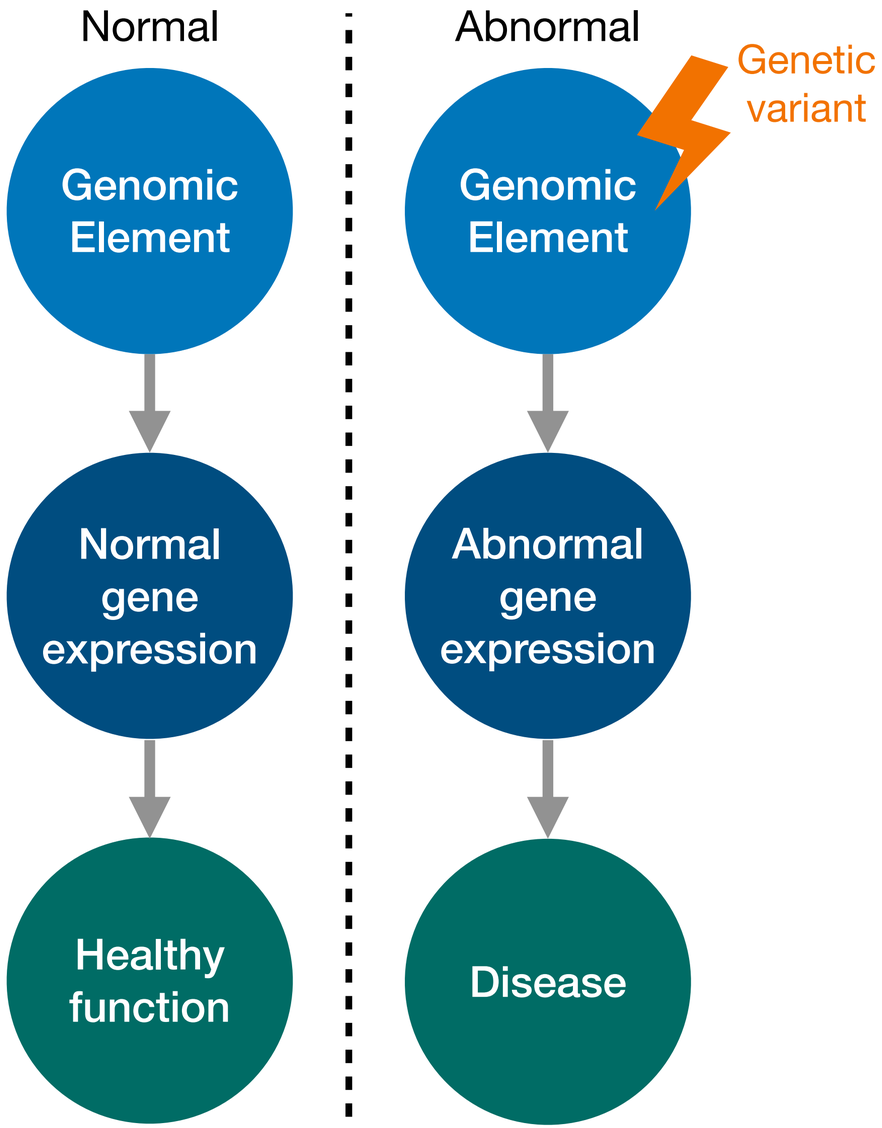
Uncovering genetic origins of human diseases

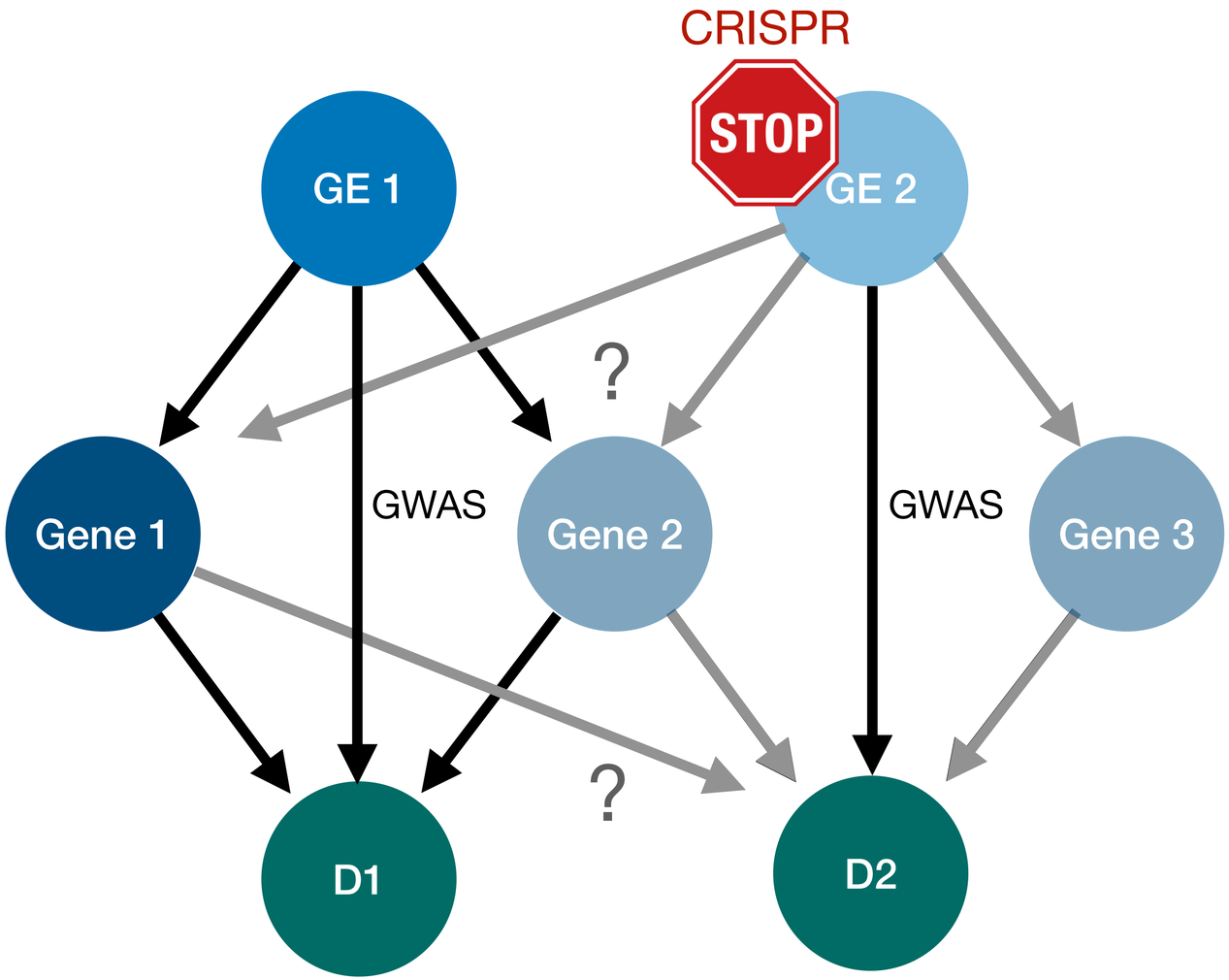




Single-cell CRISPR screens

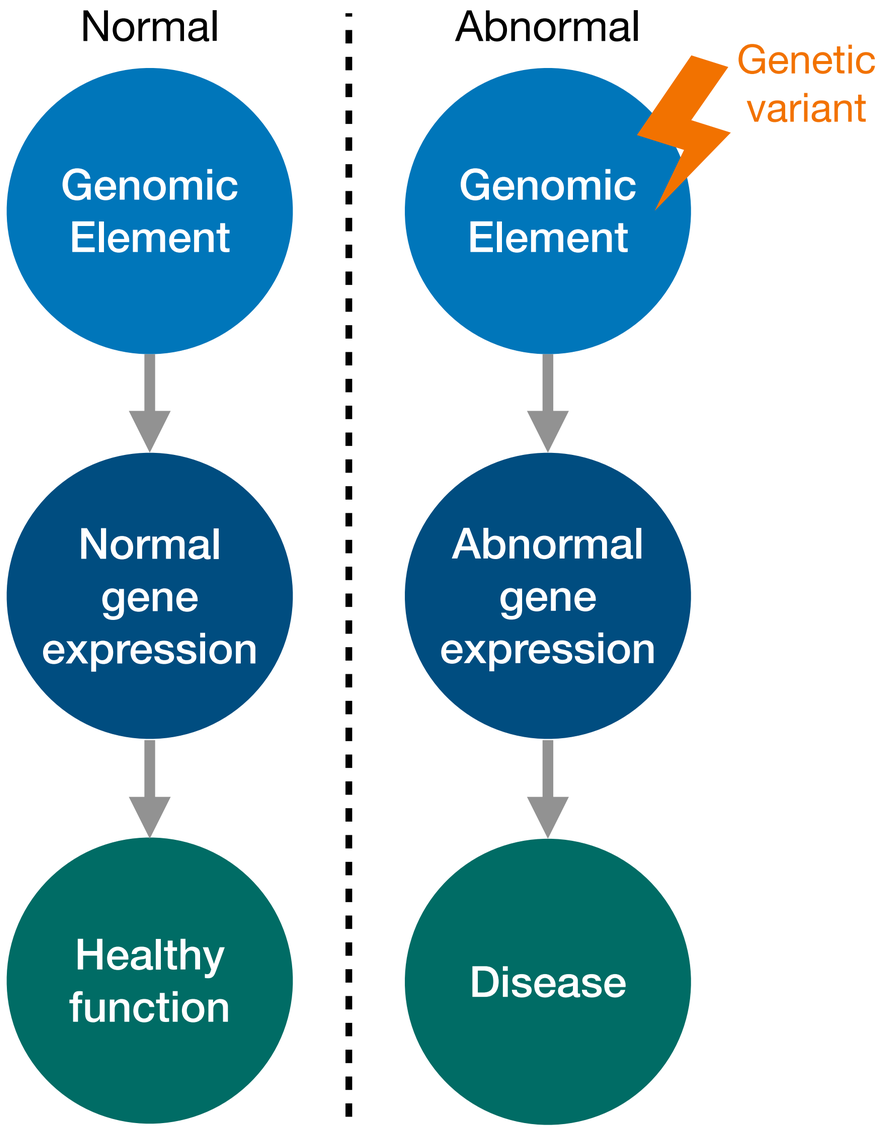
Uncovering genetic origins of human diseases

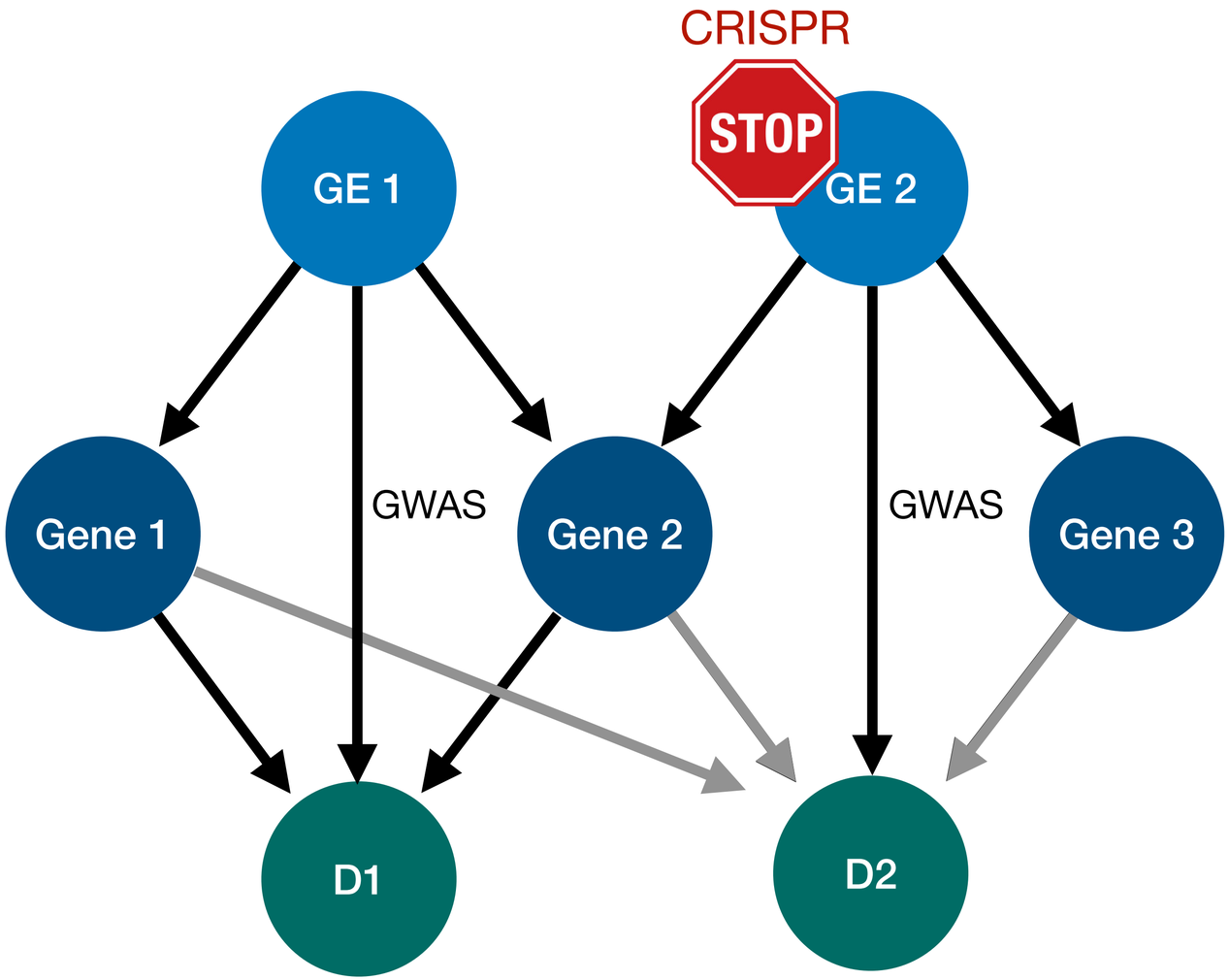




Single-cell CRISPR screens

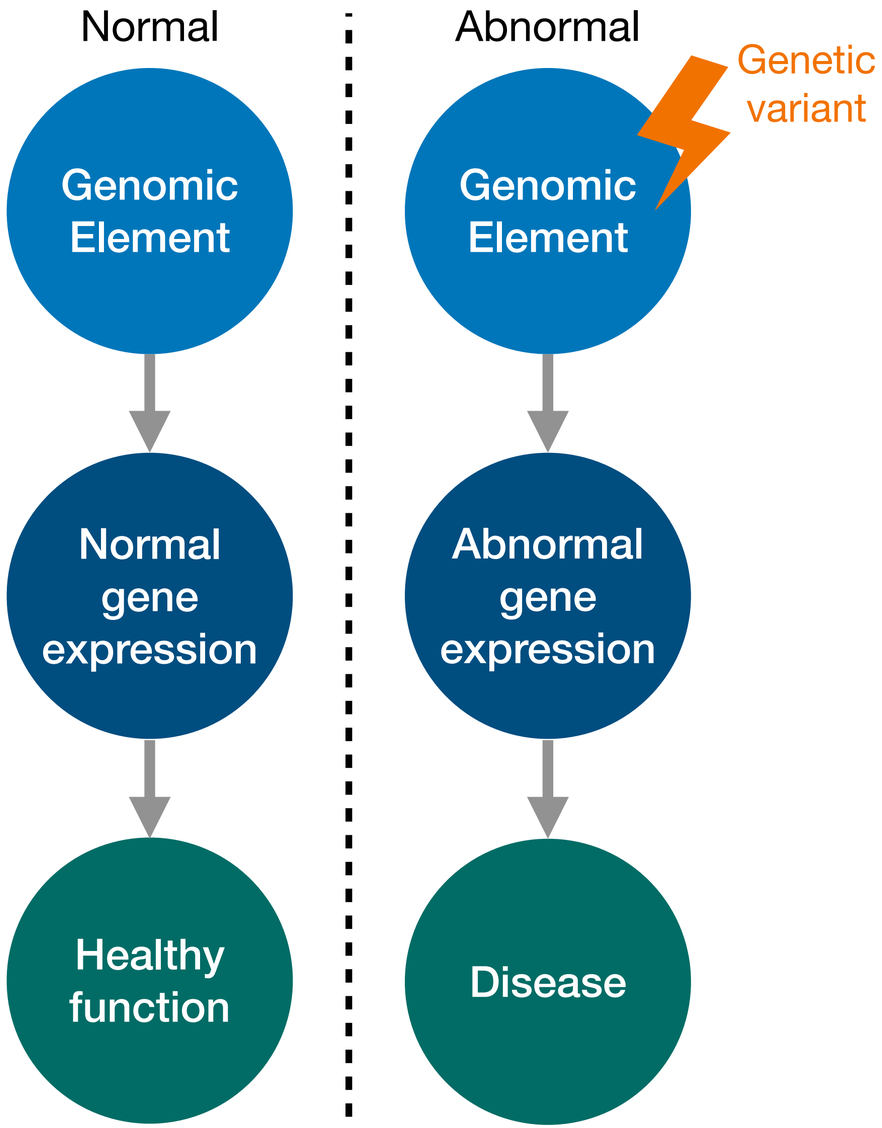
Uncovering genetic origins of human diseases

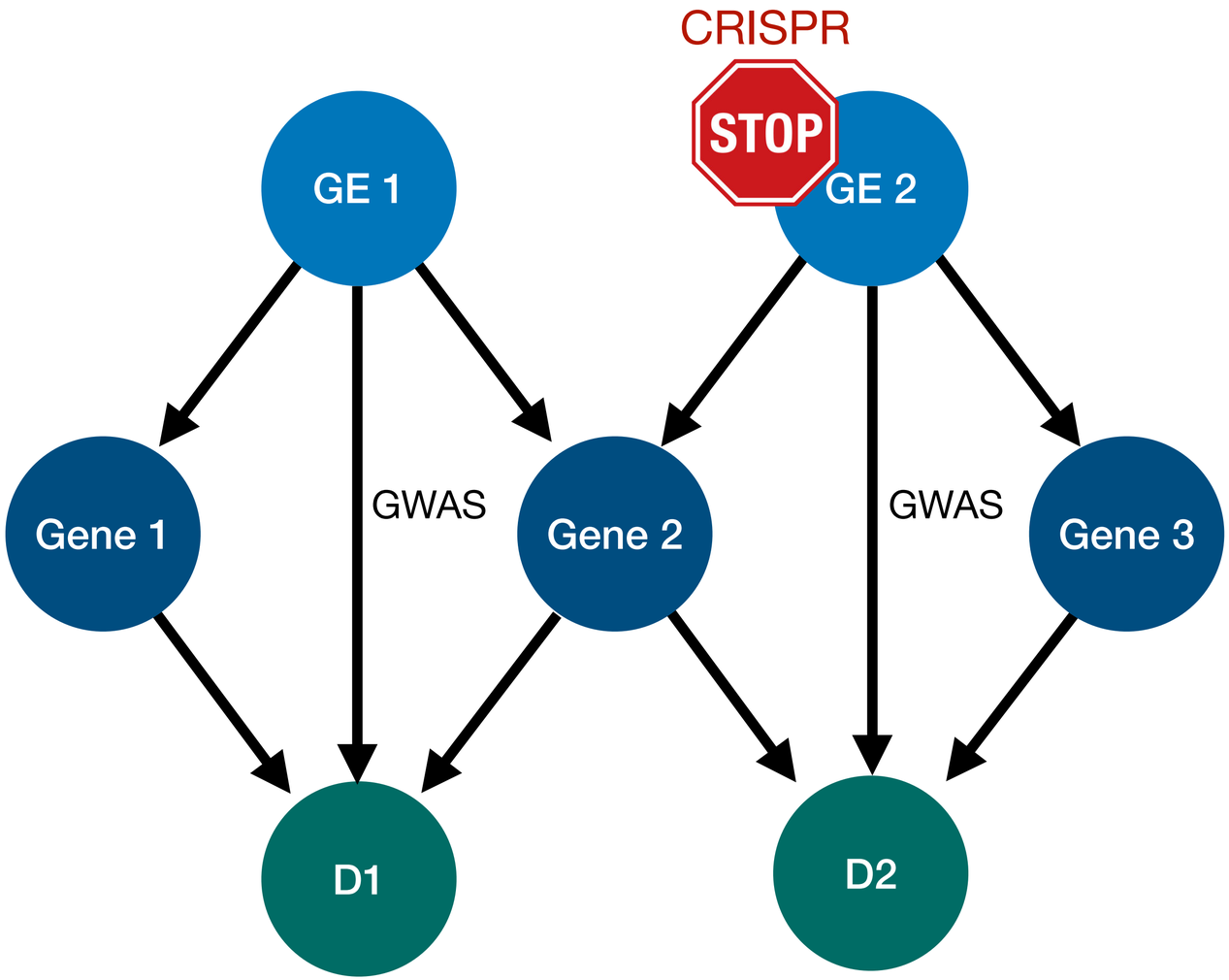




Single-cell CRISPR screens

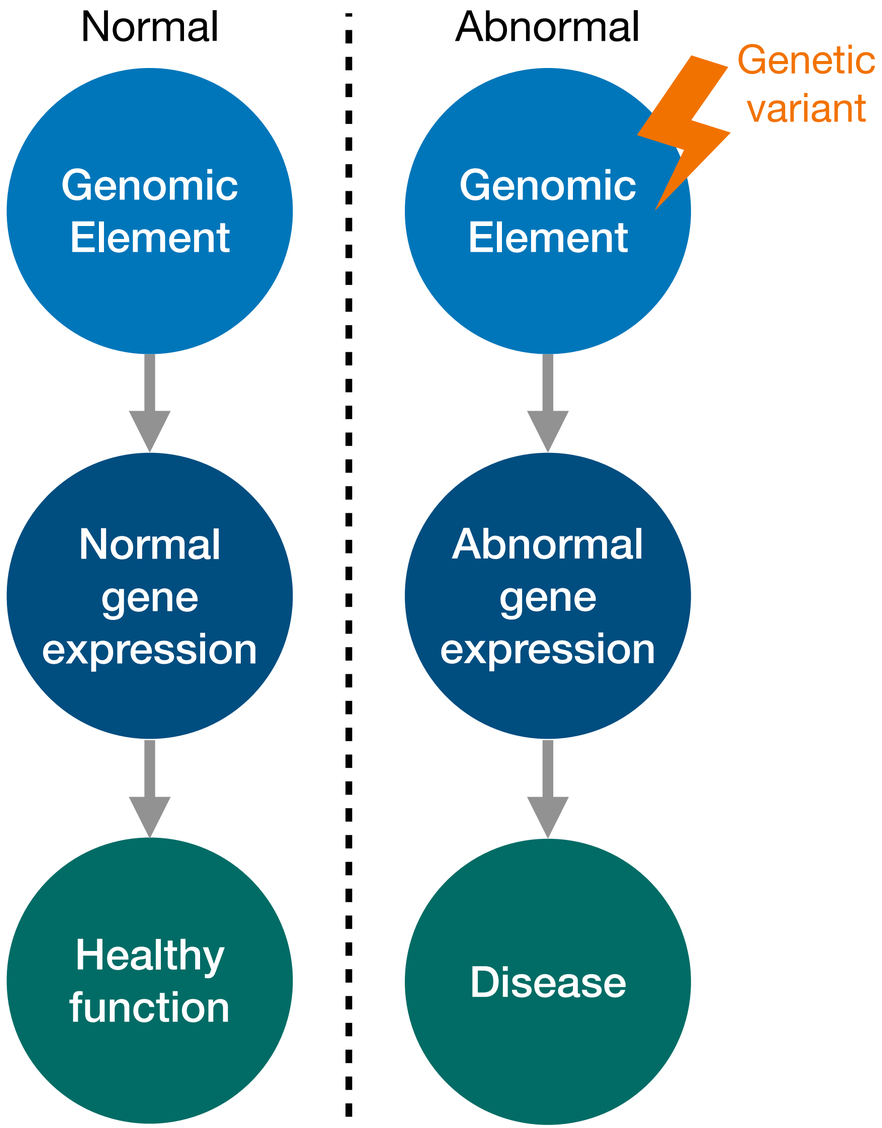
Uncovering genetic origins of human diseases

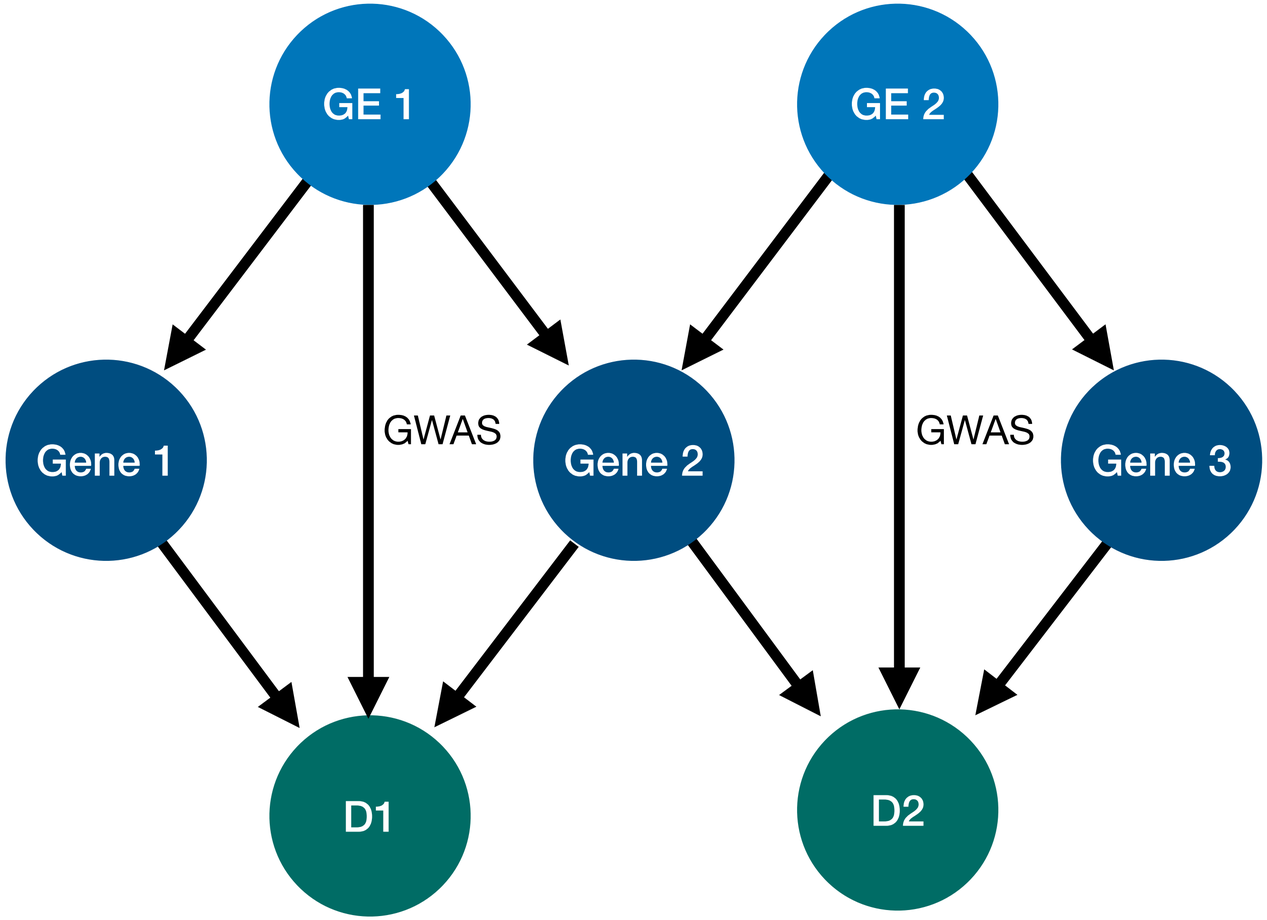




Single-cell CRISPR screens

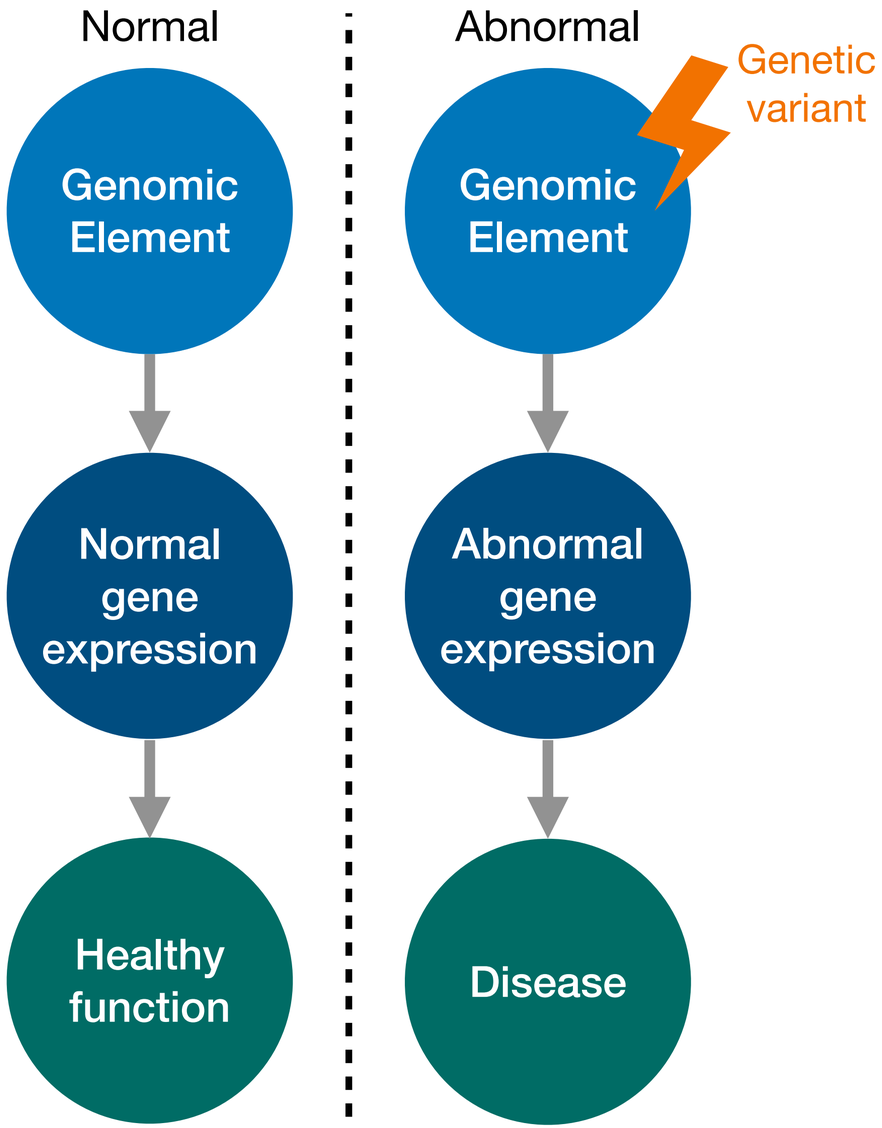
Uncovering genetic origins of human diseases

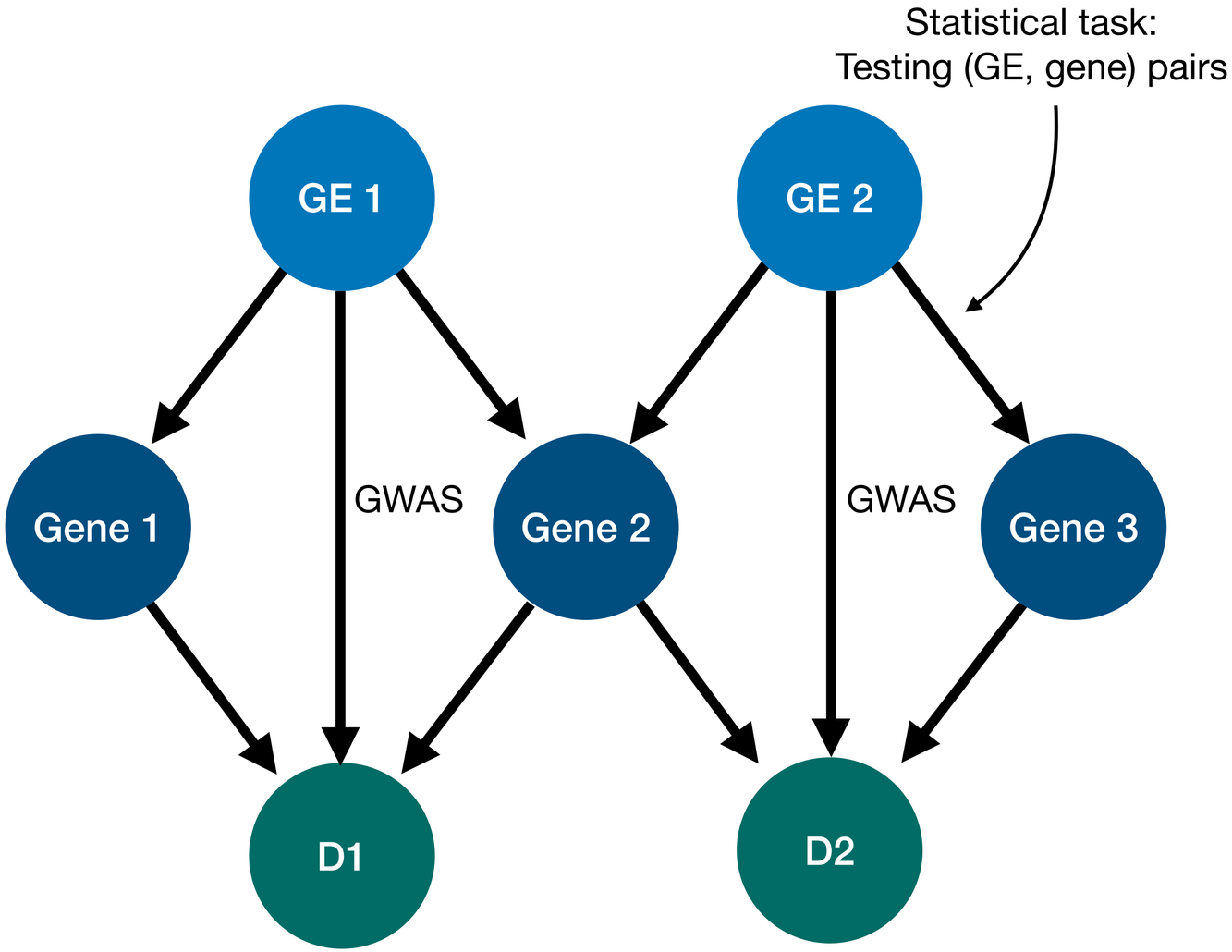




Single-cell CRISPR screens

Uncovering genetic origins of human diseases





Single-cell CRISPR screens

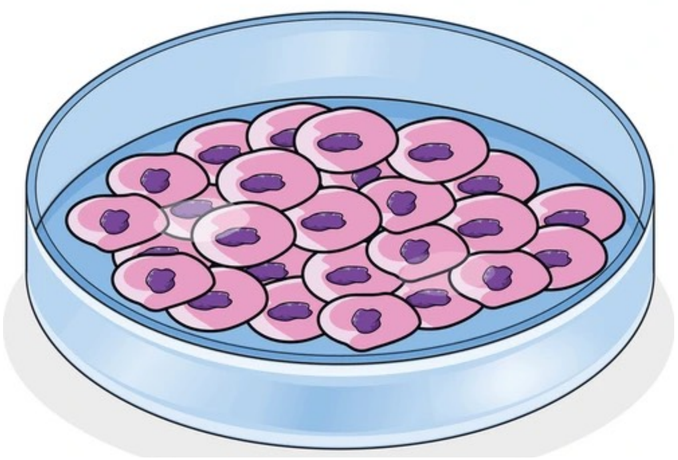
Perturbation-gene association testing

Perturbation-gene association testing

Focus on one genomic element and one gene.

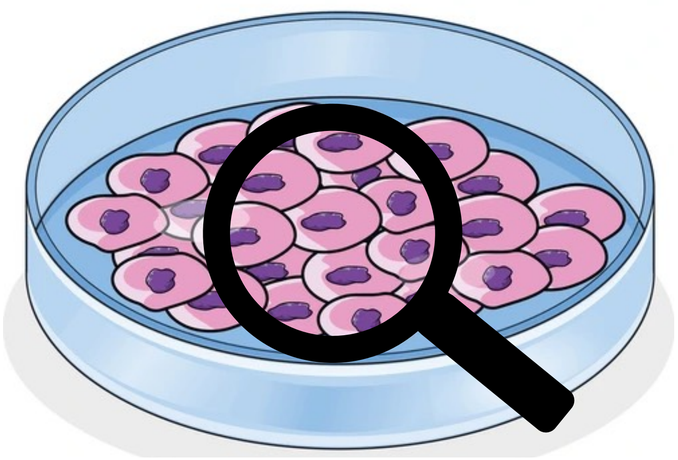
Perturbation-gene association testing

Focus on one genomic element and one gene.



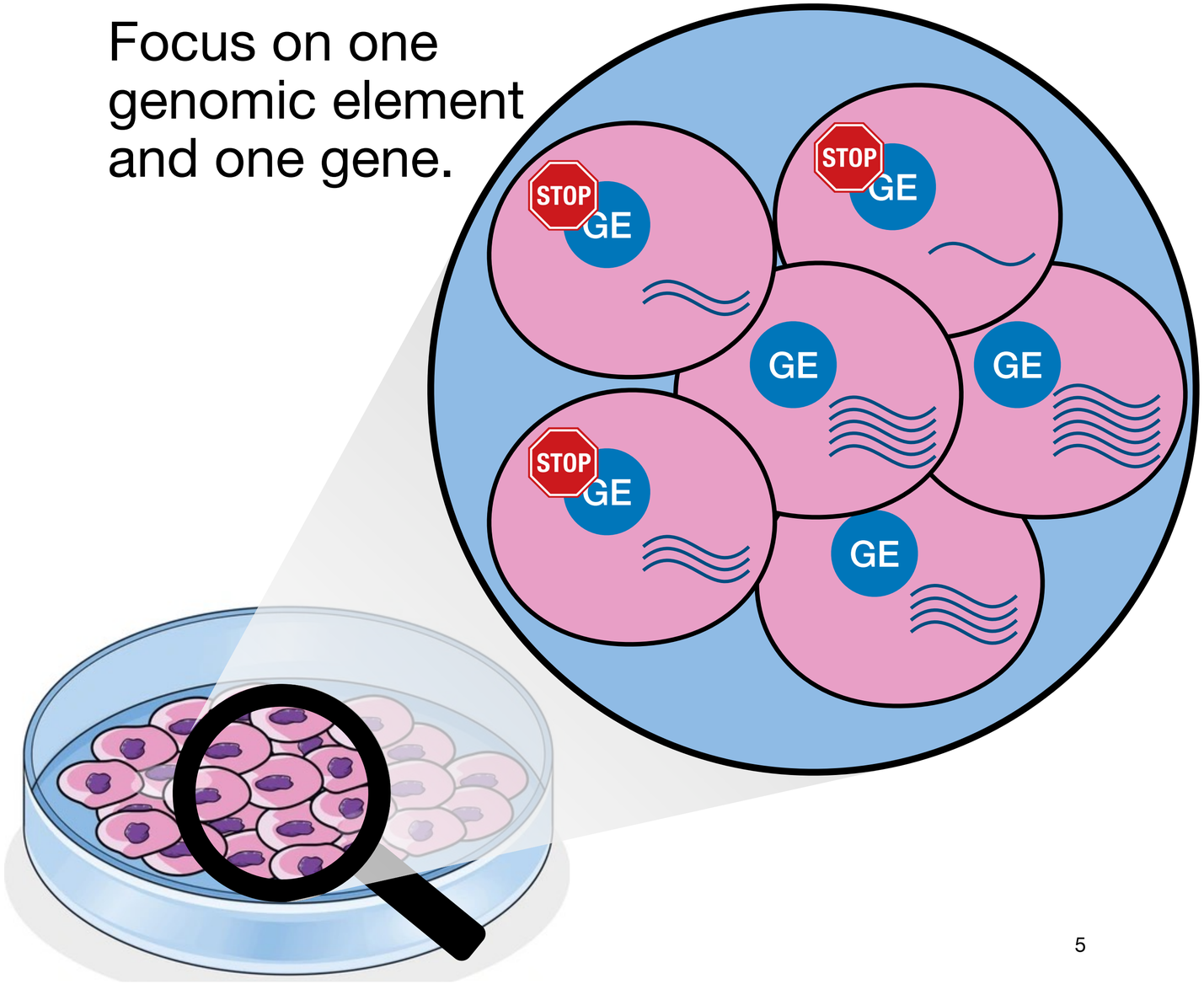
Perturbation-gene association testing

Focus on one genomic element and one gene.



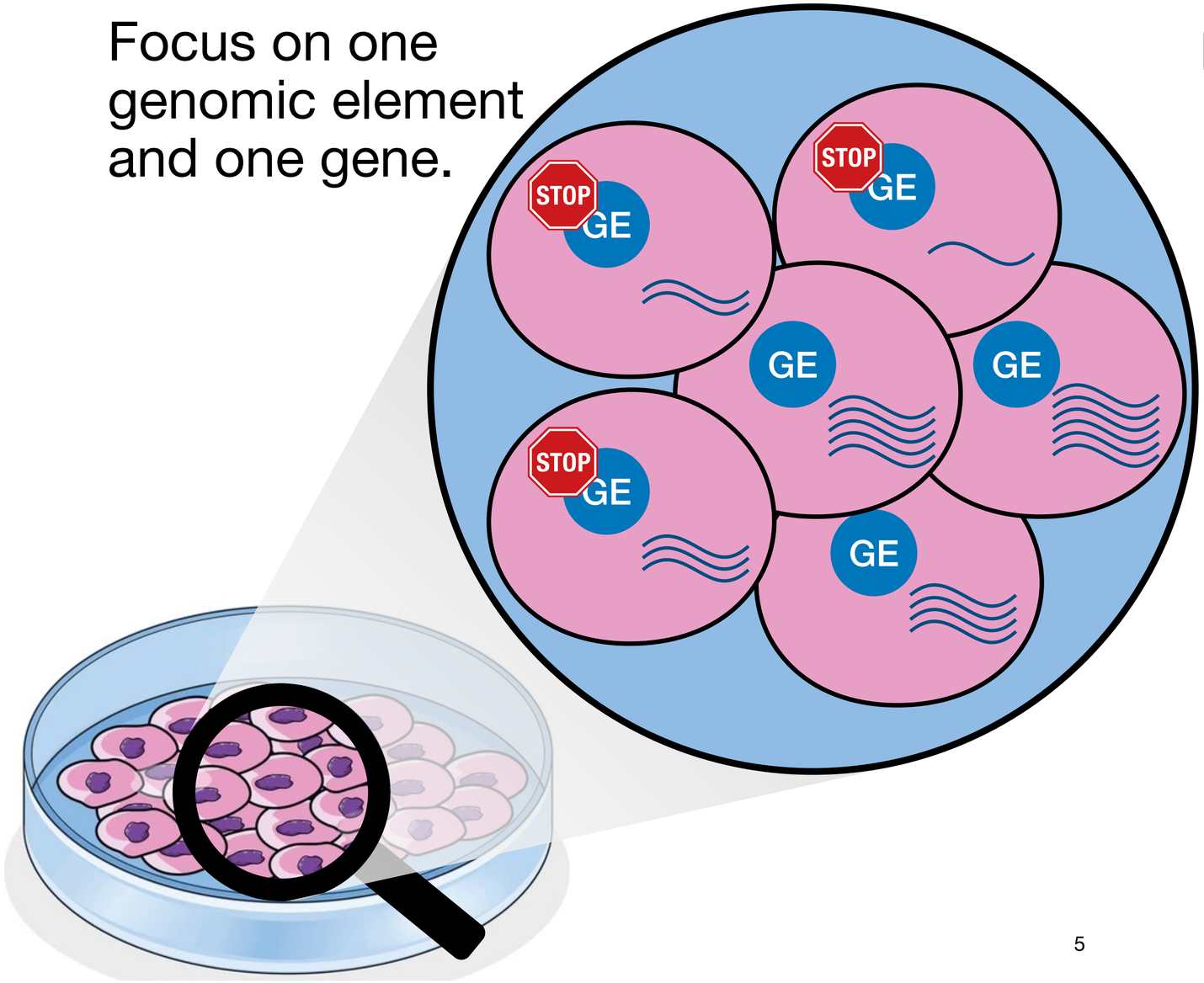
Perturbation-gene association testing

Focus on one genomic element and one gene.



Perturbation-gene association testing

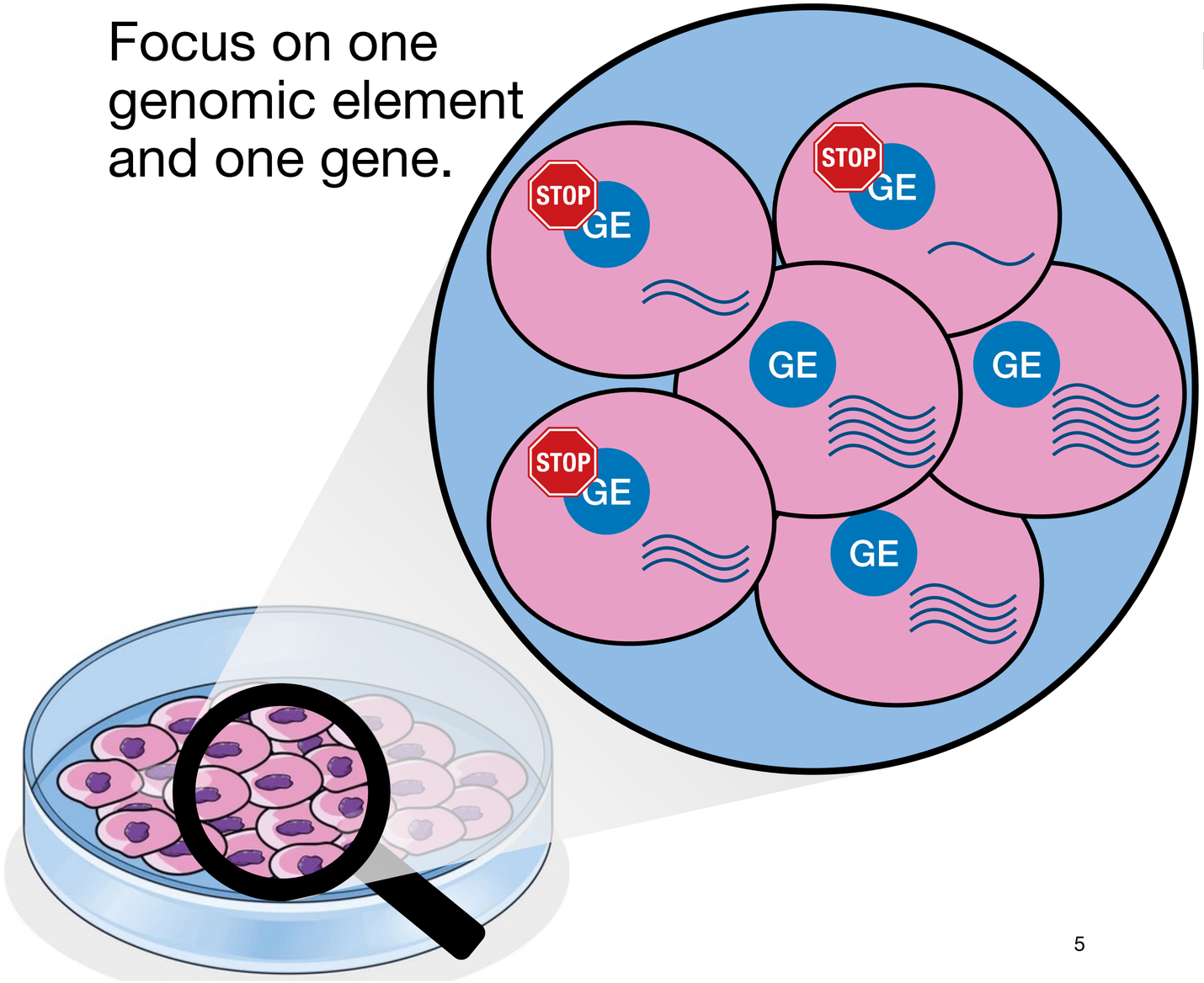
Focus on one genomic element and one gene.



For each cell $i = 1, \ldots, n$,

Perturbation-gene association testing

Focus on one genomic element and one gene.

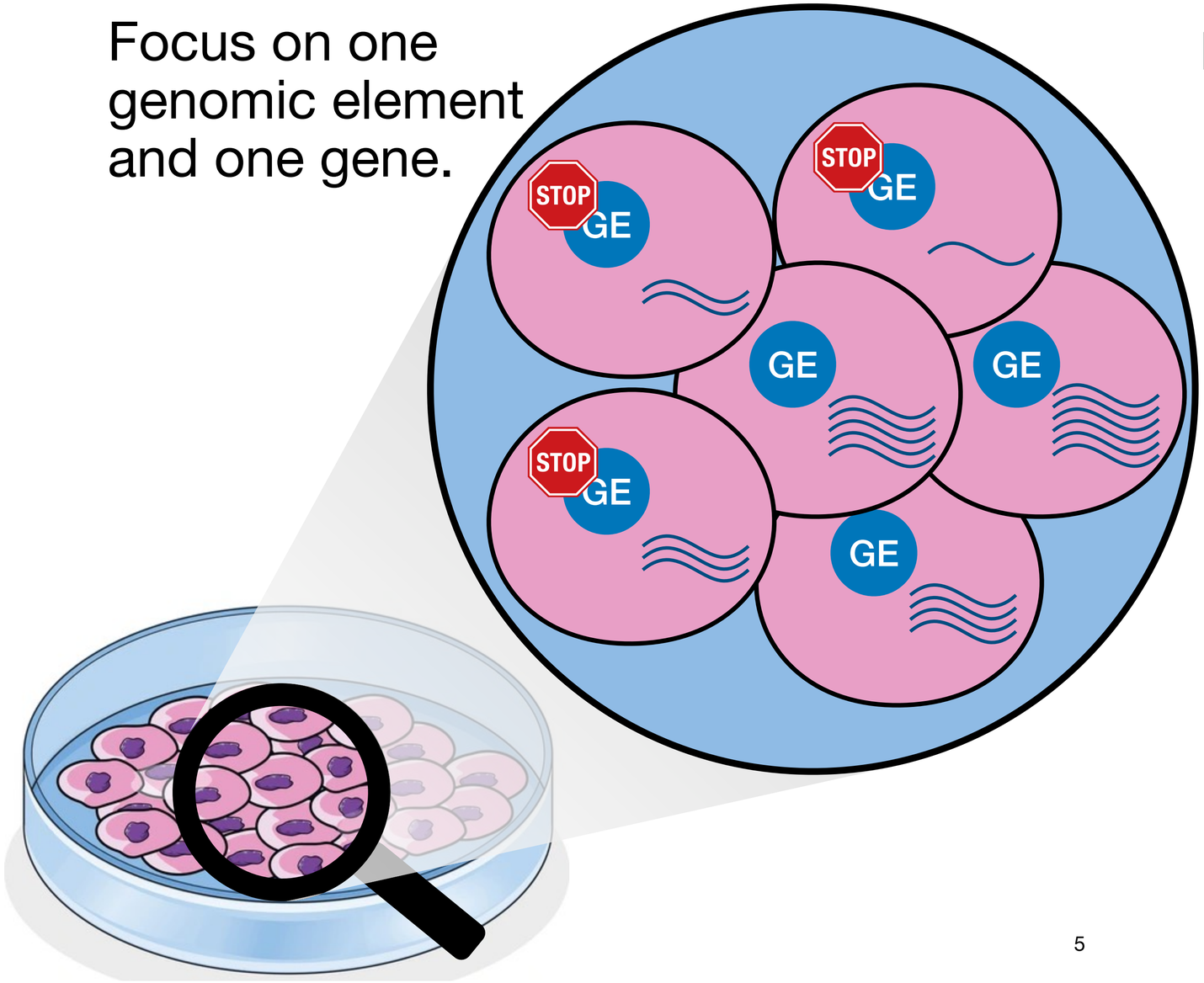


For each cell $i = 1, \ldots, n$,

- $X_i \in \{0,1\}$: perturbation presence

Perturbation-gene association testing

Focus on one genomic element and one gene.

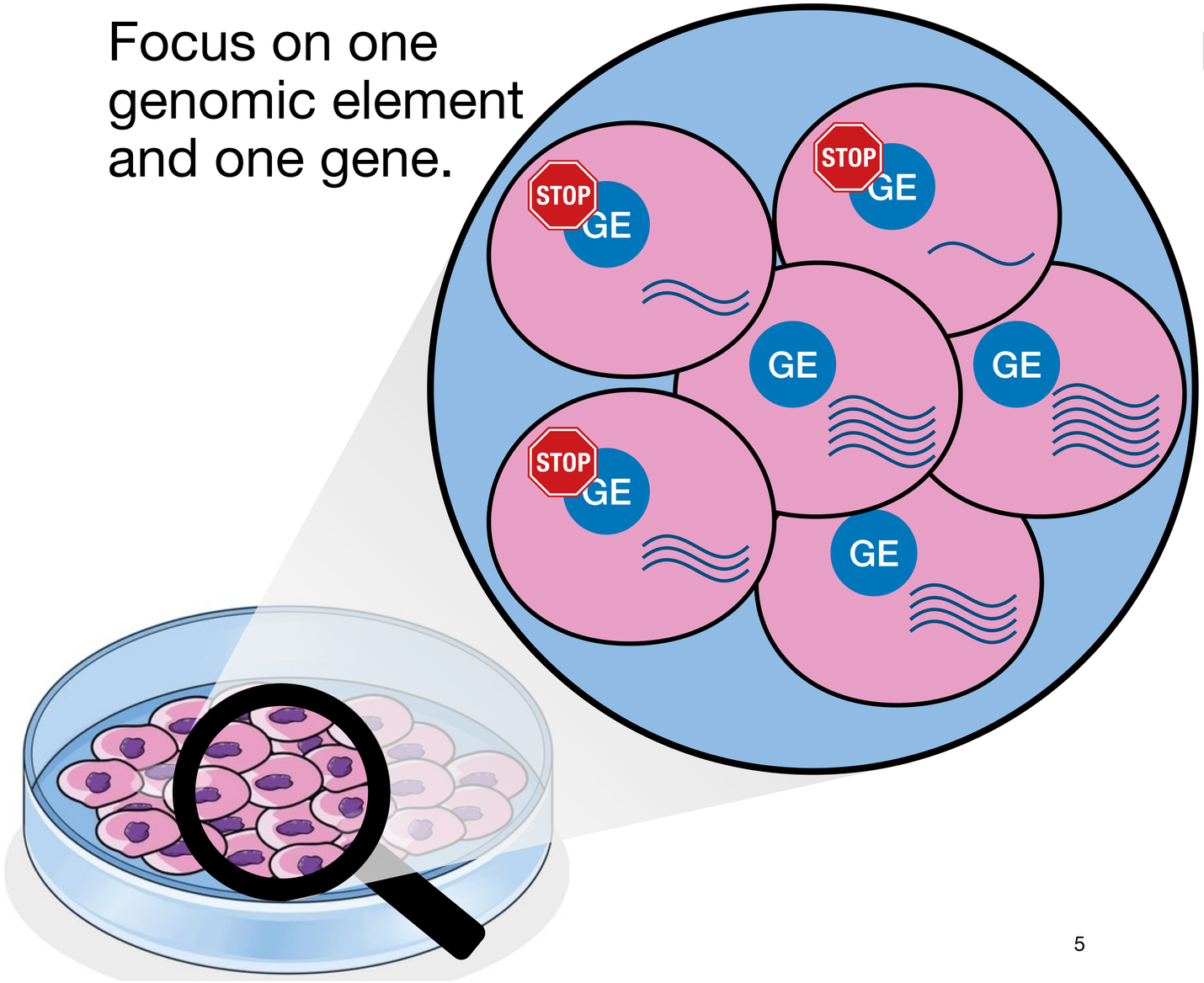


For each cell $i = 1, \ldots, n$,

- $X_i \in \{0,1\}$: perturbation presence
- $Y_i \in \mathbb{N}$: gene expression

Perturbation-gene association testing

Focus on one genomic element and one gene.

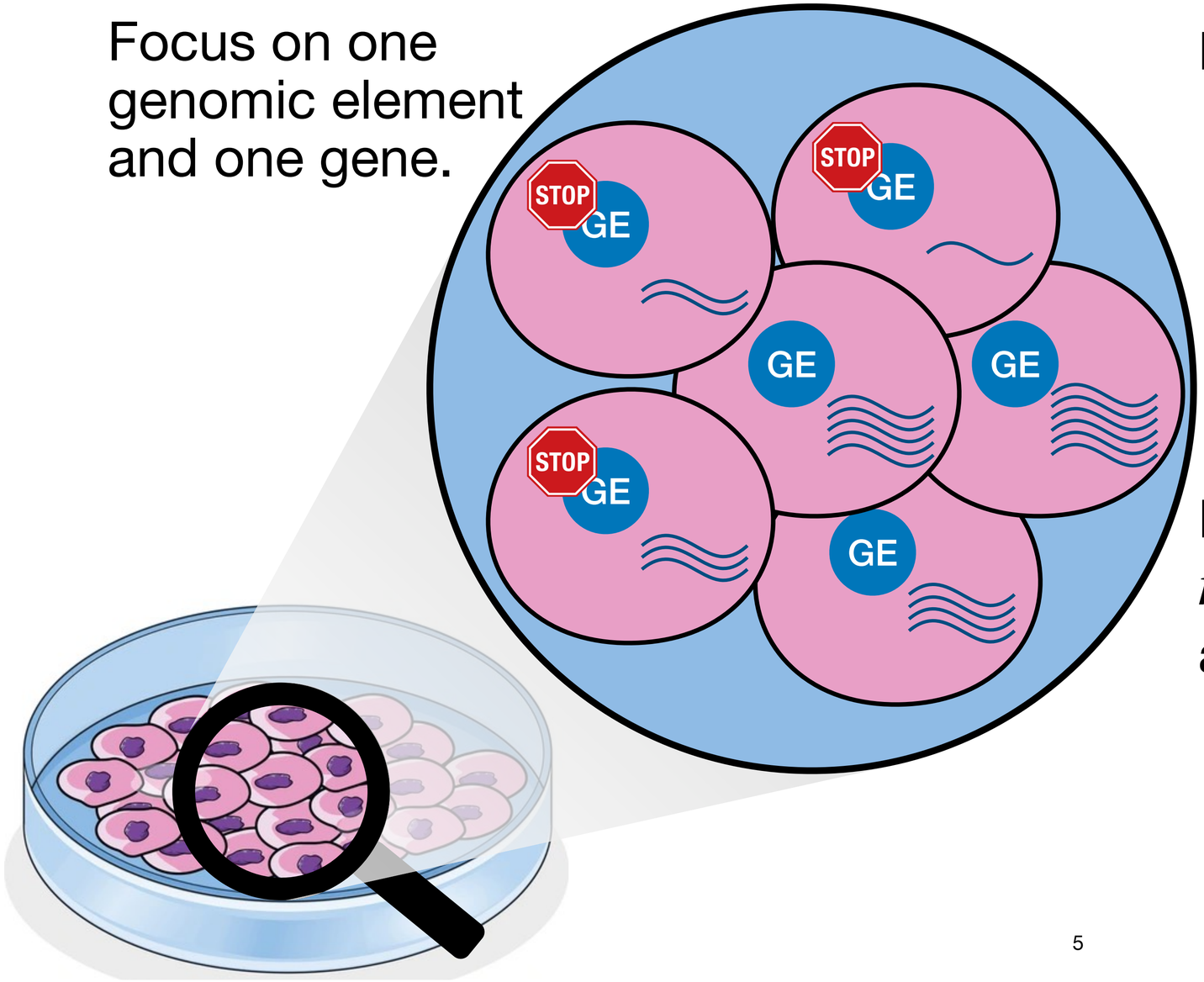


For each cell $i = 1, \ldots, n$,

- $X_i \in \{0,1\}$: perturbation presence
- $Y_i \in \mathbb{N}$: gene expression
- $Z_i \in \mathbb{R}^p$: Vector of covariates

Perturbation-gene association testing

Focus on one genomic element and one gene.



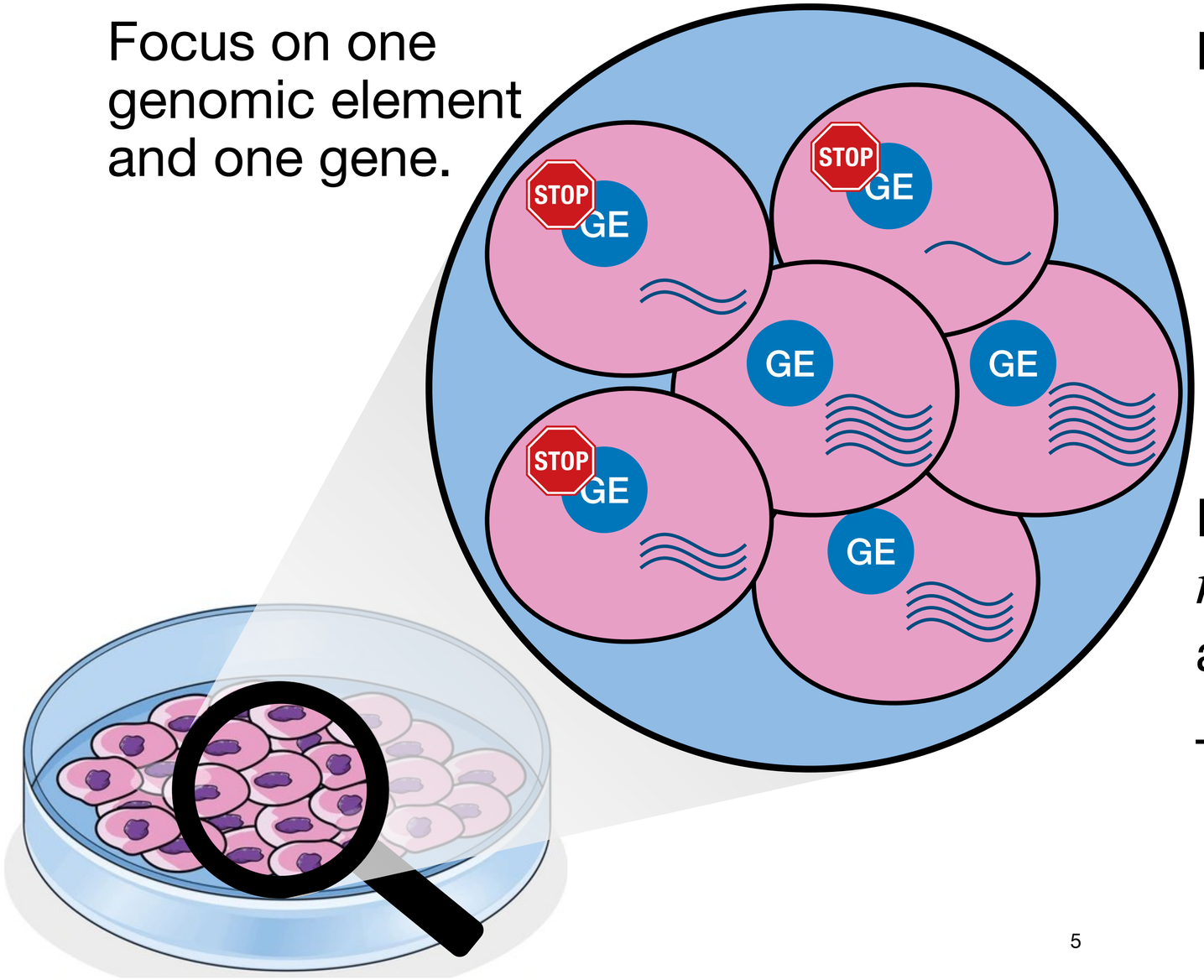
For each cell $i = 1, \ldots, n$,

- $X_i \in \{0,1\}$: perturbation presence
- $Y_i \in \mathbb{N}$: gene expression
- $Z_i \in \mathbb{R}^p$: Vector of covariates

In Gasperini (2019) data, we have $n \approx 200{,}000$ cells, $\approx 13{,}000$ genes, and $\approx 6{,}000$ genomic elements.

Perturbation-gene association testing

Focus on one genomic element and one gene.



For each cell $i = 1, \ldots, n$,

- $X_i \in \{0,1\}$: perturbation presence
- $Y_i \in \mathbb{N}$: gene expression
- $Z_i \in \mathbb{R}^p$: Vector of covariates

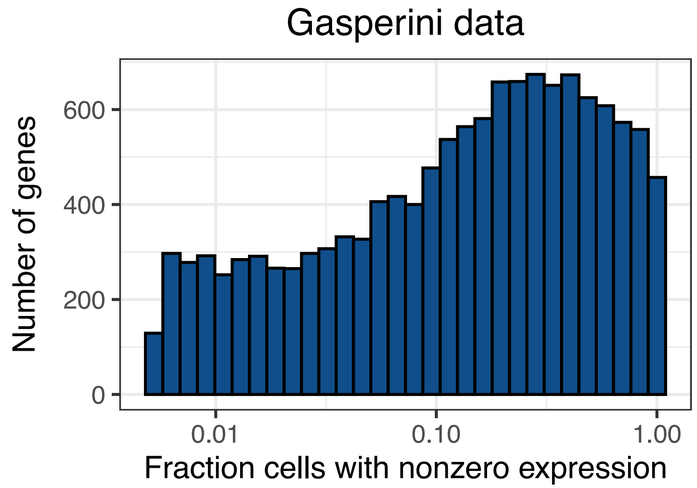
In Gasperini (2019) data, we have $n \approx 200{,}000$ cells, $\approx 13{,}000$ genes, and $\approx 6{,}000$ genomic elements.

The goal is to test

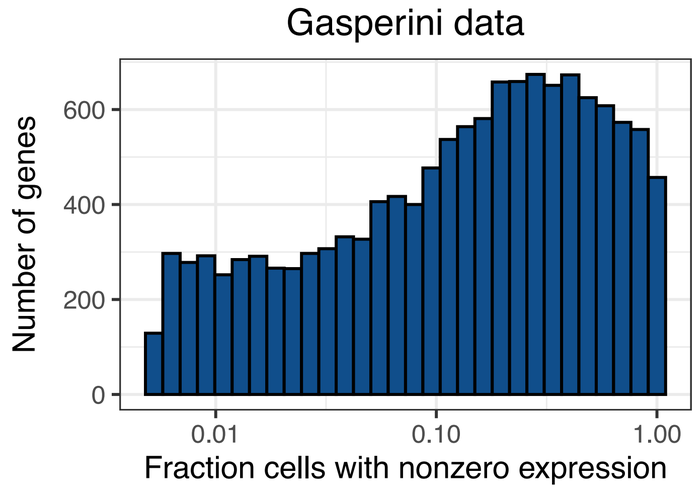
$$H_0 : X_i \perp\!\!\!\perp Y_i \mid Z_i.$$

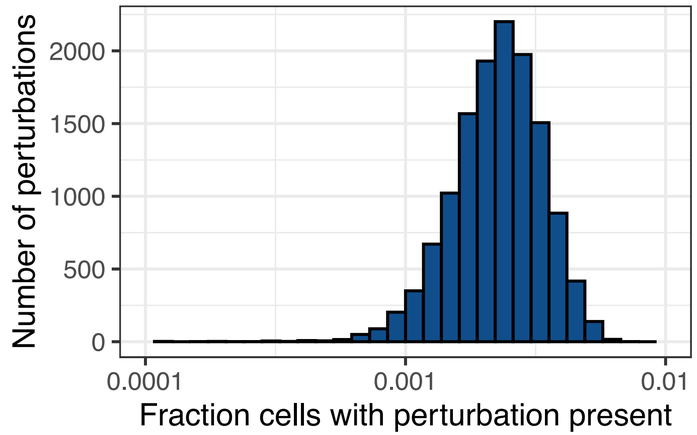
Challenge: The data are highly sparse

Challenge: The data are highly sparse

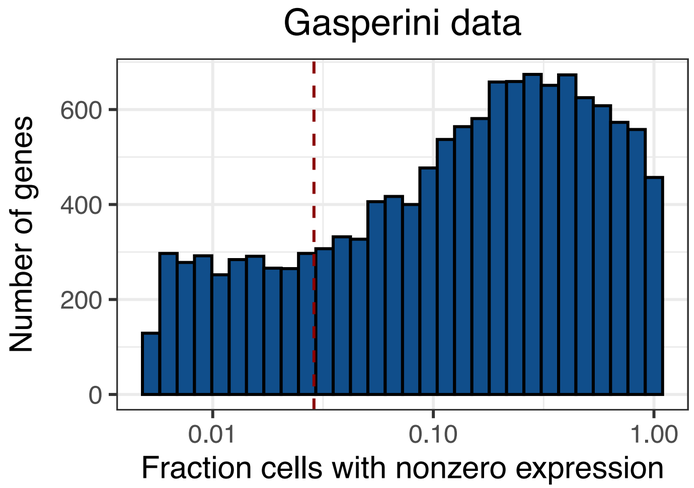


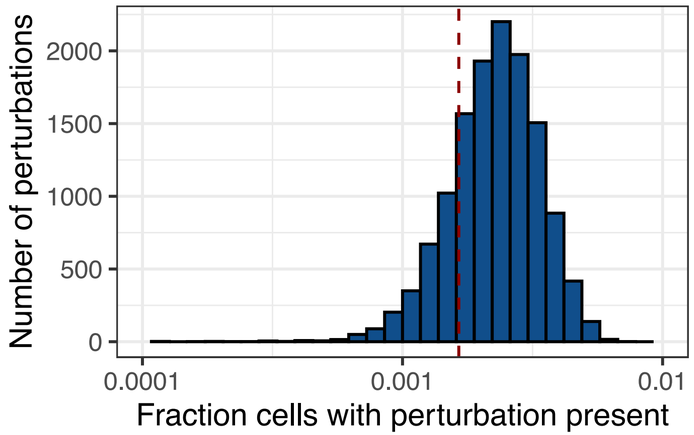
Challenge: The data are highly sparse



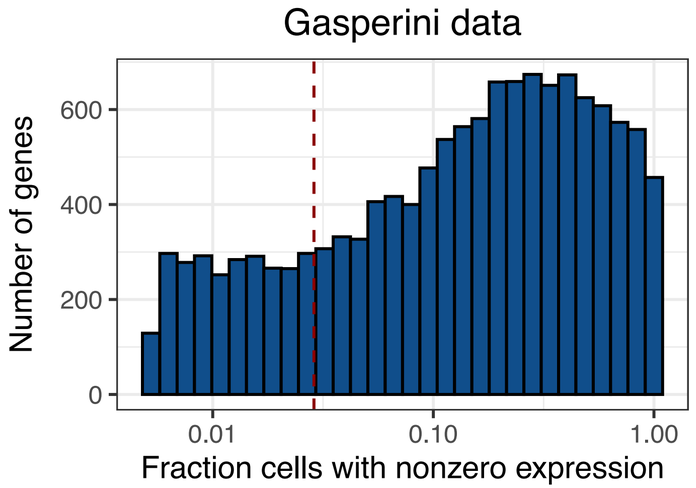


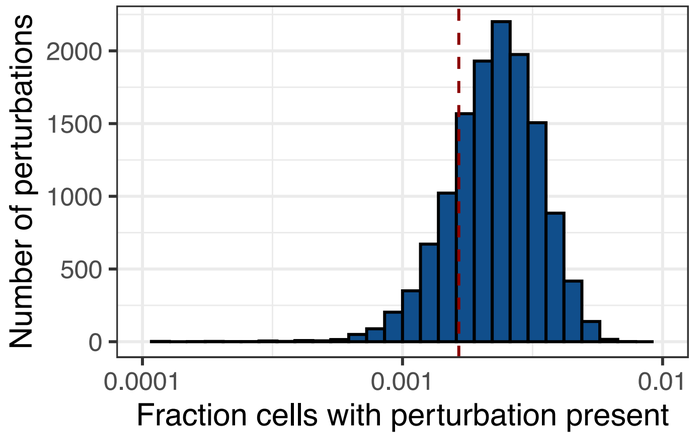
Challenge: The data are highly sparse





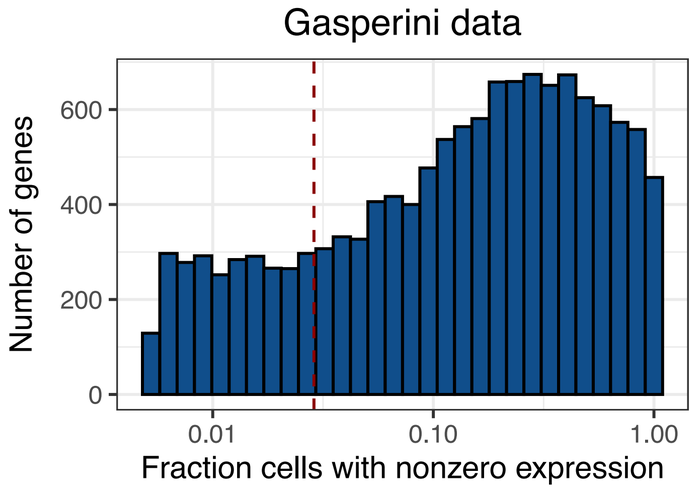
Challenge: The data are highly sparse

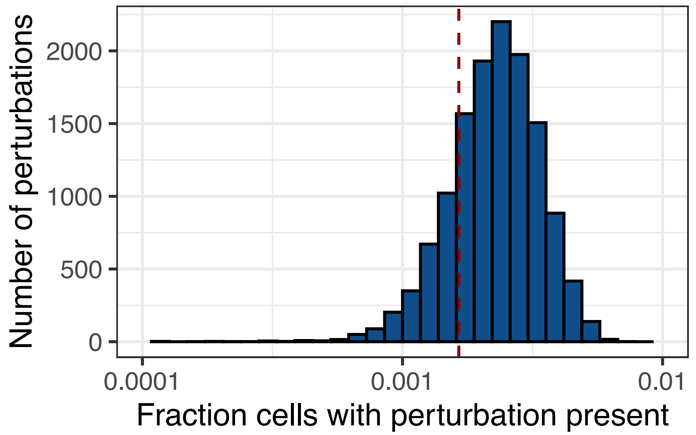




	Gene Expression					
Cell perturbed	**0**	**1**	**2**	**3**	**4**	**8**
FALSE	201027	5594	324	32	5	1
TRUE	332	9	0	0	0	0

Challenge: The data are highly sparse





	Gene Expression					
Cell perturbed	**0**	**1**	**2**	**3**	**4**	**8**
FALSE	201027	5594	324	32	5	1
TRUE	332	9	0	0	0	0

Consequence: Cannot rely on the CLT, especially in the tails.

Model for the biological data

Model for the biological data

Working model:

$$\begin{cases} X_i \mid Z_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); & \text{logit}(\pi_i) = Z_i^T \delta \\ Y_i \mid X_i, Z_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \theta); & \log \mu_i = X_i \beta + Z_i^T \gamma \end{cases}.$$

Testing strategies

Working model:

$$\begin{cases} X_i \mid Z_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); & \text{logit}(\pi_i) = Z_i^T \delta \\ Y_i \mid X_i, Z_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \theta); & \log \mu_i = X_i \beta + Z_i^T \gamma \end{cases}.$$

Three tests:

Testing strategies

Working model:

$$\begin{cases} X_i \mid Z_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); & \text{logit}(\pi_i) = Z_i^T \delta \\ Y_i \mid X_i, Z_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \theta); & \log \mu_i = X_i \beta + Z_i^T \gamma \end{cases}.$$

Three tests:

1. Negative binomial regression score test[1]

[1]Gasperini et al. (2019)

Testing strategies

Working model:

$$\begin{cases} X_i \mid Z_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); & \text{logit}(\pi_i) = Z_i^T \delta \\ Y_i \mid X_i, Z_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \theta); & \log \mu_i = X_i \beta + Z_i^T \gamma \end{cases}.$$

Three tests:

1. Negative binomial regression score test[1]
2. Generalized covariance measure (GCM) test,[2] based on normal limit of $T_n \equiv \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mathbb{E}}[X_i \mid Z_i])(Y_i - \hat{\mathbb{E}}[Y_i \mid Z_i])$

[1]Gasperini et al. (2019)

[2]Shah and Peters (2020)

Testing strategies

Working model:

$$\begin{cases} X_i \mid Z_i \overset{\text{ind}}{\sim} \text{Ber}(\pi_i); & \text{logit}(\pi_i) = Z_i^T \delta \\ Y_i \mid X_i, Z_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \theta); & \log \mu_i = X_i \beta + Z_i^T \gamma \end{cases}.$$

Three tests:

1. Negative binomial regression score test[1]
2. Generalized covariance measure (GCM) test,[2] based on normal limit of $T_n \equiv \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mathbb{E}}[X_i \mid Z_i])(Y_i - \hat{\mathbb{E}}[Y_i \mid Z_i])$
3. Distilled conditional randomization test (dCRT),[3] based on T_n and resampling from fitted distribution $X_i \mid Z_i$

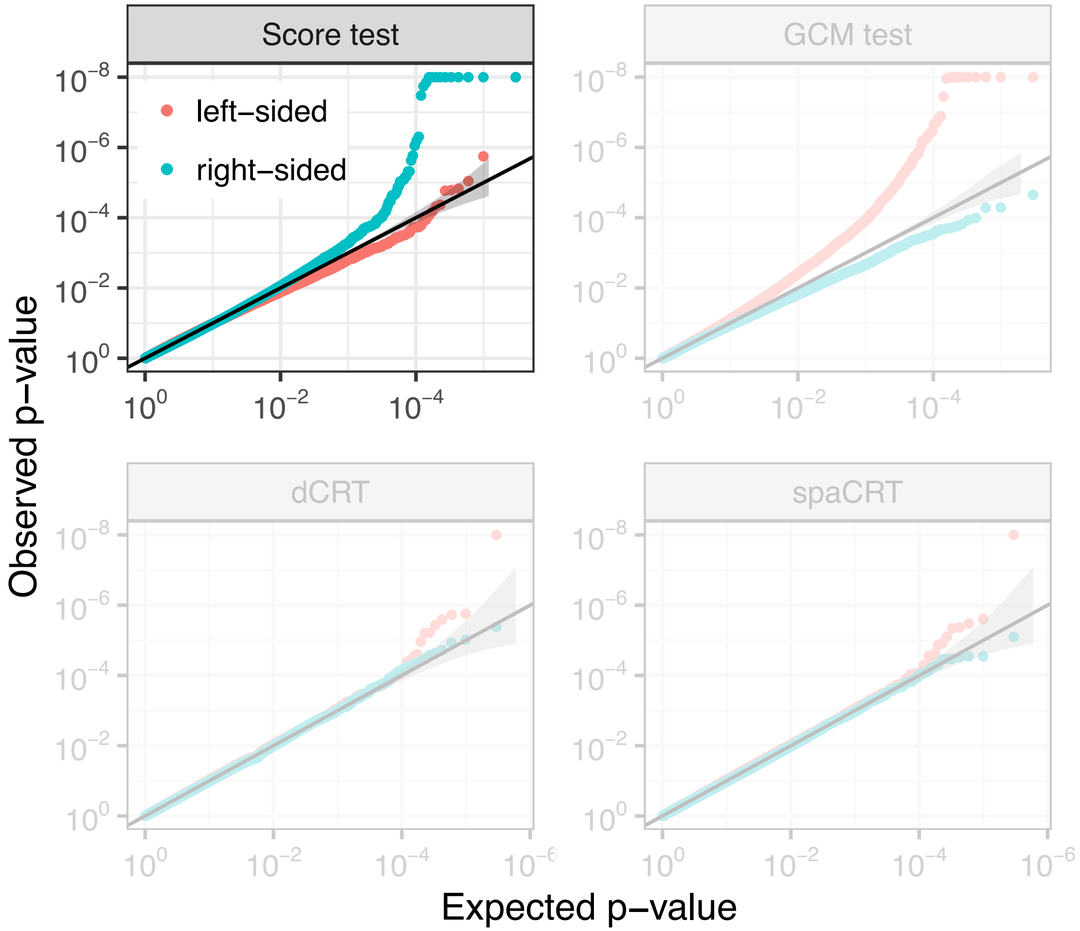
[1]Gasperini et al. (2019)

[2]Shah and Peters (2020)

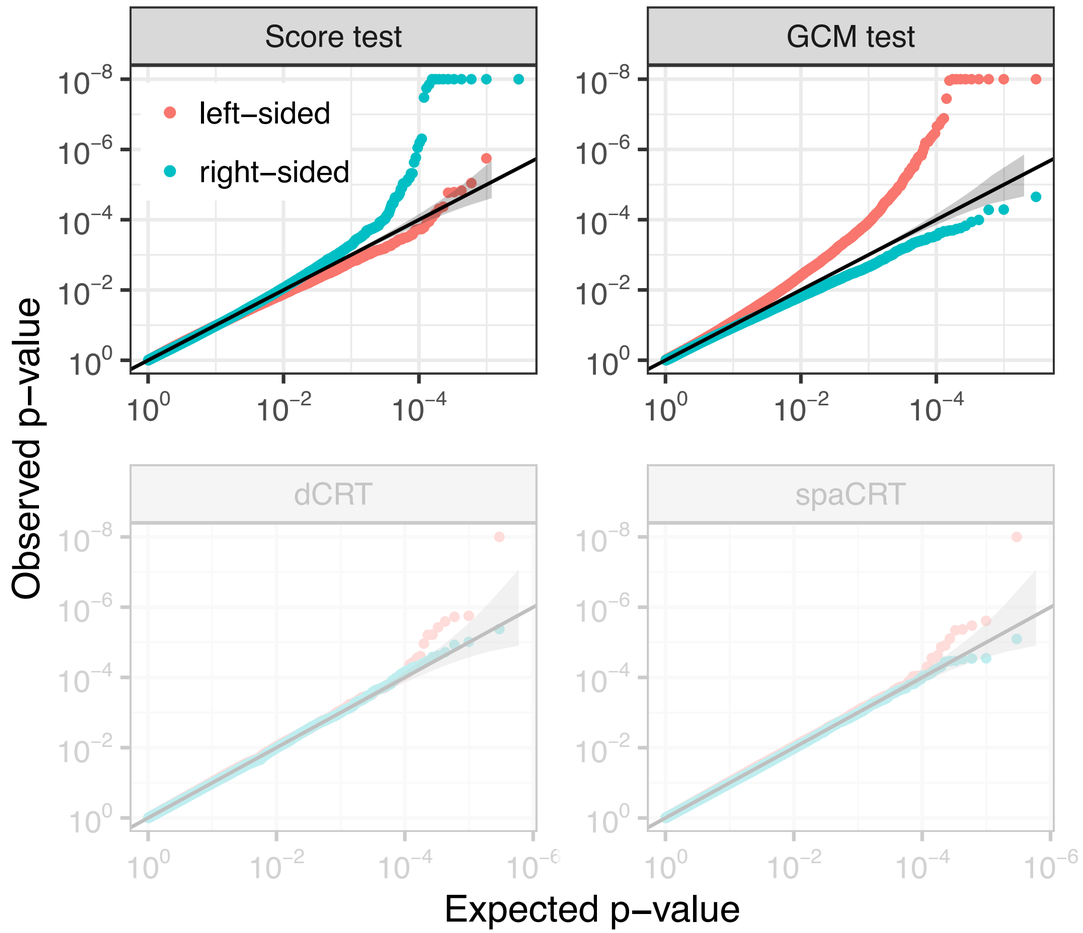
[3]Liu et al. (2022), Barry et al. (2021)

Comparing Type-I error using negative controls

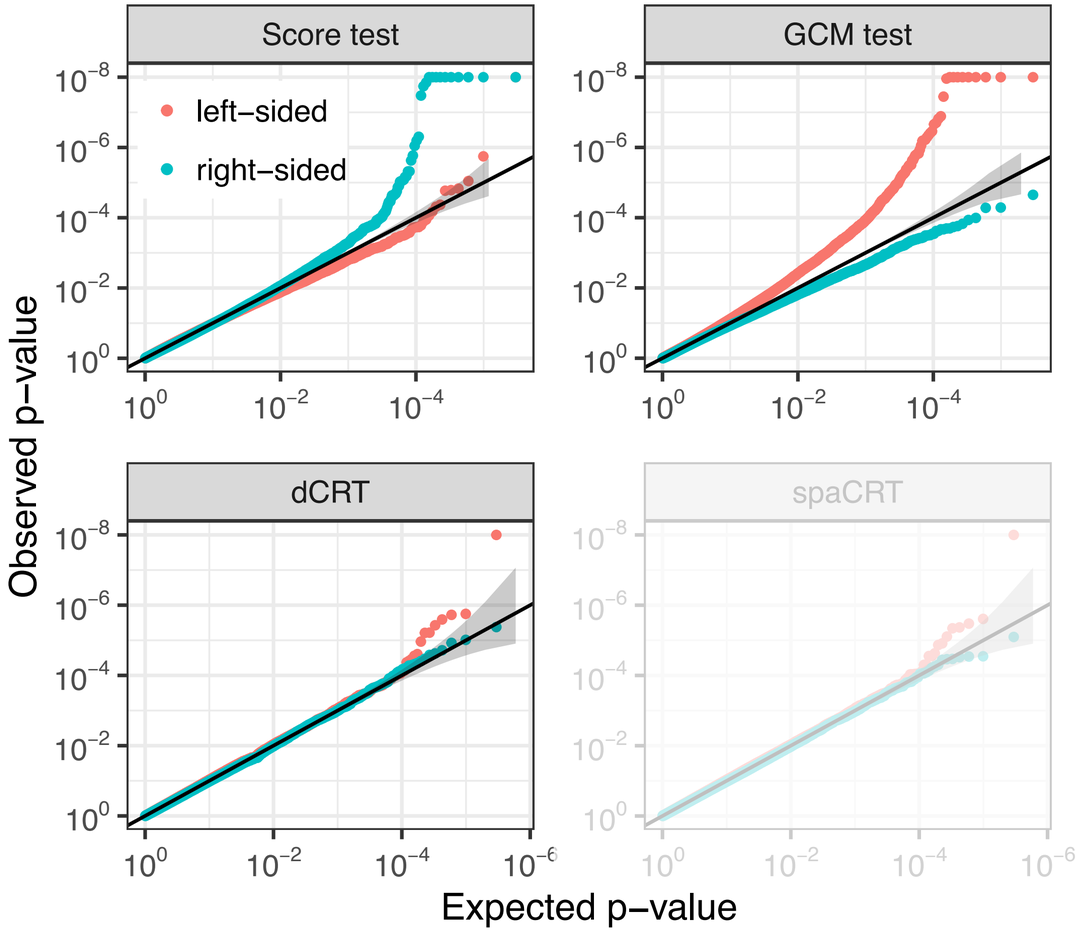
Comparing Type-I error using negative controls



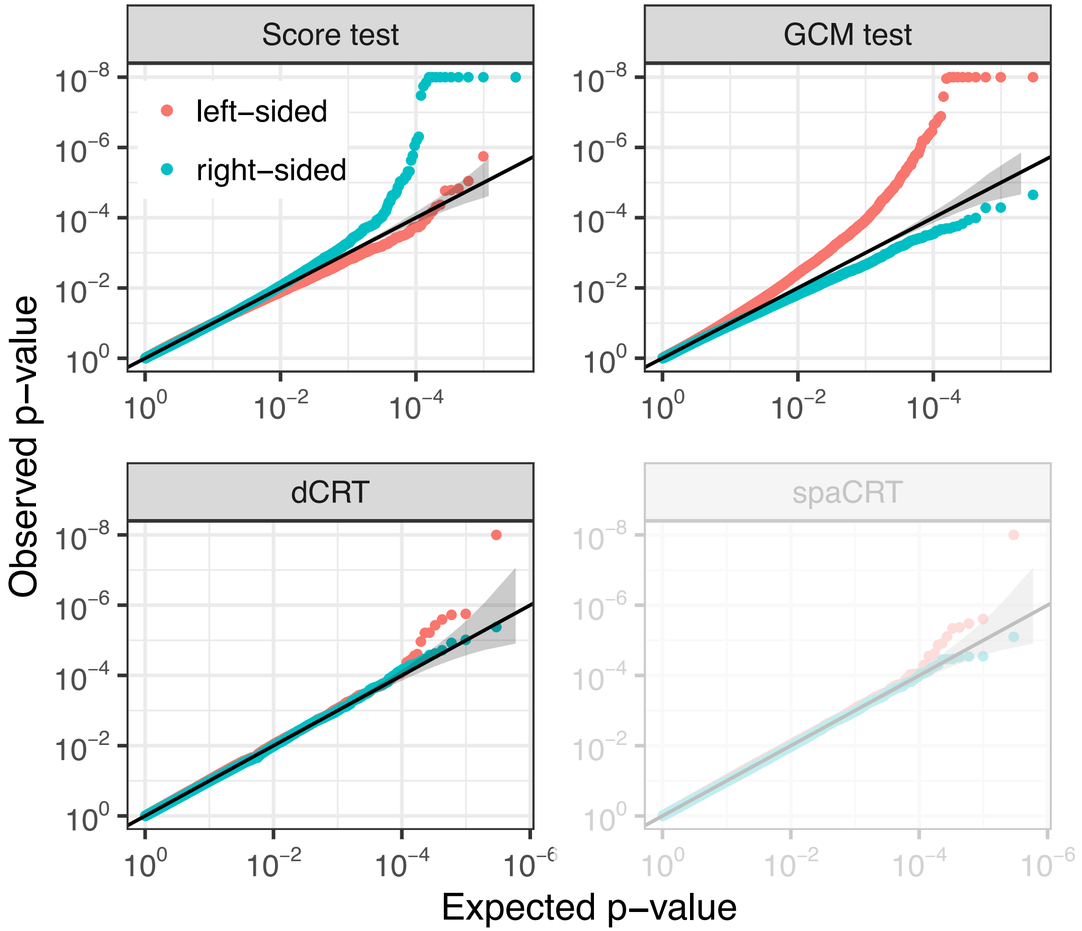
Comparing Type-I error using negative controls

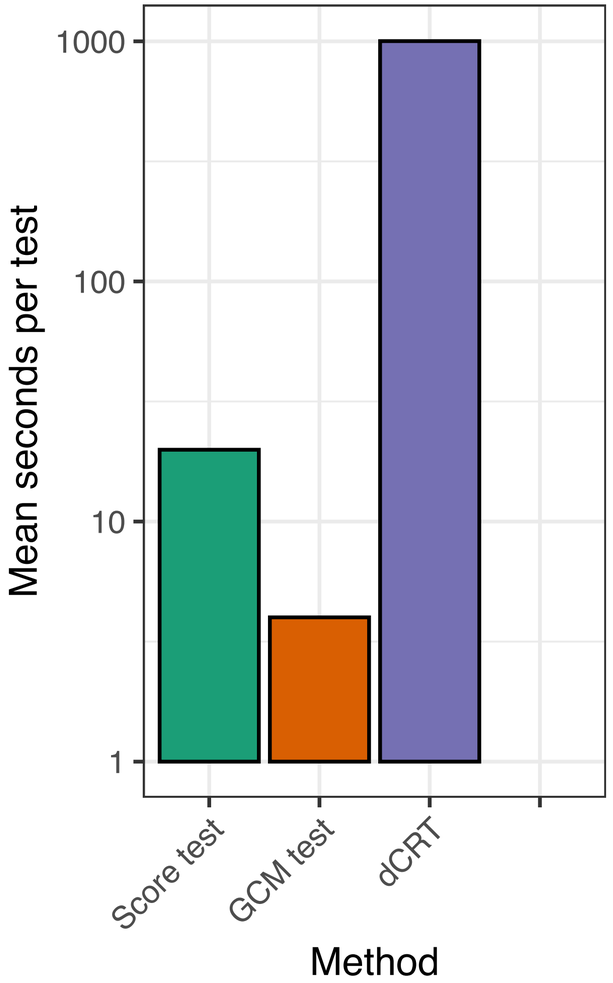


Comparing Type-I error using negative controls

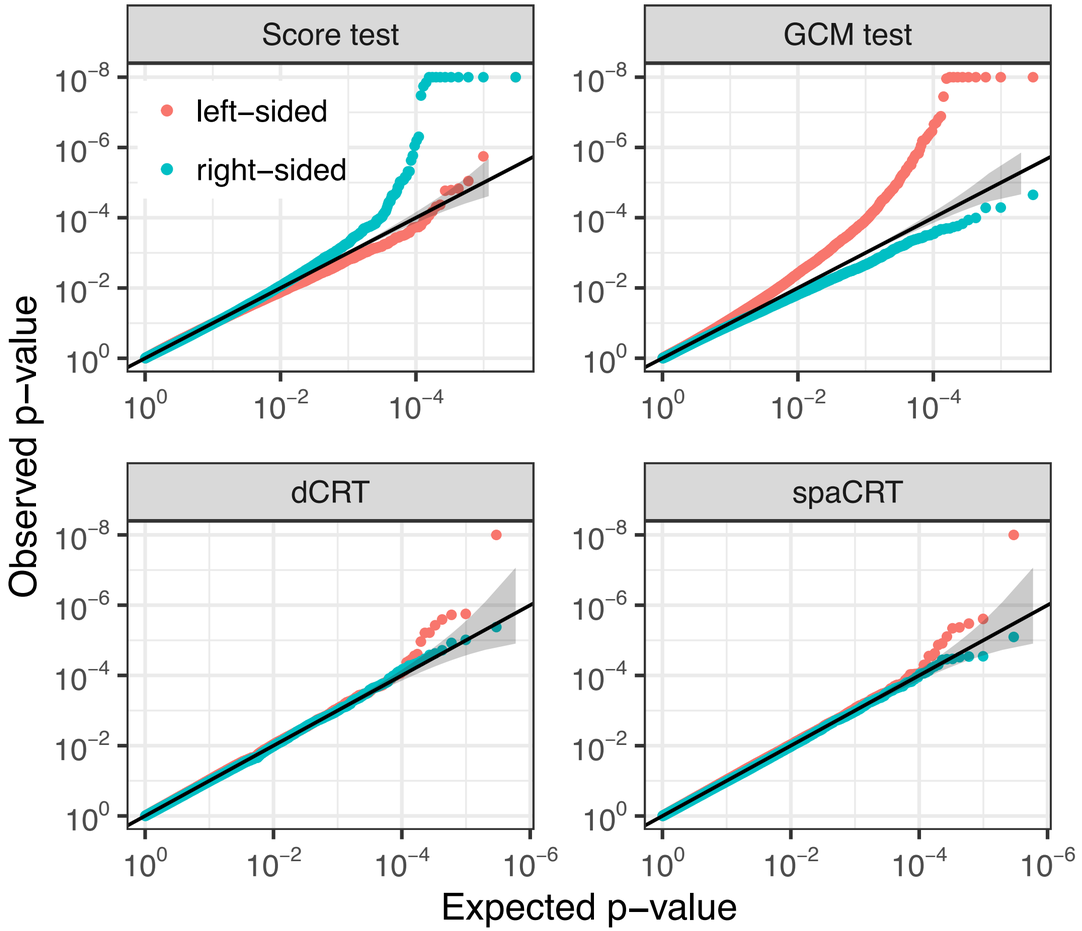


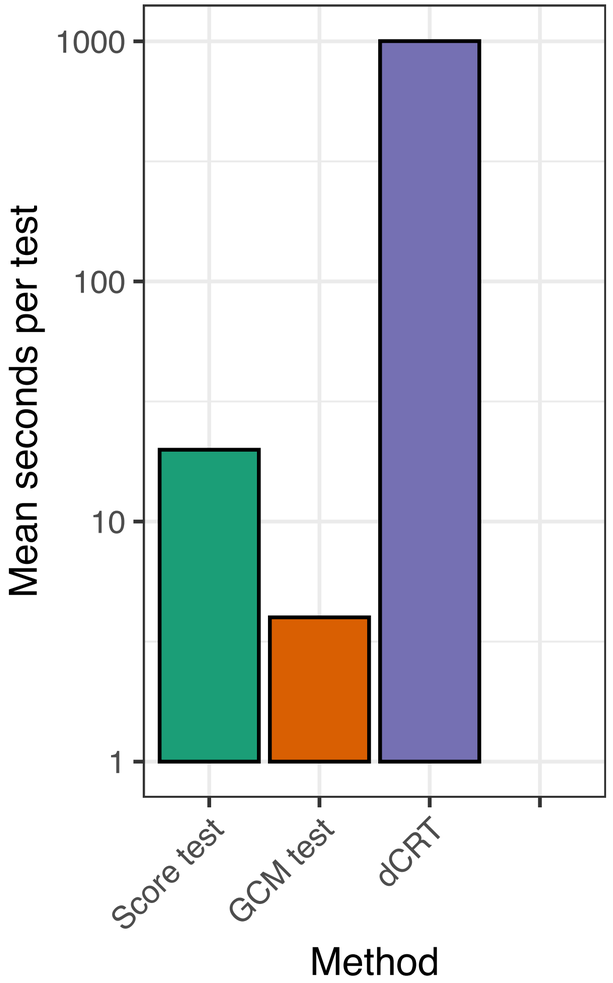
Comparing Type-I error using negative controls



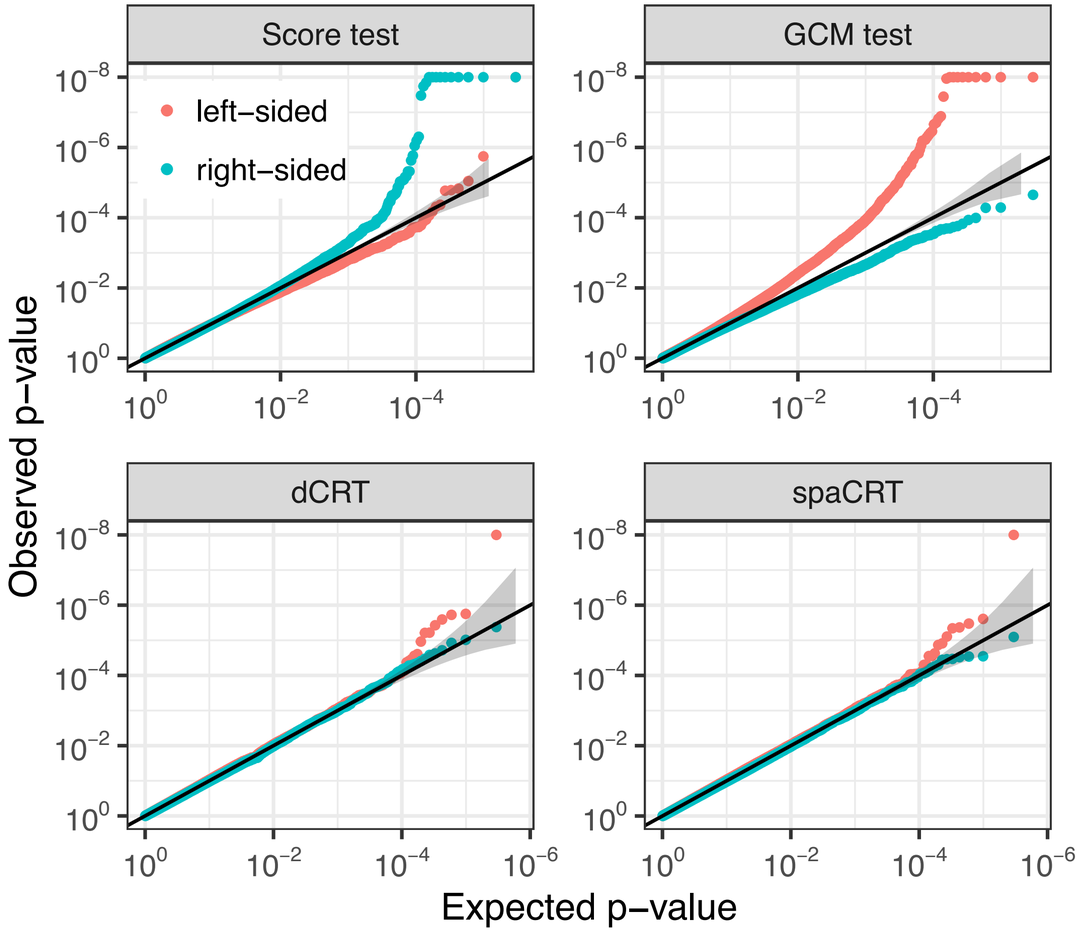


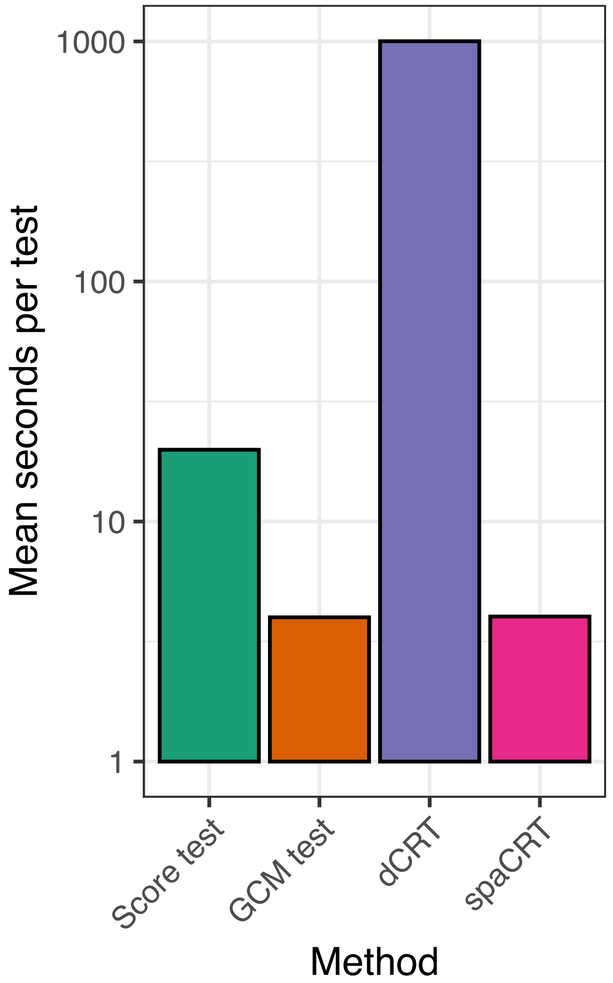
Comparing Type-I error using negative controls



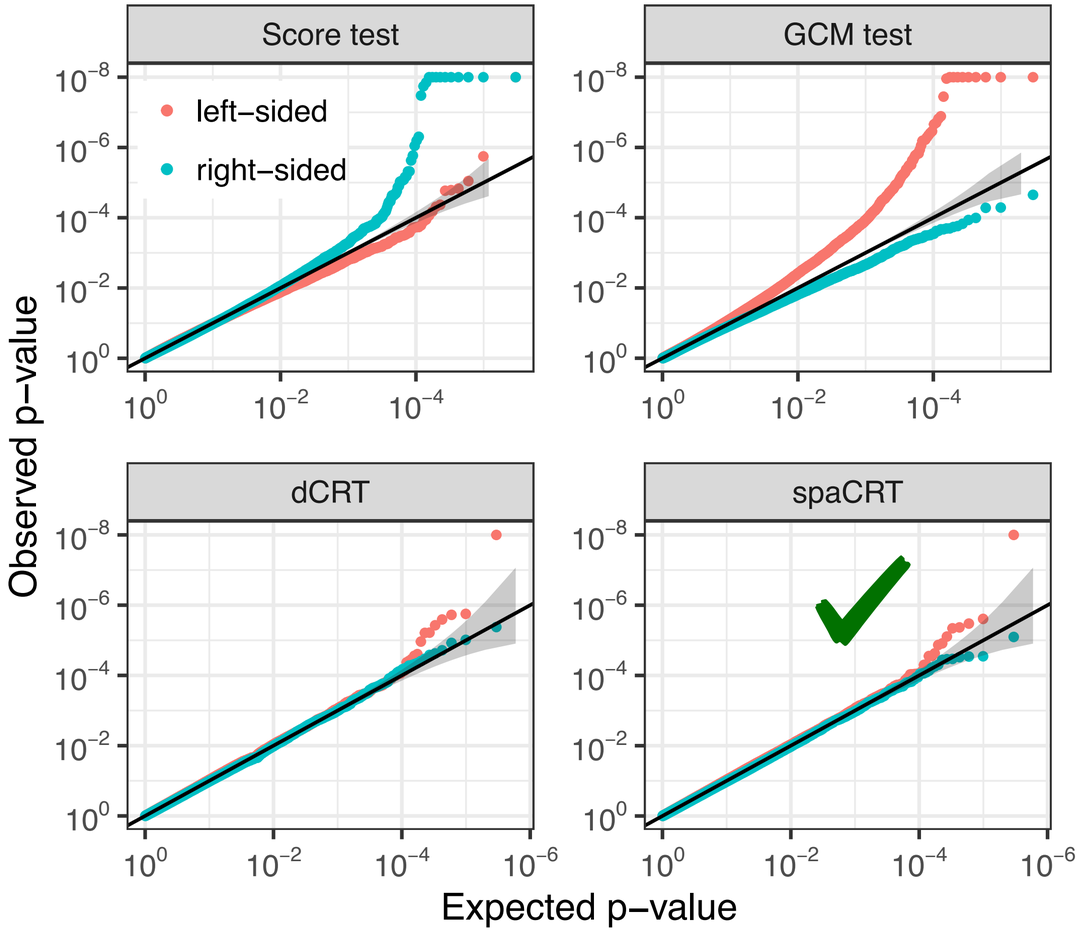


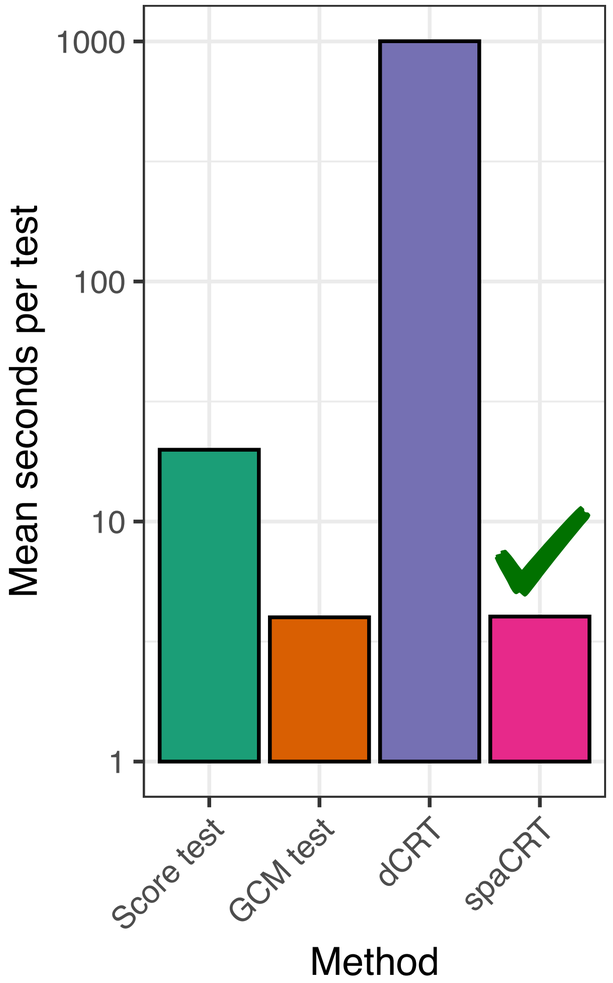
Comparing Type-I error using negative controls





Comparing Type-I error using negative controls





The distilled conditional randomization test

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.
2. Compute $T_n \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\pi}(Z_i))(Y_i - \widehat{\mu}(Z_i))$.

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.
2. Compute $T_n \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\pi}(Z_i))(Y_i - \widehat{\mu}(Z_i))$.
3. for $b = 1, \ldots, B$:

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.
2. Compute $T_n \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\pi}(Z_i))(Y_i - \widehat{\mu}(Z_i))$.
3. for $b = 1, \ldots, B$:
    a. Sample $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \text{Ber}(\widehat{\pi}(Z_i))$.

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

1. Learn $\hat{\pi}(Z_i)$ by regressing X on Z and $\hat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.
2. Compute $T_n \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i))$.
3. for $b = 1, \ldots, B$:
    a. Sample $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \text{Ber}(\hat{\pi}(Z_i))$.
    b. Compute $\tilde{T}_n^{(b)} \equiv \frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i^{(b)} - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i))$.

The distilled conditional randomization test

Suppose $X_i \overset{\text{ind}}{\sim} \text{Ber}(\pi(Z_i))$ for some function π.

dCRT algorithm

1. Learn $\hat{\pi}(Z_i)$ by regressing X on Z and $\hat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.
2. Compute $T_n \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i))$.
3. for $b = 1, \ldots, B$:
   a. Sample $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \text{Ber}(\hat{\pi}(Z_i))$.
   b. Compute $\tilde{T}_n^{(b)} \equiv \frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i^{(b)} - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i))$.
4. Output dCRT p-value $p_n^{\text{dCRT}} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} 1(\tilde{T}_n^{(b)} \geq T_n)\right)$.

Excessive computation with dCRT

Excessive computation with dCRT

- dCRT can only produce accurate p-value up to order $1/B$;

Excessive computation with dCRT

- dCRT can only produce accurate p-value up to order $1/B$;
- In reality, two important reasons can make this procedure infeasible:

Excessive computation with dCRT

- dCRT can only produce accurate p-value up to order $1/B$;
- In reality, two important reasons can make this procedure infeasible:

    1. **High multiplicity:** more than $M = 6000 \times 13000$ hypotheses in Gasperini dataset.

Excessive computation with dCRT

- dCRT can only produce accurate p-value up to order $1/B$;
- In reality, two important reasons can make this procedure infeasible:

    1. **High multiplicity:** more than $M = 6000 \times 13000$ hypotheses in Gasperini dataset.

    2. **Accurate estimate for small p-value:** requires the accuracy up to α/M if Bonferroni correction is used.

Excessive computation with dCRT

- dCRT can only produce accurate p-value up to order $1/B$;
- In reality, two important reasons can make this procedure infeasible:

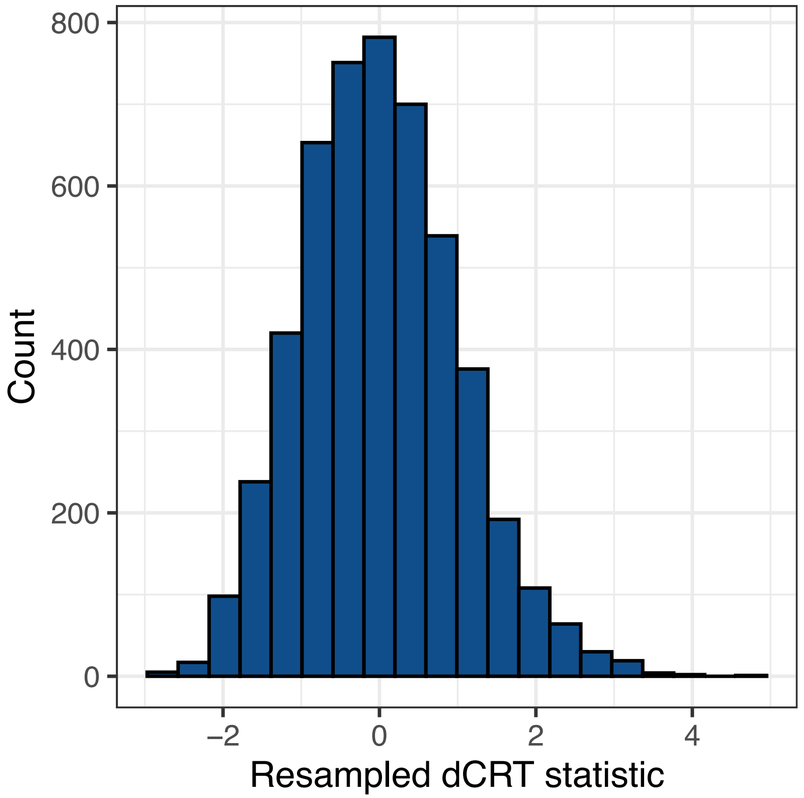
    1. **High multiplicity:** more than $M = 6000 \times 13000$ hypotheses in Gasperini dataset.

    2. **Accurate estimate for small p-value:** requires the accuracy up to α/M if Bonferroni correction is used.

No hope for resampling-based method if advanced parallelization and heuristic approximation is not available!

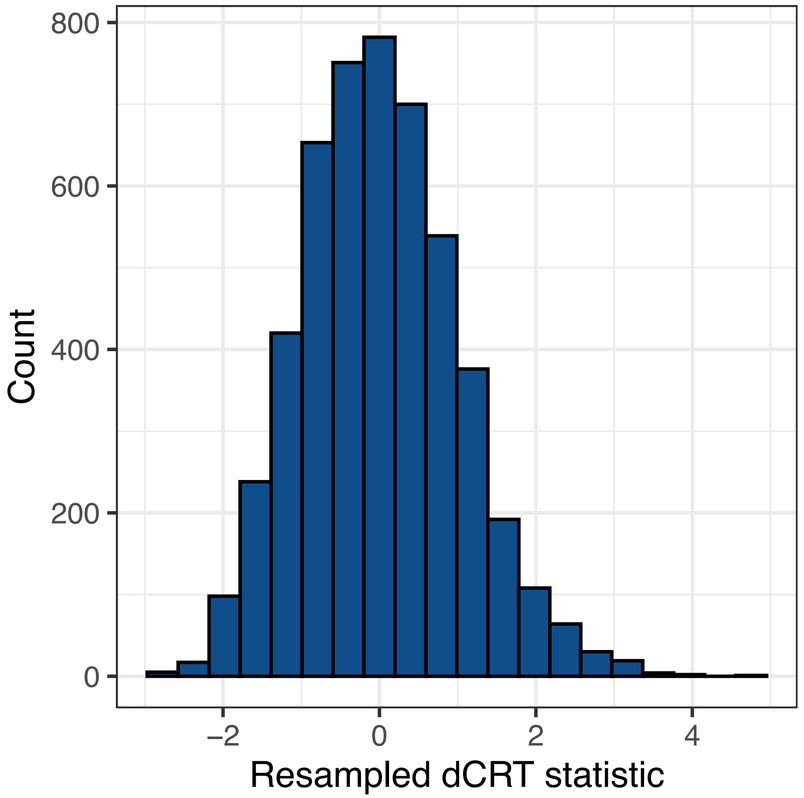
Approximating dCRT p-values

Approximating dCRT p-values



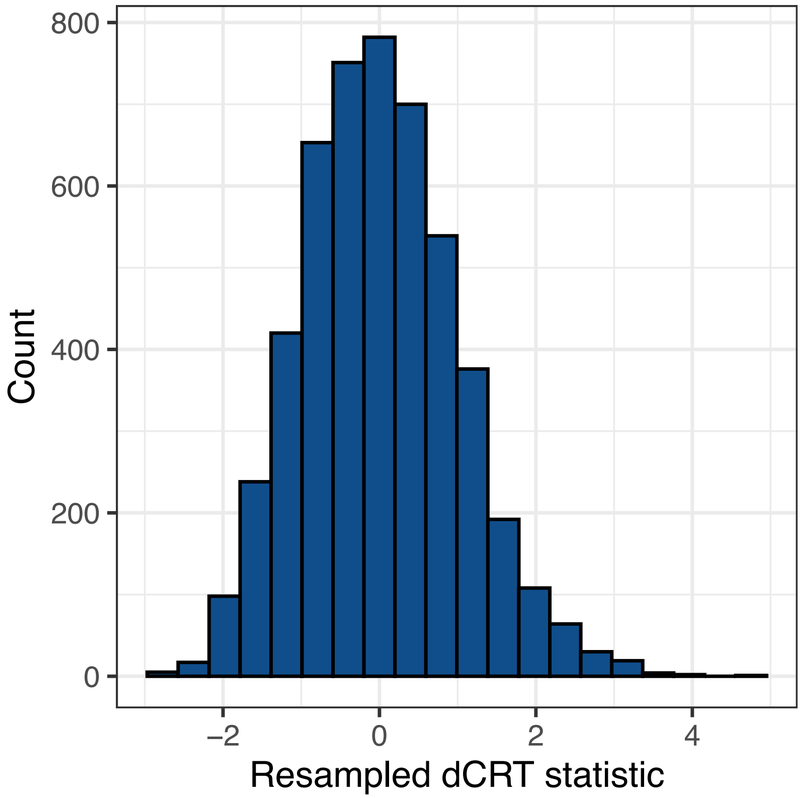
Approximating dCRT p-values

- Draw moderate number of resamples and fit parametric curve to distribution.



Approximating dCRT p-values

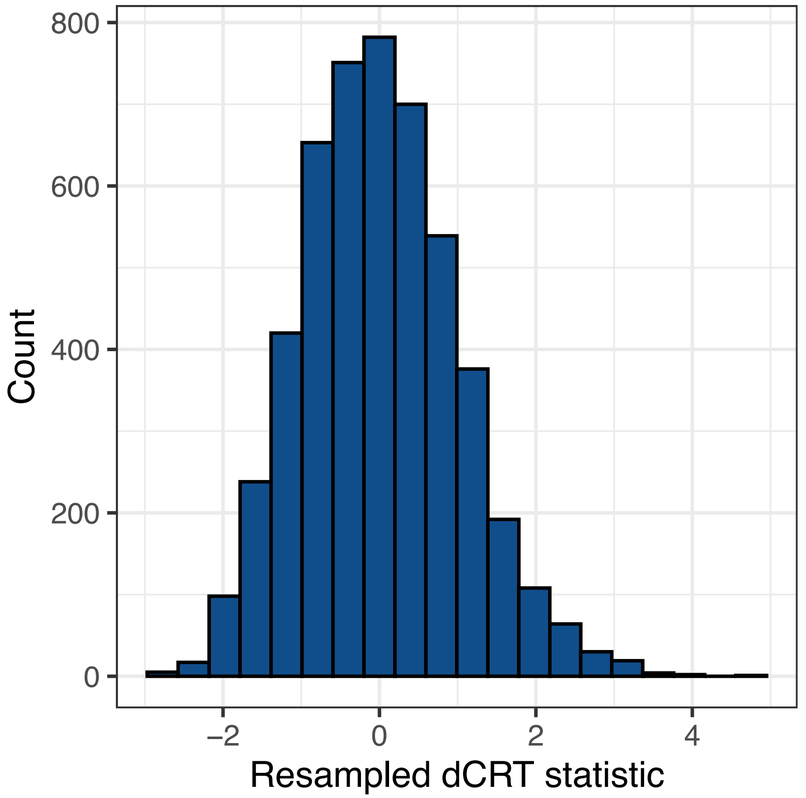
- Draw moderate number of resamples and fit parametric curve to distribution.[1]



[1]Barry et al. (2021)

Approximating dCRT p-values

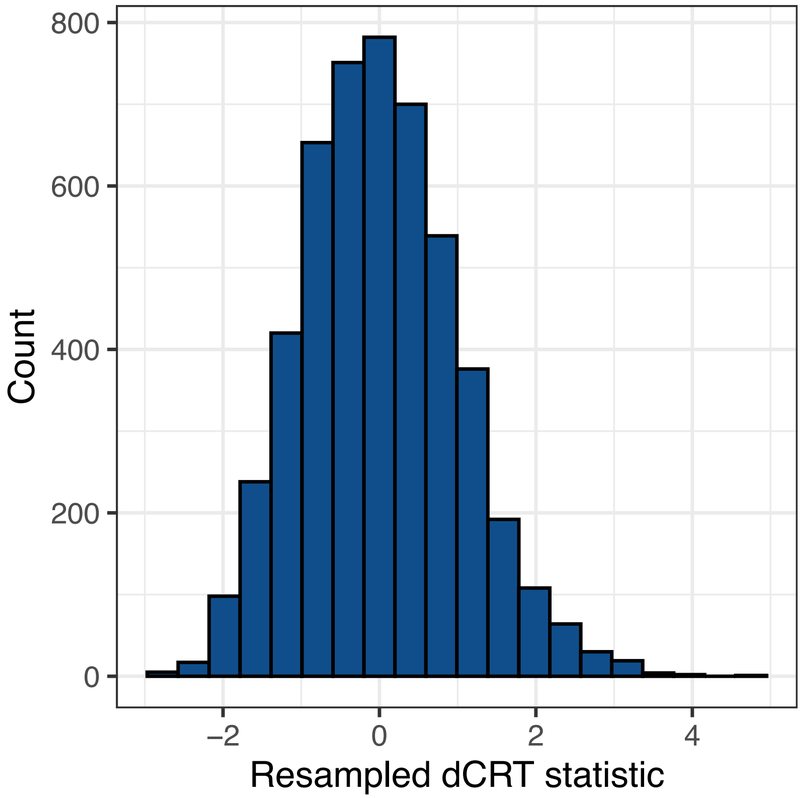
- Draw moderate number of resamples and fit parametric curve to distribution.[1]
- Draw an adaptive number of resamples based on test statistic.



[1]Barry et al. (2021)

Approximating dCRT p-values

- Draw moderate number of resamples and fit parametric curve to distribution.[1]
- Draw an adaptive number of resamples based on test statistic.[2]

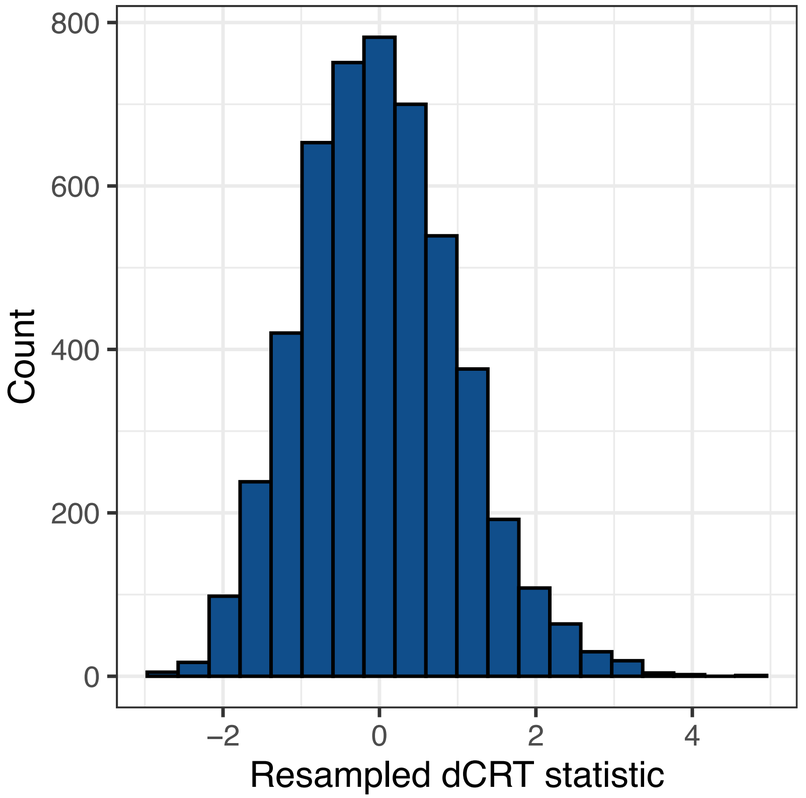


[1]Barry et al. (2021)

[2]Besag and Clifford (1991), Fay et al. (2007), Gandy (2009), Silva and Assuncao (2013), Fischer and Ramdas (2024)

Approximating dCRT p-values

- Draw moderate number of resamples and fit parametric curve to distribution.[1]
- Draw an adaptive number of resamples based on test statistic.[2]
- Use the saddlepoint approximation for sums of independent random variables.

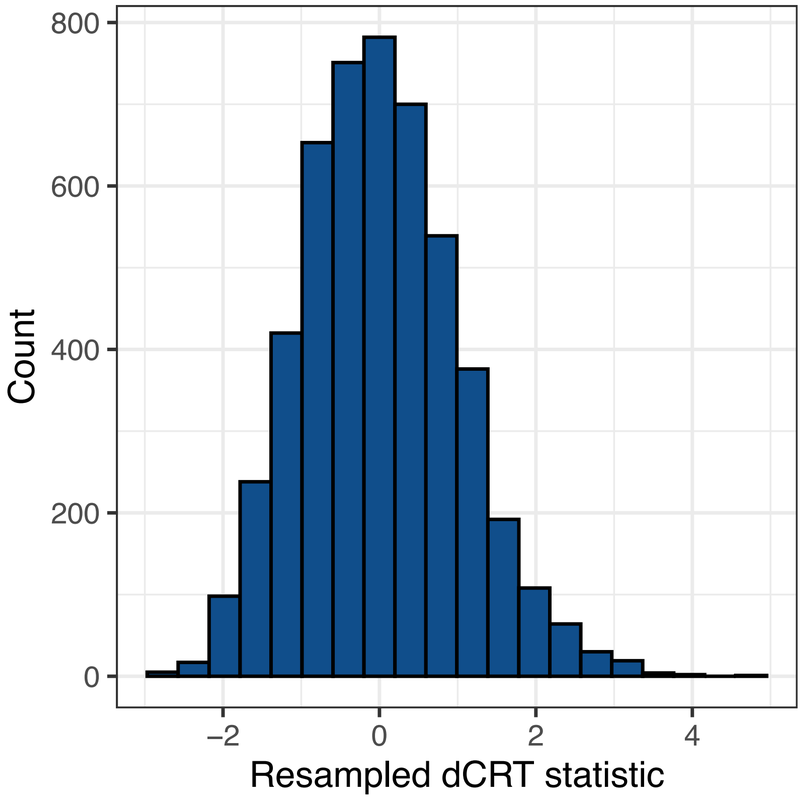


[1]Barry et al. (2021)

[2]Besag and Clifford (1991), Fay et al. (2007), Gandy (2009), Silva and Assuncao (2013), Fischer and Ramdas (2024)

Approximating dCRT p-values

- Draw moderate number of resamples and fit parametric curve to distribution.[1]
- Draw an adaptive number of resamples based on test statistic.[2]
- Use the saddlepoint approximation for sums of independent random variables.[3]



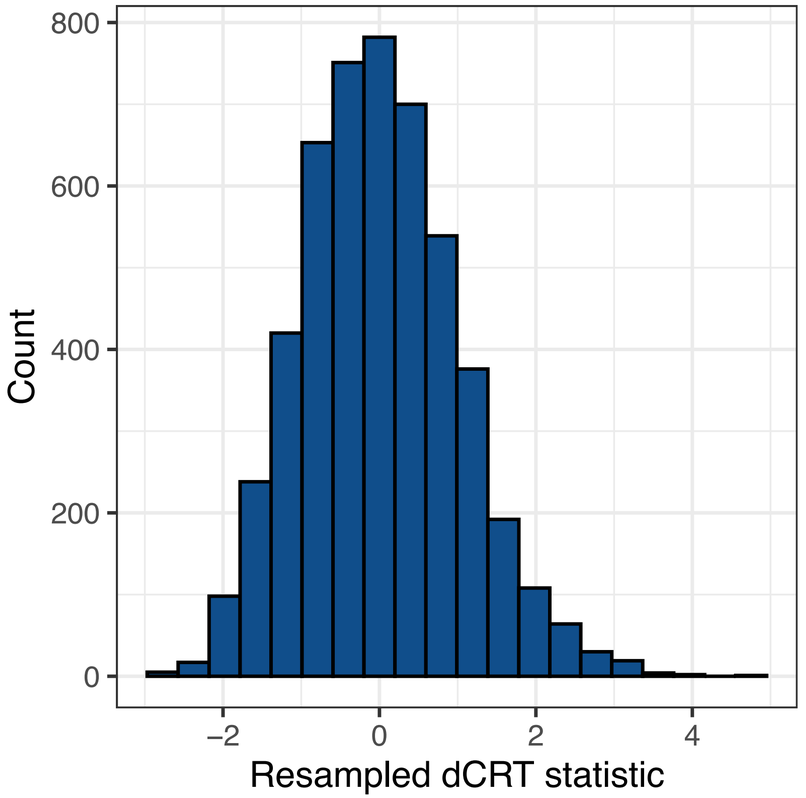
[1]Barry et al. (2021)

[2]Besag and Clifford (1991), Fay et al. (2007), Gandy (2009), Silva and Assuncao (2013), Fischer and Ramdas (2024)

[3]Lugannani and Rice (1980), Robinson (1982), Davison and Hinkley (1988), Jing et al. (1994)

Approximating dCRT p-values

- Draw moderate number of resamples and fit parametric curve to distribution.[1]
- Draw an adaptive number of resamples based on test statistic.[2]
- Use the saddlepoint approximation for sums of independent random variables.[3]



[1]Barry et al. (2021)

[2]Besag and Clifford (1991), Fay et al. (2007), Gandy (2009), Silva and Assuncao (2013), Fischer and Ramdas (2024)

[3]Lugannani and Rice (1980), Robinson (1982), Davison and Hinkley (1988), Jing et al. (1994)

dCRT p-value as conditional tail probability

dCRT p-value as conditional tail probability

We wish to approximate the tail probability

$$\mathbb{P}\left[\tilde{T}_n \geq T_n \mid X, Y, Z\right] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i)) \geq T_n \mid X, Y, Z\right].$$

dCRT p-value as conditional tail probability

We wish to approximate the tail probability

$$\mathbb{P}\left[\tilde{T}_n \geq T_n \mid X, Y, Z\right] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i)) \geq T_n \mid X, Y, Z\right].$$

Define

$$W_{in} = (\tilde{X}_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i)); \quad w_n = T_n; \quad \mathcal{F}_n = \sigma(X, Y, Z).$$

dCRT p-value as conditional tail probability

We wish to approximate the tail probability

$$\mathbb{P}\left[\tilde{T}_n \geq T_n \mid X, Y, Z\right] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{X}_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i)) \geq T_n \mid X, Y, Z\right].$$

Define

$$W_{in} = (\tilde{X}_i - \hat{\pi}(Z_i))(Y_i - \hat{\mu}(Z_i)); \quad w_n = T_n; \quad \mathscr{F}_n = \sigma(X, Y, Z).$$

Therefore, we wish to approximate

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathscr{F}_n\right],$$

where W_{in} are independent and mean-zero conditionally on $\mathscr{F}_n$ and $w_n \in \mathscr{F}_n$.

Saddlepoint approximations

Saddlepoint approximations

In unconditional case, Lugannani-Rice formula[1] gives approximation $\hat{p}_n^{\mathsf{LR}}$ of

$$p_n = \mathbb{P}_{\mathsf{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w\right]$$

such that

[1]Lugannani and Rice (1980)

Saddlepoint approximations

In unconditional case, Lugannani-Rice formula[1] gives approximation $\hat{p}_n^{\mathsf{LR}}$ of

$$p_n = \mathbb{P}_{\mathsf{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w\right]$$

such that

$$|\hat{p}_n^{\mathsf{LR}} - p_n| \leq p_n \cdot O(n^{-1}),$$

[1]Lugannani and Rice (1980)

Saddlepoint approximations

In unconditional case, Lugannani-Rice formula[1] gives approximation $\hat{p}_n^{\text{LR}}$ of

$$p_n = \mathbb{P}_{\text{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w\right]$$

such that

$$|\hat{p}_n^{\text{LR}} - p_n| \leq p_n \cdot O(n^{-1}),$$

as long as

1. $\mathbb{E}[\exp(sW_{in})] < \infty$ for all $s \in (-\epsilon, \epsilon)$ and all i, n;
2. W_{in} have sufficiently smooth distributions.

[1]Lugannani and Rice (1980)

Challenges for SPA application to resampling

Challenges for SPA application to resampling

1. Conditioning. We are interested in the conditional probability

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathscr{F}_n\right].$$

   The extra randomness in $\mathscr{F}_n$ must be accounted for.

Challenges for SPA application to resampling

1. Conditioning. We are interested in the conditional probability

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathscr{F}_n\right].$$

   The extra randomness in $\mathscr{F}_n$ must be accounted for.

2. Lack of smoothness. Resampling distributions typically do not meet smoothness assumptions required for existing SPA results.

Challenges for SPA application to resampling

1. Conditioning. We are interested in the conditional probability

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathscr{F}_n\right].$$

   The extra randomness in $\mathscr{F}_n$ must be accounted for.

2. Lack of smoothness. Resampling distributions typically do not meet smoothness assumptions required for existing SPA results.

Existing applications of SPA to resampling distributions have ignored these.

SPA for conditional probabilities

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

(a) $\mathbb{P}[\,|W_{in}| \geq t \mid \mathscr{F}_n] \leq \theta_n \exp(-\beta t)$ for all $t > 0$, with $\theta_n \in \mathscr{F}_n$, $\theta_n = O_{\mathbb{P}}(1)$

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

(a) $\mathbb{P}[\,|W_{in}| \geq t \mid \mathscr{F}_n] \leq \theta_n \exp(-\beta t)$ for all $t > 0$, with $\theta_n \in \mathscr{F}_n$, $\theta_n = O_{\mathbb{P}}(1)$

(b) $W_{in} \in [-\nu_{in}, \nu_{in}]$ almost surely, with $\nu_{in} \in \mathscr{F}_n$, $\frac{1}{n}\sum_{i=1}^{n} \nu_{in}^4 = O_{\mathbb{P}}(1)$

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

(a) $\mathbb{P}[\,|W_{in}| \geq t \mid \mathscr{F}_n] \leq \theta_n \exp(-\beta t)$ for all $t > 0$, with $\theta_n \in \mathscr{F}_n$, $\theta_n = O_{\mathbb{P}}(1)$

(b) $W_{in} \in [-\nu_{in}, \nu_{in}]$ almost surely, with $\nu_{in} \in \mathscr{F}_n$, $\frac{1}{n}\sum_{i=1}^{n} \nu_{in}^4 = O_{\mathbb{P}}(1)$

For $w_n \in \mathscr{F}_n$ such that $w_n \xrightarrow{\mathbb{P}} 0$, LR approx.

$\hat{p}_n^{\mathsf{LR}}$ of $p_n = \mathbb{P}_{\mathsf{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathscr{F}_n\right]$

satisfies $|\hat{p}_n^{\mathsf{LR}} - p_n| \leq p_n \cdot o_{\mathbb{P}}(1)$.

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

(a) $\mathbb{P}[\,|W_{in}| \geq t \mid \mathscr{F}_n] \leq \theta_n \exp(-\beta t)$ for all $t > 0$, with $\theta_n \in \mathscr{F}_n$, $\theta_n = O_{\mathbb{P}}(1)$

(b) $W_{in} \in [-\nu_{in}, \nu_{in}]$ almost surely, with $\nu_{in} \in \mathscr{F}_n$, $\frac{1}{n}\sum_{i=1}^{n} \nu_{in}^4 = O_{\mathbb{P}}(1)$

For $w_n \in \mathscr{F}_n$ such that $w_n \xrightarrow{\mathbb{P}} 0$, LR approx.

$\hat{p}_n^{\mathsf{LR}}$ of $p_n = \mathbb{P}_{\mathsf{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathscr{F}_n\right]$

satisfies $|\hat{p}_n^{\mathsf{LR}} - p_n| \leq p_n \cdot o_{\mathbb{P}}(1)$.

- Explicit conditioning

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

(a) $\mathbb{P}[\,|W_{in}| \geq t \mid \mathcal{F}_n] \leq \theta_n \exp(-\beta t)$ for all $t > 0$, with $\theta_n \in \mathcal{F}_n$, $\theta_n = O_{\mathbb{P}}(1)$

(b) $W_{in} \in [-\nu_{in}, \nu_{in}]$ almost surely, with $\nu_{in} \in \mathcal{F}_n$, $\frac{1}{n}\sum_{i=1}^{n} \nu_{in}^4 = O_{\mathbb{P}}(1)$

For $w_n \in \mathcal{F}_n$ such that $w_n \xrightarrow{\mathbb{P}} 0$, LR approx.

$\hat{p}_n^{\mathrm{LR}}$ of $p_n = \mathbb{P}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right]$

satisfies $|\hat{p}_n^{\mathrm{LR}} - p_n| \leq p_n \cdot o_{\mathbb{P}}(1)$.

- Explicit conditioning
- No smoothness assumption

SPA for conditional probabilities

Theorem (Niu et al., 2024, informal)

Suppose at least one of the following holds:

(a) $\mathbb{P}[\,|W_{in}| \geq t \mid \mathcal{F}_n] \leq \theta_n \exp(-\beta t)$ for all $t > 0$, with $\theta_n \in \mathcal{F}_n$, $\theta_n = O_{\mathbb{P}}(1)$

(b) $W_{in} \in [-\nu_{in}, \nu_{in}]$ almost surely, with $\nu_{in} \in \mathcal{F}_n$, $\frac{1}{n}\sum_{i=1}^{n} \nu_{in}^4 = O_{\mathbb{P}}(1)$

For $w_n \in \mathcal{F}_n$ such that $w_n \xrightarrow{\mathbb{P}} 0$, LR approx.

$\hat{p}_n^{\mathrm{LR}}$ of $p_n = \mathbb{P}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right]$

satisfies $|\hat{p}_n^{\mathrm{LR}} - p_n| \leq p_n \cdot o_{\mathbb{P}}(1)$.

- Explicit conditioning
- No smoothness assumption
- Cost: Lack of rate

The spaCRT

The spaCRT

spaCRT algorithm (Niu et al., 2024)

The spaCRT

spaCRT algorithm (Niu et al., 2024)

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.

The spaCRT

spaCRT algorithm (Niu et al., 2024)

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.

2. Let $\hat{s}_n$ be the solution of the saddlepoint equation in s:

$$\sum_{i=1}^{n} \left\{ X_i - \text{expit}\left(\text{logit}(\hat{\pi}(Z_i)) + s(Y_i - \hat{\mu}(Z_i))\right) \right\} \left\{ Y_i - \hat{\mu}(Z_i) \right\} = 0.$$

The spaCRT

spaCRT algorithm (Niu et al., 2024)

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.

2. Let $\hat{s}_n$ be the solution of the saddlepoint equation in s:

$$\sum_{i=1}^{n} \left\{ X_i - \text{expit}\left(\text{logit}(\hat{\pi}(Z_i)) + s(Y_i - \hat{\mu}(Z_i))\right) \right\} \left\{ Y_i - \hat{\mu}(Z_i) \right\} = 0.$$

3. Letting $\tilde{\pi}_i = \text{expit}\left(\text{logit}(\hat{\pi}(Z_i)) + \hat{s}_n(Y_i - \hat{\mu}(Z_i))\right)$ and $\hat{\pi}_i = \hat{\pi}(Z_i)$, define

$$\lambda_n = \hat{s}_n \sqrt{\sum_{i=1}^{n} (Y_i - \hat{\mu}(Z_i))^2 \tilde{\pi}(1-\tilde{\pi})}; \quad r_n = \text{sgn}(\hat{s}_n) \sqrt{2 \sum_{i=1}^{n} \left\{ X_i \log \frac{\tilde{\pi}}{\hat{\pi}} + (1 - X_i) \log \frac{1-\tilde{\pi}}{1-\hat{\pi}} \right\}}.$$

The spaCRT

spaCRT algorithm (Niu et al., 2024)

1. Learn $\widehat{\pi}(Z_i)$ by regressing X on Z and $\widehat{\mu}(Z_i) \approx \mathbb{E}[Y_i \mid Z_i]$ by regressing Y on Z.

2. Let $\hat{s}_n$ be the solution of the saddlepoint equation in s:

$$\sum_{i=1}^{n} \left\{ X_i - \text{expit}\left(\text{logit}(\hat{\pi}(Z_i)) + s(Y_i - \hat{\mu}(Z_i))\right) \right\} \left\{ Y_i - \hat{\mu}(Z_i) \right\} = 0.$$

3. Letting $\tilde{\pi}_i = \text{expit}\left(\text{logit}(\hat{\pi}(Z_i)) + \hat{s}_n(Y_i - \hat{\mu}(Z_i))\right)$ and $\hat{\pi}_i = \hat{\pi}(Z_i)$, define

$$\lambda_n = \hat{s}_n \sqrt{\sum_{i=1}^{n} (Y_i - \hat{\mu}(Z_i))^2 \tilde{\pi}(1-\tilde{\pi})}; \quad r_n = \text{sgn}(\hat{s}_n) \sqrt{2 \sum_{i=1}^{n} \left\{ X_i \log \frac{\tilde{\pi}}{\hat{\pi}} + (1 - X_i) \log \frac{1-\tilde{\pi}}{1-\hat{\pi}} \right\}}.$$

4. Return $p_n^{\text{spaCRT}} = 1 - \Phi(r_n) + \phi(r_n)\left(\frac{1}{\lambda_n} - \frac{1}{r_n}\right)$.

spaCRT is a good approximation to dCRT

spaCRT is a good approximation to dCRT

Theorem (Niu et al., 2024, informal)

spaCRT is a good approximation to dCRT

Theorem (Niu et al., 2024, informal)

Suppose $\hat{\pi}$ and $\hat{\mu}$ consistent for $\pi(Z_i) = \mathbb{P}[X_i = 1 \mid Z_i]$ and $\mu(Z_i) = \mathbb{E}[Y_i \mid Z_i]$, and $\sup_n \mathbb{E}[Y_{in}^4] < \infty$. Then,

$$|p_n^{\text{spaCRT}} - p_n^{\text{dCRT}}| \leq p_n^{\text{dCRT}} \cdot o_{\mathbb{P}}(1).$$

spaCRT is a good approximation to dCRT

Theorem (Niu et al., 2024, informal)

Suppose $\hat{\pi}$ and $\hat{\mu}$ consistent for $\pi(Z_i) = \mathbb{P}[X_i = 1 \mid Z_i]$ and $\mu(Z_i) = \mathbb{E}[Y_i \mid Z_i]$, and $\sup_n \mathbb{E}[Y_{in}^4] < \infty$. Then,

$$|p_n^{\text{spaCRT}} - p_n^{\text{dCRT}}| \leq p_n^{\text{dCRT}} \cdot o_{\mathbb{P}}(1).$$

Furthermore, the spaCRT and dCRT level-α tests are asymptotically equivalent.

spaCRT is a good approximation to dCRT

Theorem (Niu et al., 2024, informal)

Suppose $\hat{\pi}$ and $\hat{\mu}$ consistent for $\pi(Z_i) = \mathbb{P}[X_i = 1 \mid Z_i]$ and $\mu(Z_i) = \mathbb{E}[Y_i \mid Z_i]$, and $\sup_n \mathbb{E}[Y_{in}^4] < \infty$. Then,

$$|p_n^{\text{spaCRT}} - p_n^{\text{dCRT}}| \leq p_n^{\text{dCRT}} \cdot o_{\mathbb{P}}(1).$$

Furthermore, the spaCRT and dCRT level-α tests are asymptotically equivalent.

Corollary (Niu et al., 2024, informal)

If dCRT controls Type-I error, then under the above assumptions, so does spaCRT.

spaCRT is a good approximation to dCRT

Theorem (Niu et al., 2024, informal)

Suppose $\hat{\pi}$ and $\hat{\mu}$ consistent for $\pi(Z_i) = \mathbb{P}[X_i = 1 \mid Z_i]$ and $\mu(Z_i) = \mathbb{E}[Y_i \mid Z_i]$, and $\sup_n \mathbb{E}[Y_{in}^4] < \infty$. Then,

$$|p_n^{\text{spaCRT}} - p_n^{\text{dCRT}}| \leq p_n^{\text{dCRT}} \cdot o_{\mathbb{P}}(1).$$

Furthermore, the spaCRT and dCRT level-α tests are asymptotically equivalent.

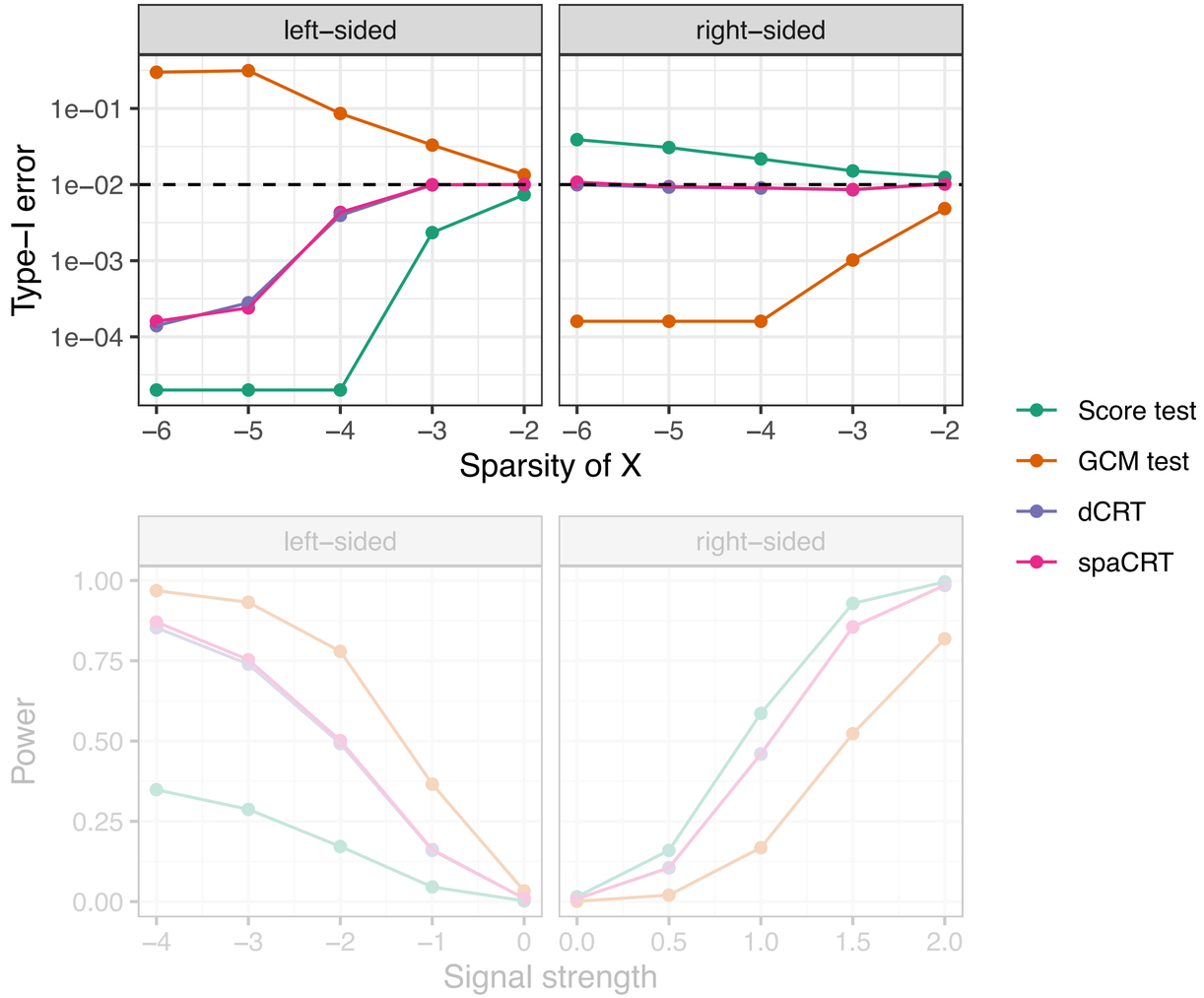
Corollary (Niu et al., 2024, informal)

If dCRT controls Type-I error,[1] then under the above assumptions, so does spaCRT.

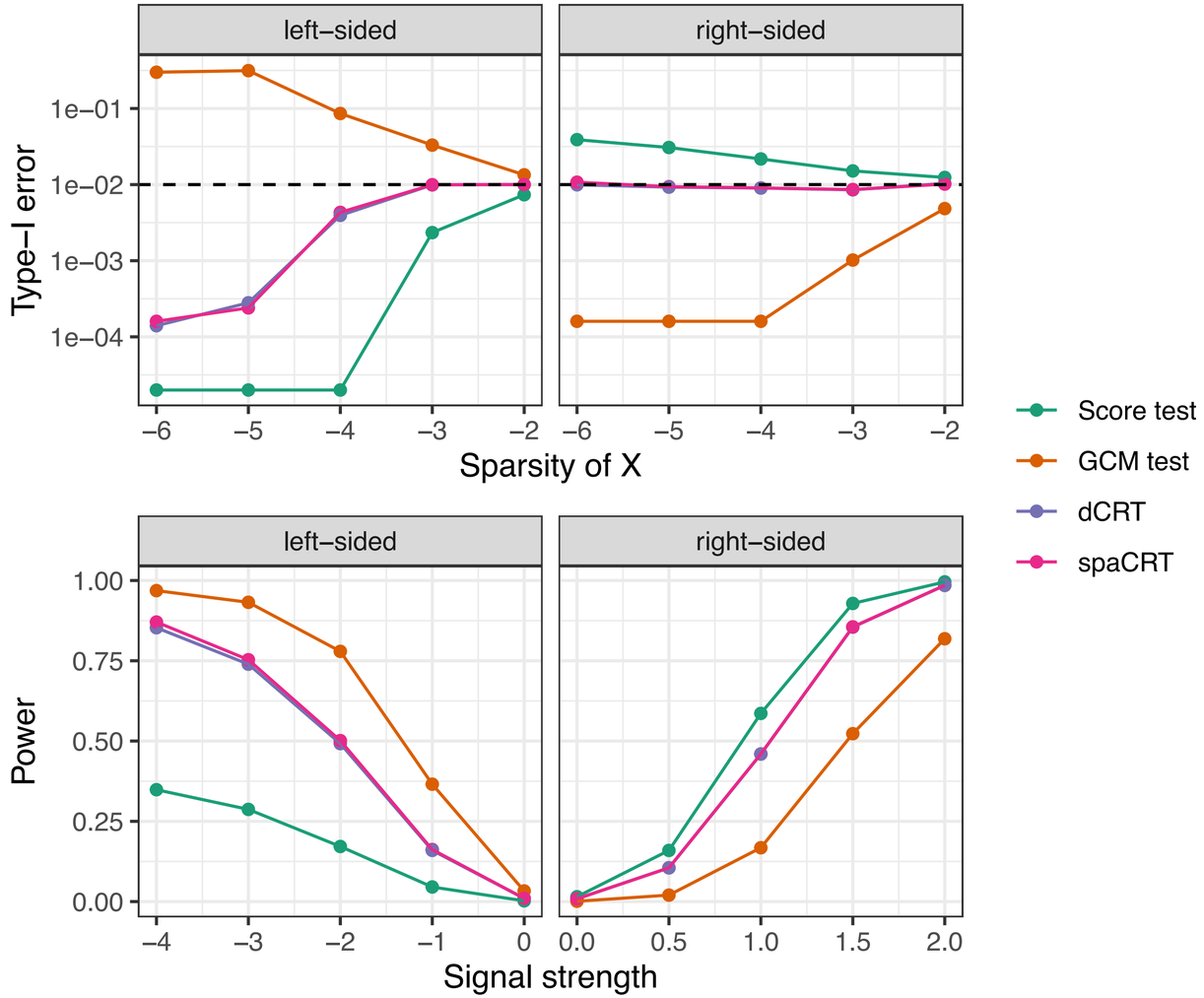
[1] Niu et al. (2022)

Numerical simulation results

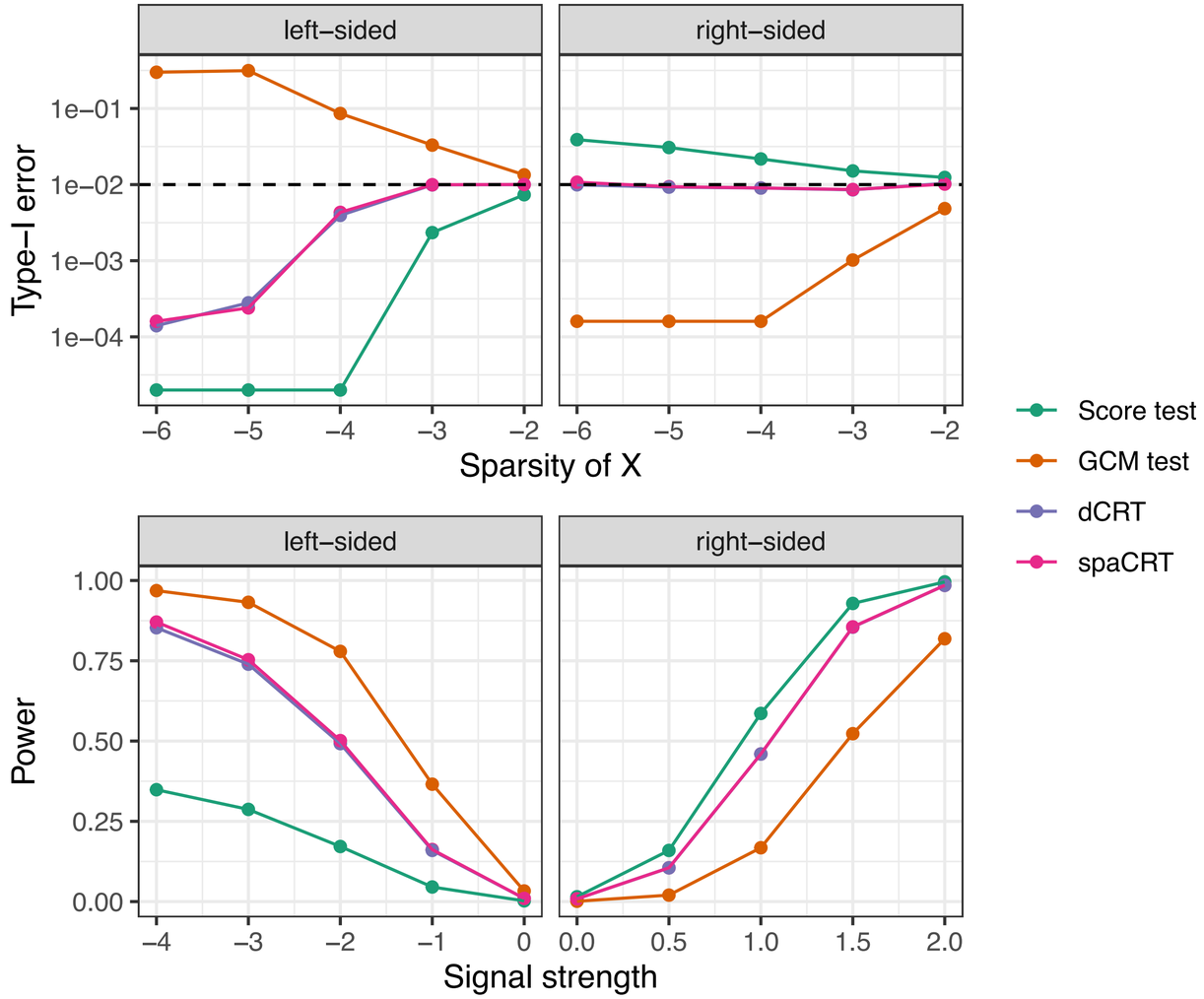
Numerical simulation results

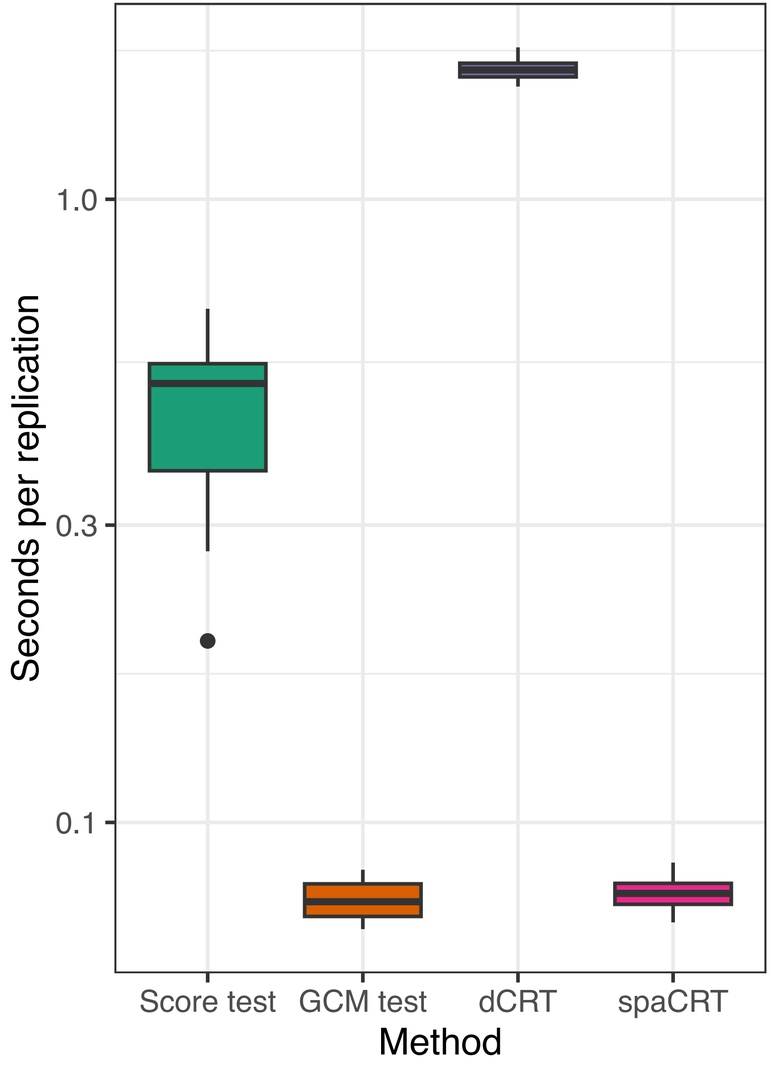


Numerical simulation results



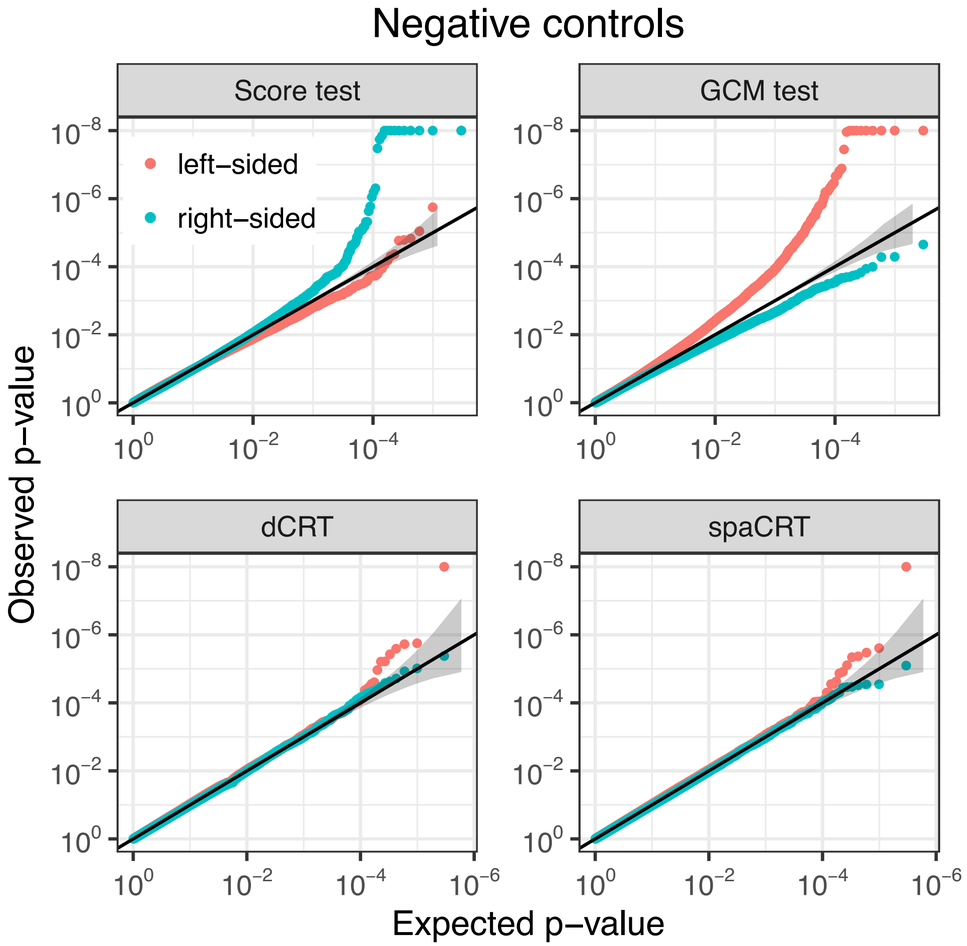
Numerical simulation results



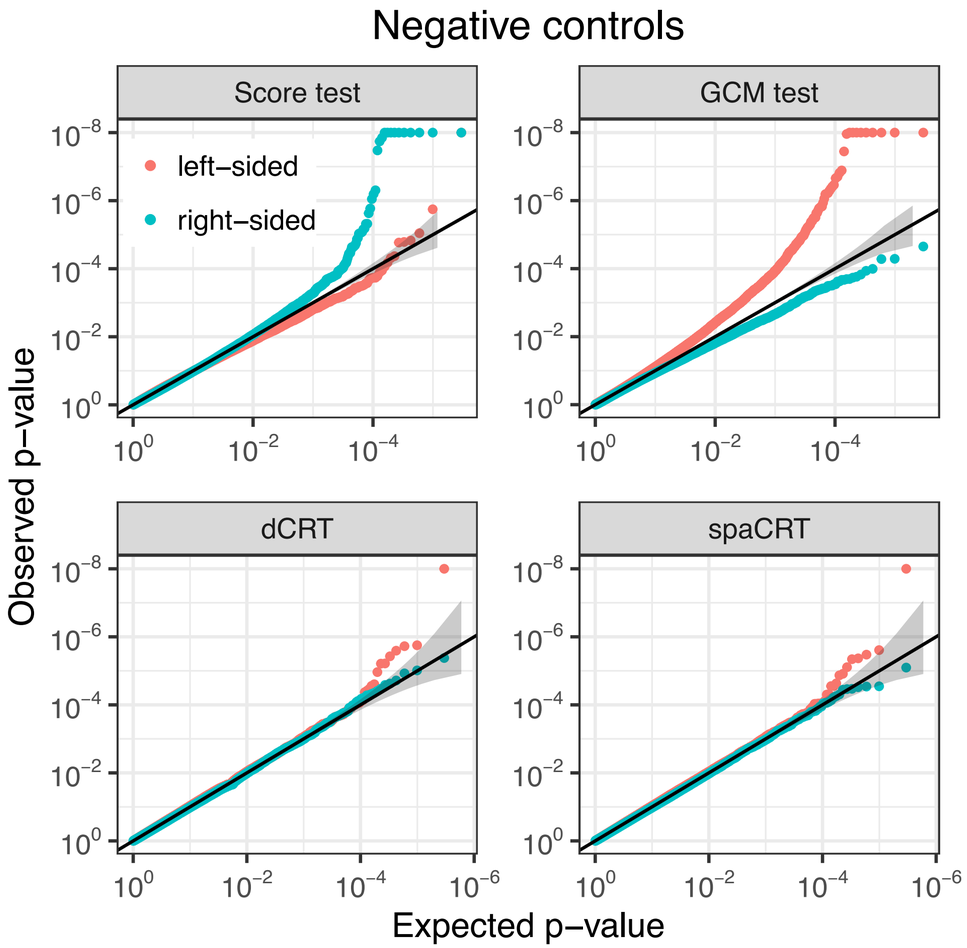


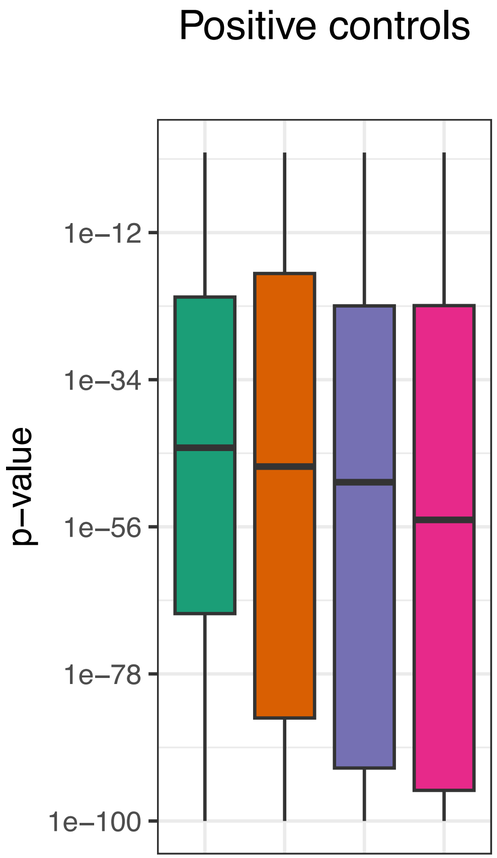
Results on Gasperini data analysis

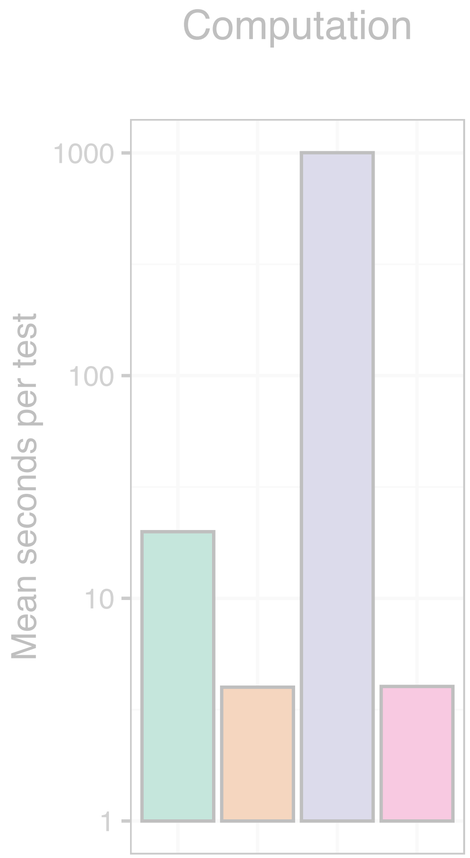
Results on Gasperini data analysis

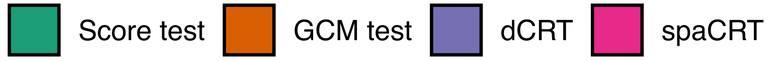


Results on Gasperini data analysis

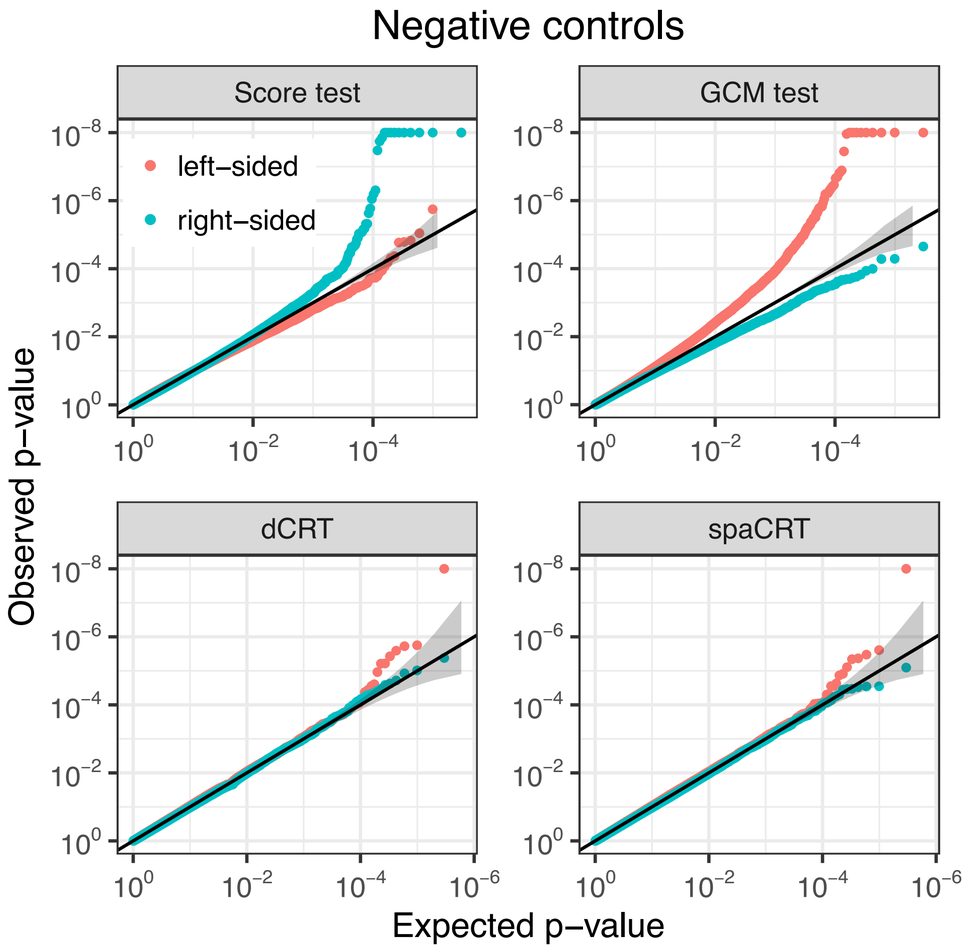


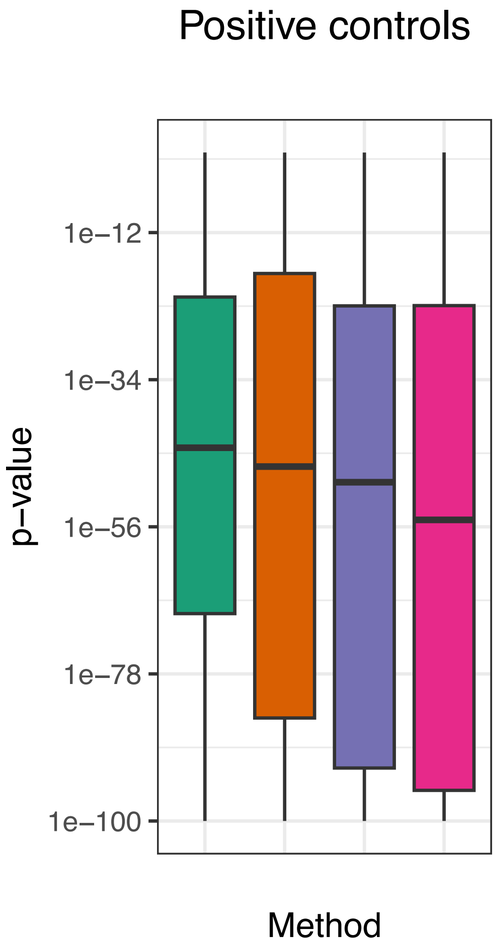


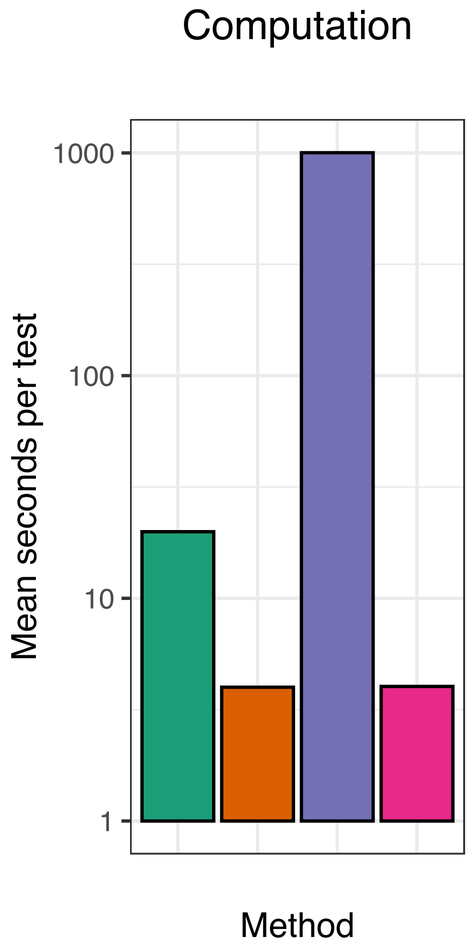


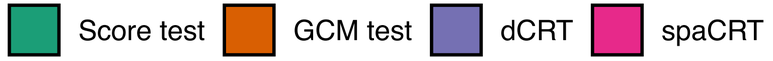


Results on Gasperini data analysis









Discussion

Discussion

Takeaways

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.

	Statistical performance	Computational performance
CLT-based		

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗
SPA-based	✓	✓

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.
- We established theoretical foundation for SPAs for resampling-based procedures.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗
SPA-based	✓	✓

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.
- We established theoretical foundation for SPAs for resampling-based procedures.
- We applied SPA to accelerate the dCRT.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗
SPA-based	✓	✓

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.
- We established theoretical foundation for SPAs for resampling-based procedures.
- We applied SPA to accelerate the dCRT.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗
SPA-based	✓	✓

Limitations

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.
- We established theoretical foundation for SPAs for resampling-based procedures.
- We applied SPA to accelerate the dCRT.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗
SPA-based	✓	✓

Limitations

- We have not yet theoretically justified finite-sample improvement of dCRT.

Discussion

Takeaways

- CLT-based inference can fail in genomics applications with sparse data.
- We established theoretical foundation for SPAs for resampling-based procedures.
- We applied SPA to accelerate the dCRT.

	Statistical performance	Computational performance
CLT-based	✗	✓
Resampling-based	✓	✗
SPA-based	✓	✓

Limitations

- We have not yet theoretically justified finite-sample improvement of dCRT.
- Current results do not cover normalized test statistics or large-deviation regime.

References

1. The saddlepoint approximation for averages of conditionally independent random variables. arXiv, 2024.
2. Computationally efficient and statistically accurate conditional independence testing with spaCRT. arXiv, 2024.