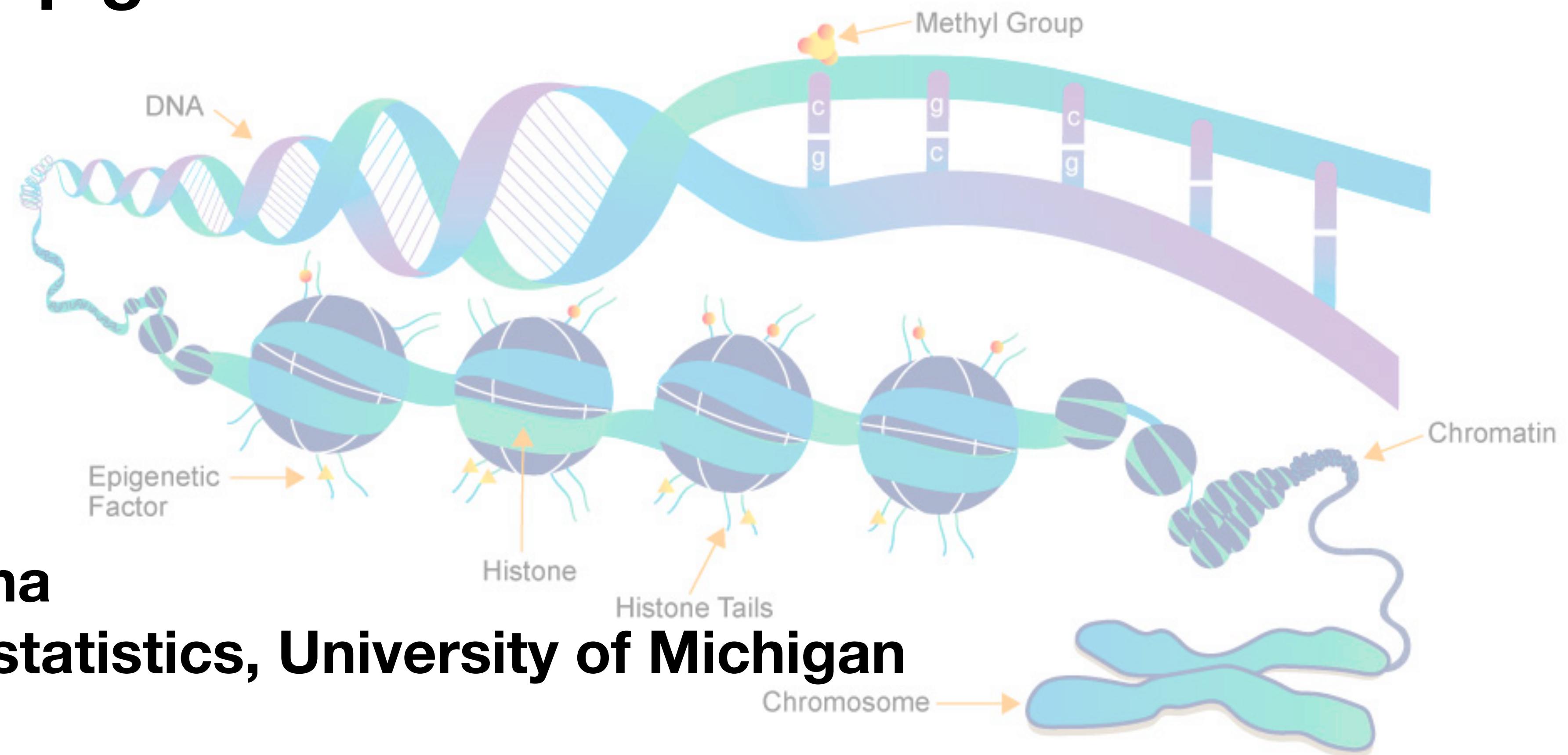


Learning association and directionality

Applications in epigenomics

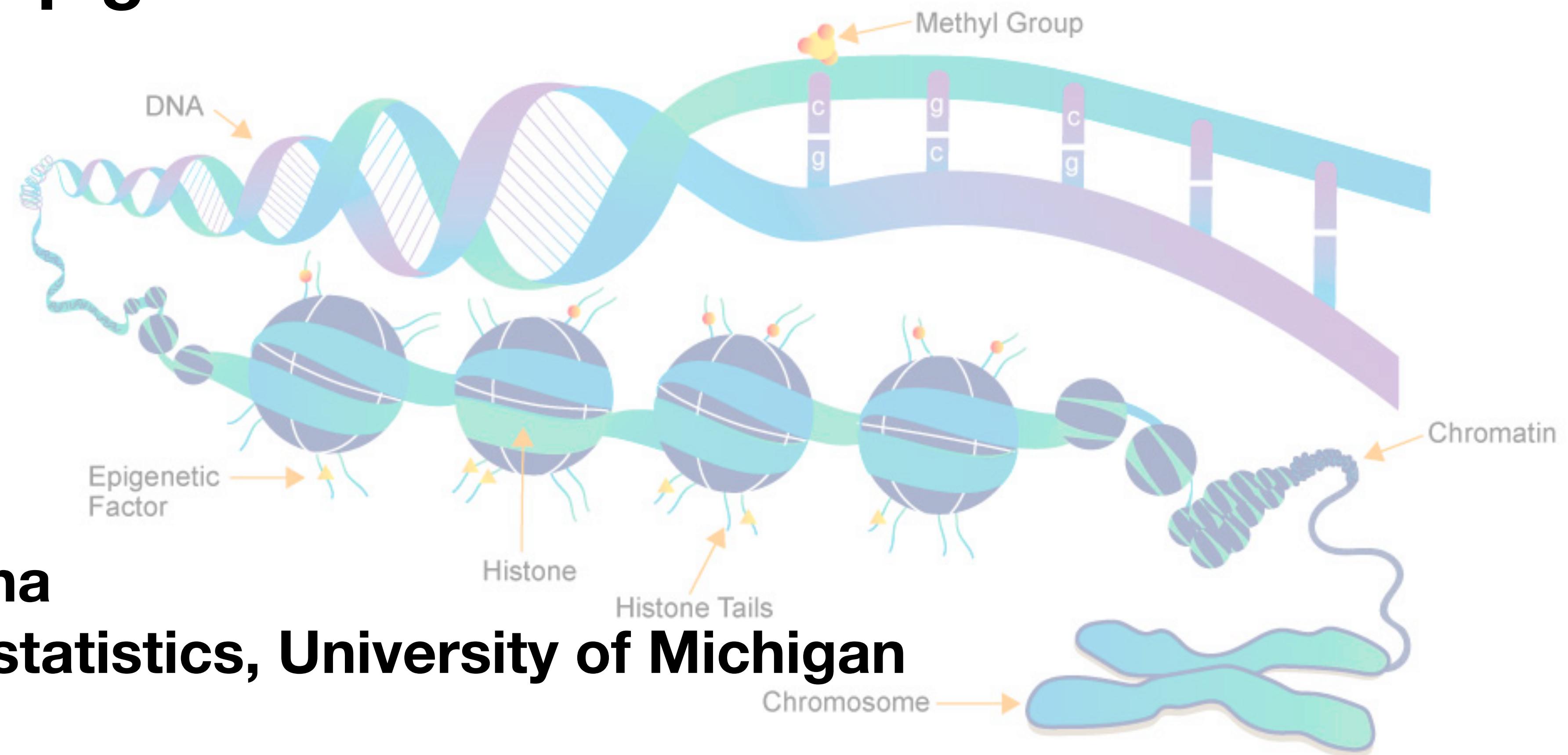


Soumik Purkayastha
Department of Biostatistics, University of Michigan

January 22, 2024

Learning association and directionality

Applications in epigenomics

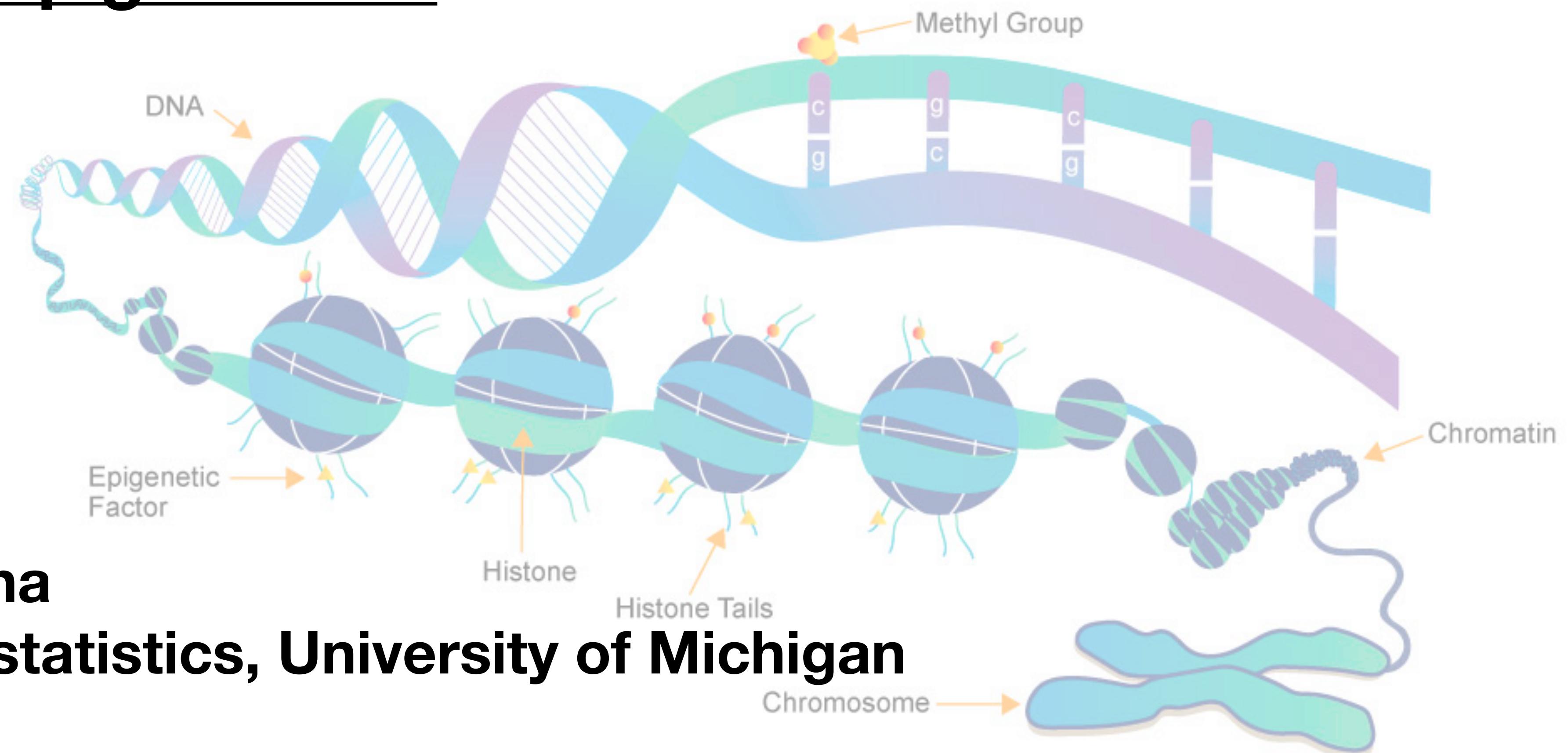


Soumik Purkayastha
Department of Biostatistics, University of Michigan

January 22, 2024

Learning association and directionality

Applications in epigenomics



Soumik Purkayastha
Department of Biostatistics, University of Michigan

January 22, 2024

About my work

- Indian Statistical Institute: 2017 - 19
 - M.Stat. with specialization in Biostatistics.
 - Dissertation: Bregman divergences.

About my work

- Indian Statistical Institute: 2017 - 19
 - M.Stat. with specialization in Biostatistics.
 - Dissertation: Bregman divergences.
- University of Michigan: 2019 - 2024 (expected)
 - M. S. in Biostatistics (RA for **Abecasis** lab).
 - Ph.D. in Biostatistics (RA for **Song** lab).
 - Dissertation: Information theory, statistics and epigenomics.

About my work

- Indian Statistical Institute: 2017 - 19
 - M.Stat. with specialization in Biostatistics.
 - Dissertation: Bregman divergences.
- University of Michigan: 2019 - 2024 (expected)
 - M. S. in Biostatistics (RA for **Abecasis** lab).
 - Ph.D. in Biostatistics (RA for **Song** lab).
 - Dissertation: Information theory, statistics and epigenomics.

About my work

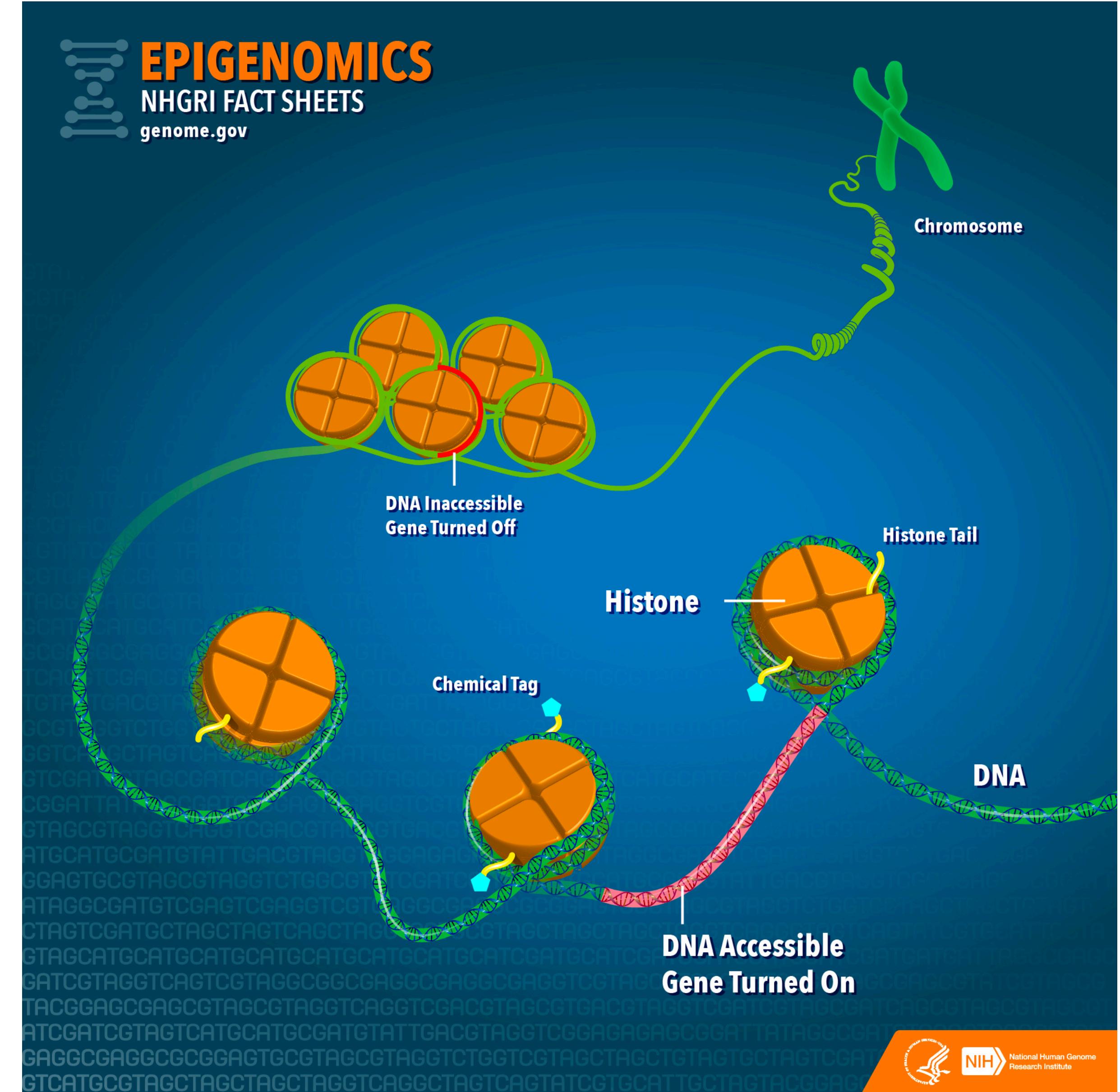
- Indian Statistical Institute: 2017 - 19
 - M.Stat. with specialization in Biostatistics.
 - Dissertation: Bregman divergences.
- University of Michigan: 2019 - 2024 (expected)
 - M. S. in Biostatistics (RA for **Abecasis** lab).
 - Ph.D. in Biostatistics (RA for **Song** lab).
 - Dissertation: Information theory, statistics and epigenomics.

Epigenomics

Motivating problem

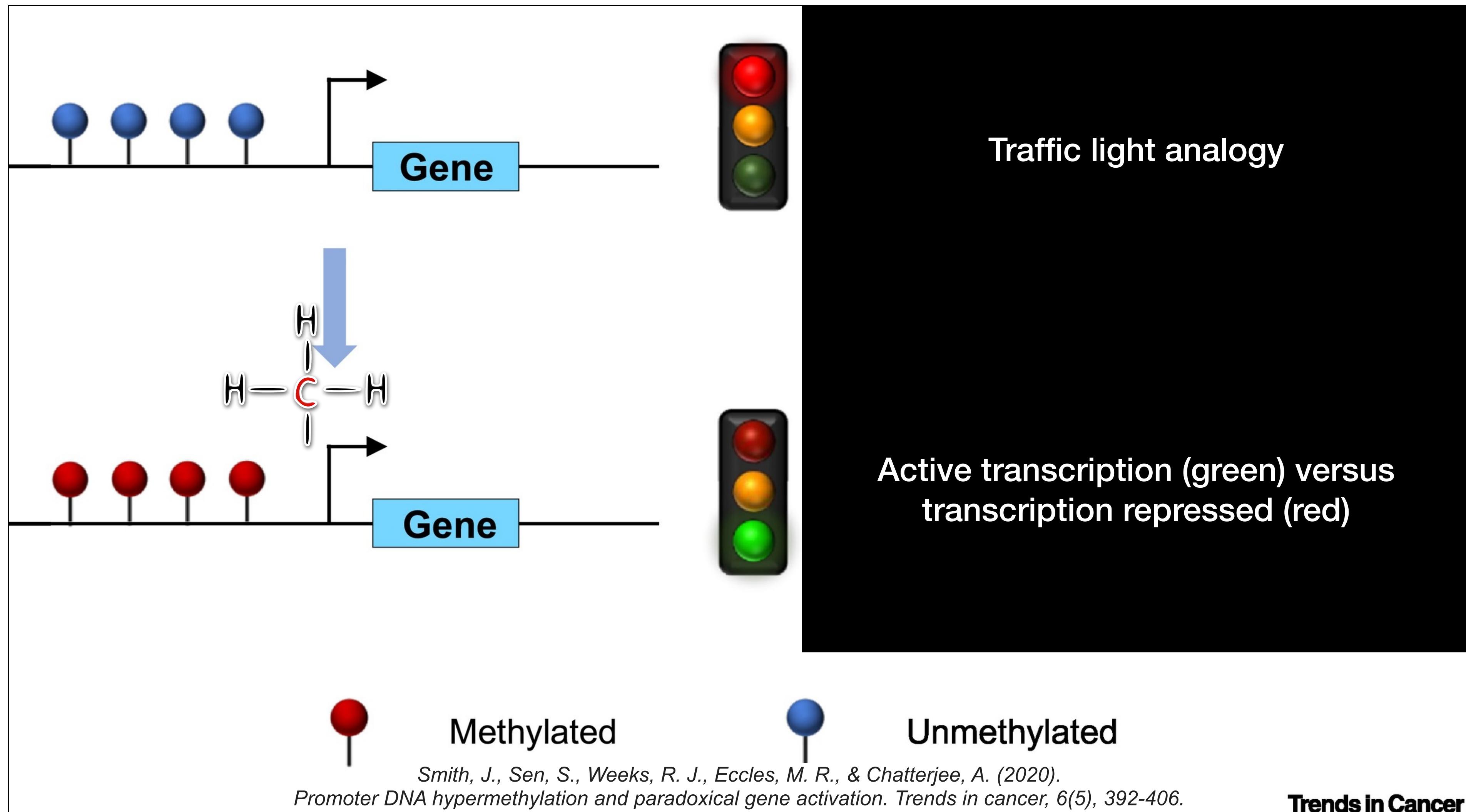
Epigenomics

- Genes regulate how to make proteins.
- Regulates gene expression.
- Impacted by lifestyle, social and environmental determinants.

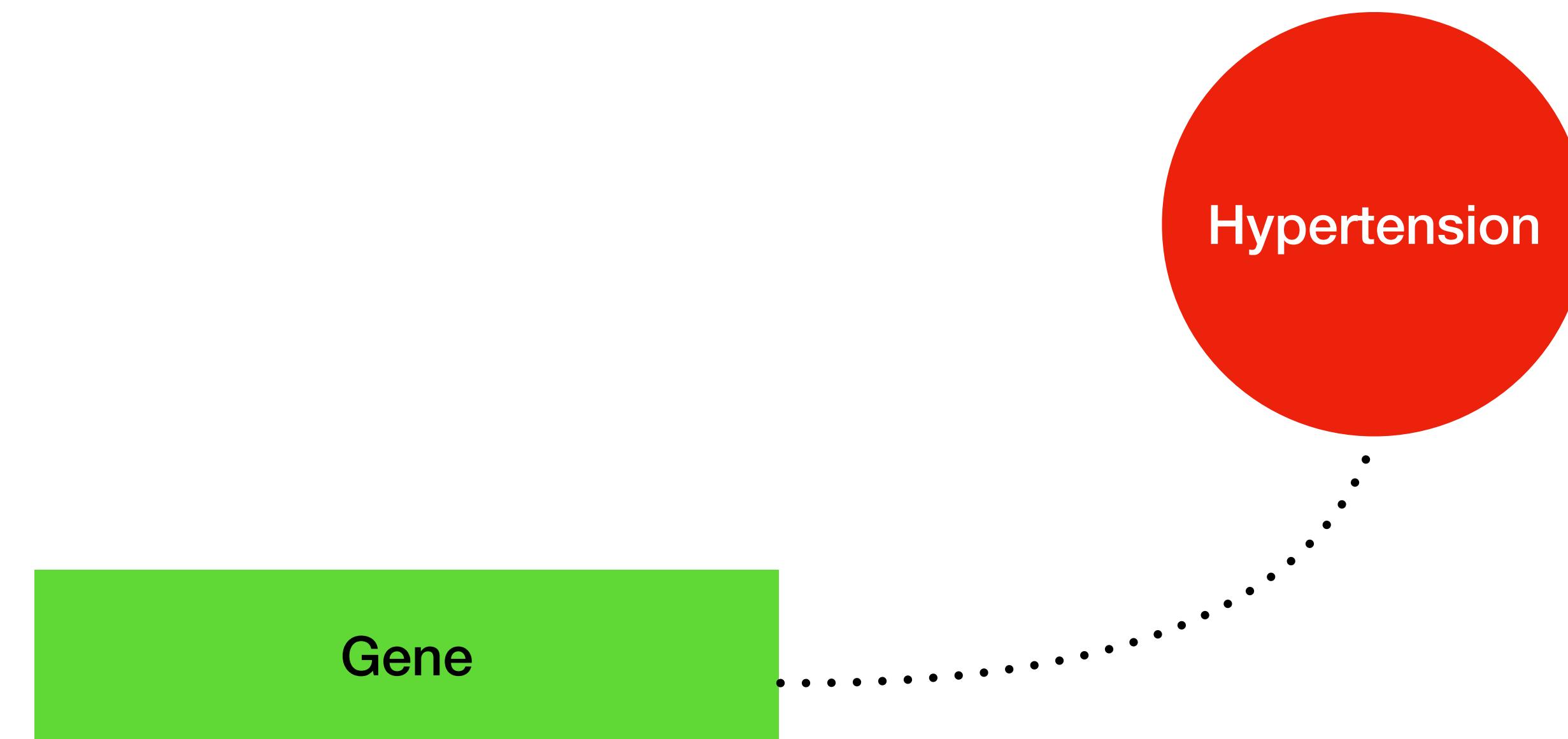


<https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet>

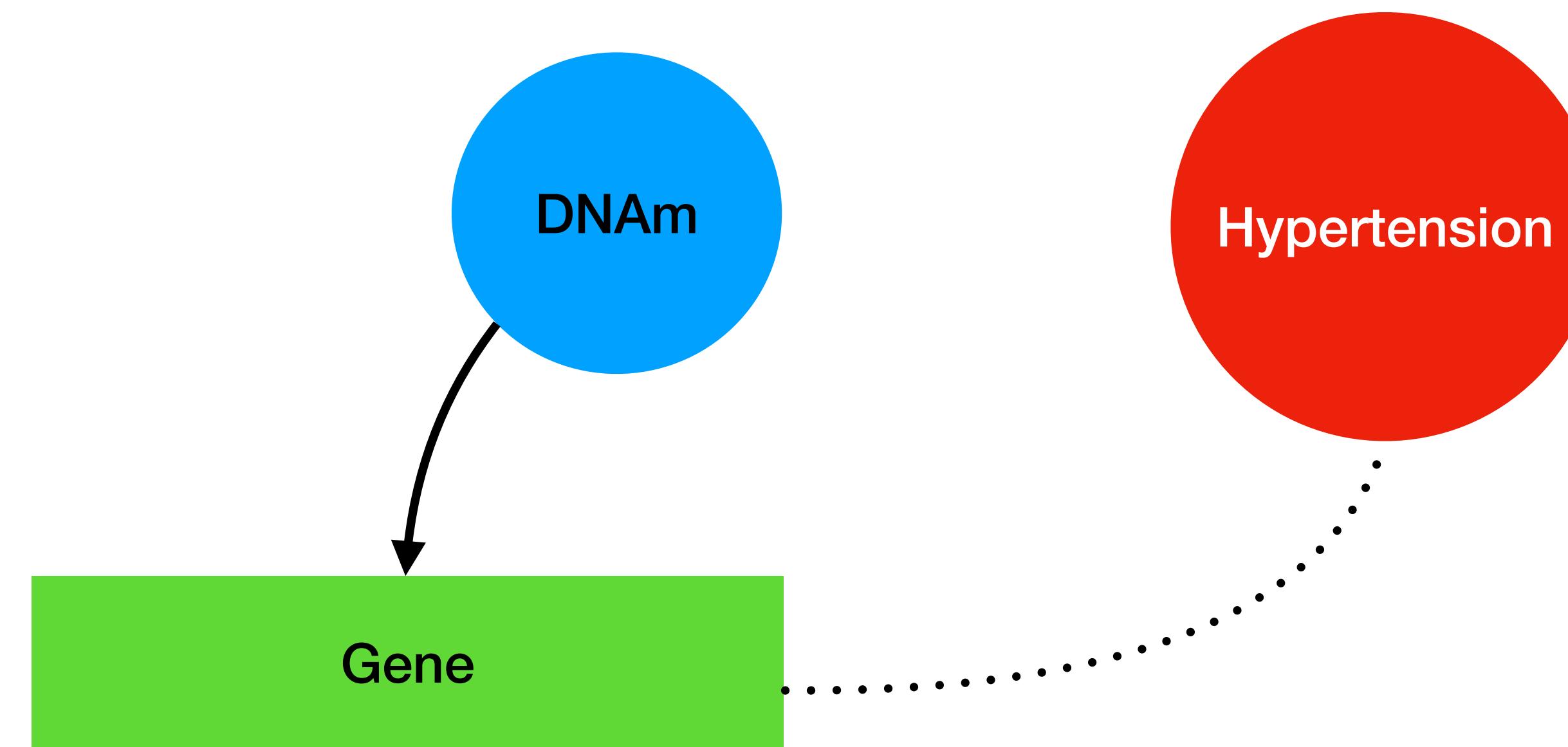
DNA methylation (DNAm)



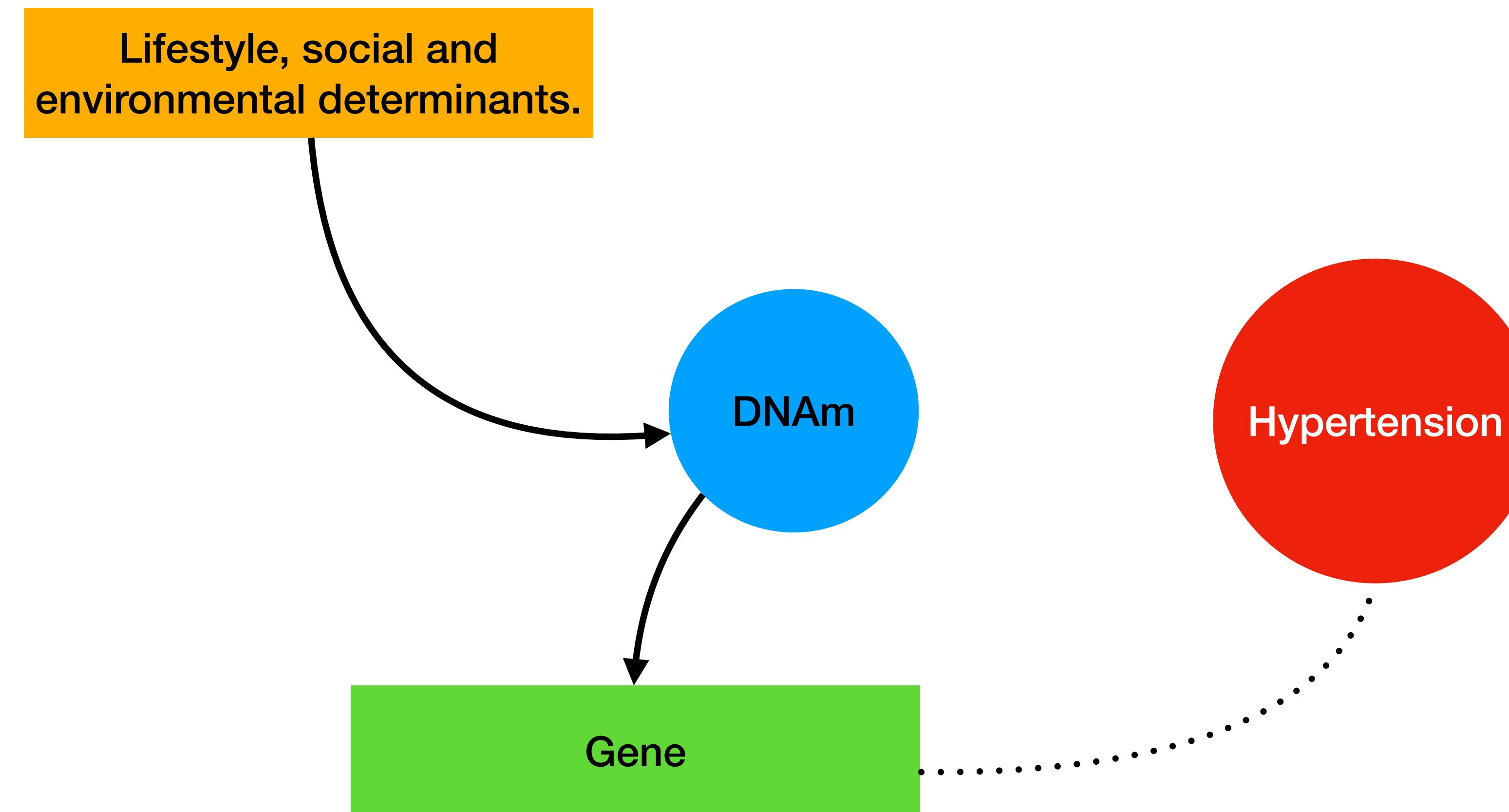
DNAm and cardiovascular diseases



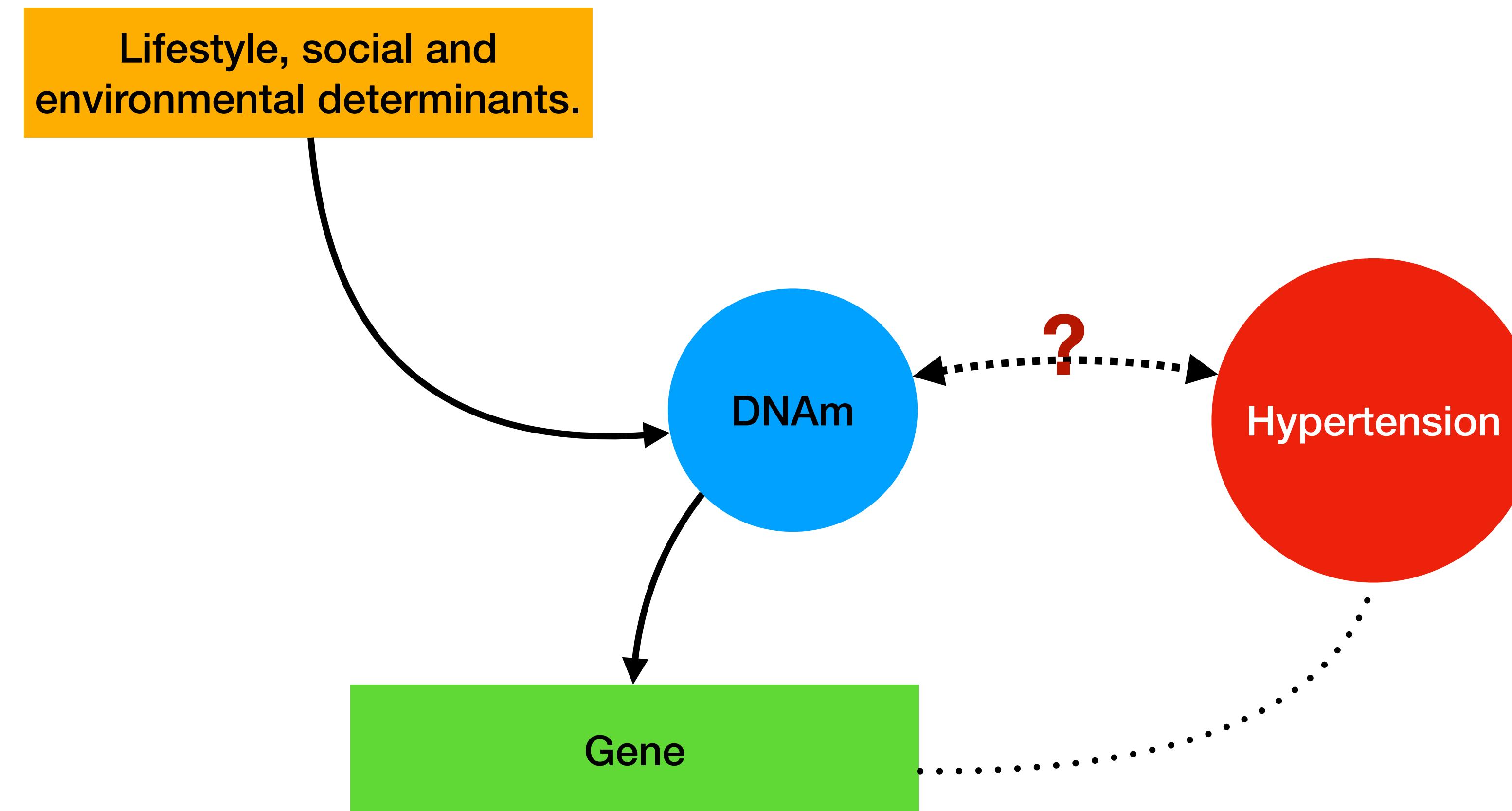
DNAm and cardiovascular diseases



DNAm and cardiovascular diseases



DNAm and cardiovascular diseases



Epigenetics and blood pressure

Leader or follower?

- Conventional knowledge:
DNAm sites are associated
with BP.

Epigenetics and blood pressure

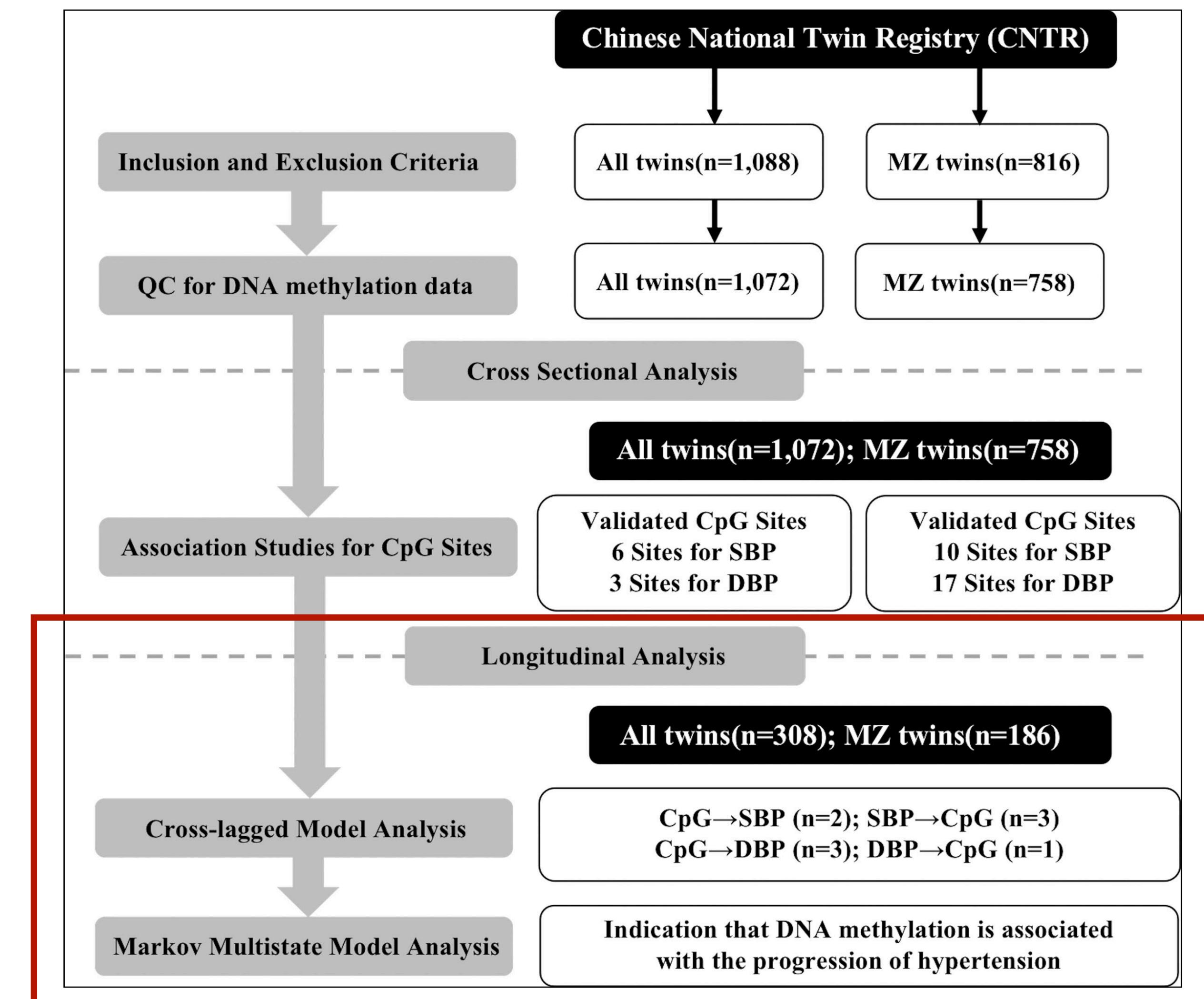
Leader or follower?

- Conventional knowledge:
DNAm sites are associated
with BP.
- Very few studies establish a
direction or ordering between
DNAm and BP.

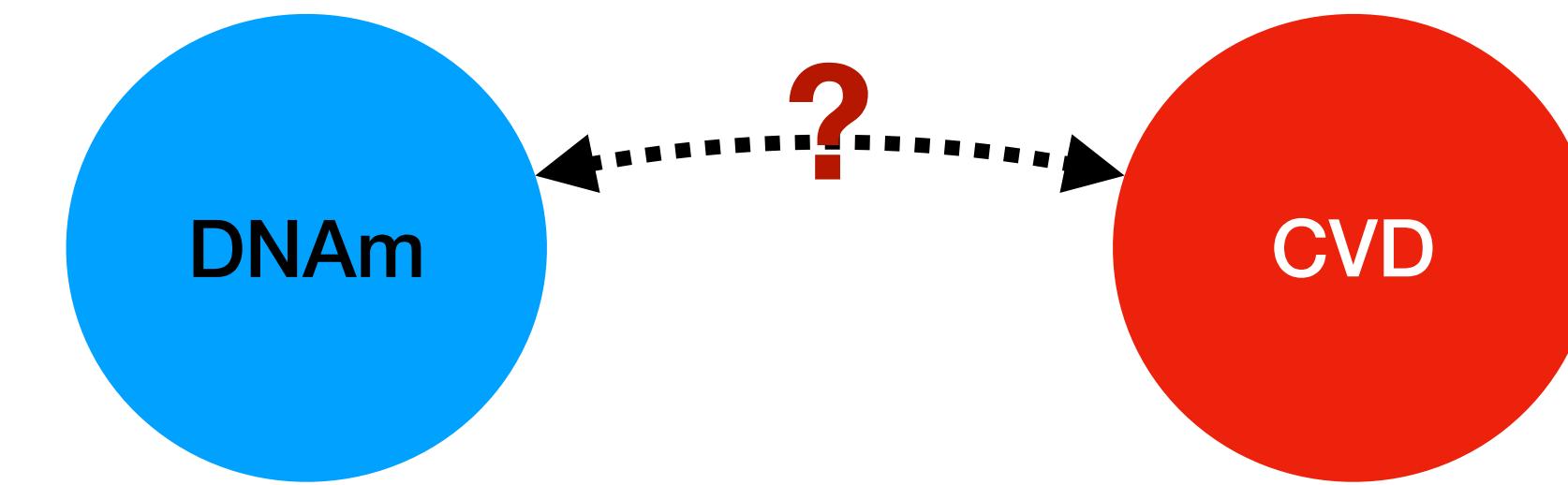
Epigenetics and blood pressure

Leader or follower?

- Conventional knowledge: DNAm sites are associated with BP.
- Very few studies establish a direction or ordering between DNAm and BP.
- Hong et al. (2023): directionality from a predictive angle.



Does DNAm drive BP or vice-versa?



What data?

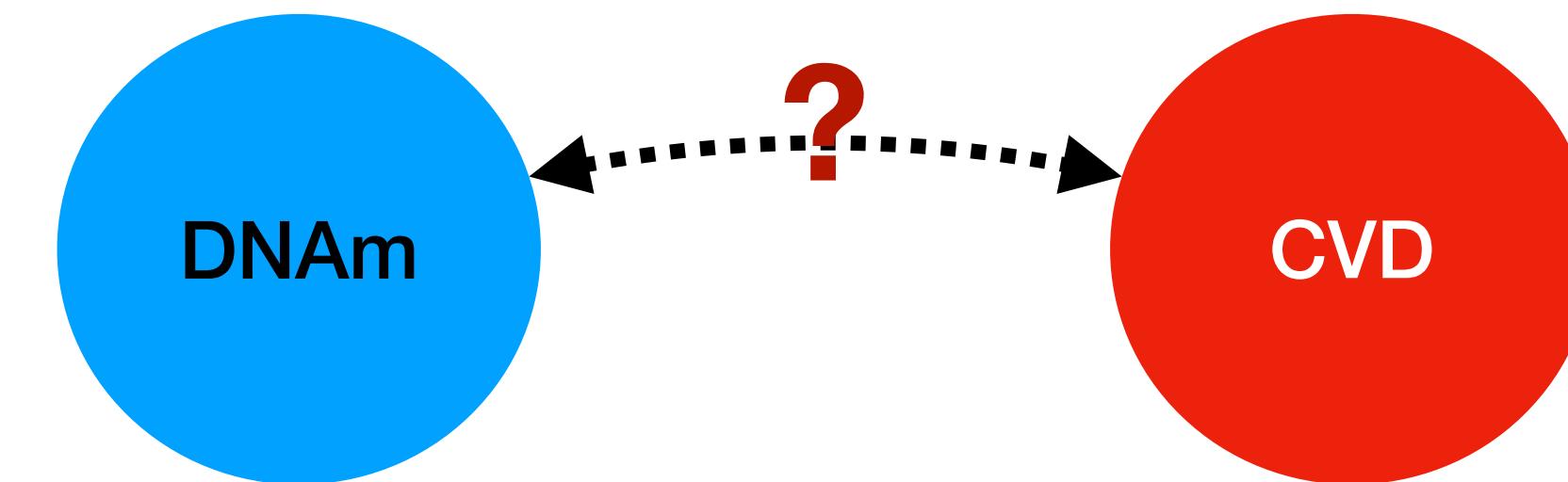
Which genes?

Biological considerations?

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

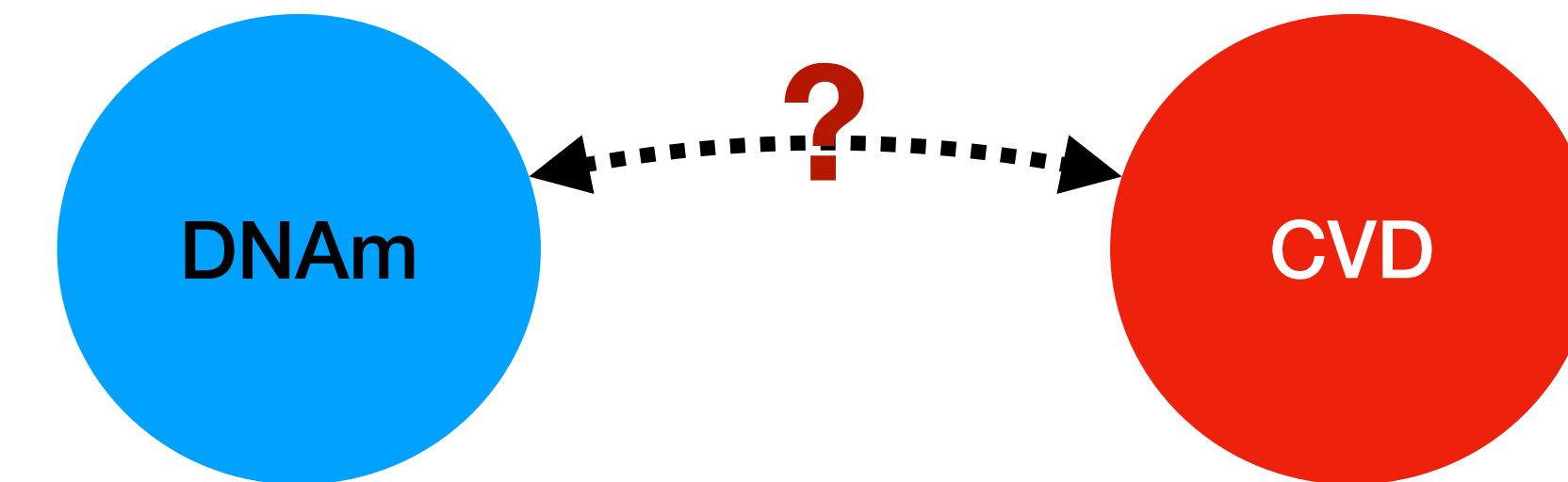
Perng, W., [Tang, L.](#), Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. *Pediatric research*, 85(3), 262-268.



Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Perng, W., Tang, L., Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. *Pediatric research*, 85(3), 262-268.



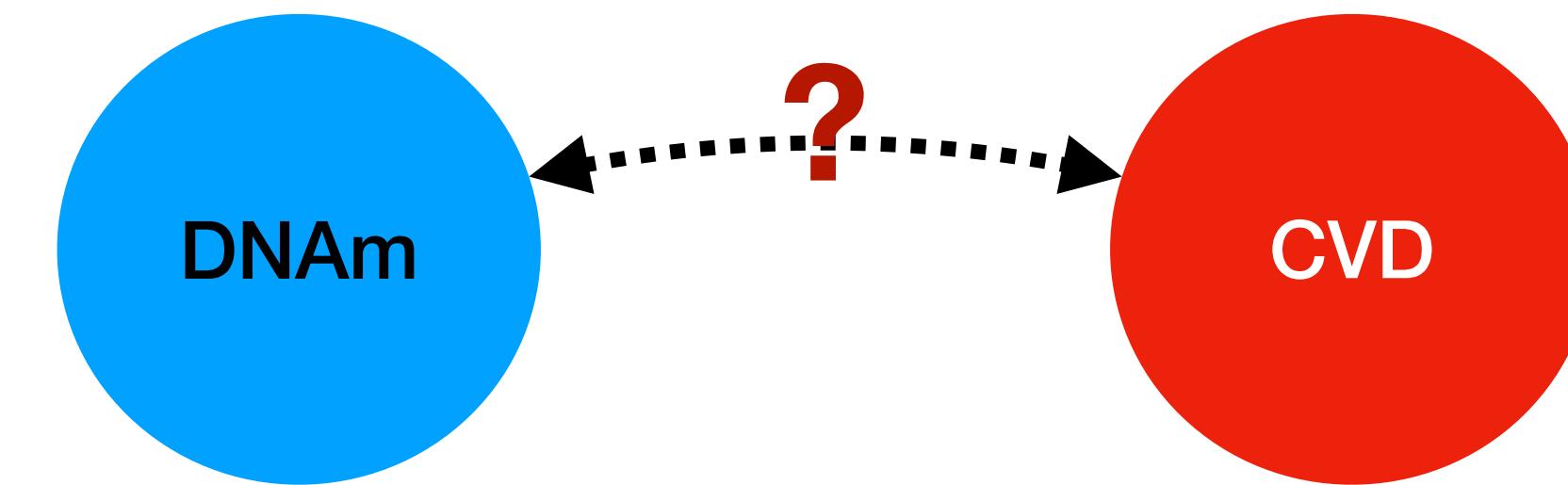
Candidate gene analysis: only examine DNAm for genes that are known to be associated with blood pressure.

ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Perng, W., Tang, L., Song, P. X., Tellez-Rojo, M. M., Cantoral, A., & Peterson, K. E. (2019). Metabolomic profiles and development of metabolic risk during the pubertal transition: a prospective study in the ELEMENT Project. *Pediatric research*, 85(3), 262-268.



Candidate gene analysis: only examine DNAm for genes that are known to be associated with blood pressure.

ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

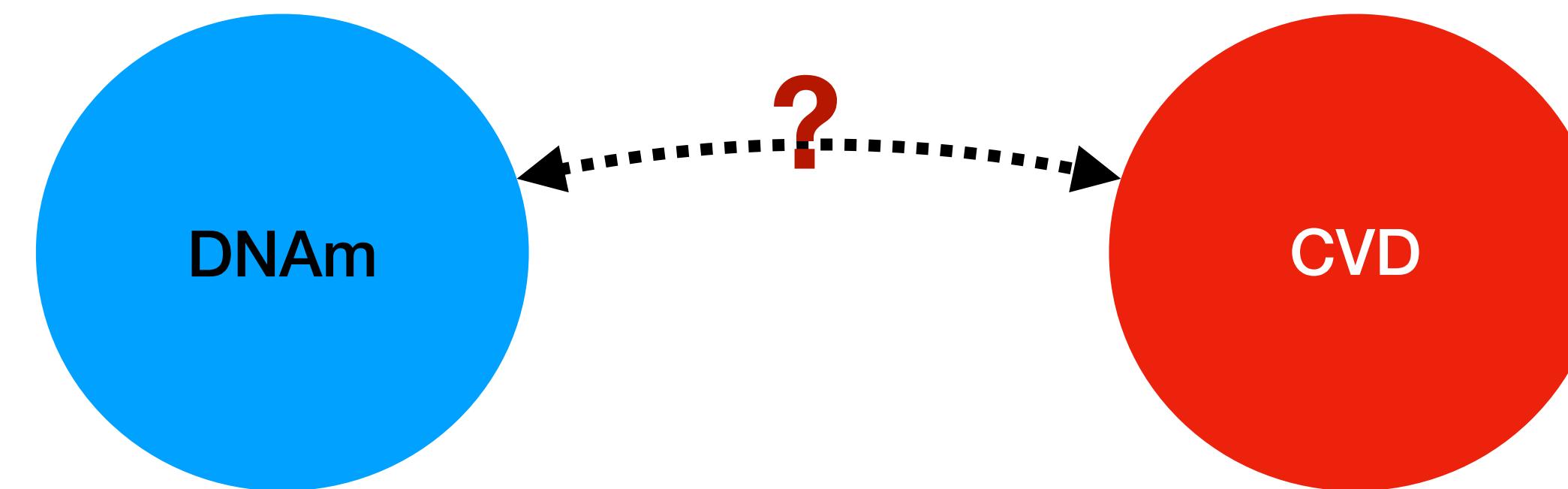
Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Candidate gene analysis: ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

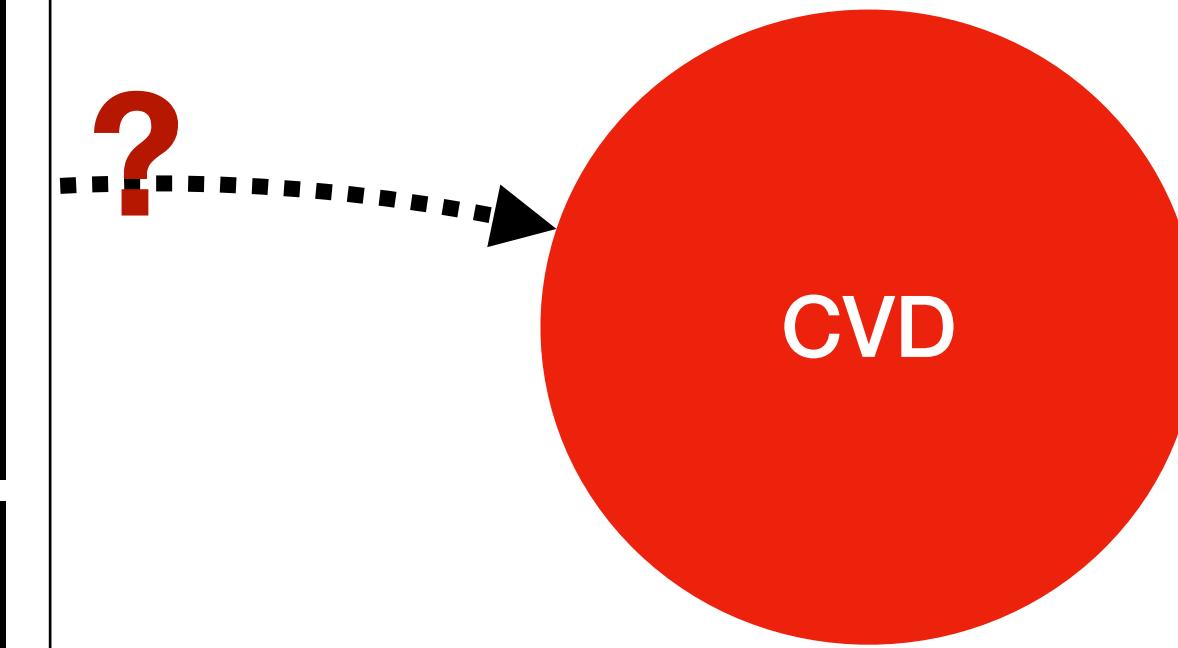
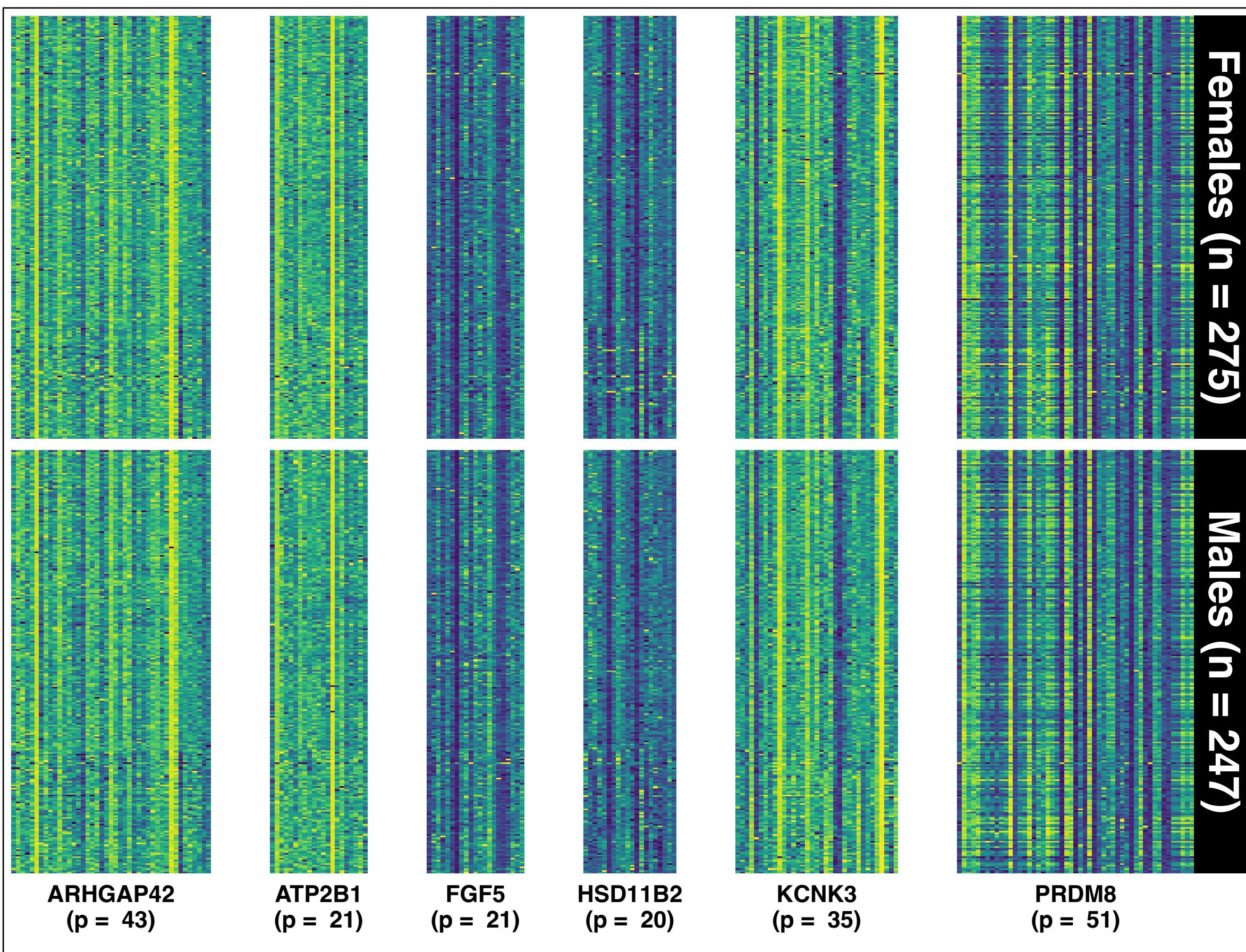


Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

Candidate gene analysis: ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.

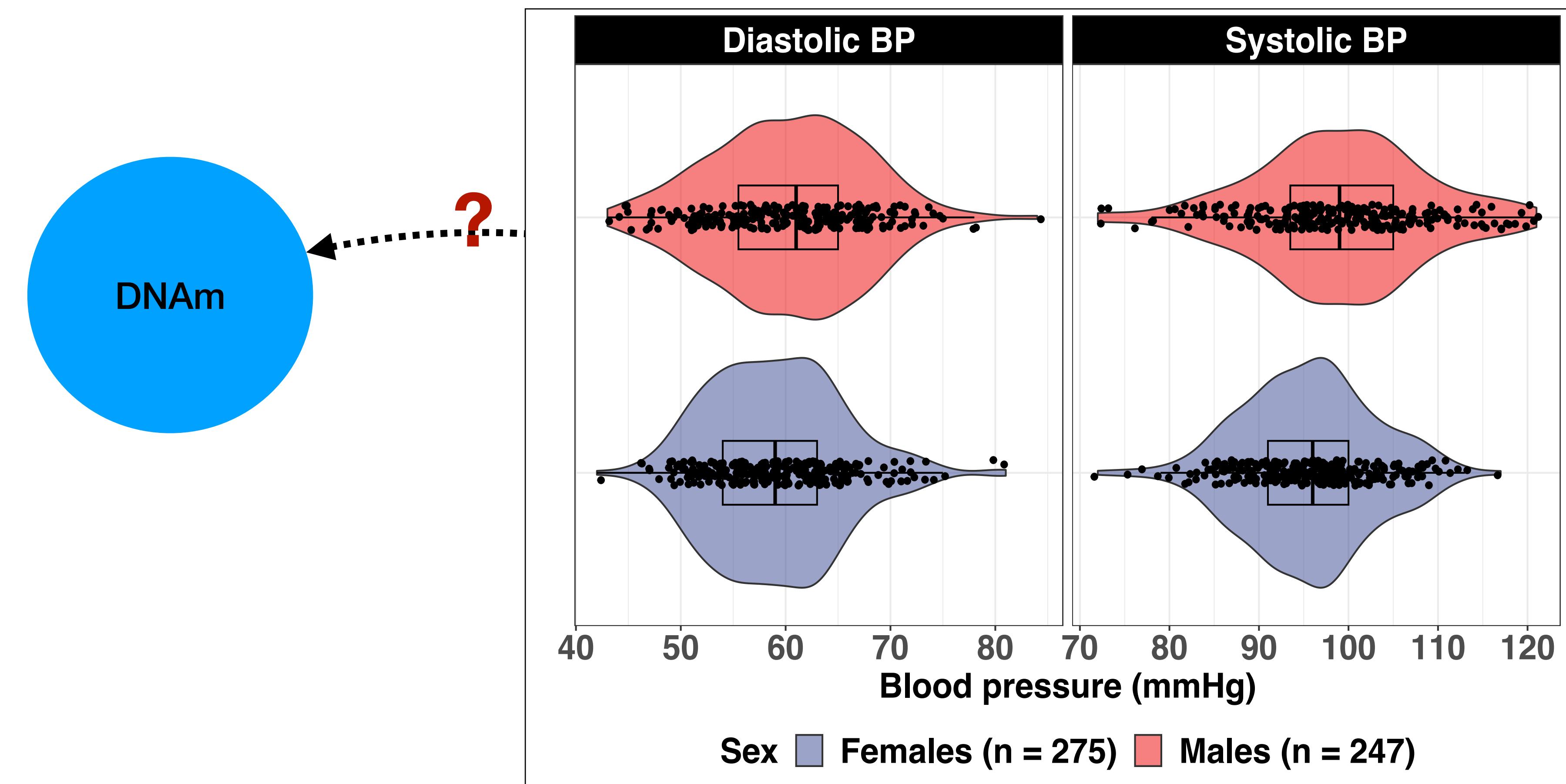


Does DNAm drive BP or vice-versa?

Data: Early Life Exposures in Mexico to Environmental Toxicants (ELEMENT) cohort study.

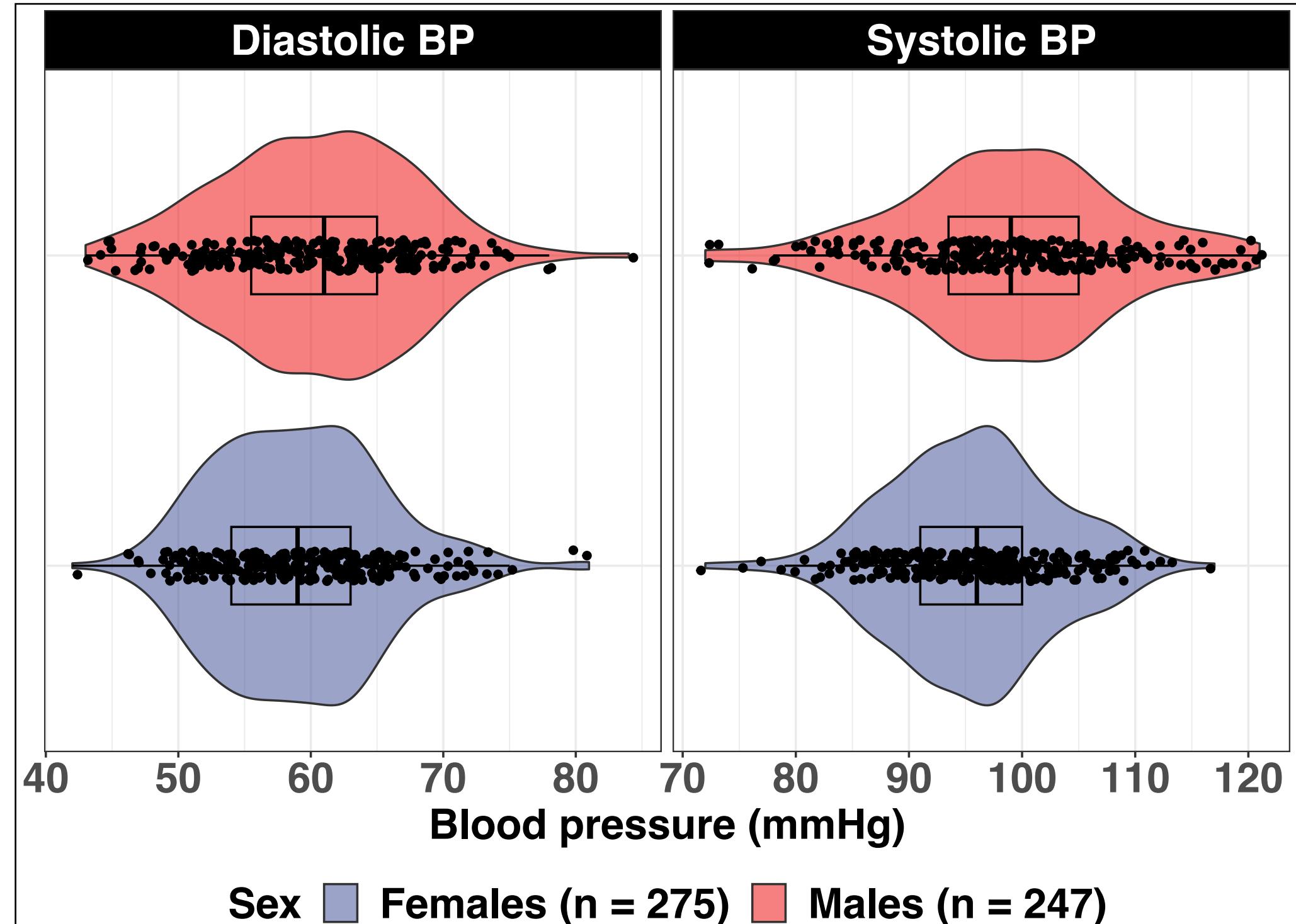
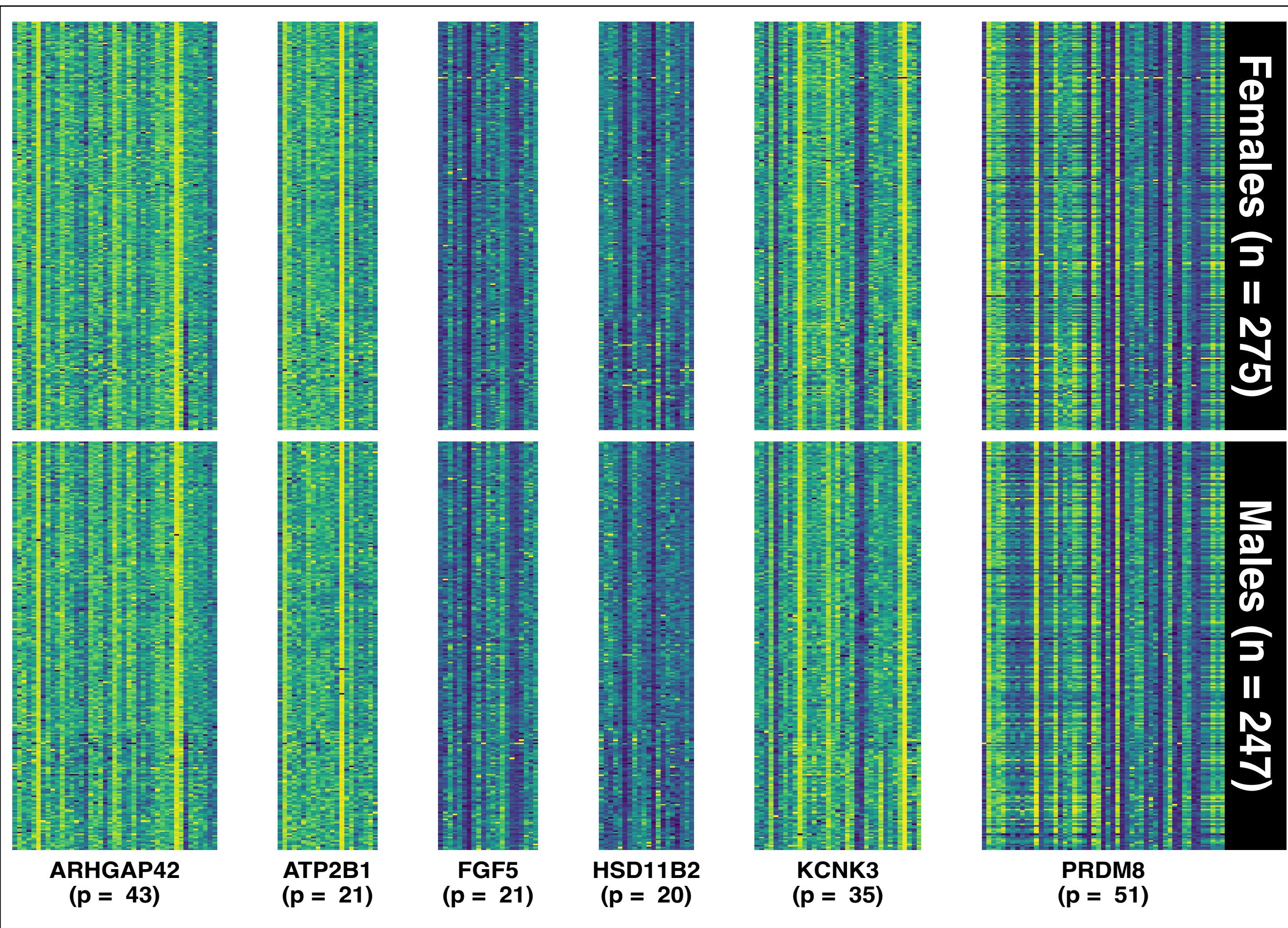
Candidate gene analysis: ARHGAP42, ATP2B1, FGF5, HSD11B2, KCNK3, PRDM8

Biological considerations: n = 522 children aged 10 - 18 years, stratified by sex.



Does DNAm drive BP?

Key challenges



- Many CpG sites.
- Mildly correlated CpG sites.
- Different developmental processes in boys and girls aged 10 - 18.

Methods

Shannon's information theory

Shannon's information theory





Shannon's information theory

Association and direction through distributions

$$(X, Y) \sim f_{XY} \quad X \sim f_X \quad Y \sim f_Y$$



Shannon's information theory

Association and direction through distributions

$$(X, Y) \sim f_{XY} \quad X \sim f_X \quad Y \sim f_Y$$

Mutual information: measure of association

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{f_{XY}}{f_X f_Y} \right) \right]$$



Shannon's information theory

Association and direction through distributions

$$(X, Y) \sim f_{XY} \quad X \sim f_X \quad Y \sim f_Y$$

Mutual information: measure of association

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{f_{XY}}{f_X f_Y} \right) \right]$$

Entropy: measure of randomness

$$H(X, Y) = E_{XY}[-\log(f_{XY})]$$

$$H(X) = E[-\log(f_X)]$$

$$H(X | Y) + H(Y) = H(X, Y)$$



Shannon's information theory

Association and direction through distributions

$$(X, Y) \sim f_{XY} \quad X \sim f_X \quad Y \sim f_Y$$

Mutual information: measure of association

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{f_{XY}}{f_X f_Y} \right) \right]$$

Entropy: measure of randomness

$$H(X, Y) = E_{XY}[-\log(f_{XY})]$$

$$H(X) = E[-\log(f_X)]$$

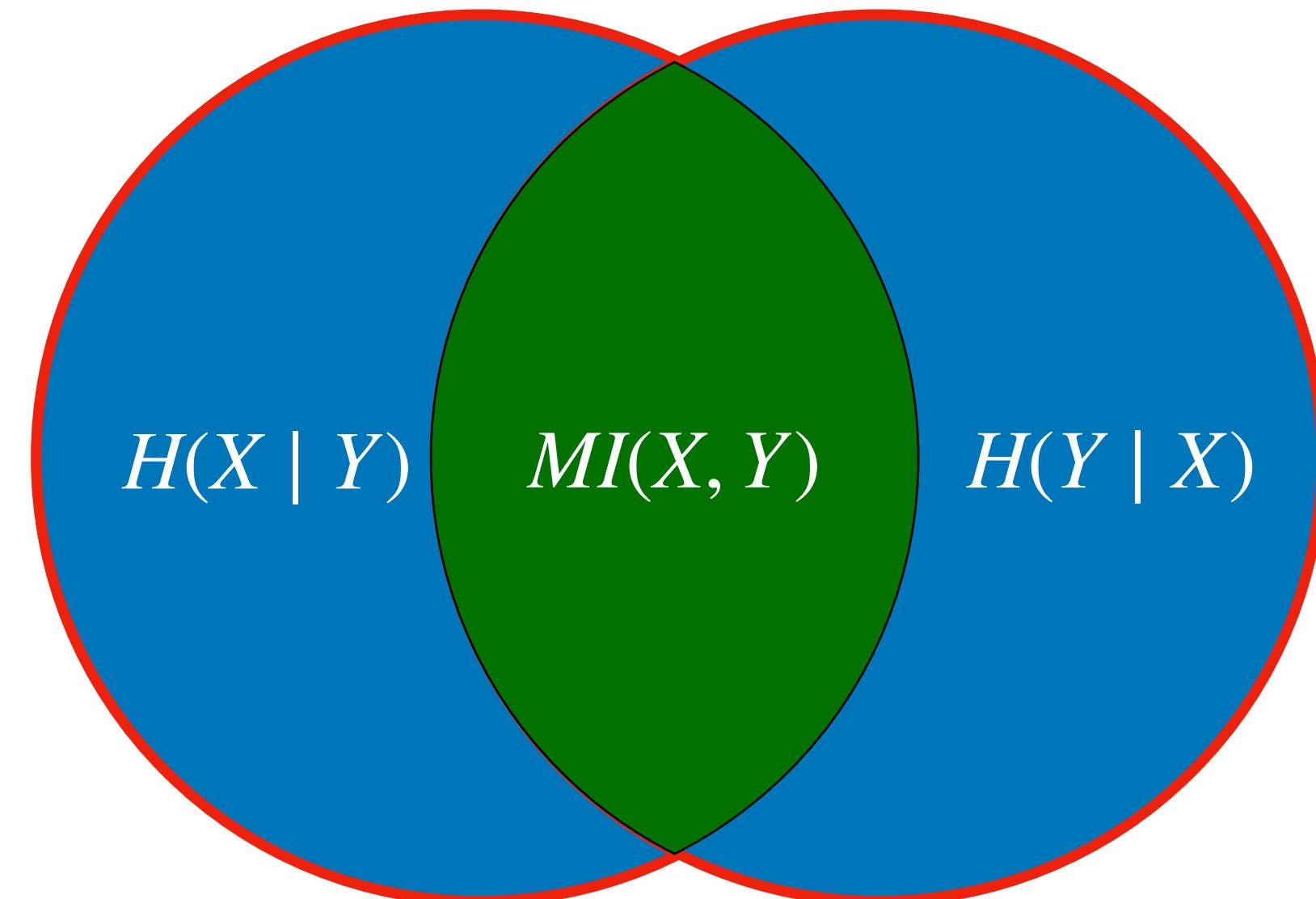
$$H(X | Y) + H(Y) = H(X, Y)$$

Entropy decomposition equation

Entropy decomposition equation

Attempt to study association and directionality

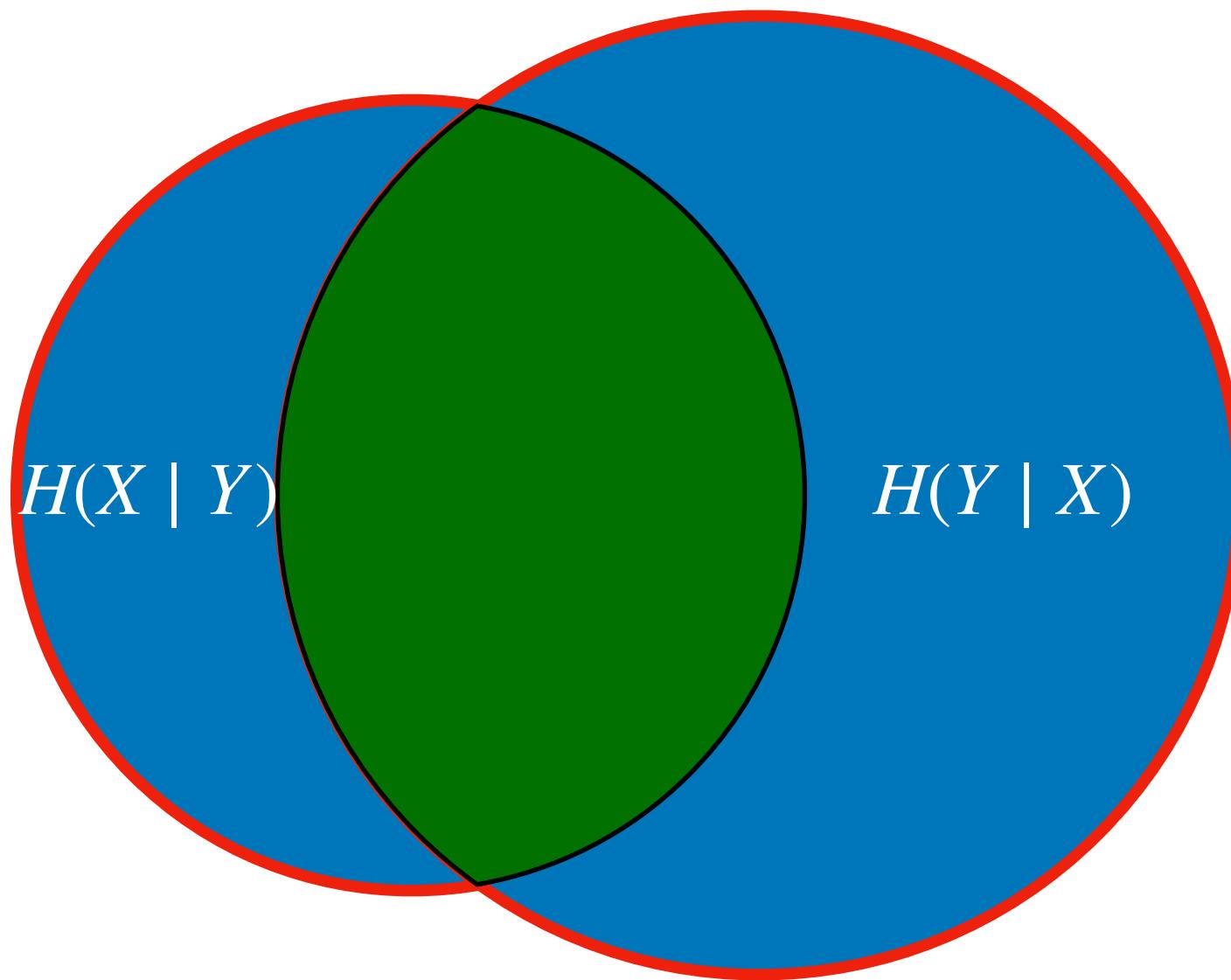
$$H(X, Y) = H(X \mid Y) + H(Y \mid X) + MI(X, Y)$$



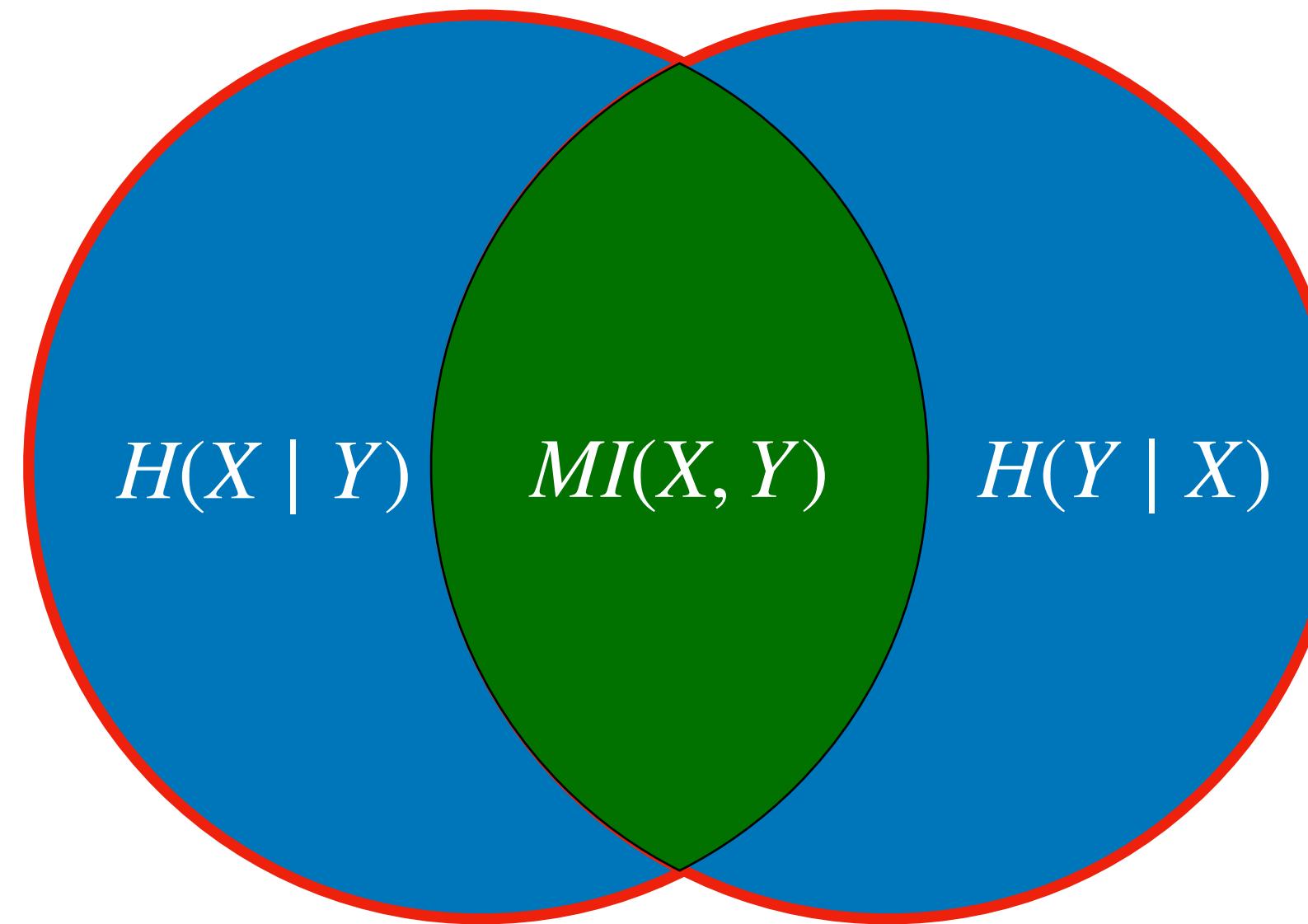
Entropy decomposition equation

Attempt to study association and directionality

$$H(X, Y) = H(X \mid Y) + H(Y \mid X) + MI(X, Y)$$



$$H(X \mid Y) < H(Y \mid X)$$

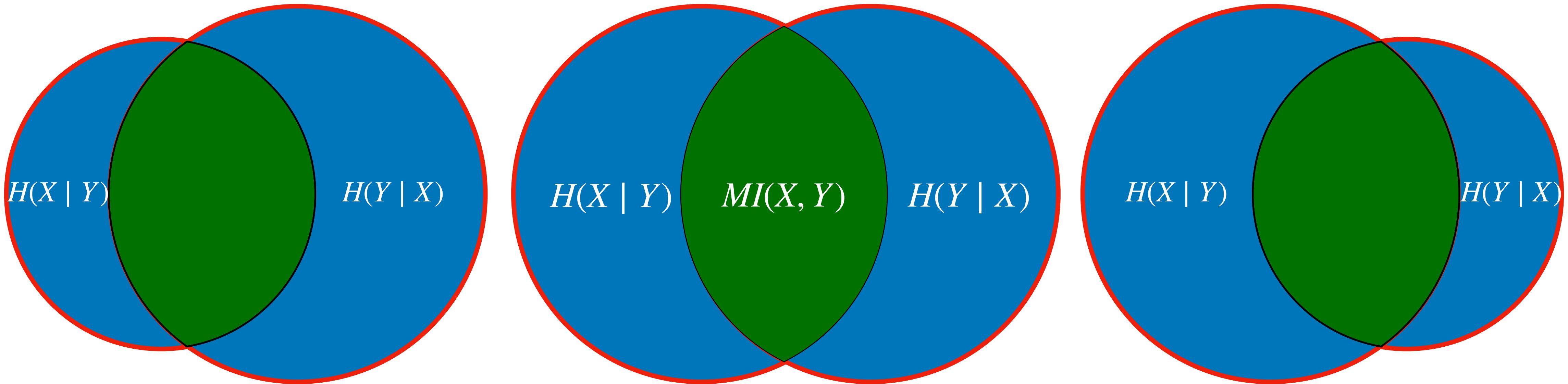


$$H(X \mid Y) > H(Y \mid X)$$

Entropy decomposition equation

Attempt to study association and directionality

Plan 1: Use $MI(X, Y)$ to capture (symmetric) association.

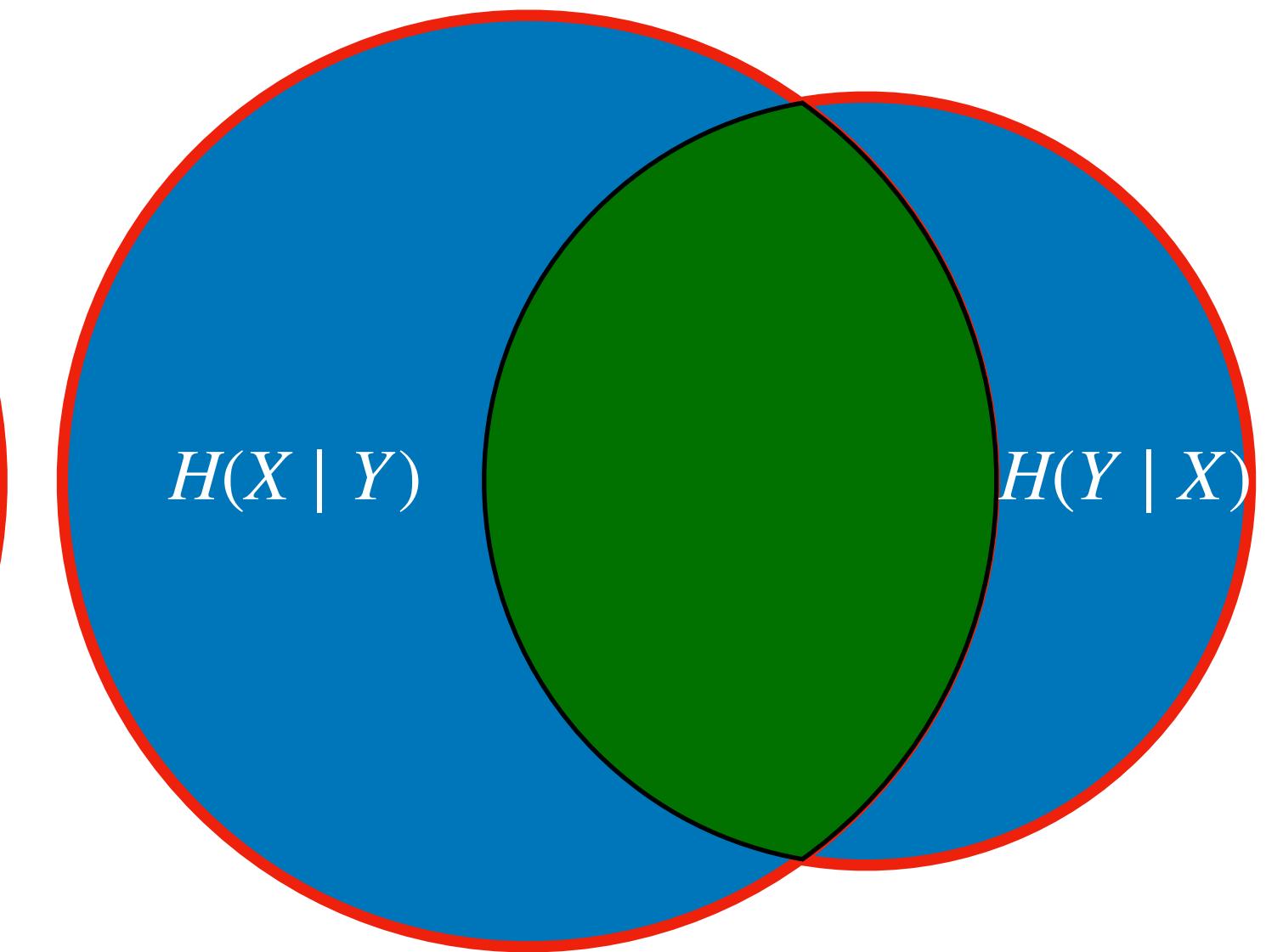
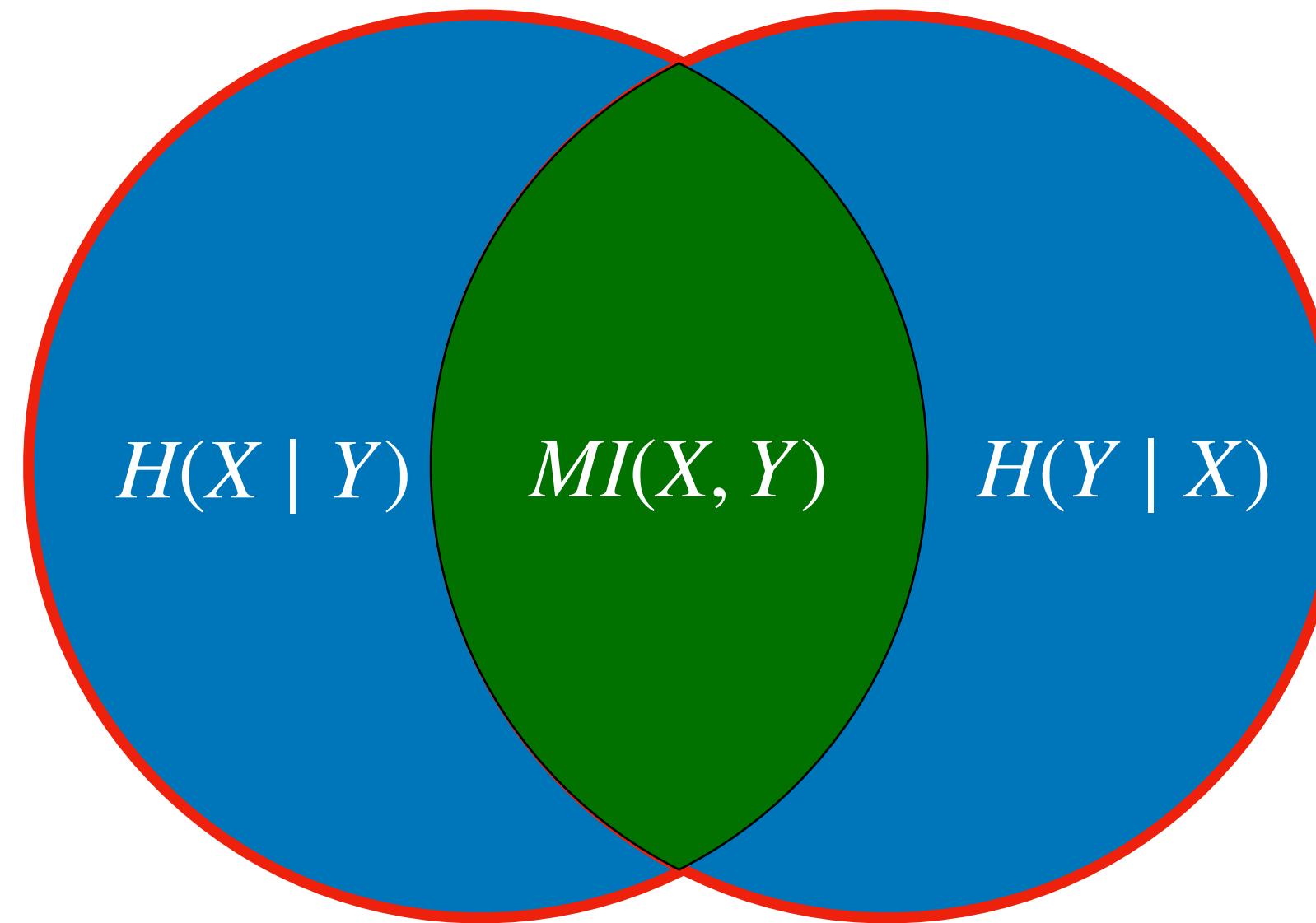
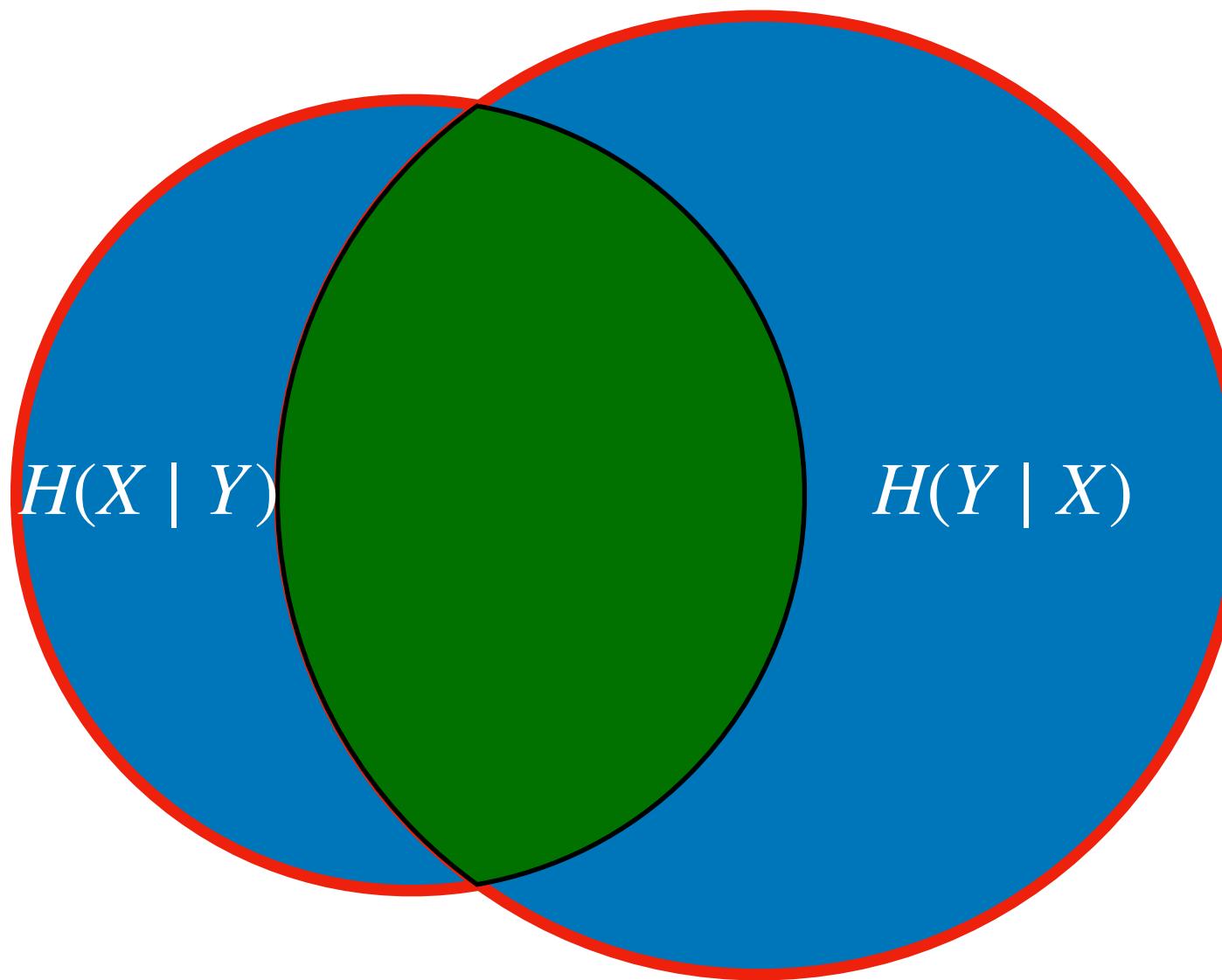


Plan 2: Use $H(X | Y)$ and $H(Y | X)$ to capture asymmetry/directionality.

Entropy decomposition equation

Attempt to study association and directionality

Plan 1: Use $MI(X, Y)$ to capture (symmetric) association.



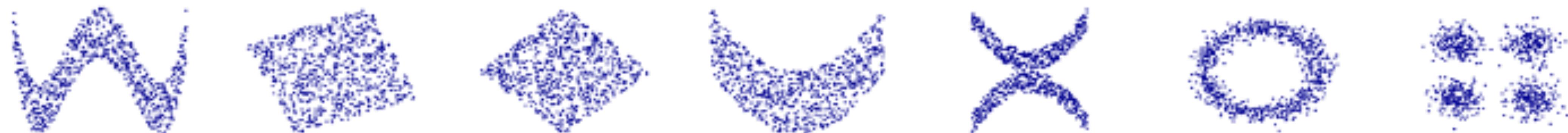
MI is a powerful measure of association

MI is self-equitable



Pearson, Spearman, Kendall:
“These are ‘good’ data to capture association ”

MI captures association across all patterns



Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{f_{XY}}{f_X f_Y} \right) \right]$$

- $MI = KL(f_{XY} \parallel f_X \otimes f_Y)$
- $MI = 0 \iff X \perp Y$

Benefits and hurdles of MI

Need a fast, scalable and accurate estimator

$$MI(X, Y) = E_{XY} \left[\log \left(\frac{f_{XY}}{f_X f_Y} \right) \right]$$

- $MI = KL(f_{XY} \| f_X \otimes f_Y)$
- $MI = 0 \iff X \perp Y$

Need \hat{f}_{XY}, \hat{f}_X , and \hat{f}_Y : bandwidth tuning!

Table: Mean (SD) computation time (in seconds) of estimators of MI for bivariate data of varying sample size (n) for $s = 100$ iterations.

	Sample size (n)		
	1000	2500	5000
Empirical copula-based MI	4.360 (0.356)	5.368 (0.308)	64.040 (0.254)
Jackknifed MI	3.150 (0.107)	18.446 (0.116)	62.454 (4.601)

fastMI

Scalable and accurate estimation of MI

- Want: \hat{f}_{XY} , \hat{f}_X , and \hat{f}_Y , without tuning to get faster estimate \hat{MI}

fastMI

Scalable and accurate estimation of MI

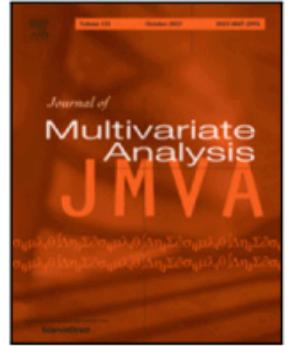
- Want: \hat{f}_{XY} , \hat{f}_X , and \hat{f}_Y , without tuning to get faster estimate \hat{MI}
- Sklar's copula and MI :
 - c_{XY} is the copula density function.
 - $MI = E[\log(c_{XY})]$
- Use Fourier transformation trick to estimate c_{XY} without tuning.

Contents lists available at [ScienceDirect](#)

 ELSEVIER

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



fastMI: A fast and consistent copula-based nonparametric estimator of mutual information

Soumik Purkayastha ¹, Peter X.-K. Song ^{*,1}

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Purkayastha, S., & Song, P. X. K. (2023). *fastMI: A fast and consistent copula-based nonparametric estimator of mutual information*. *Journal of Multivariate Analysis*, 105270.

fastMI

Scalable and accurate estimation of MI

- Want: \hat{f}_{XY}, \hat{f}_X , and \hat{f}_Y , without tuning to get faster estimate \hat{MI}
- Sklar's copula and MI :
 - c_{XY} is the copula density function.
 - $MI = E[\log(c_{XY})]$
- Use Fourier transformation trick to estimate c_{XY} without tuning.

Contents lists available at [ScienceDirect](#)

Journal of Multivariate Analysis

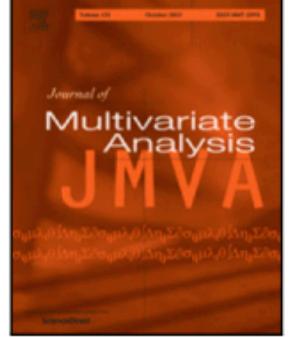
journal homepage: www.elsevier.com/locate/jmva

 ELSEVIER

fastMI: A fast and consistent copula-based nonparametric estimator of mutual information

Soumik Purkayastha ¹, Peter X.-K. Song ^{*,1}

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA



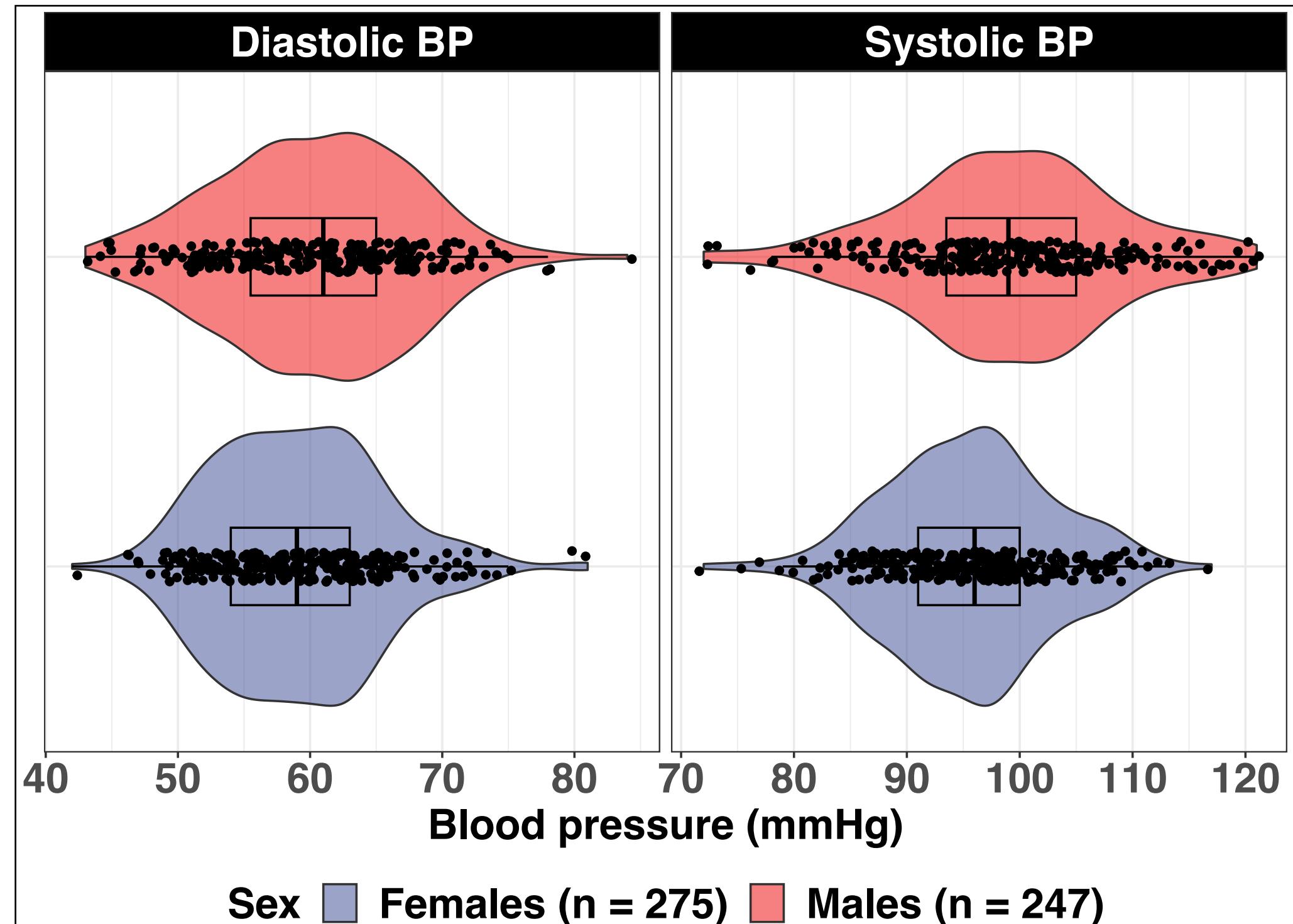
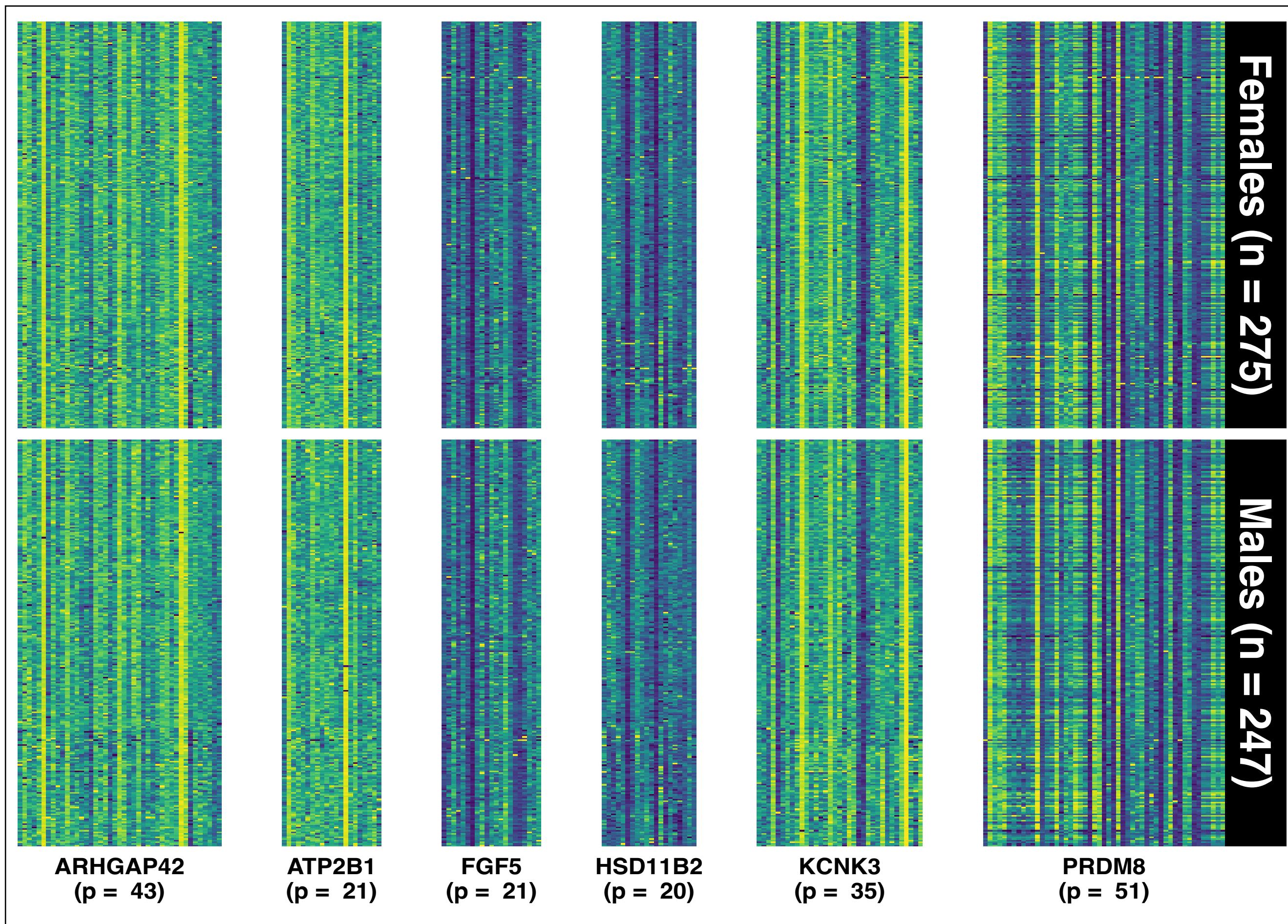
Purkayastha, S., & Song, P. X. K. (2023). *fastMI: A fast and consistent copula-based nonparametric estimator of mutual information*. *Journal of Multivariate Analysis*, 105270.

	Sample size (n)	1000	2500	5000
Empirical copula-based MI		4.360 (0.356)	5.368 (0.308)	64.040 (0.254)
Jackknifed MI		3.150 (0.107)	18.446 (0.116)	62.454 (4.601)
fastMI		1.199 (0.135)	2.964 (0.181)	5.952 (0.125)

p-value plots of fastMI

$MI(SBP, DNAm) = ?$

$MI(DBP, DNAm) = ?$



- Genes: 6. Each gene has many CpG sites
- Compute pairwise MI
- Test if $H_0 : MI = 0$.

p-value plots of fastMI

$MI(SBP, DNAm) \neq 0$

$MI(DBP, DNAm) \neq 0$

- Powerful test of independence.
- Faster computation leads to scalable estimator!
- fastMI unearths associations between DNAm and BP across all six candidate genes at various CpG sites.

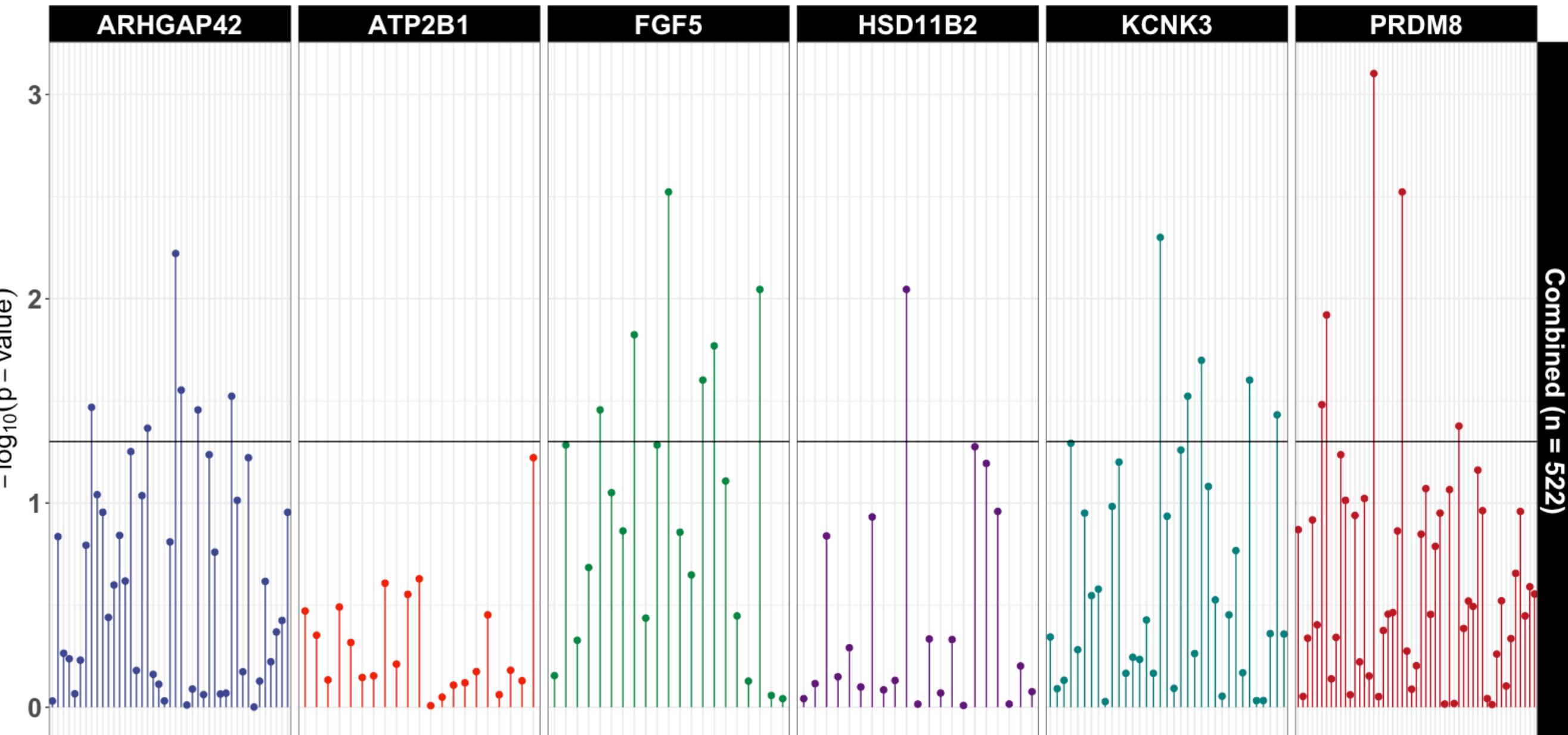
p-value plots of fastMI

$MI(SBP, DNAm) \neq 0$

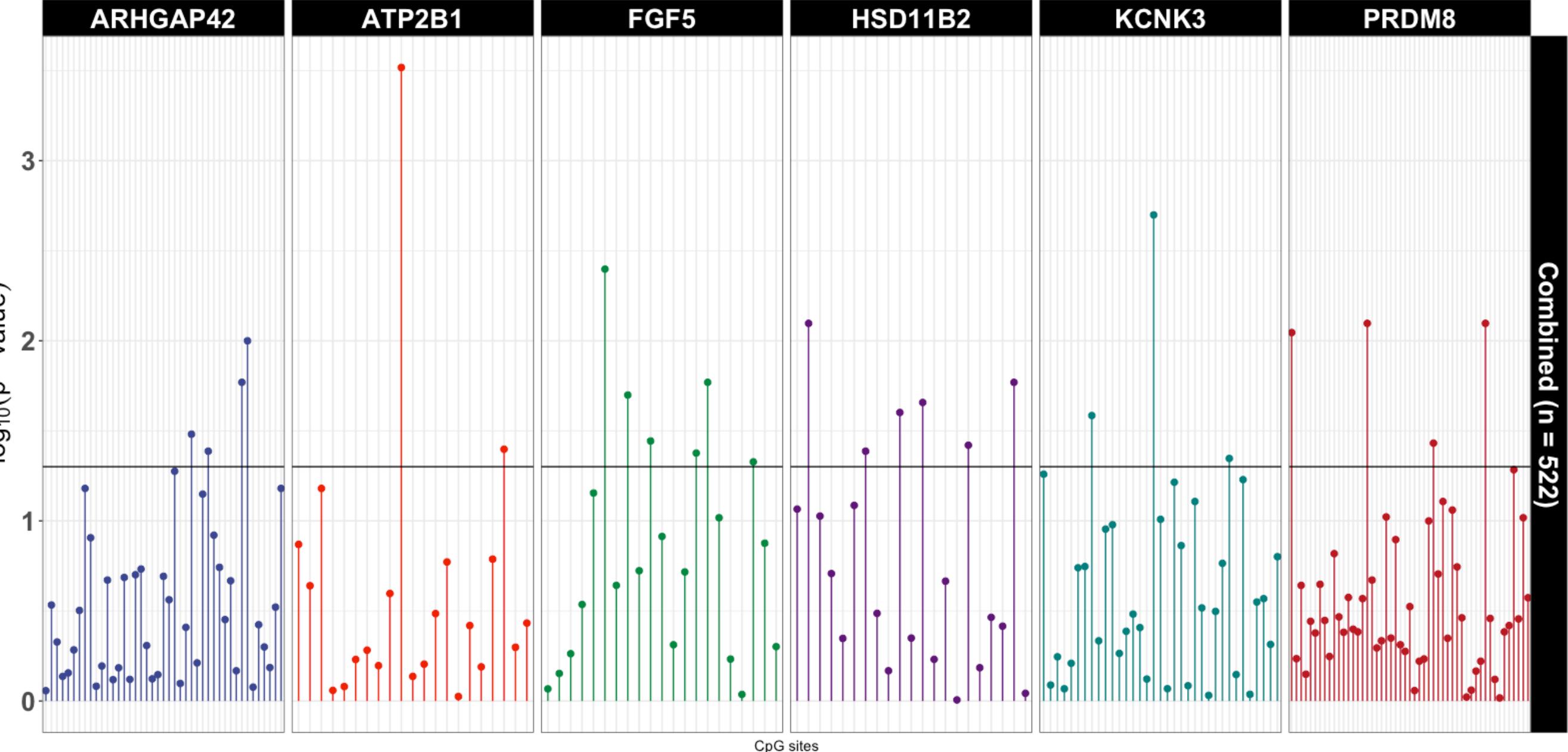
$MI(DBP, DNAm) \neq 0$

- Powerful test of independence.
- Faster computation leads to scalable estimator!
- fastMI unearths associations between DNAm and BP across all six candidate genes at various CpG sites.

Examining association of CPG site with systolic blood pressure.



Examining association of CPG site with diastolic blood pressure.



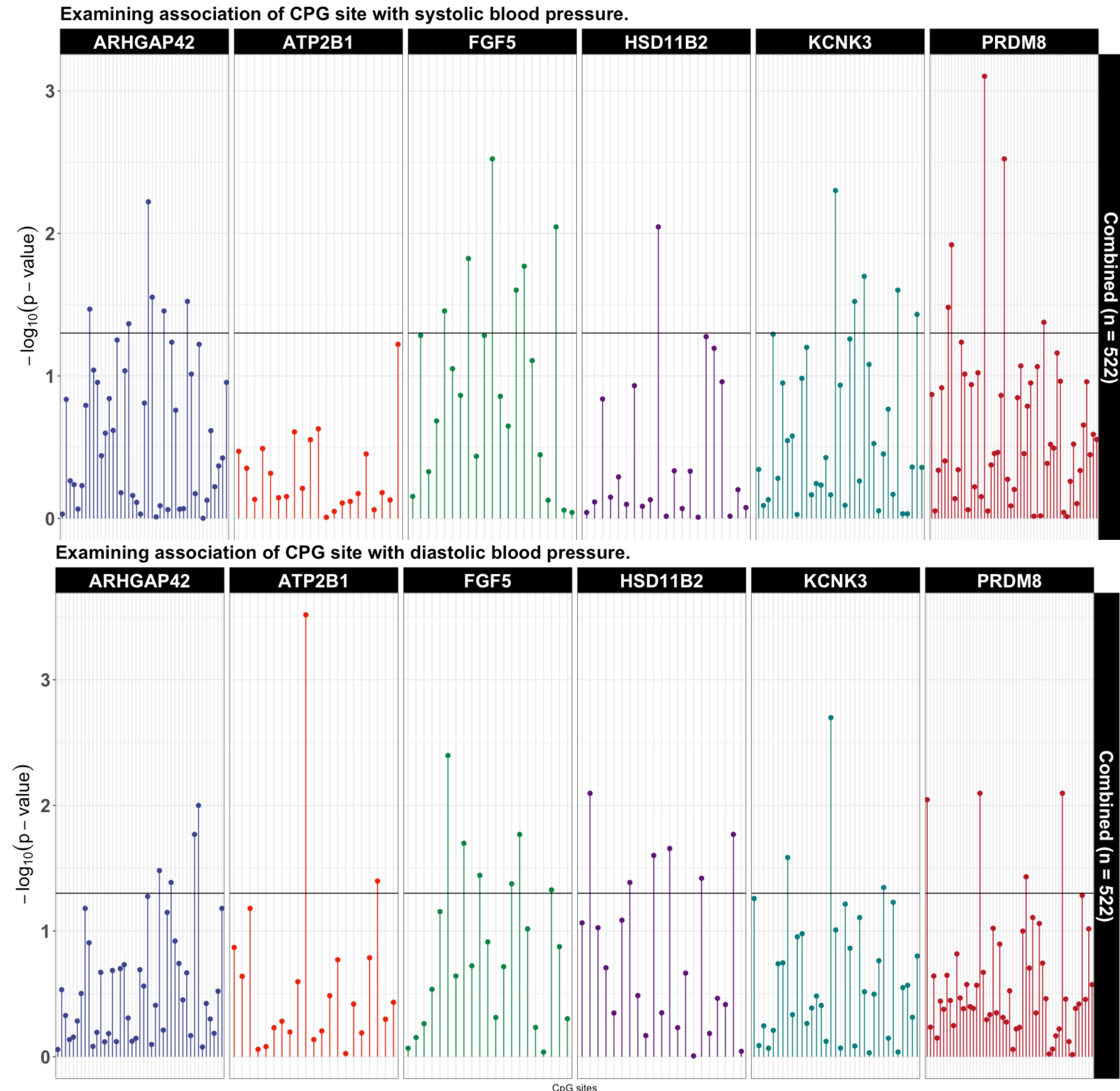
p-value plots of fastMI

$MI(SBP, DNAm) \neq 0$

$MI(DBP, DNAm) \neq 0$

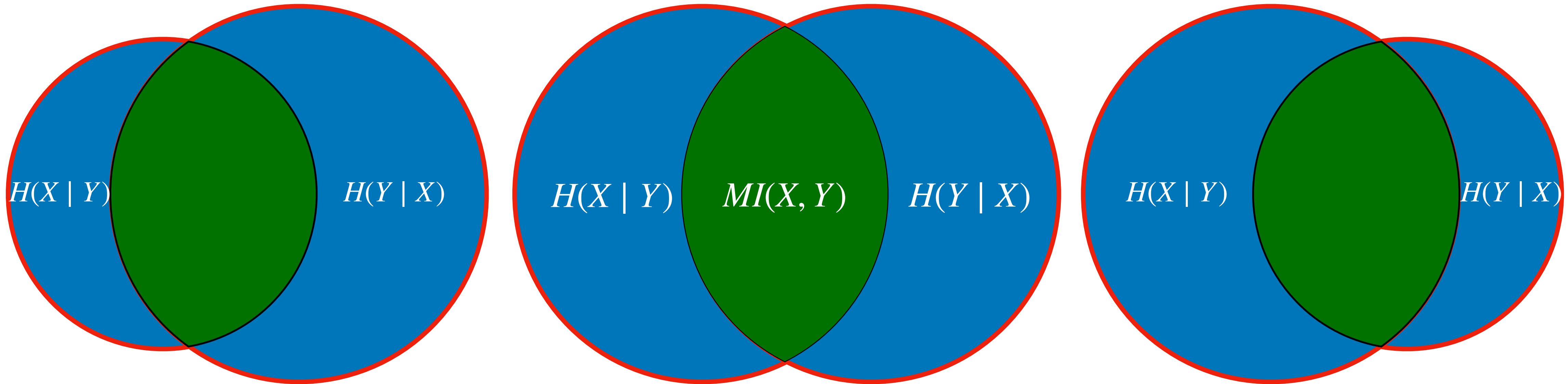
- Powerful test of independence.
- Faster computation leads to scalable estimator!
- fastMI unearths associations between DNAm and BP across all six candidate genes at various CpG sites.

But what about directionality?



Entropy decomposition equation

Attempt to study association and directionality



Plan 2: Use $H(X | Y)$ and $H(Y | X)$ to capture asymmetry/directionality.

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

- Allow for covariate adjustment:

$$Y = g(X, \mathbf{Z})$$

- Allow for noise contamination:

$$Y = g(X) + \epsilon, \text{ with } X \perp \epsilon.$$

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Subject to **identifiability constraints**:
*“GEMs reveals distributional discrepancy between
exposure-outcome that are captured using the entropy analytic”*

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Subject to **identifiability constraints**:
“GEMs reveals distributional discrepancy between exposure-outcome that are captured using the entropy analytic”

What identifiability conditions?

Generative exposure mapping (GEM)

Entropy captures directionality!

- Exposure $X \sim f_X$ from an experiment.
- Outcome $Y \sim f_Y$, generated by X being mapped through **bijective** $g(\cdot)$

$$Y = g(X)$$

Subject to **identifiability constraints**:
“GEMs reveals distributional discrepancy between exposure-outcome that are captured using the entropy analytic”

Impose orthogonality condition on $g(\cdot)$ and f_X

Asymmetry in GEMs using entropy

Approve or disprove $X \rightarrow Y$

- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.

Asymmetry in GEMs using entropy

Approve or disprove $X \rightarrow Y$

- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.
- Population: if identifiability conditions hold

$$C_{X>Y} := H(X) - H(Y) > 0$$

Asymmetry in GEMs using entropy

Approve or disprove $X \rightarrow Y$

- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.
- Population: if identifiability conditions hold

$$C_{X>Y} := H(X) - H(Y) > 0$$

- Sample: if $\hat{C}_{X>Y} > 0$, confirm hypothesis of direction induced by GEM.

Asymmetry in GEMs using entropy

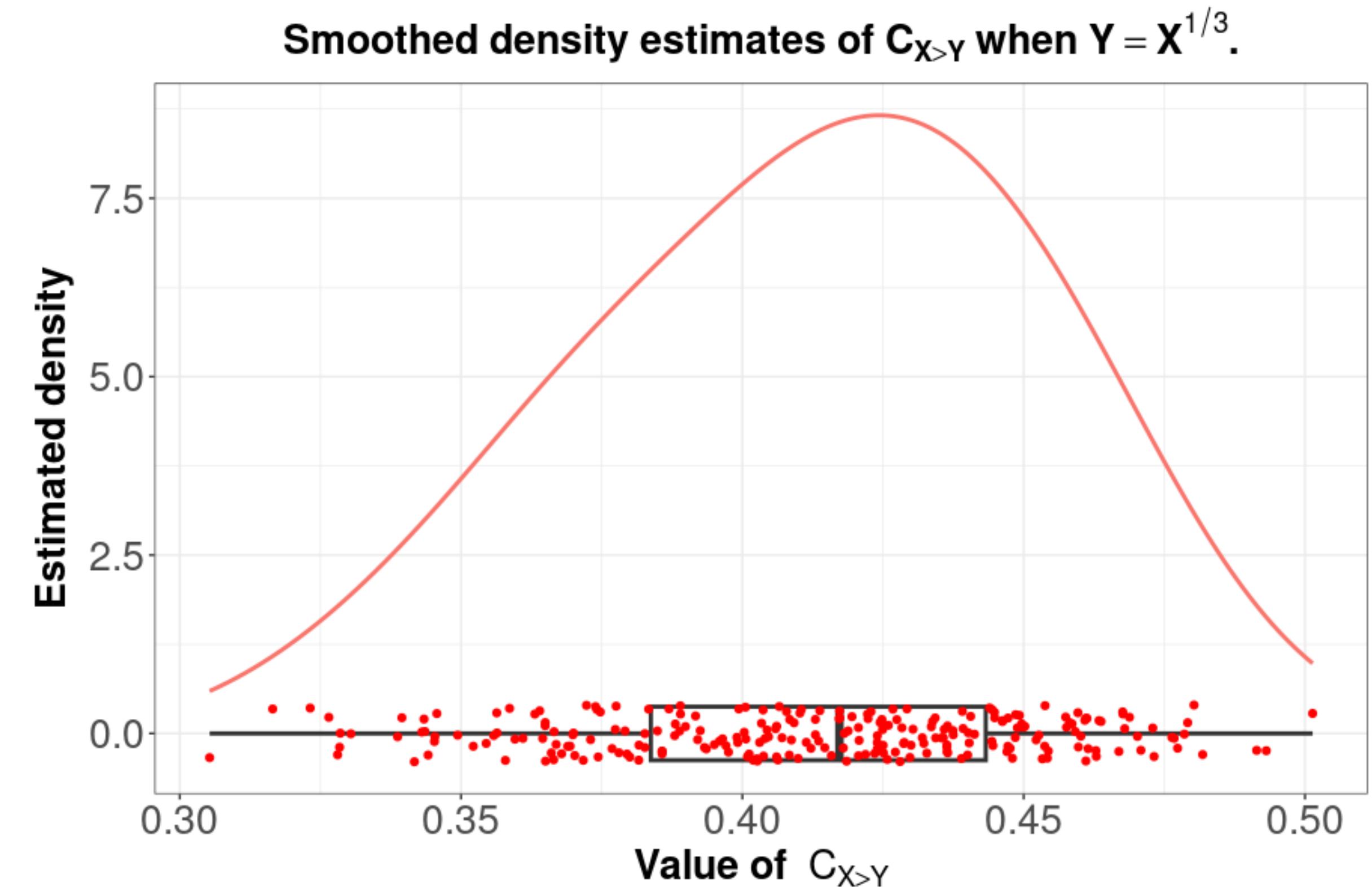
Approve or disprove $X \rightarrow Y$

- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
 - $X \sim U(0,1)$.
 - Generate $Y = g(X)$.
 - **Simulated behavior of $C_{X>Y}$?**

Asymmetry in GEMs using entropy

Approve or disprove $X \rightarrow Y$

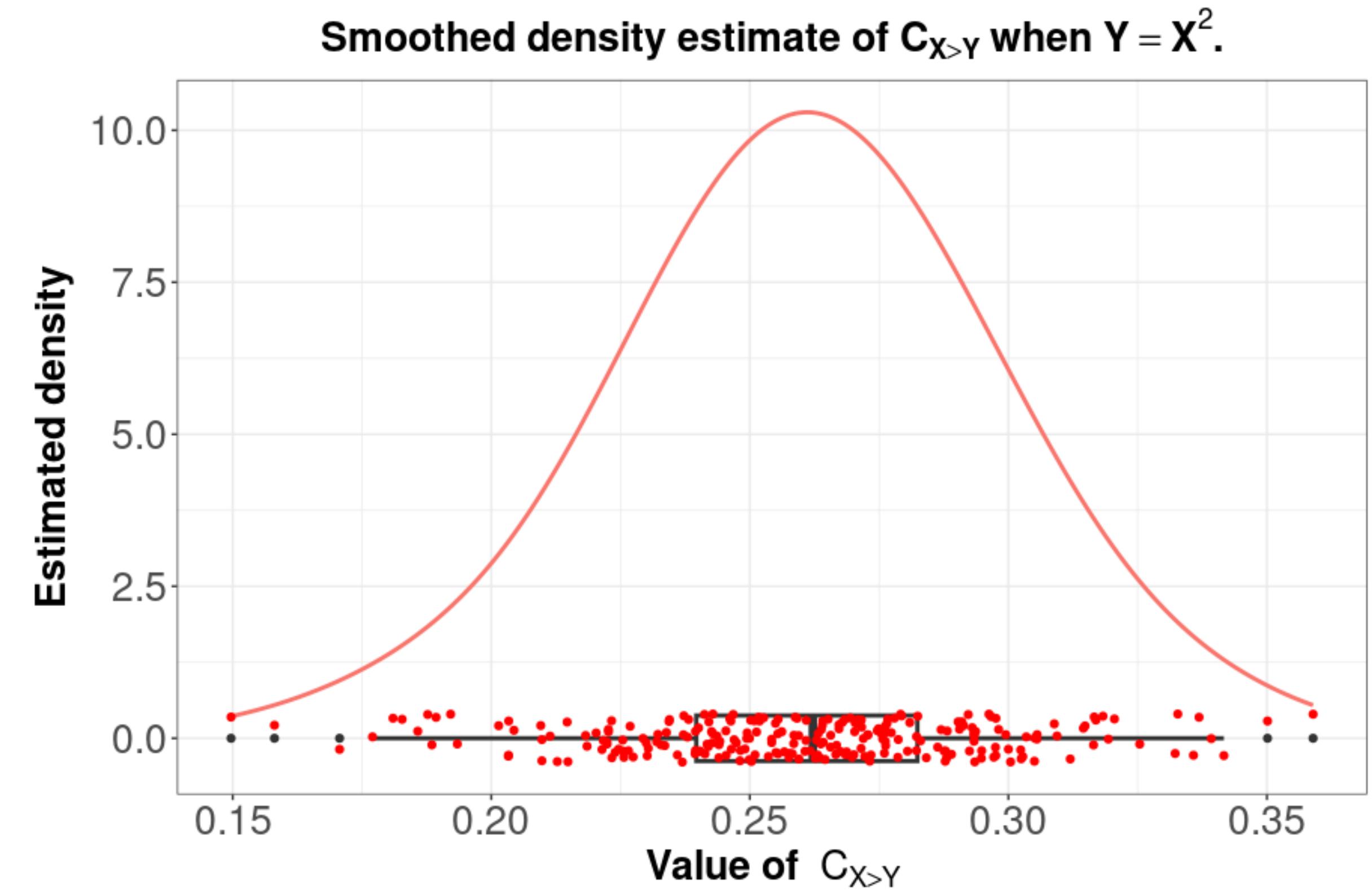
- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
 - $X \sim U(0,1)$.
 - Generate $Y = g(X)$.
- Simulated behavior of $C_{X>Y}$?



Asymmetry in GEMs using entropy

Approve or disprove $X \rightarrow Y$

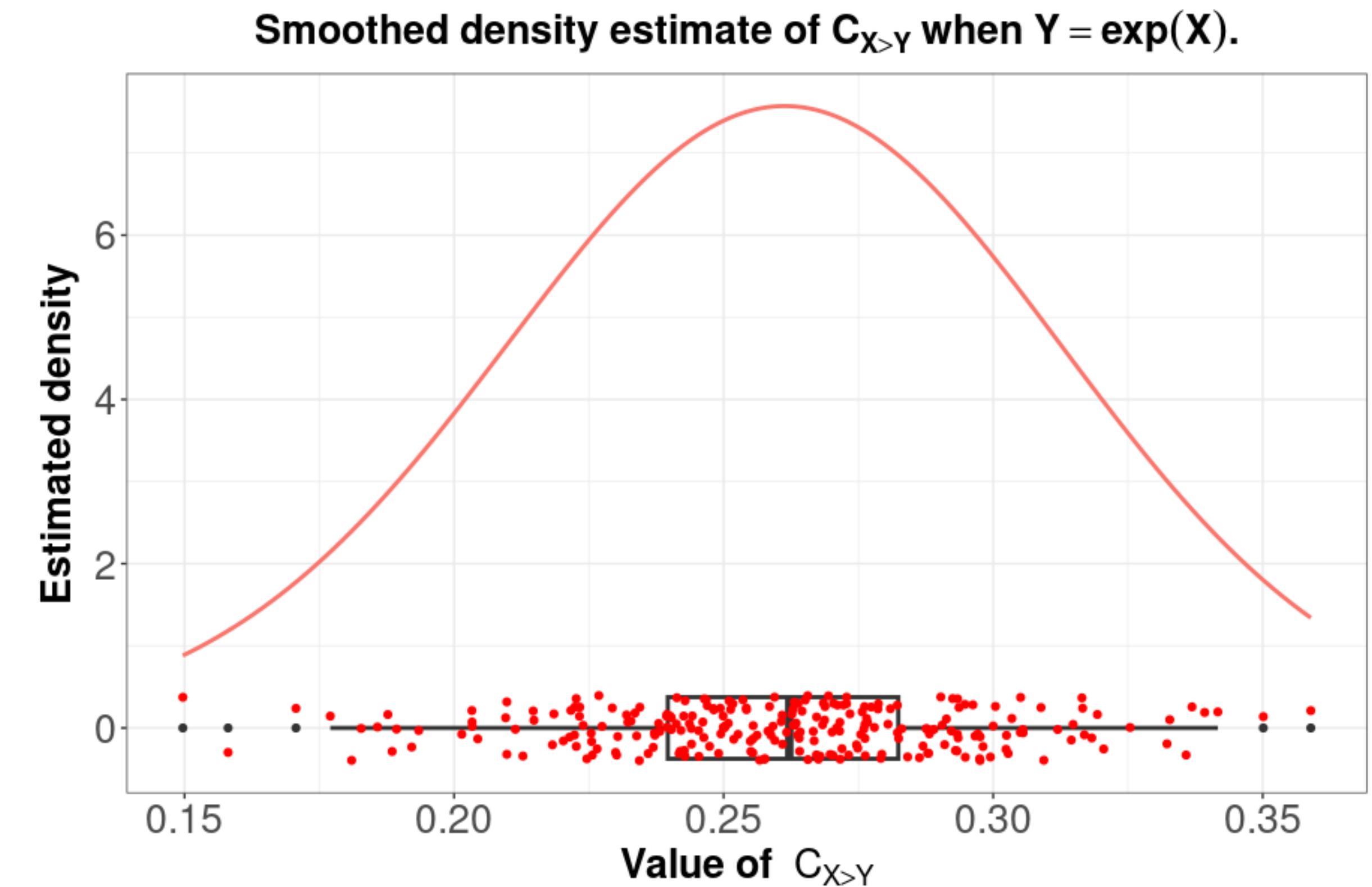
- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
 - $X \sim U(0,1)$.
 - Generate $Y = g(X)$.
- Simulated behavior of $C_{X>Y}$?



Asymmetry in GEMs using entropy

Approve or disprove $X \rightarrow Y$

- Hypothesis: $X \rightarrow Y$.
- GEM-induced direction $Y = g(X)$.
- Experiment:
 - $X \sim U(0,1)$.
 - Generate $Y = g(X)$.
- Simulated behavior of $C_{X>Y}$?



Asymmetry coefficient $C_{X>Y}$

Advantages and challenges

Strong asymmetry:

- GEM + identifiability assumptions $\implies C_{X>Y} > 0$.
- Prove or disprove $X \rightarrow Y$ using $\hat{C}_{X>Y}$

Asymmetry coefficient $C_{X>Y}$

Advantages and challenges

Strong asymmetry:

- GEM + identifiability assumptions $\implies C_{X>Y} > 0$.
- Prove or disprove $X \rightarrow Y$ using $\hat{C}_{X>Y}$

Weak asymmetry: what if GEM is absent? What if identifiability conditions don't hold?

- $C_{X>Y} = H(X) - H(Y) = H(X|Y) - H(Y|X)$
- Better predictor selection using $\hat{C}_{X>Y}$

Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

Need to estimate \hat{f}_X and \hat{f}_Y : **infinite-dimensional nuisance parameters.**

Data splitting and cross fitting

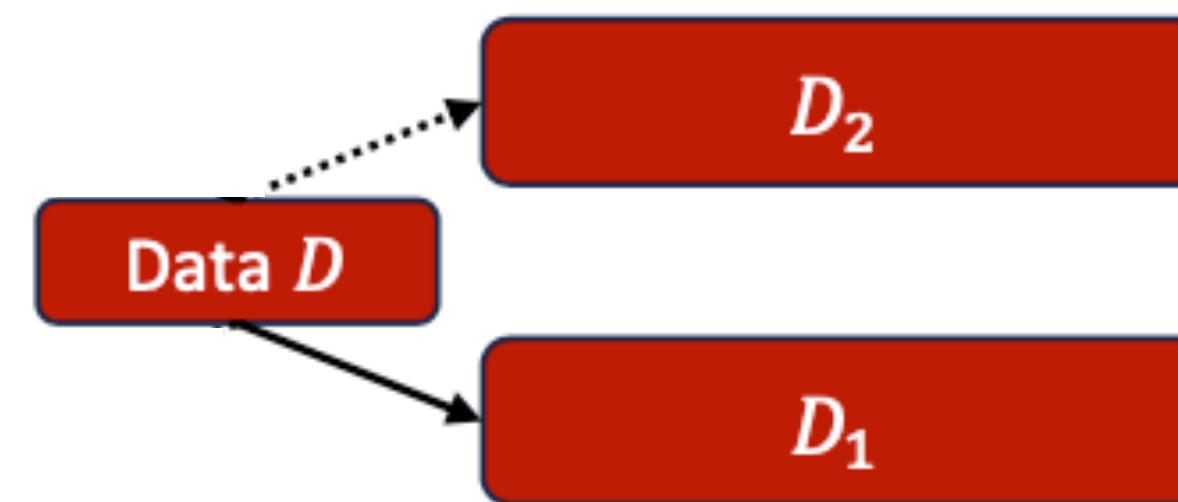
Estimation and inference for $C_{X>Y}$

Data splitting and cross-fitting reduces bias and permits inference.

Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

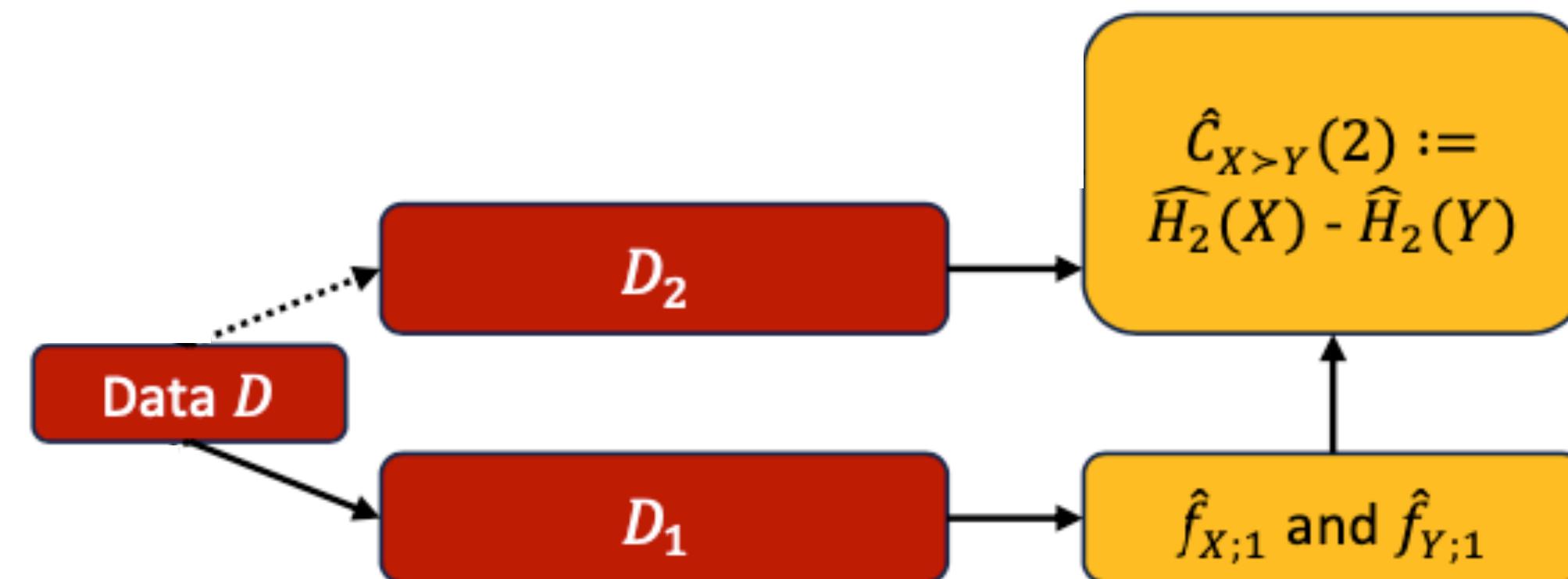
$$D = \left\{ (X_i, Y_i) \right\}_{i=1}^n \cup \left\{ (X_i, Y_i) \right\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

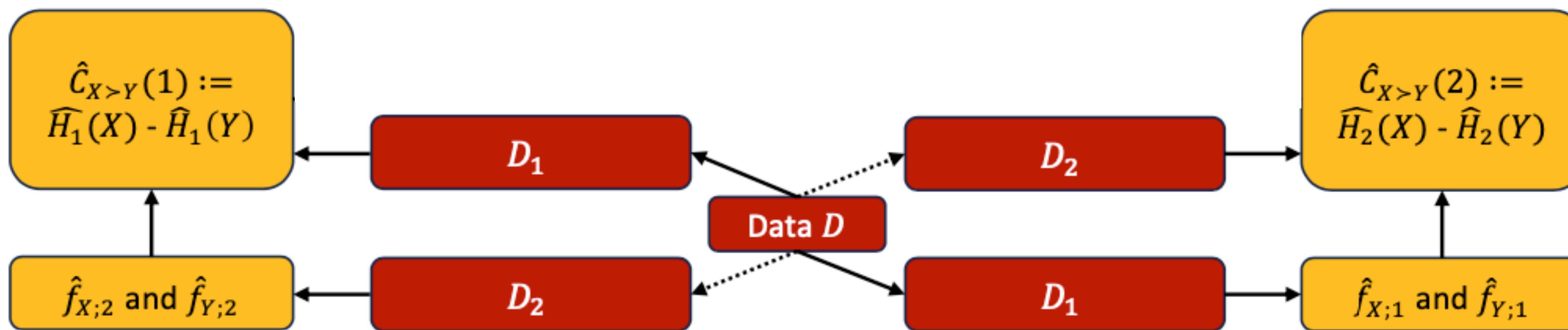
$$D = \{(X_i, Y_i)\}_{i=1}^n \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

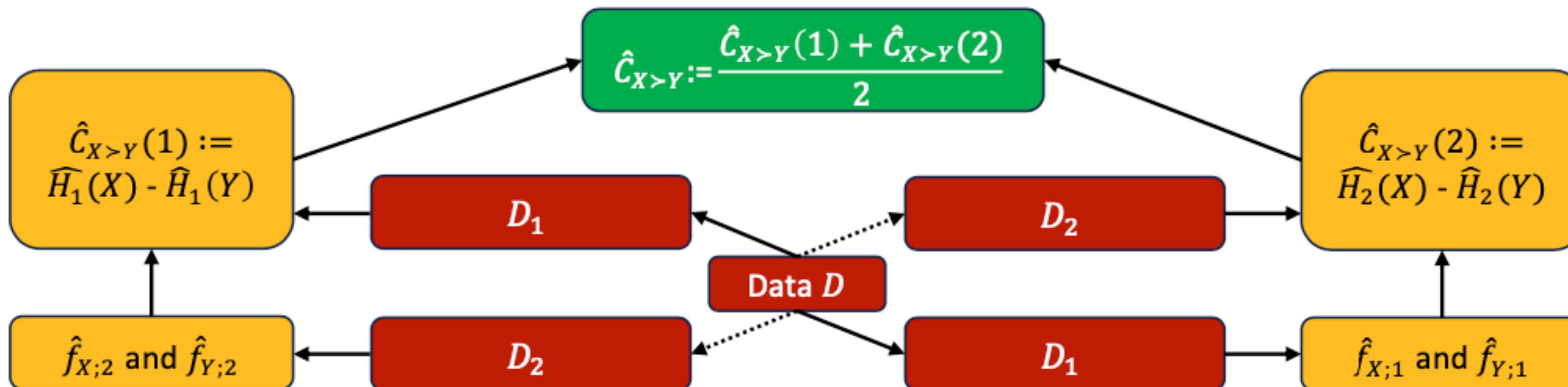
$$D = \{(X_i, Y_i)\}_{i=1}^n \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^n \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$



Data splitting and cross fitting

Estimation and inference for $C_{X>Y}$

$$D = \{(X_i, Y_i)\}_{i=1}^n \cup \{(X_i, Y_i)\}_{i=n+1}^{2n} = D_1 \cup D_2$$

$\hat{C}_{X>Y}$ has a **limiting distribution** subject to **regularity conditions**

$$\sqrt{n} \left(\hat{C}_{X>Y} - C_{X>Y} \right) \rightarrow N(0, \sigma_C^2), \text{ as } n \rightarrow \infty.$$

$$\sigma_C^2 = V[\log(f_X(X)) + \log(f_Y(Y))]$$

Estimated by Monte-Carlo methods with estimated \hat{f}_X and \hat{f}_Y

Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ for what range of σ ?

- Noise-perturbed GEM (NPGEM)

$Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ for what range of σ ?

- Noise-perturbed GEM (NPGEM)

$Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

- Coefficient will work as long as $H(Y^*) \leq H(X)$.
- Establish “critical value” σ_{CRIT}

$$\sigma \leq \sigma_{CRIT} = \frac{\exp(2C_{X>Y}) - 1}{I(Y)}.$$

Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ for what range of σ ?

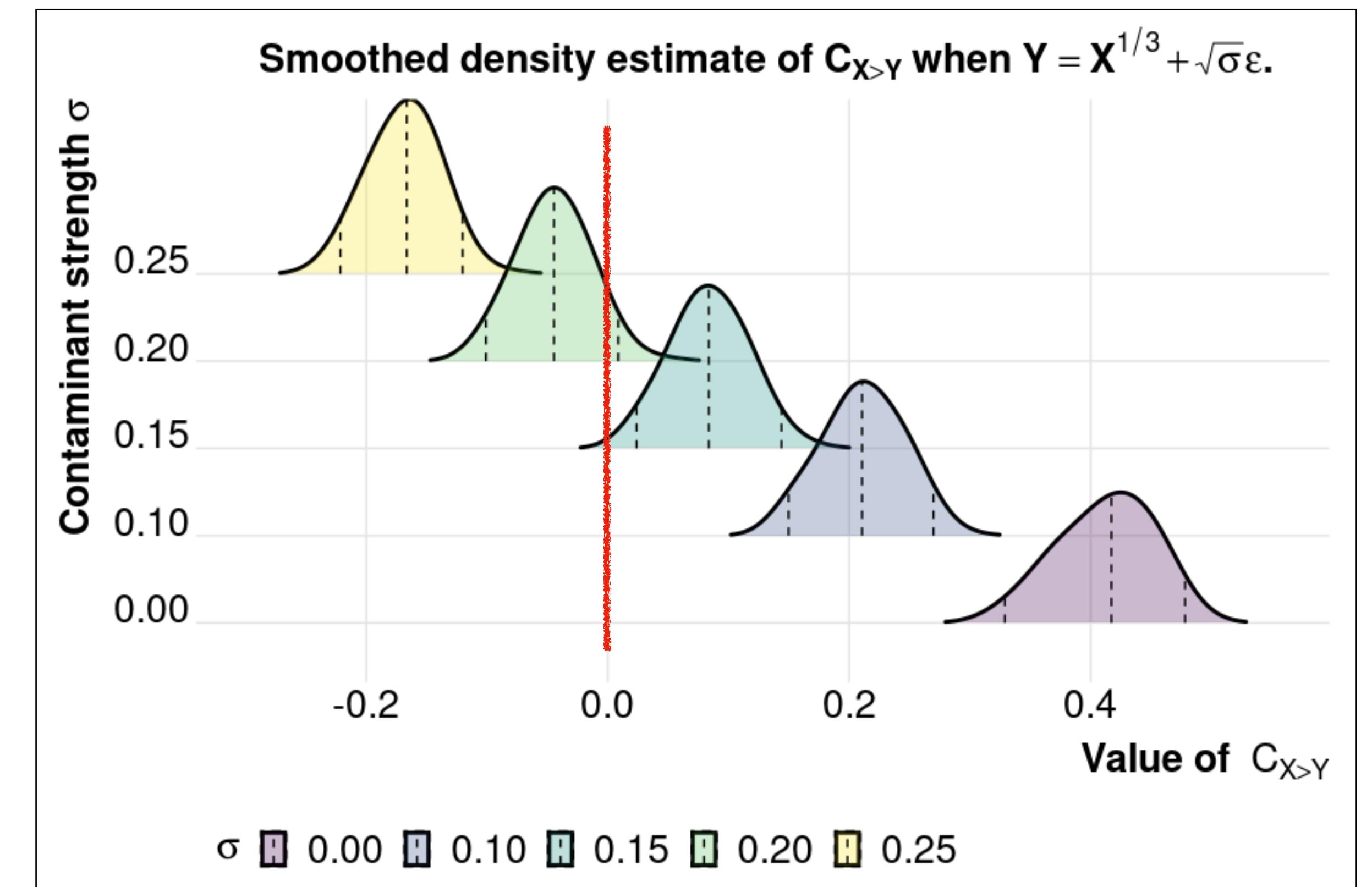
- Noise-perturbed GEM (NPGEM)

$Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

- Coefficient will work as long as $H(Y^*) \leq H(X)$.

- Establish “critical value” σ_{CRIT}

$$\sigma \leq \sigma_{CRIT} = \frac{\exp(2C_{X>Y}) - 1}{I(Y)}.$$



Coefficient can tolerate contaminated outcomes!

$H(Y + \sqrt{\sigma}\epsilon) < H(X)$ for what range of σ ?

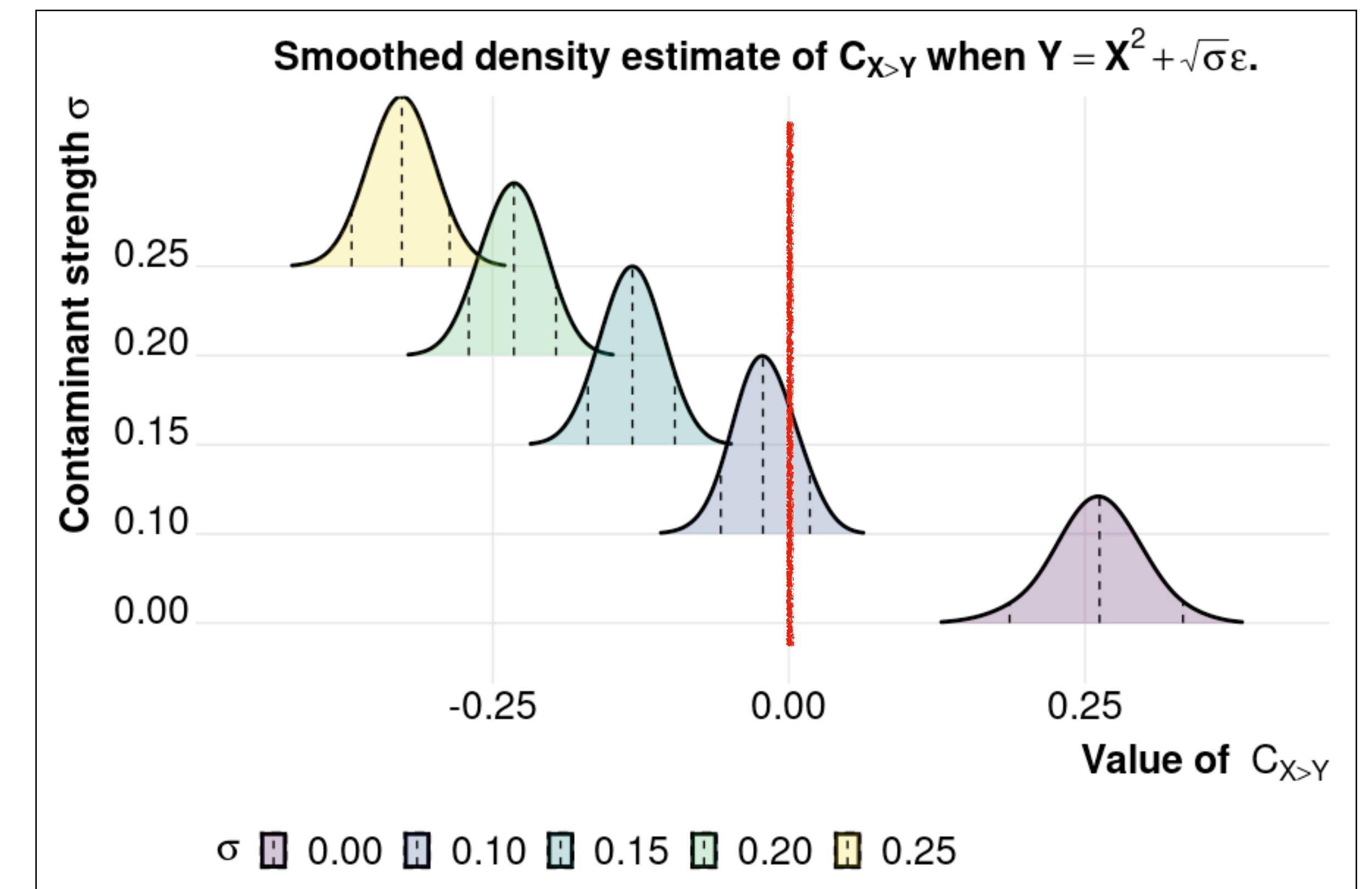
- Noise-perturbed GEM (NPGEM)

$Y^* = g(X) + \sqrt{\sigma}\epsilon$, with $\epsilon \sim N(0,1)$ and $X \perp \epsilon$.

- Coefficient will work as long as $H(Y^*) \leq H(X)$.

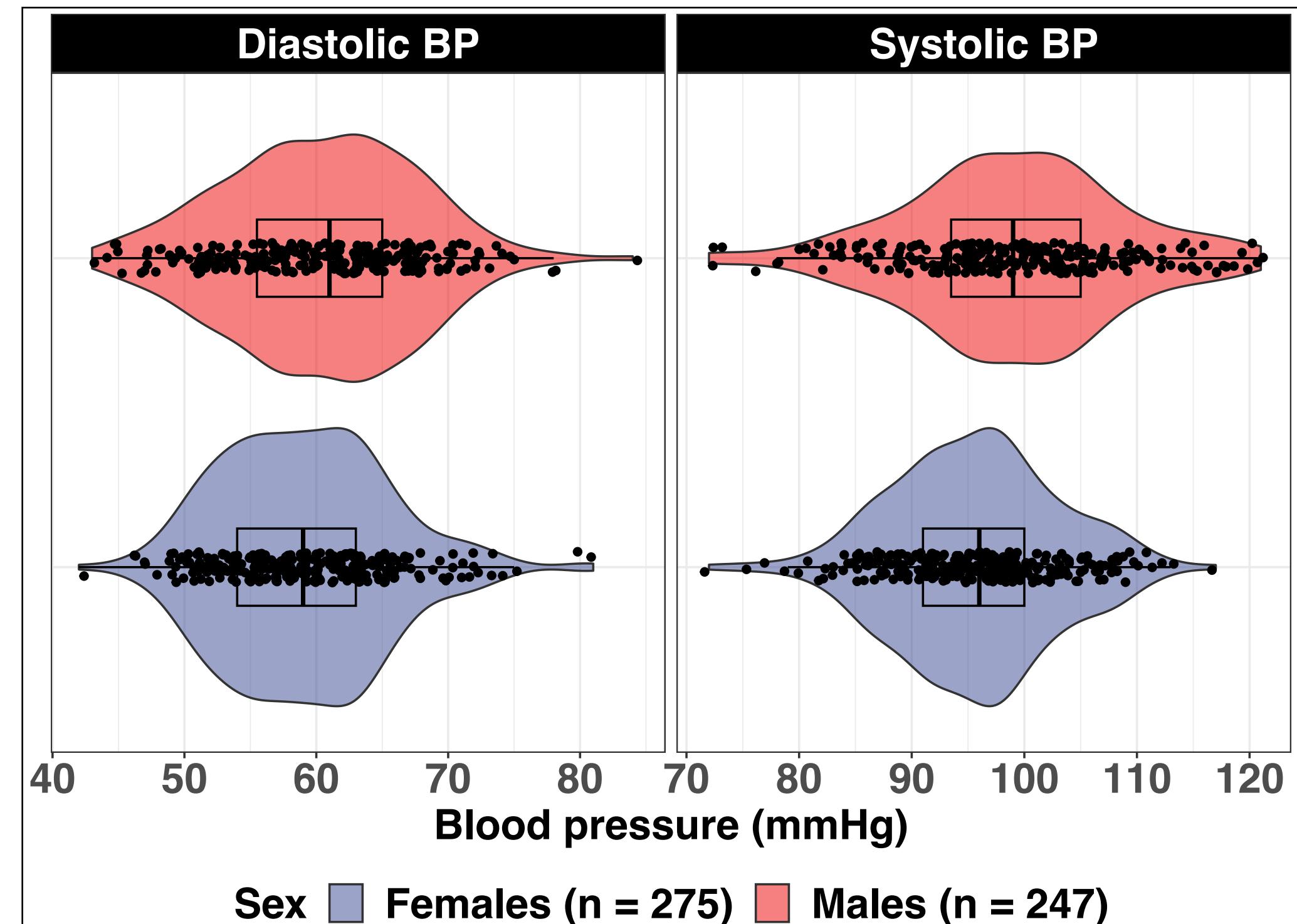
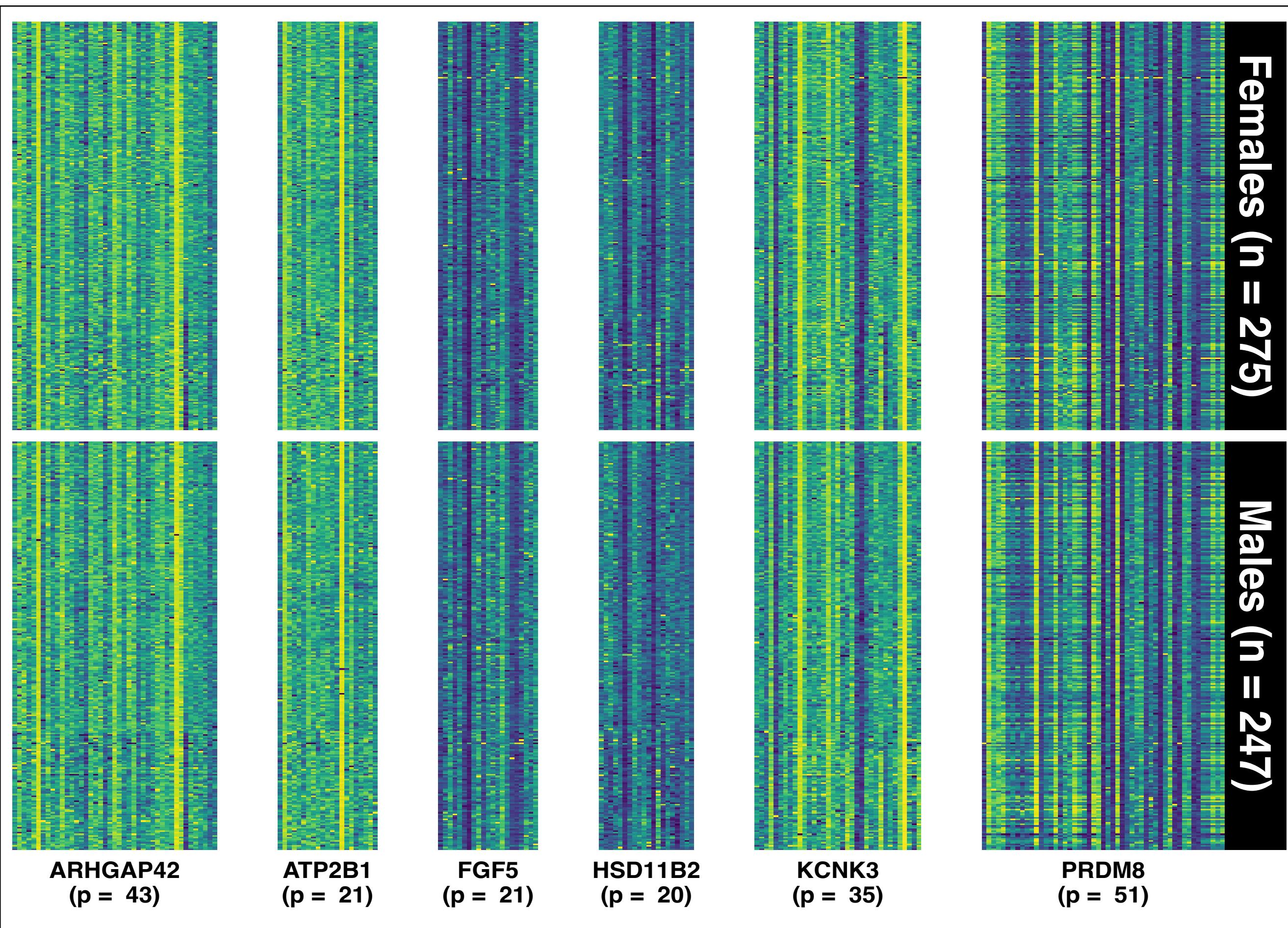
- Establish “critical value” σ_{CRIT}

$$\sigma \leq \sigma_{CRIT} = \frac{\exp(2C_{X>Y}) - 1}{I(Y)}.$$



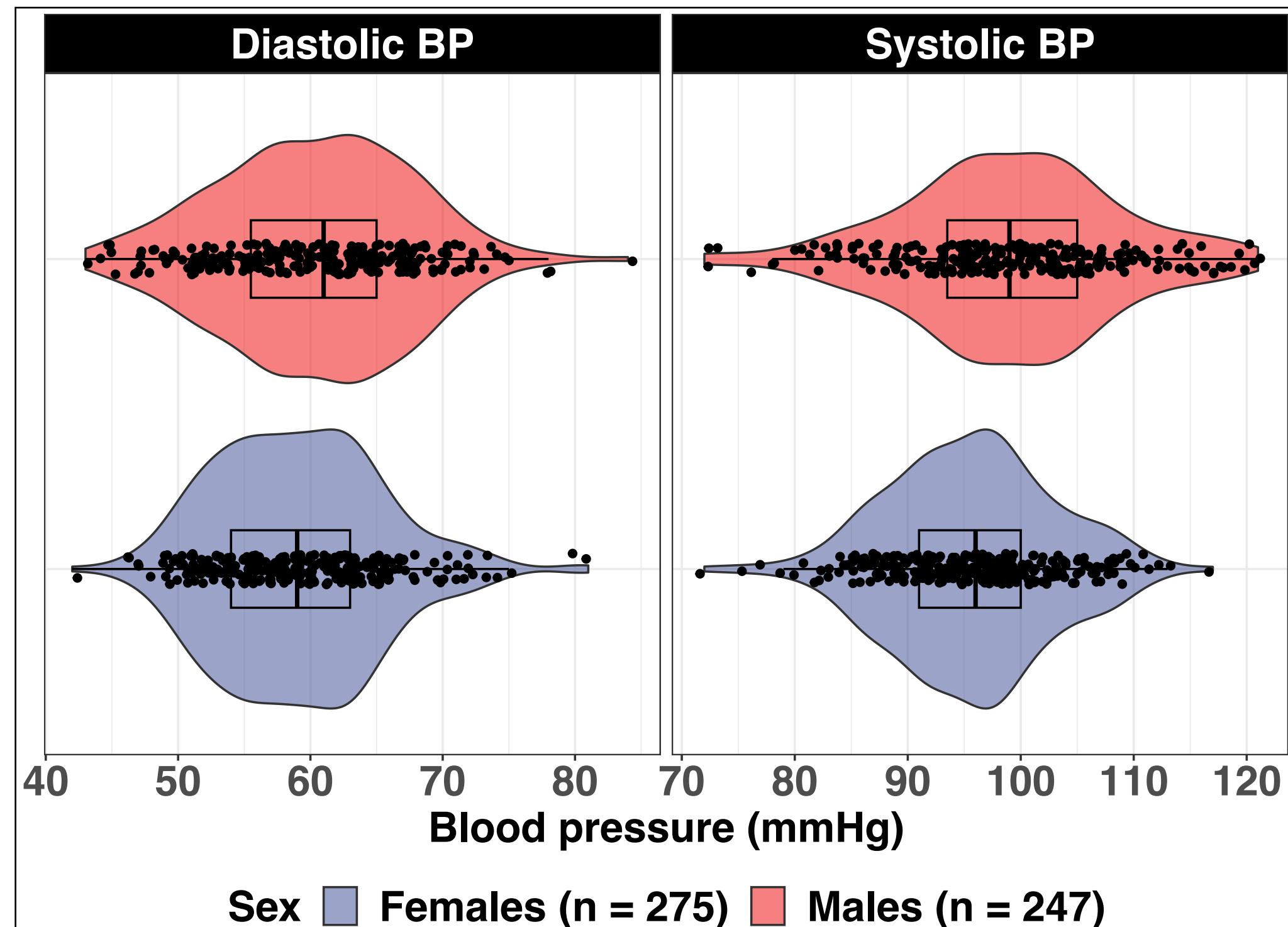
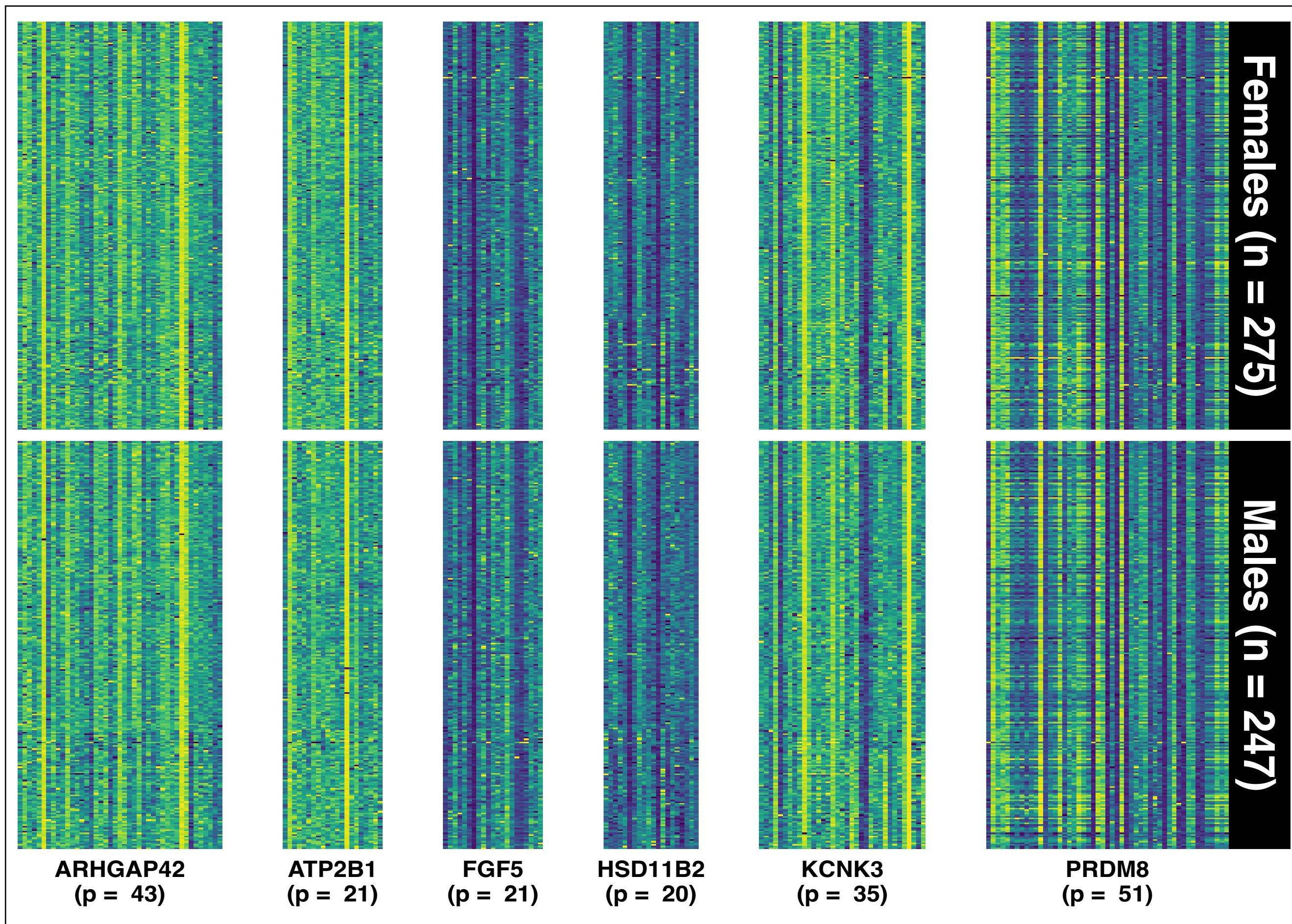
Adjusting for Z

Low-dimensional confounders



Adjusting for Z

Low-dimensional confounders

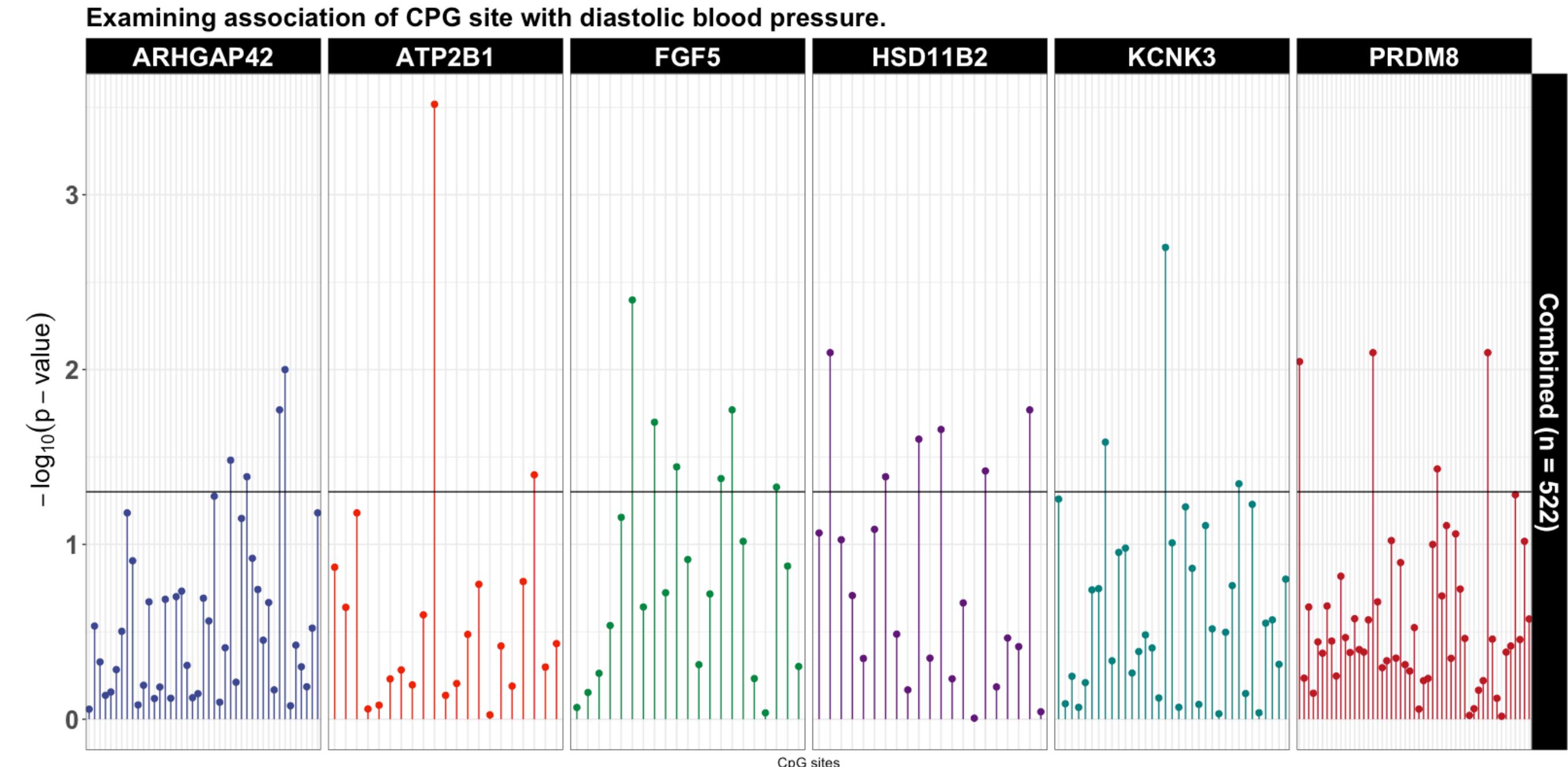
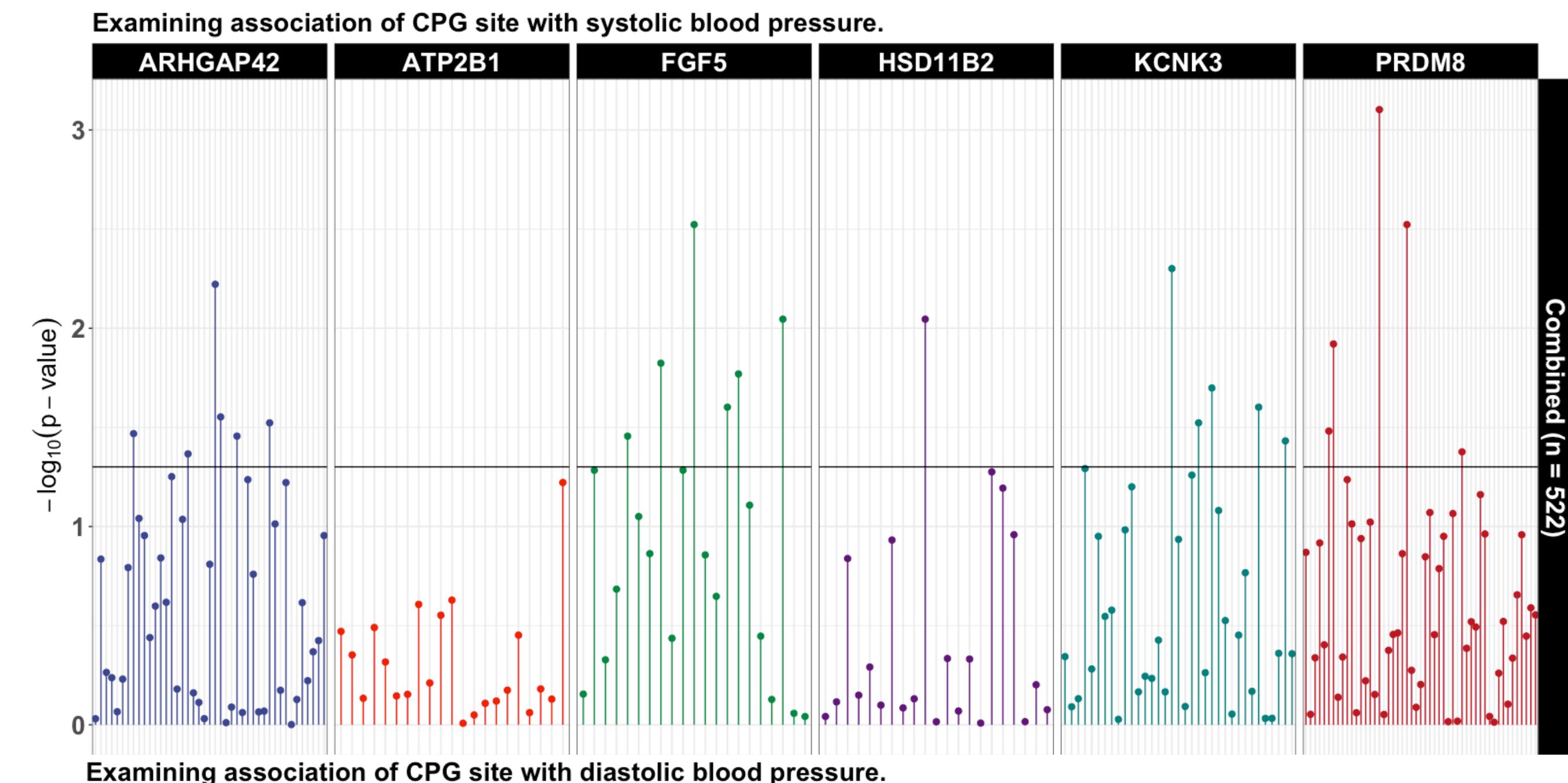


Directionality in X and Y conditional on Z :

$$C_{X>Y|Z} := H(X \mid Z = z) - H(Y \mid Z = z)$$

BP → *DNAm*?
Use $\hat{C}_{X>Y}$ for clues!

But what about directionality?



$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

But what about directionality?

Correlated CpG sites: Aggregated mean DNAm for a given gene.

$X : BP$ (either diastolic or systolic)

$Y : DNAm$ for a given gene

Obtain $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z : sex of study participant.

$BP \rightarrow DNAm$?

Use $\hat{C}_{X>Y}$ for clues!

Correlated CpG sites: Aggregated mean DNAm for a given gene.

$X : BP$ (either diastolic or systolic)

$Y : DNAm$ for a given gene

Obtain $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z : sex of study participant.

Forest plot of $\hat{C}_{X>Y}$ (95% CI)

- Estimates for female group in blue
- Estimates for male group in red
- Unadjusted combined sample in green

$BP \rightarrow DNAm$?

Use $\hat{C}_{X>Y}$ for clues!

Correlated CpG sites: Aggregated mean DNAm for a given gene.

$X : BP$ (either diastolic or systolic)

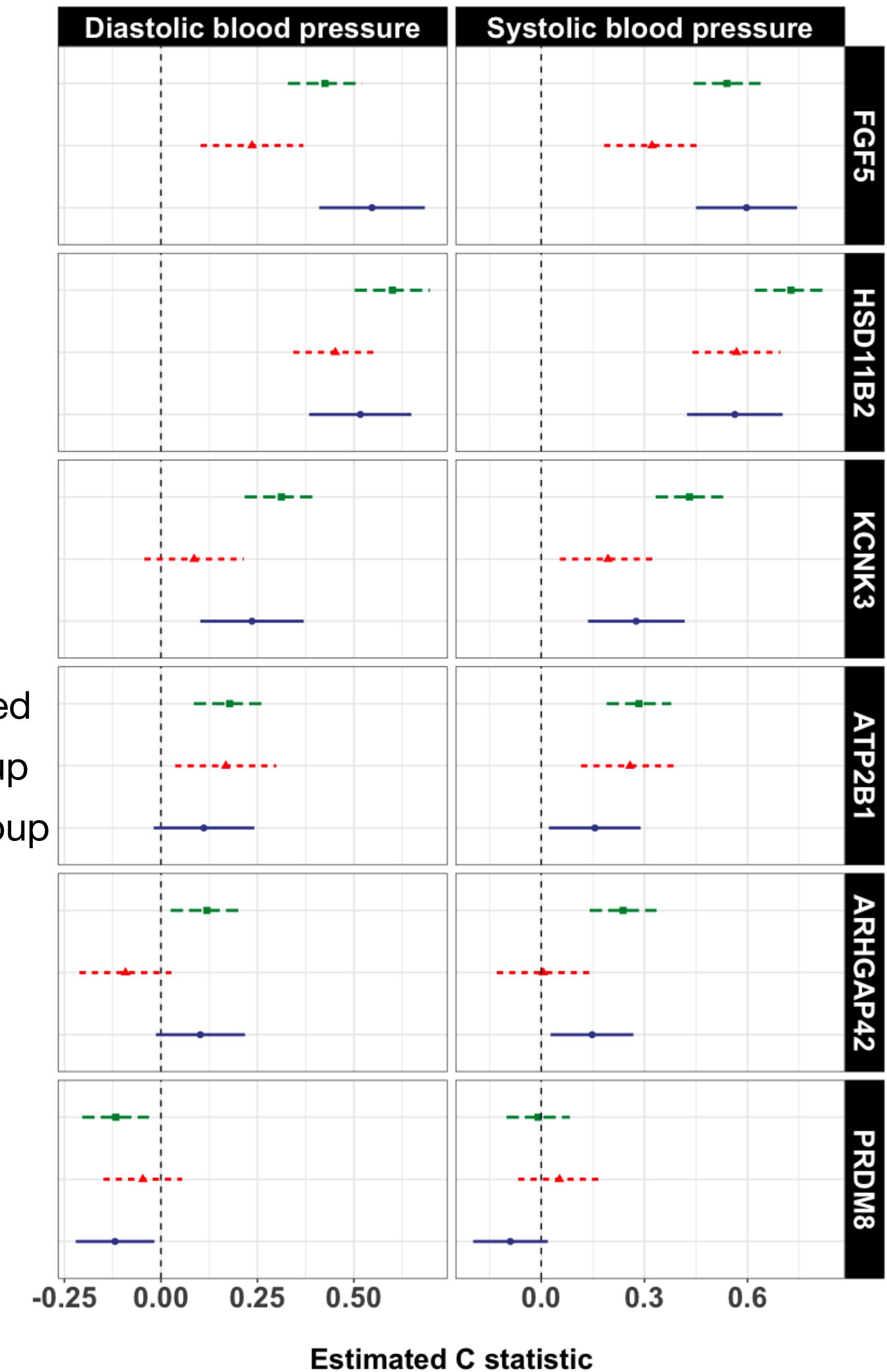
$Y : DNAm$ for a given gene

Obtain $\hat{C}_{X>Y}$ (95% CI)

Stratify by Z : sex of study participant.

Forest plot of $\hat{C}_{X>Y}$ (95% CI)

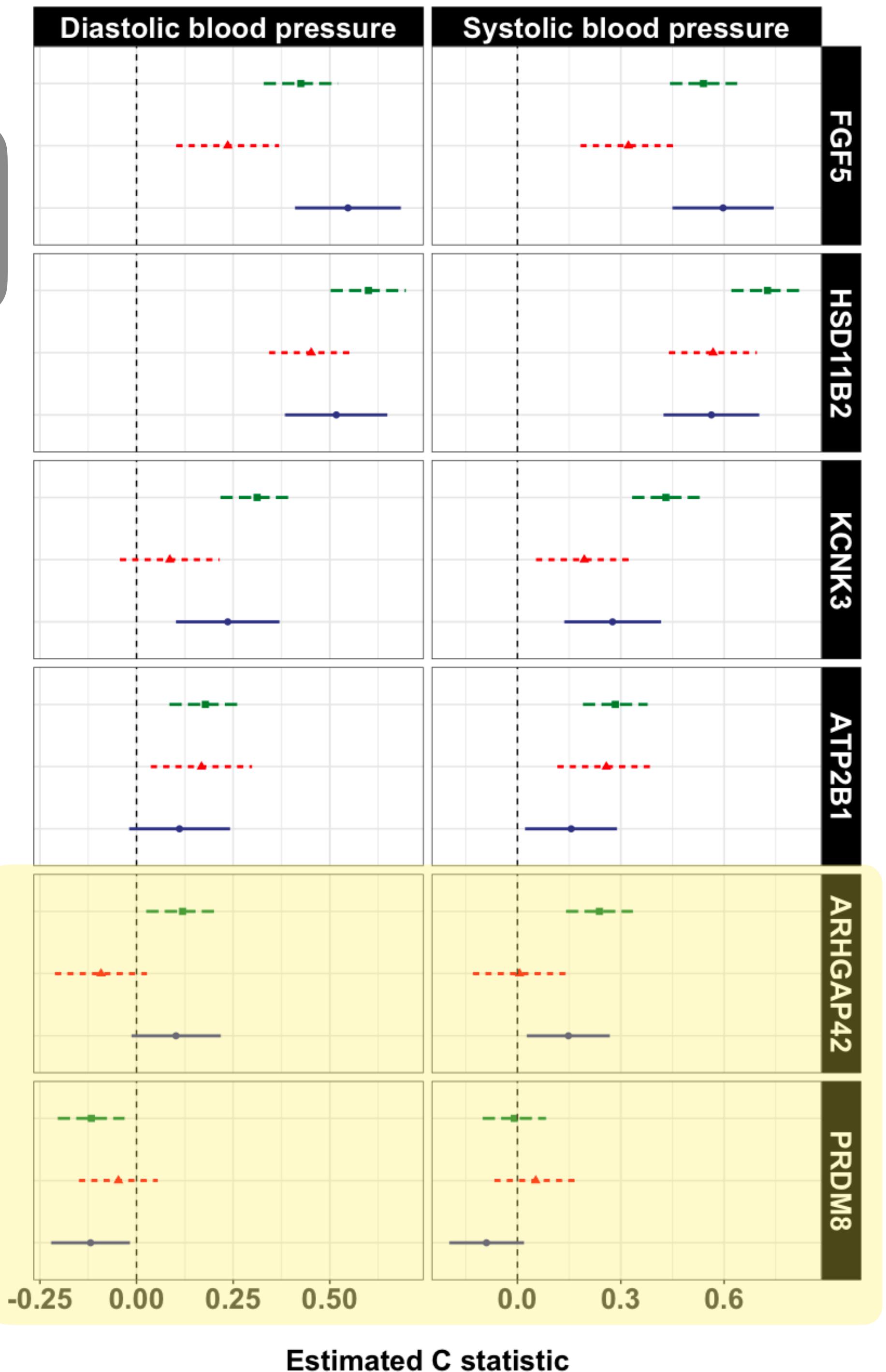
- Estimates for female group in blue
- Estimates for male group in red
- Unadjusted combined sample in green



$BP \rightarrow DNAm?$
Use $\hat{C}_{X>Y}$ for clues!

1. Weak/no signal: *PRDM8* and *ARHGAP42*

- Forest plot of $\hat{C}_{X>Y}$ (95% CI)
- Estimates for female group in blue
- Estimates for male group in red
- Unadjusted combined sample in green

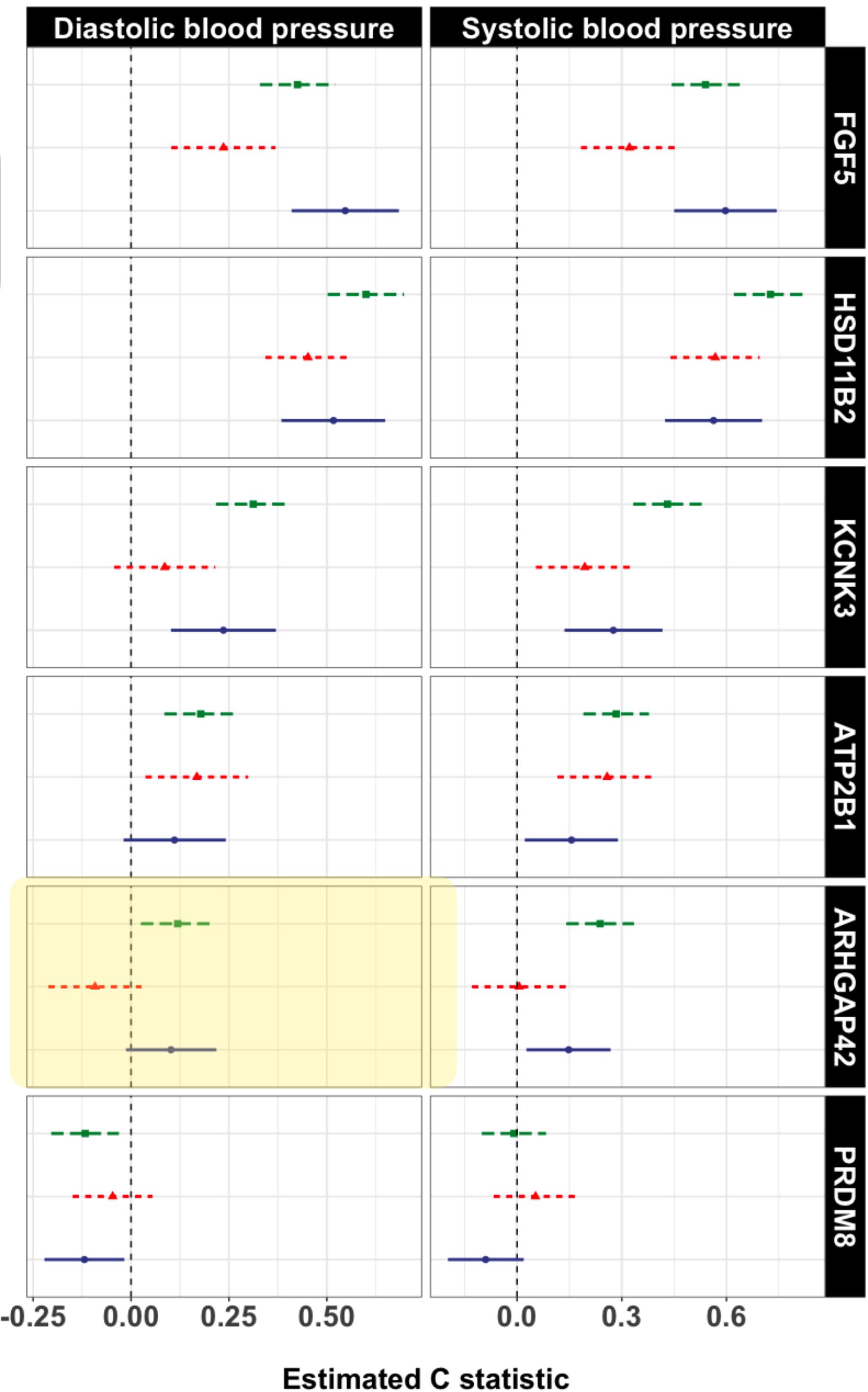


$BP \rightarrow DNAm?$

Use $\hat{C}_{X>Y}$ for clues!

- Forest plot of $\hat{C}_{X>Y}$ (95% CI)
- Estimates for female group in blue
 - Estimates for male group in red
 - Unadjusted combined sample in green

1. Weak/no signal: *PRDM8* and *ARHGAP42*
2. Simpson's paradox: *DBP* and *ARHGAP42*

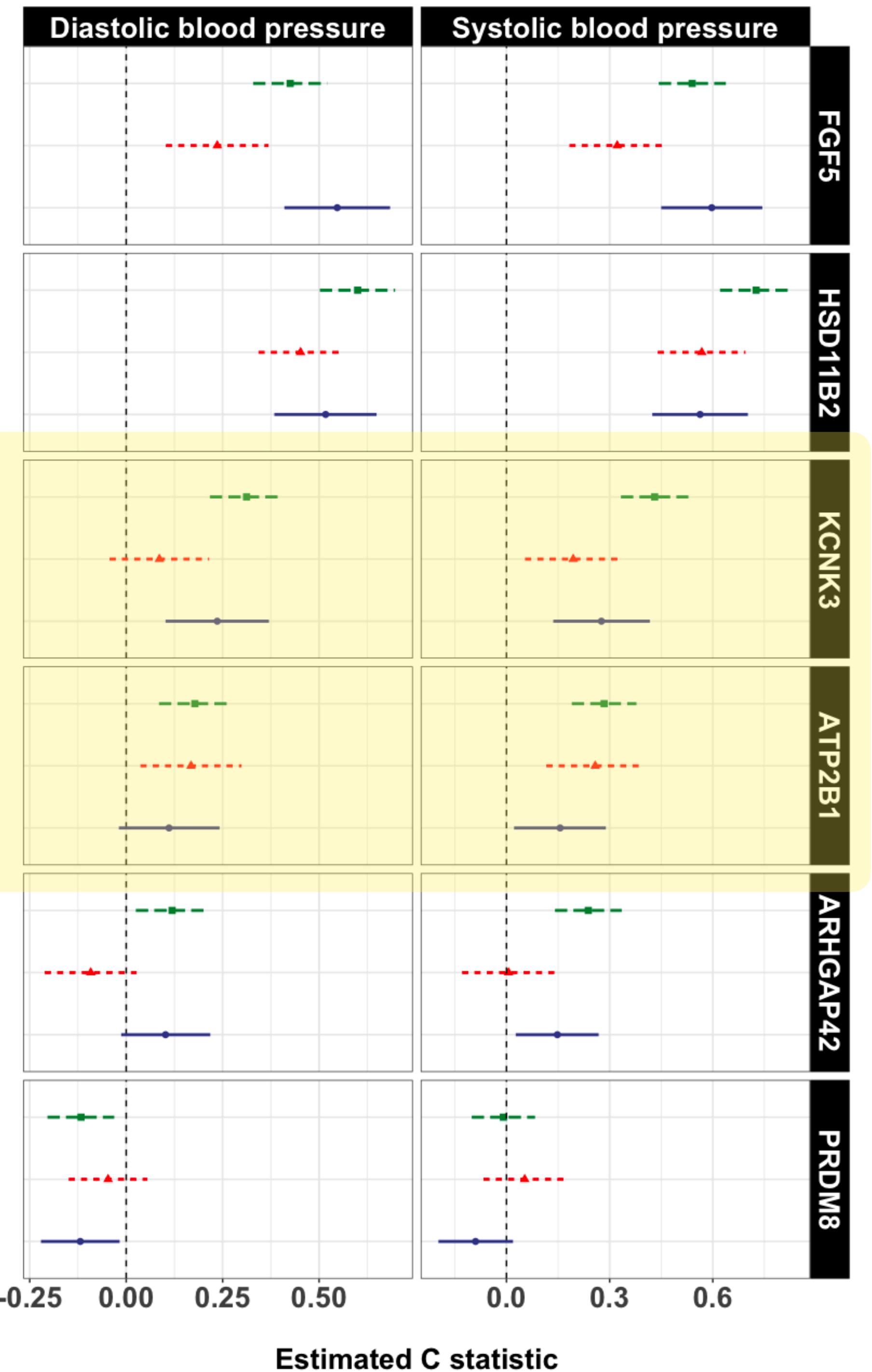


$BP \rightarrow DNAm$?

Use $\hat{C}_{X>Y}$ for clues!

- Forest plot of $\hat{C}_{X>Y}$ (95% CI)
- Estimates for female group in blue
 - Estimates for male group in red
 - Unadjusted combined sample in green

1. Weak/no signal: *PRDM8* and *ARHGAP42*
2. Simpson's paradox: *DBP* and *ARHGAP42*
3. Some positive signals:
 1. Unadjusted *DBP* \rightarrow *ATP2B1*
 2. Unadjusted *SBP* \rightarrow *ATP2B1*
 3. Unadjusted *DBP* \rightarrow *KCNK3*
 4. Unadjusted *SBP* \rightarrow *KCNK3*

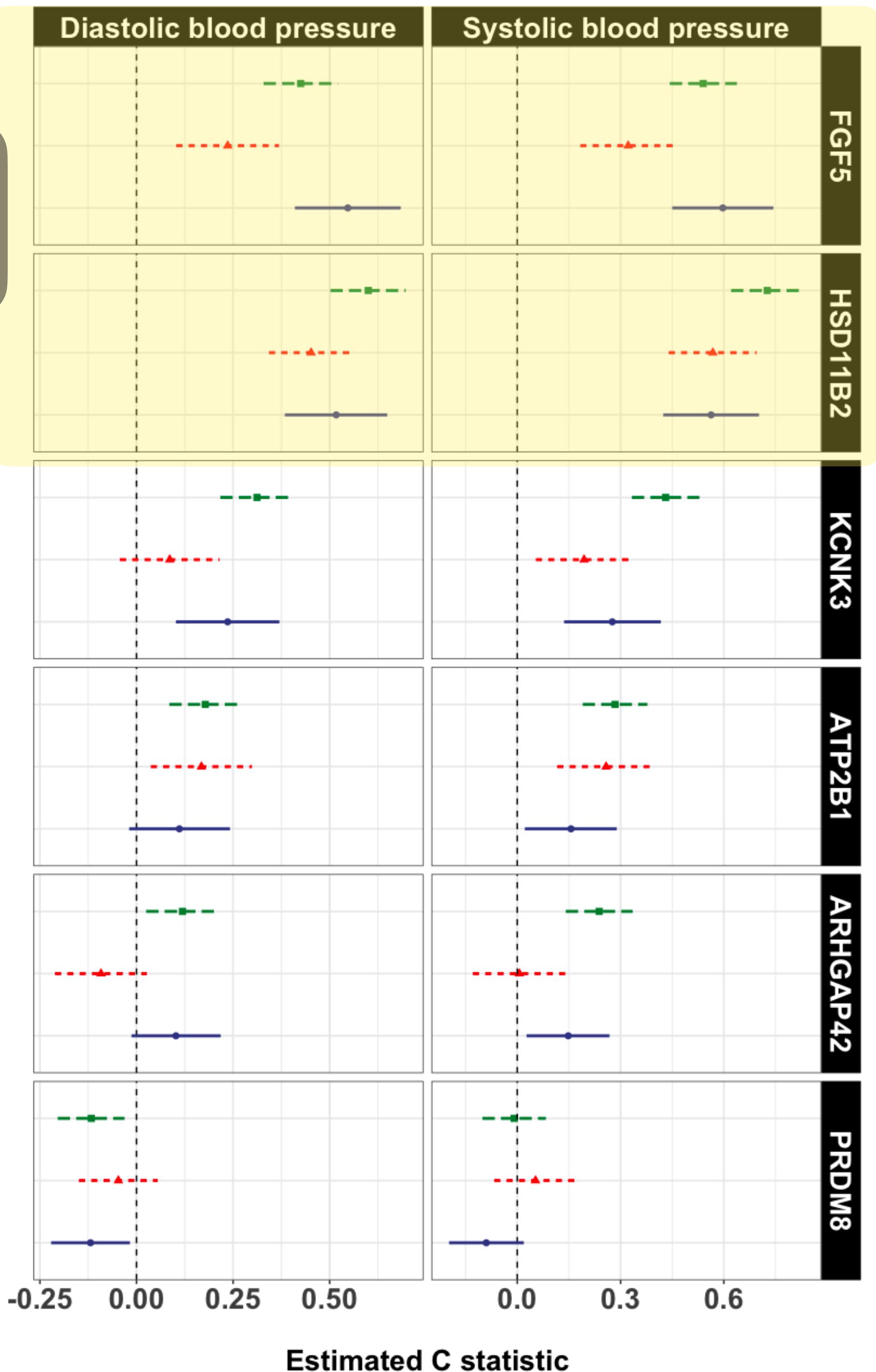


$BP \rightarrow DNAm?$

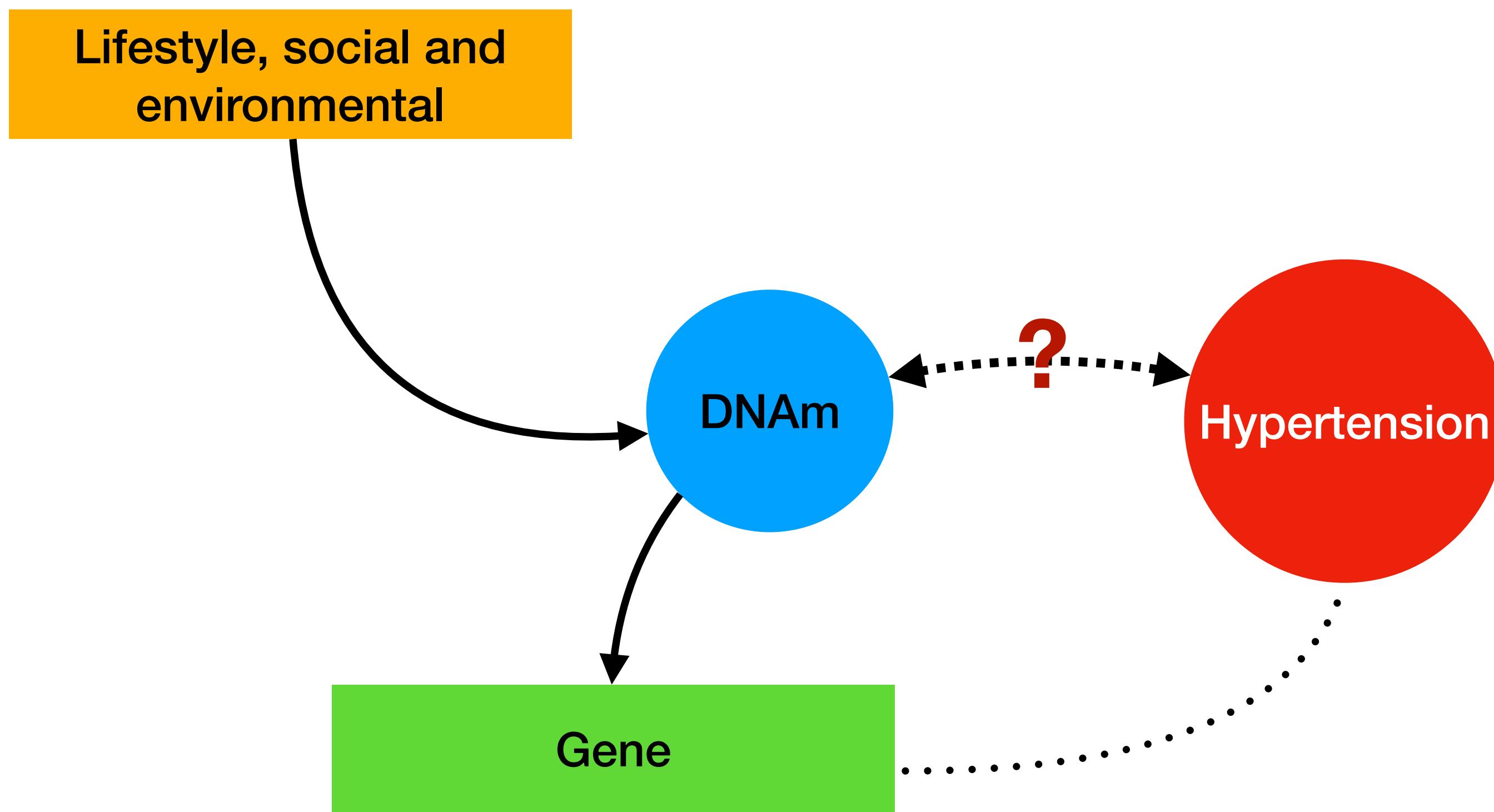
Use $\hat{C}_{X>Y}$ for clues!

- Forest plot of $\hat{C}_{X>Y}$ (95% CI)
- Estimates for female group in blue
 - Estimates for male group in red
 - Unadjusted combined sample in green

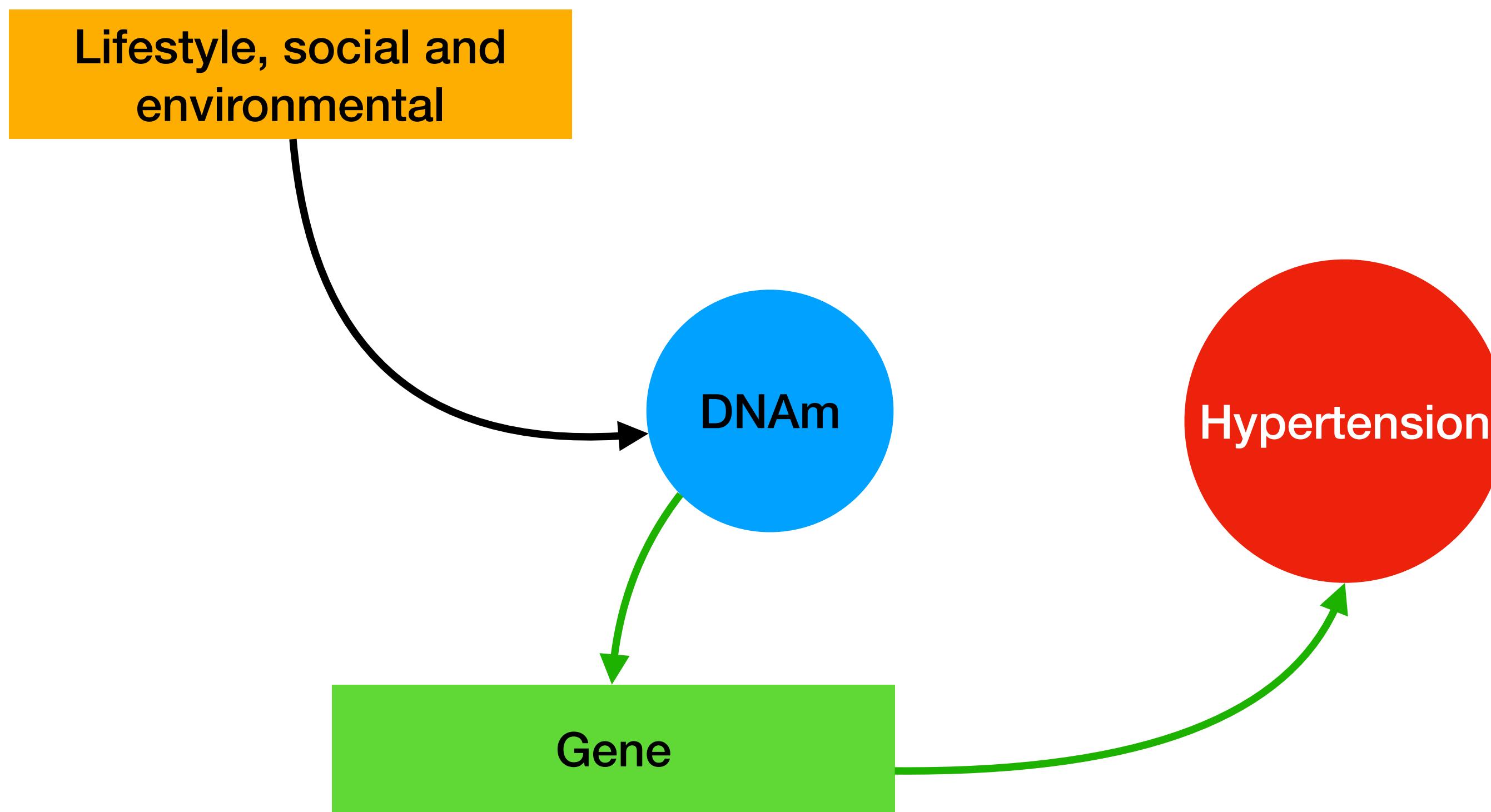
1. Weak/no signal: *PRDM8* and *ARHGAP42*
2. Simpson's paradox: *DBP* and *ARHGAP42*
3. Some positive signals:
 1. Unadjusted *DBP* \rightarrow *ATP2B1*
 2. Unadjusted *SBP* \rightarrow *ATP2B1*
 3. Unadjusted *DBP* \rightarrow *KCNK3*
 4. Unadjusted *SBP* \rightarrow *KCNK3*
4. All positive:
 1. *SBP* \rightarrow *KCNK3*
 2. *DBP* \rightarrow *HSD11B2*; *SBP* \rightarrow *HSD11B2*
 3. *DBP* \rightarrow *FGF5*; *SBP* \rightarrow *FGF5*



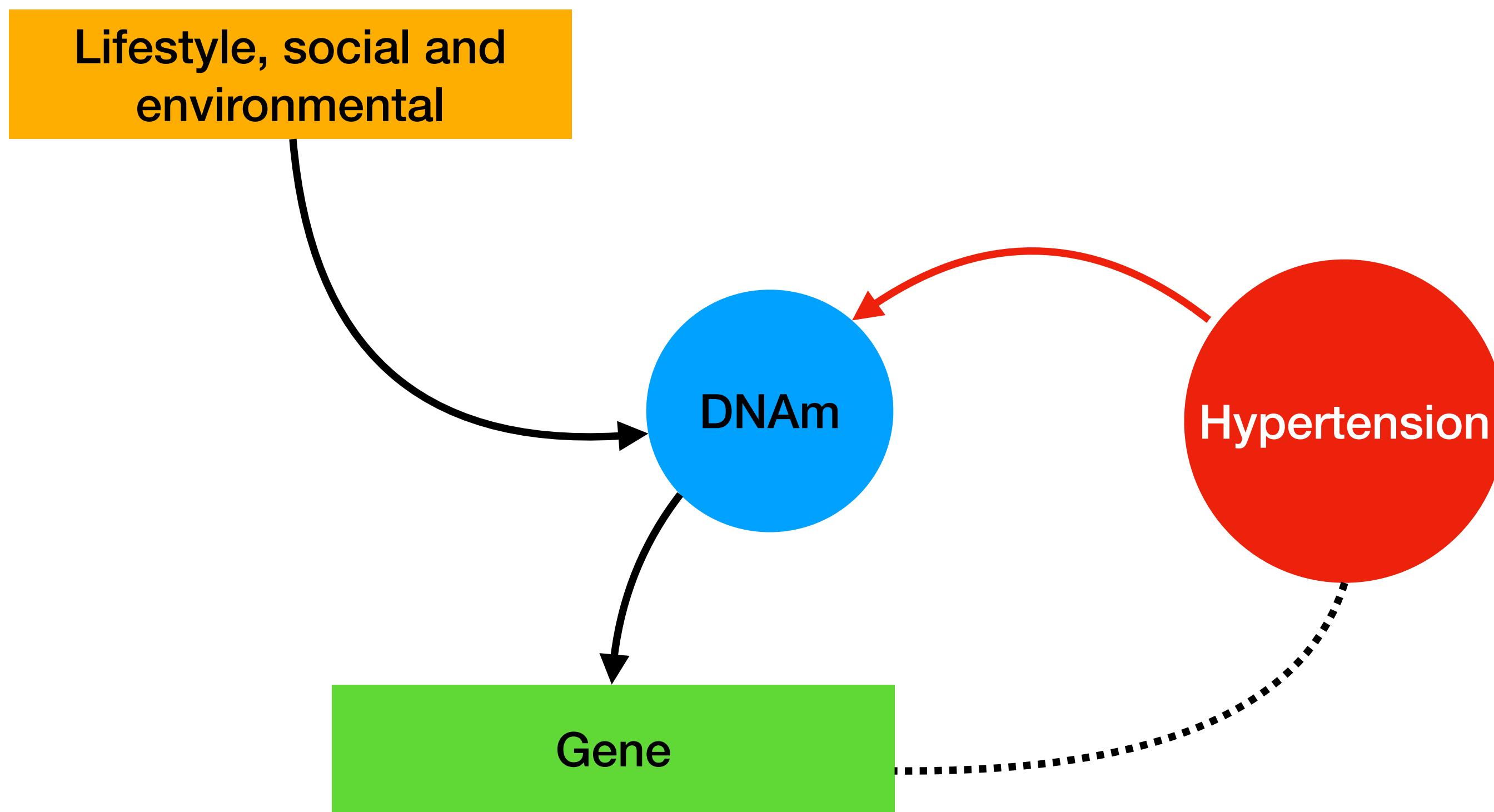
$BP \rightarrow DNAm?$



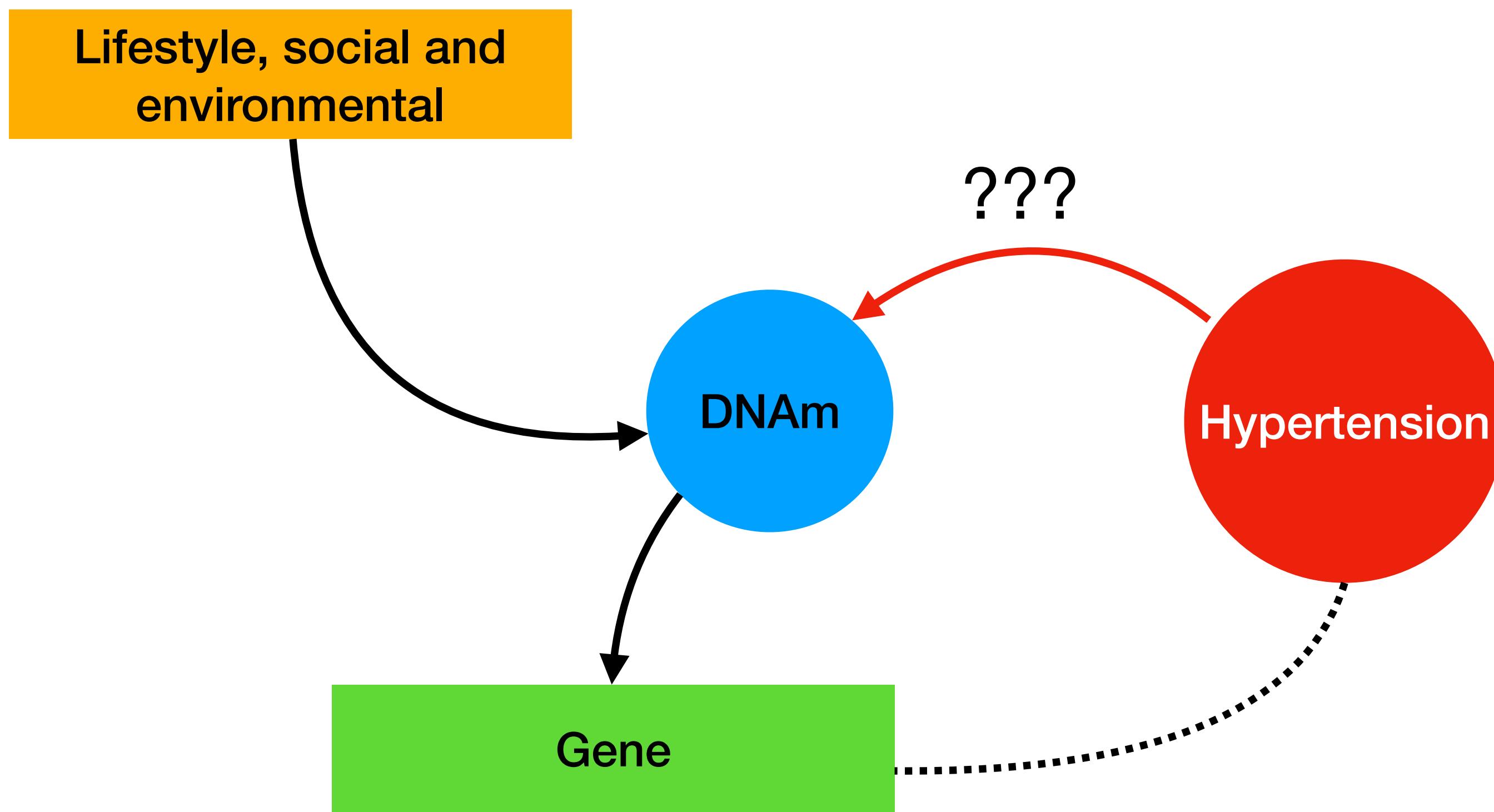
$BP \rightarrow DNAm?$



$BP \rightarrow DNAm?$

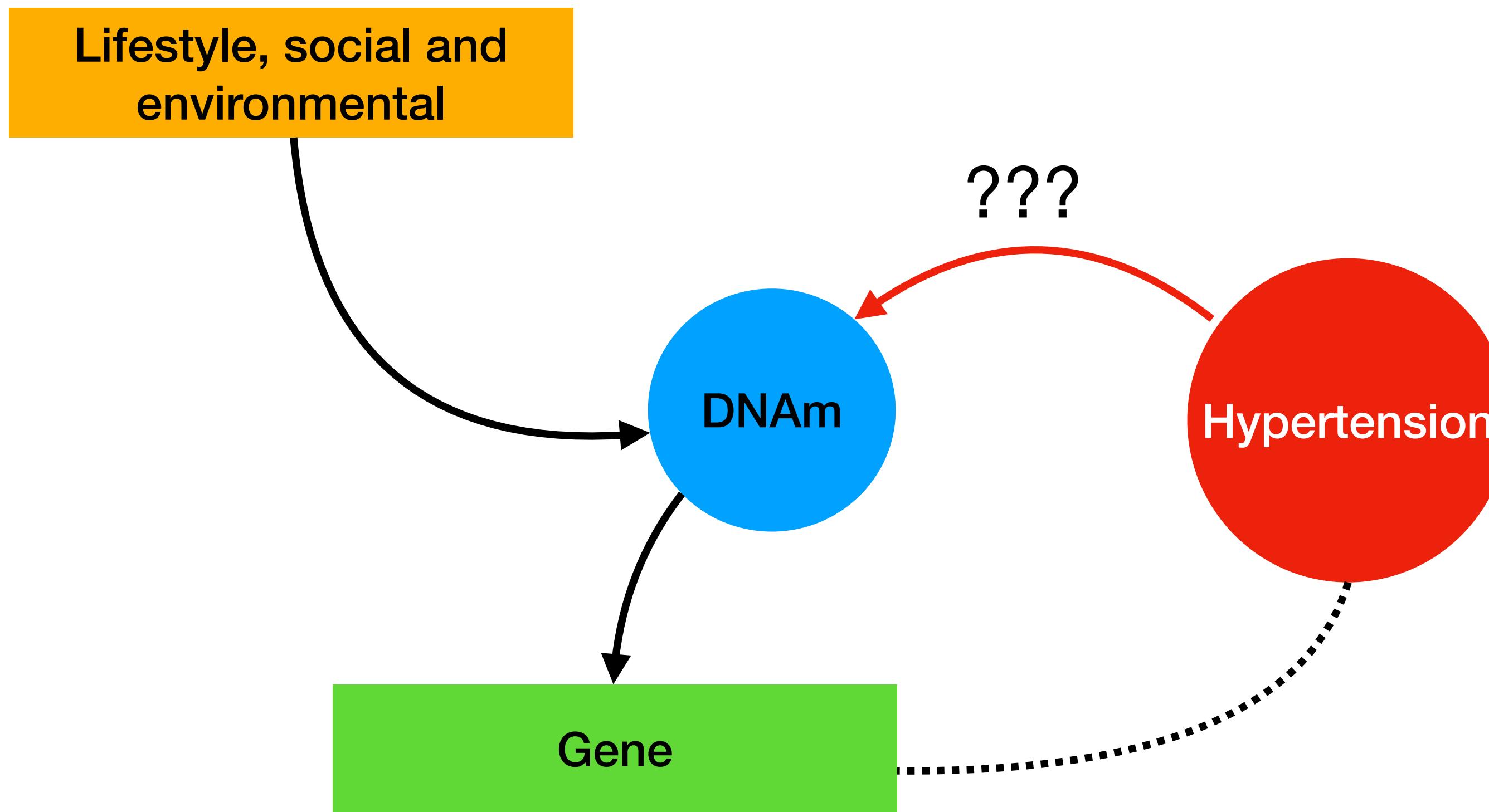


$BP \rightarrow DNAm?$



BP → *DNAm*?

Epigenetic Changes in
Response to Hypertension

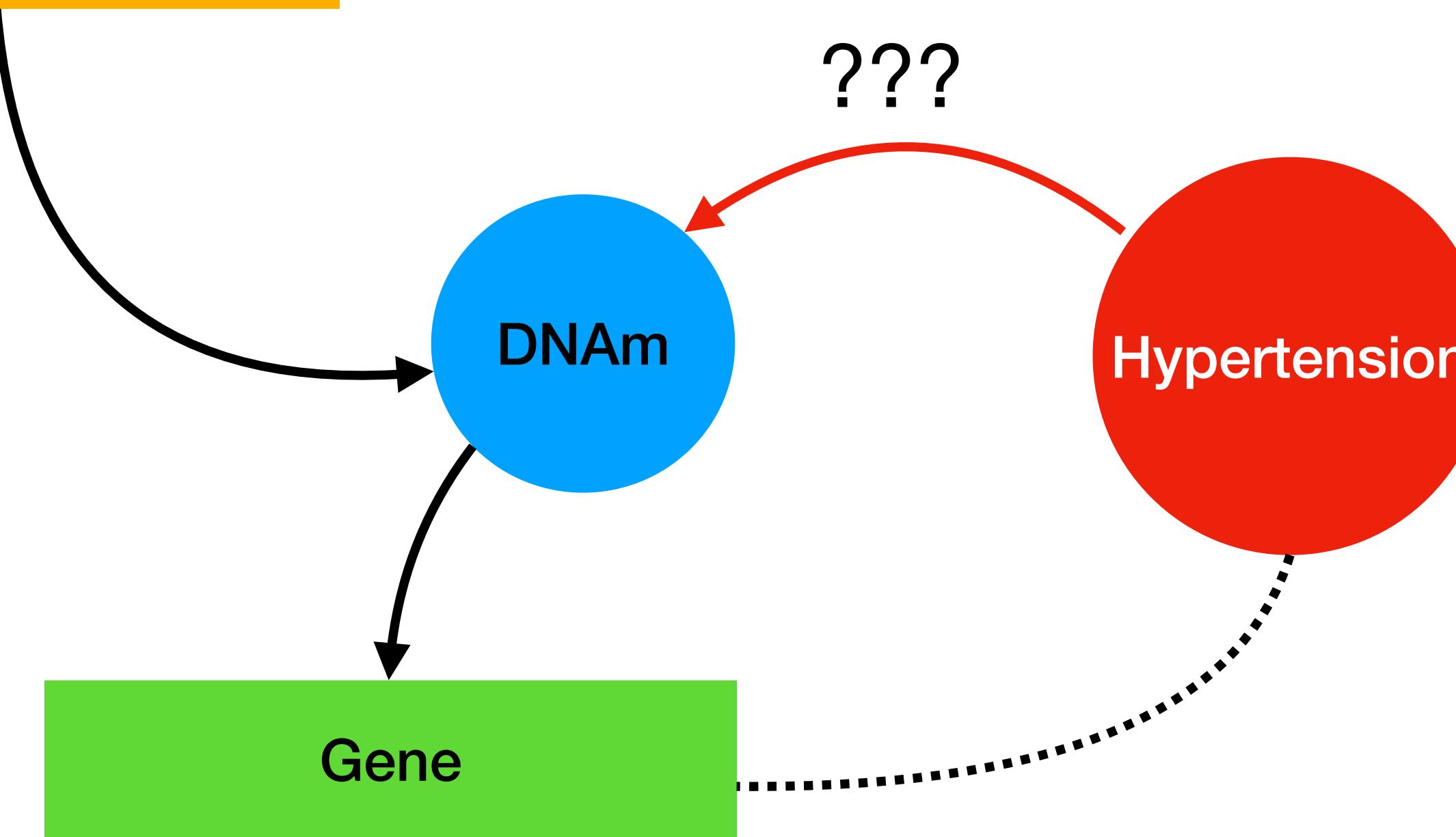


$BP \rightarrow DNAm?$

Epigenetic Changes in
Response to Hypertension

Lifestyle, social and
environmental

Impact on Endothelial Cells

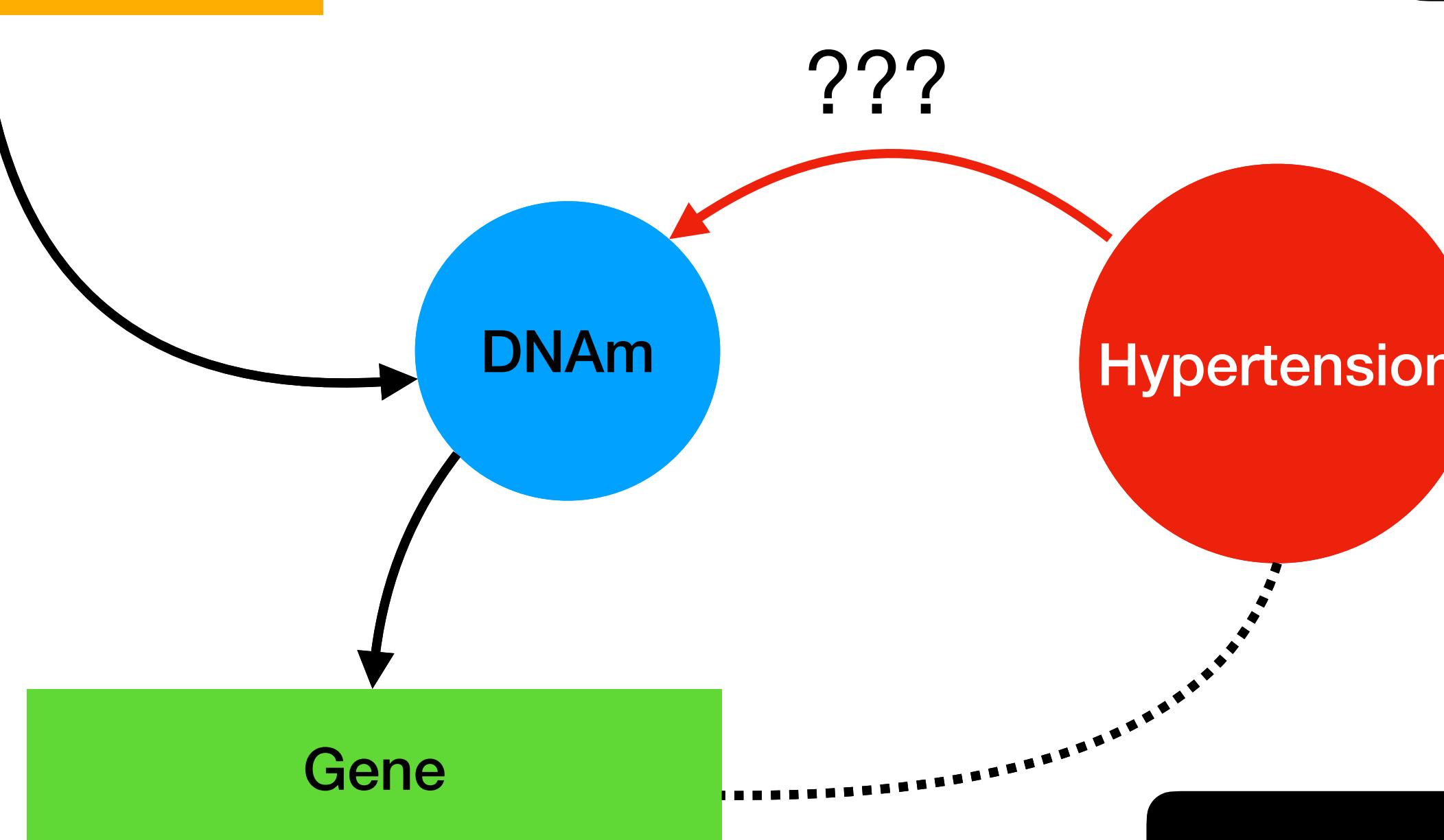


BP → *DNAm*?

Epigenetic Changes in
Response to Hypertension

Lifestyle, social and
environmental

Impact on Endothelial Cells



$BP \rightarrow DNAm?$

Epigenetic Changes in Response to Hypertension

Lifestyle, social and environmental

Impact on Endothelial Cells

Hypertension

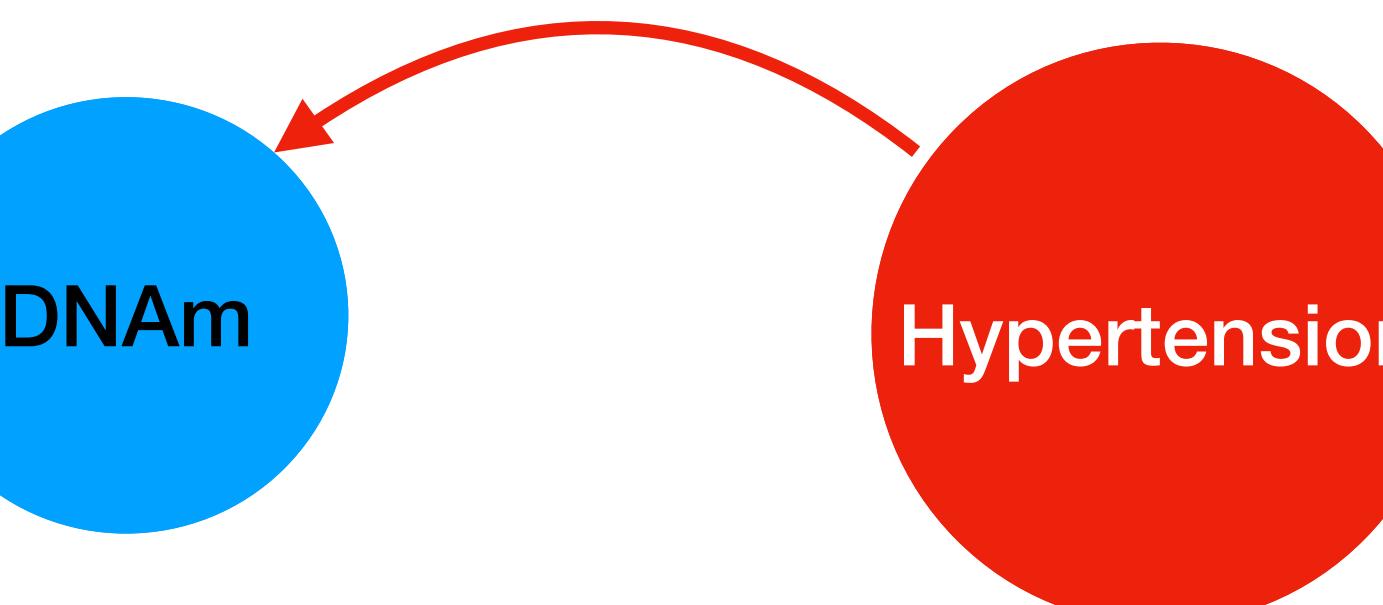
DNAm

Gene

Inflammation and Oxidative Stress

Effects on Immune System Genes

???



$BP \rightarrow DNAm?$

Epigenetic Changes in Response to Hypertension

Lifestyle, social and environmental

Impact on Hormonal Regulation

Effects on Immune System Genes

Impact on Endothelial Cells

Hypertension

DNAm

Gene

???

Inflammation and Oxidative Stress

$BP \rightarrow DNAm?$

Epigenetic Changes in Response to Hypertension

Lifestyle, social and environmental

Impact on Hormonal Regulation

Effects on Immune System Genes

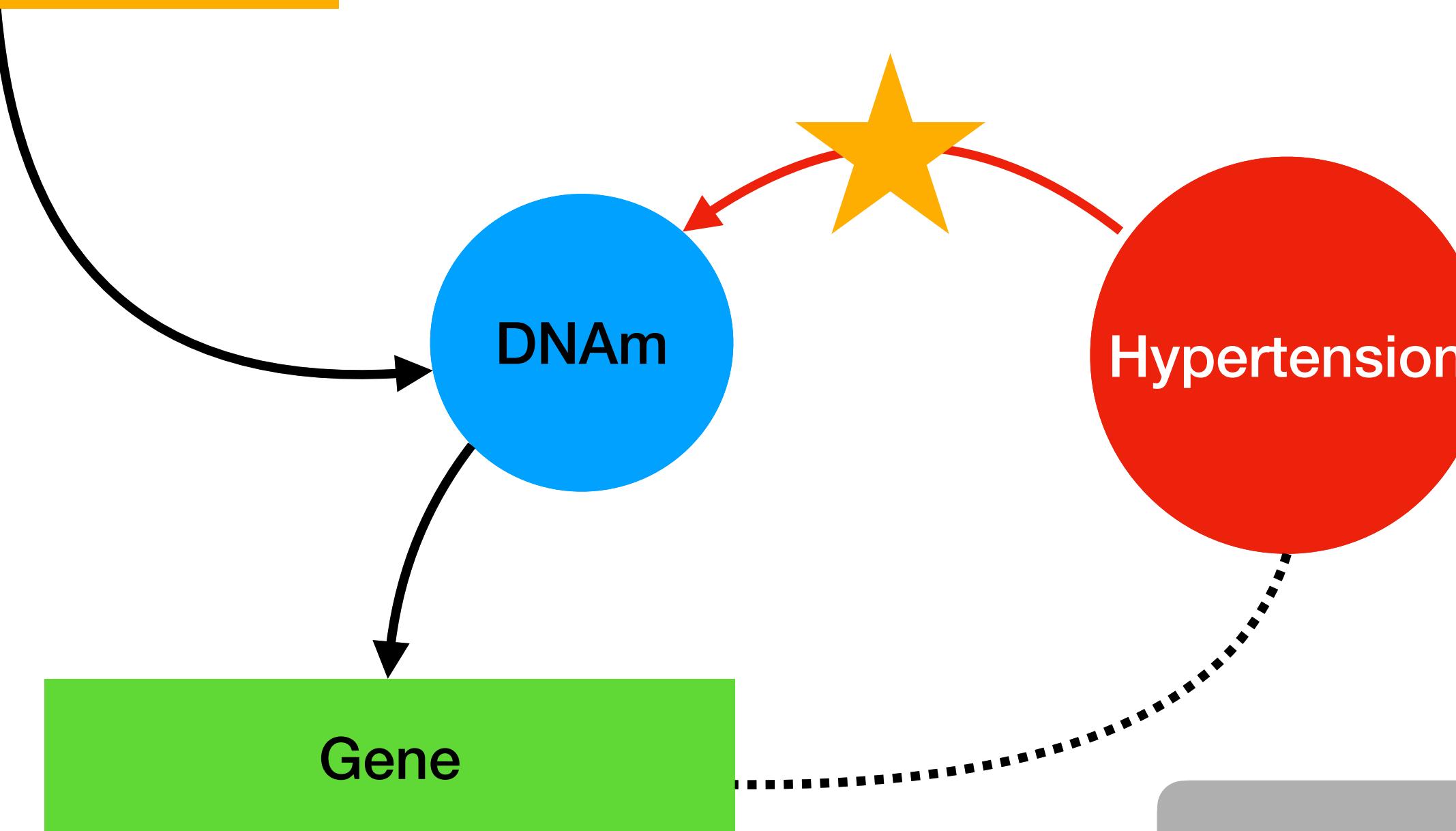
Impact on Endothelial Cells

Hypertension

DNAm

Gene

Inflammation and Oxidative Stress



Toolkit to study association and direction

Summary

Components of new toolkit

Technical strengths of toolkit

Scientific question examined

Toolkit to study association and direction

Summary

- **fastMI to study association.**
- **GEMs and asymmetry coefficient to study directionality.**

Technical strengths of toolkit

Scientific question examined

Toolkit to study association and direction

Summary

Components of new toolkit

- **Fast estimation for large n**
- **Reduced estimation error in simulations.**
- **Technical guarantees of data splitting.**

Scientific question examined

Toolkit to study association and direction

Summary

Components of new toolkit

Technical strengths of toolkit

- **Directionality in BP variation and epigenetic biomarkers established for the ELEMENT cohort.**

Toolkit to study association and direction

Summary

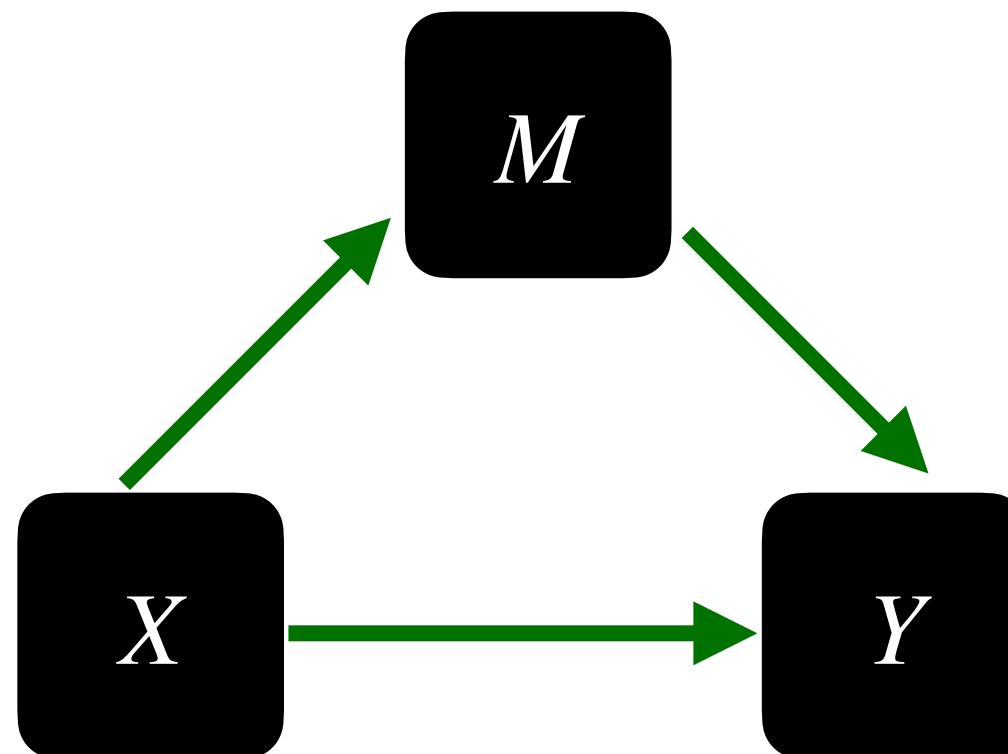
- **fastMI to study association.**
- **GEMs and asymmetry coefficient to study directionality.**
- **Fast estimation for large n**
- **Reduced estimation error in simulations.**
- **Technical guarantees of data splitting.**
- **Directionality in BP variation and epigenetic biomarkers established for the ELEMENT cohort.**

Future work

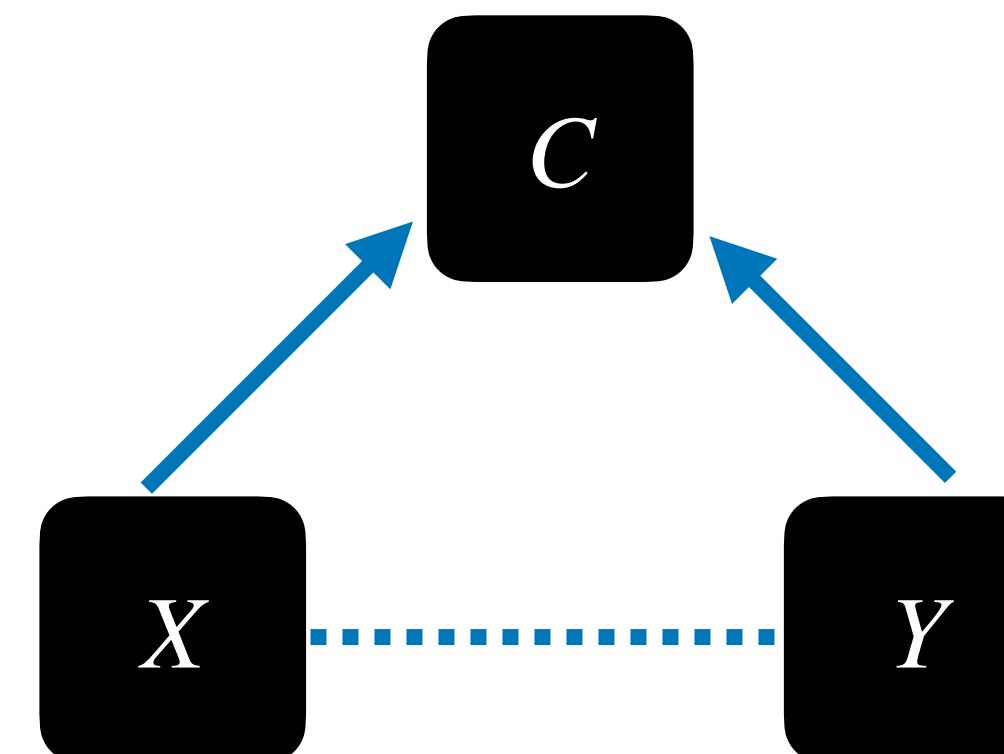
Future work I

Mediator or collider?

Extend GEM framework to detect third-variable DAGs



M mediates
path from X to Y

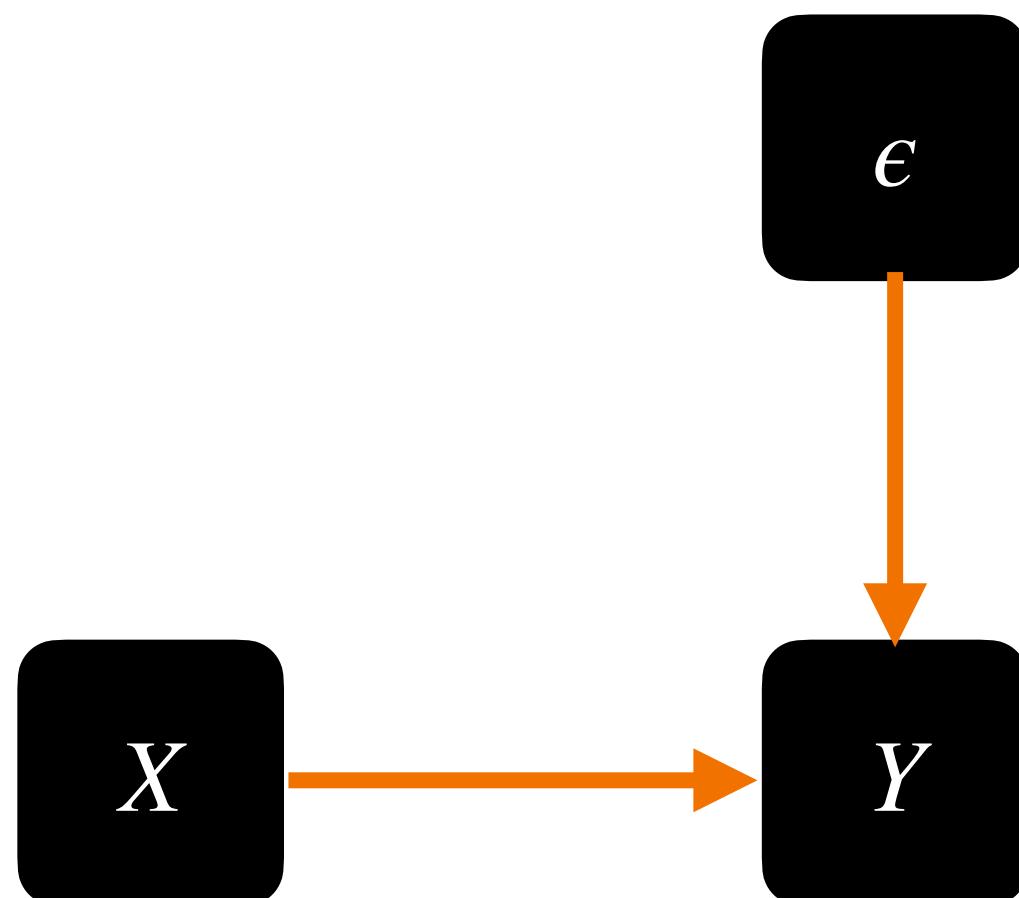


C is a collider
for X, Y

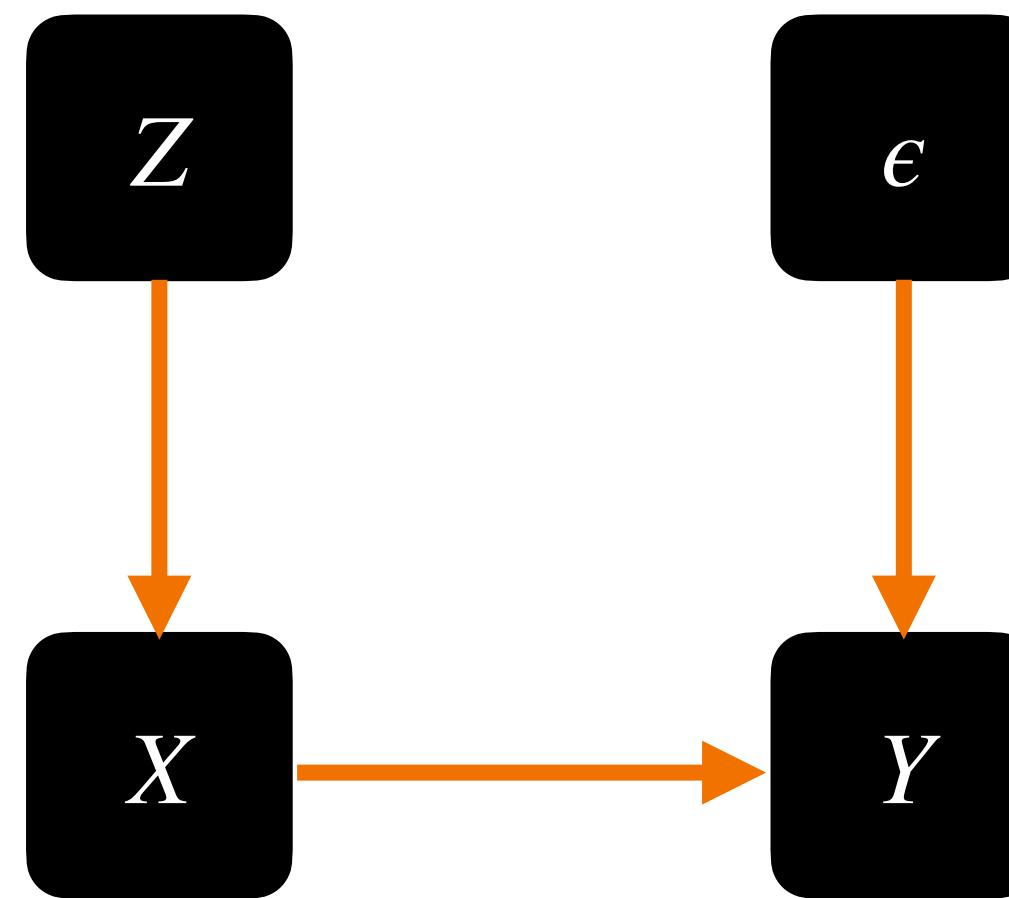
Future work II

Application to instrumental variables (IV)?

Extend NPGEM framework to handle exogeneity.



$X \perp \epsilon$

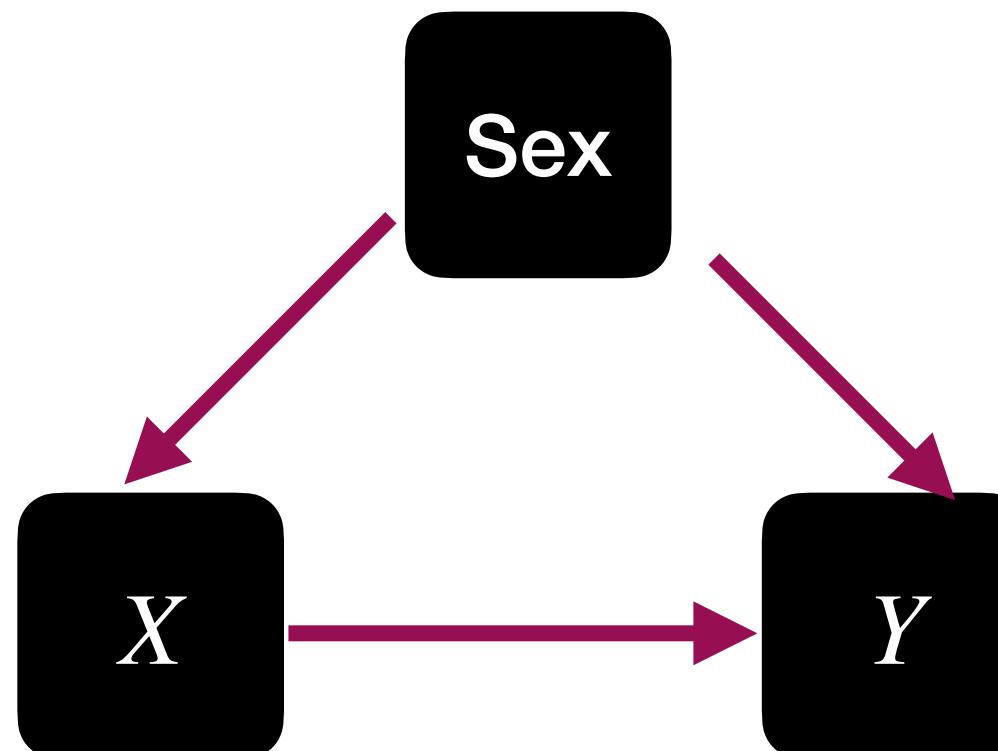


Check if Z is IV?

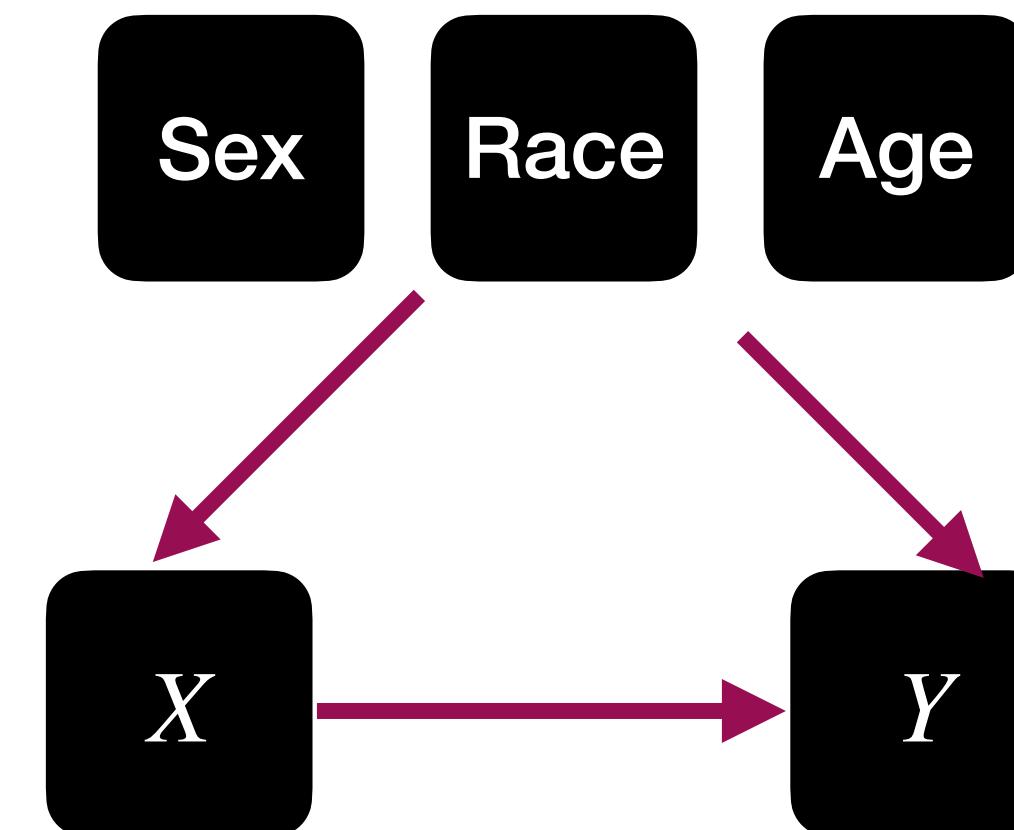
Future work III

High-dimensional confounder Z?

Estimating $C_{X>Y|Z}$ for high-dimensional Z is challenging!



Doable for few Z



Challenging as d increases

Acknowledgements



Peter Song



Karen Peterson



Jackie Goodrich

*Dissertation work:
learning from epigenetic data*

Acknowledgements



Bhramar Mukherjee

*Analyzing infectious diseases
using spatiotemporal models*



Peter Song

*Analyzing infectious diseases
using spatiotemporal models*



Cathie Spino

*Imaging and gene-based biomarkers
of wound healing in diabetes foot ulcers*



Peter Song



Thanks!

soumikp@umich.edu

soumikp.github.io

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

2.

$\hat{c}_{\mathbf{Z}}$ must minimize $MISE = \mathbb{E} \left[\int \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}) - c_{\mathbf{Z}}(\mathbf{z}) \right\}^2 d\mathbf{z} \right]$

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

2.

$\hat{c}_{\mathbf{Z}}$ must minimize $MISE = \mathbb{E} \left[\int \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}) - c_{\mathbf{Z}}(\mathbf{z}) \right\}^2 d\mathbf{z} \right]$

3.

$$\hat{\phi}_{\mathbf{t}} := \mathcal{F}(\hat{c}_{\mathbf{Z}}) \quad \phi_{\mathbf{t}} := \mathcal{F}(c_{\mathbf{Z}})$$

$$MISE \text{ in Fourier space } \mathbb{E} \left[\int \left\{ \hat{\phi}_{\mathbf{t}}(\mathbf{t}) - \phi_{\mathbf{t}}(\mathbf{t}) \right\}^2 d\mathbf{t} \right]$$

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

2.

$\hat{c}_{\mathbf{Z}}$ must minimize $MISE = \mathbb{E} \left[\int \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}) - c_{\mathbf{Z}}(\mathbf{z}) \right\}^2 d\mathbf{z} \right]$

4.

$$\hat{\mathcal{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^n \exp(i\mathbf{t}'\mathbf{Z}_j)$$

$\hat{\phi}_{\mathbf{t}}$ depends on empirical characteristic function $\hat{\mathcal{C}}$

3.

$$\hat{\phi}_{\mathbf{t}} := \mathcal{F}(\hat{c}_{\mathbf{Z}}) \quad \phi_{\mathbf{t}} := \mathcal{F}(c_{\mathbf{Z}})$$

$MISE$ in Fourier space $\mathbb{E} \left[\int \left\{ \hat{\phi}_{\mathbf{t}}(\mathbf{t}) - \phi_{\mathbf{t}}(\mathbf{t}) \right\}^2 dt \right]$

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

2.

$\hat{c}_{\mathbf{Z}}$ must minimize $MISE = \mathbb{E} \left[\int \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}) - c_{\mathbf{Z}}(\mathbf{z}) \right\}^2 d\mathbf{z} \right]$

4. $\hat{\mathcal{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^n \exp(i\mathbf{t}'\mathbf{Z}_j)$

$\hat{\phi}_{\mathbf{t}}$ depends on empirical characteristic function $\hat{\mathcal{C}}$

3.

$\hat{\phi}_{\mathbf{t}} := \mathcal{F}(\hat{c}_{\mathbf{Z}}) \quad \phi_{\mathbf{t}} := \mathcal{F}(c_{\mathbf{Z}})$

$MISE$ in Fourier space $\mathbb{E} \left[\int \left\{ \hat{\phi}_{\mathbf{t}}(\mathbf{t}) - \phi_{\mathbf{t}}(\mathbf{t}) \right\}^2 d\mathbf{t} \right]$

5. Antitransform $\hat{\phi}_{\mathbf{t}}$ to get $\hat{c}_{\mathbf{Z}}$

$\hat{c}_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-1} \int \hat{\phi}_{\mathbf{Z}}(\mathbf{t}) \exp(-i\mathbf{t}'\mathbf{z}) dt$

Appendix I

Fourier transform-based copula density estimation

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

1. $\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$

Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

2.

$\hat{c}_{\mathbf{Z}}$ must minimize $MISE = \mathbb{E} \left[\int \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}) - c_{\mathbf{Z}}(\mathbf{z}) \right\}^2 d\mathbf{z} \right]$

4. $\hat{\mathcal{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^n \exp(i\mathbf{t}'\mathbf{Z}_j)$

$\hat{\phi}_{\mathbf{t}}$ depends on empirical characteristic function $\hat{\mathcal{C}}$

3.

$\hat{\phi}_{\mathbf{t}} := \mathcal{F}(\hat{c}_{\mathbf{Z}}) \quad \phi_{\mathbf{t}} := \mathcal{F}(c_{\mathbf{Z}})$

$MISE$ in Fourier space $\mathbb{E} \left[\int \left\{ \hat{\phi}_{\mathbf{t}}(\mathbf{t}) - \phi_{\mathbf{t}}(\mathbf{t}) \right\}^2 d\mathbf{t} \right]$

5. Antitransform $\hat{\phi}_{\mathbf{t}}$ to get $\hat{c}_{\mathbf{Z}}$

$\hat{c}_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-1} \int \hat{\phi}_{\mathbf{Z}}(\mathbf{t}) \exp(-i\mathbf{t}'\mathbf{z}) dt$

6.

$\text{fastMI} = n^{-1} \sum_{j=1}^n \log \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}_i) \right\}$

Appendix II

Asymmetry in GEMS reflects underlying directionality Low-level imprint of Neyman-Rubin causal (NRC) model?

- Implicitly assume $X \rightarrow Y$ in NRC.
 - Impact of changing X on Y ?
 - Error-based model on Y .
 - SUTVA and random assignment assumptions.

- GEM to approve/disprove $X \rightarrow Y$
 - Use Shannon's entropy analytic
 - f_Y depends on f_X and ∇g
 - Identifiability assumption:**orthogonality.**

Appendix III

Identifiability condition for GEMs

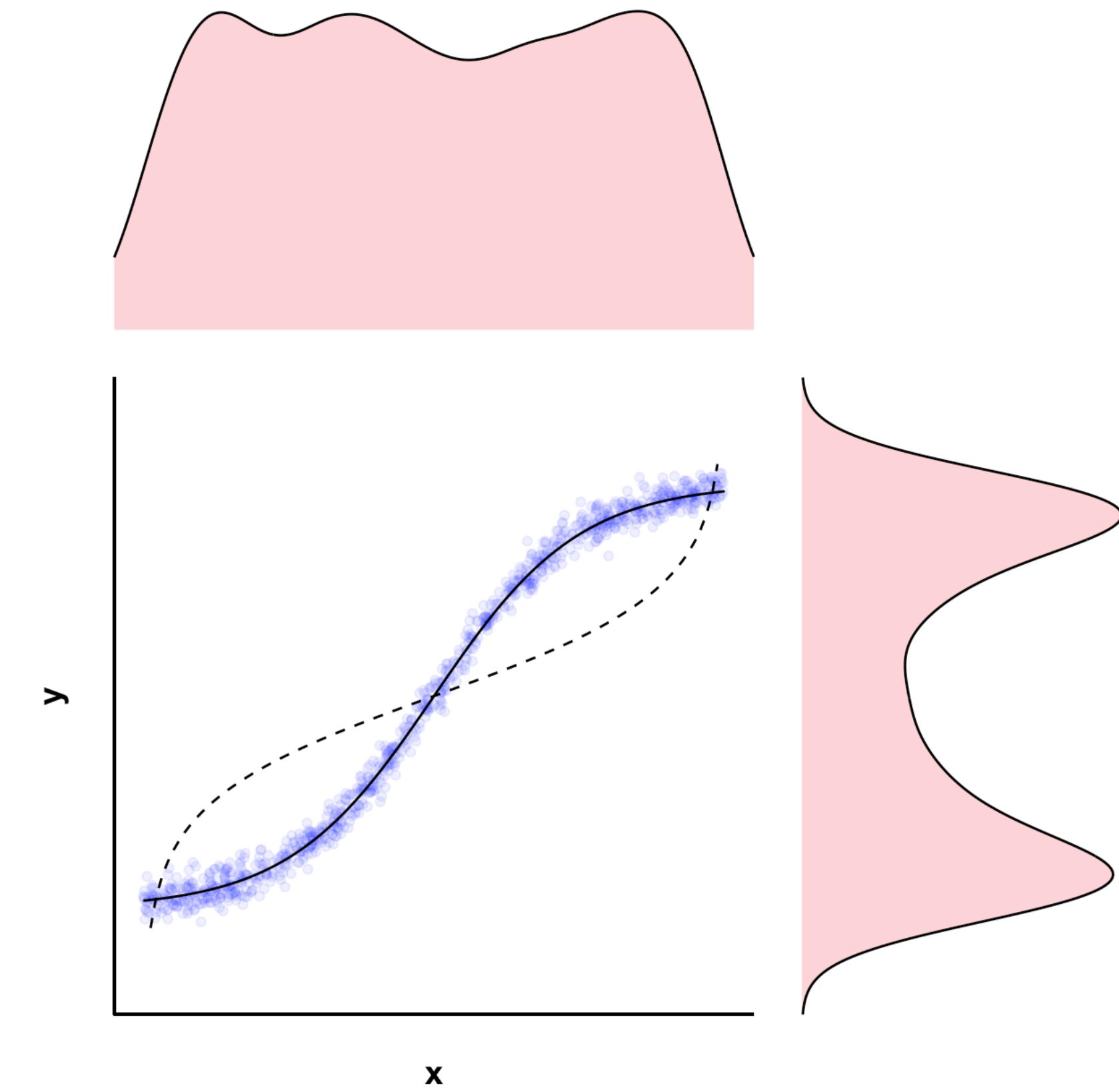
Orthogonality of f_X and g induces discrepancy in f_Y and g^{-1}

$$\int \log(|\nabla g(x)|) f_X(x) dx = \int \log(|\nabla g(x)|) dx$$

$$\int \log(|\nabla g^{-1}(y)|) f_Y(y) dy \geq \int \log(|\nabla g^{-1}(y)|) dy$$

Easier to retrieve Y from X through g ? Or get X from Y through g^{-1} ?

Entropy captures this discrepancy.



Appendix IV

Another application of the GEM framework

Auto-MPG dataset: confirming known pathways

Plot A: Displacement → MPG

- North America: 2.590 (2.514, 2.666)
- Rest: 0.767 (0.636, 0.898)
- Combined: 1.908 (1.886, 1.930)

Plot B: Horsepower → MPG

- North America: 1.681 (1.559, 1.803)
- Rest: 0.749 (0.645, 0.853)
- Combined: 1.332 (1.308, 1.356)

Plot C: Weight → MPG

- North America: 4.631 (4.558, 4.704)
- Rest: 3.664 (3.518, 3.810)
- Combined: 4.269 (4.247, 4.292)

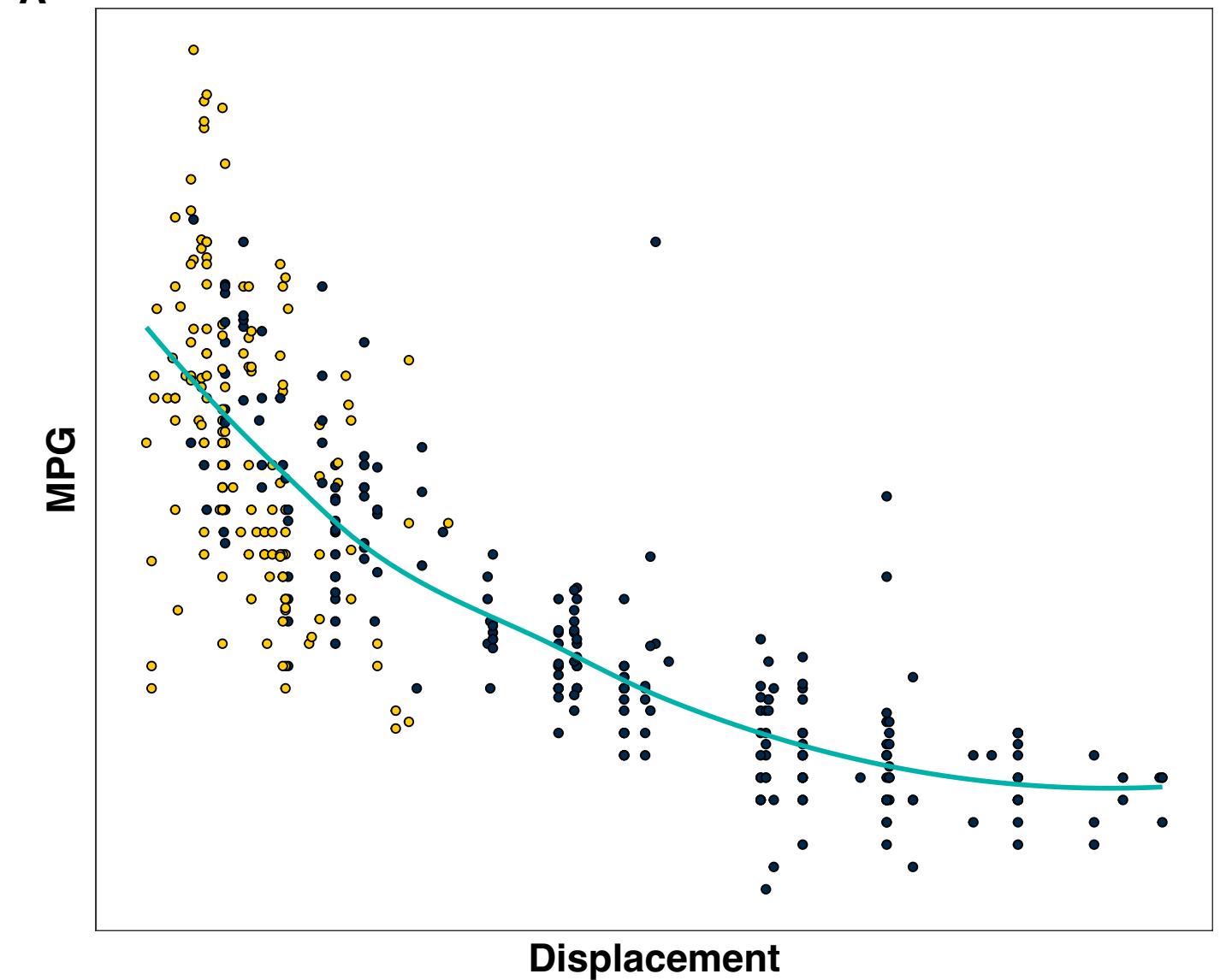
Plot D: Horsepower → Acceleration

- North America: 2.798 (2.693, 2.904)
- Rest: 2.200 (2.066, 2.334)
- Combined: 2.574 (2.550, 2.598)

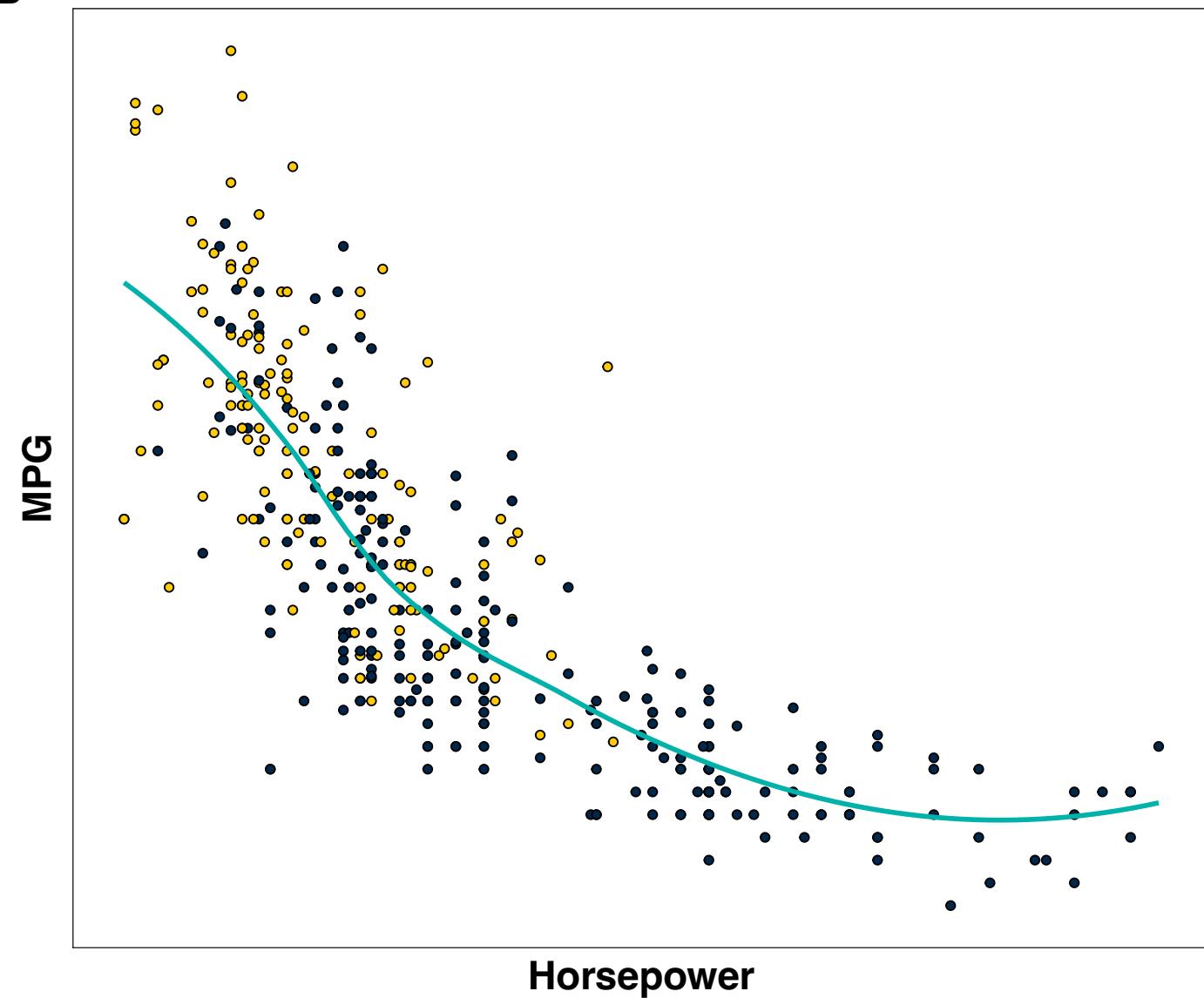
Pairwise scatterplots of features from the Auto MPG dataset of mechanical attributes of n = 390 cars.

Attributes: fuel consumption in miles per gallon (MPG), displacement, horsepower, weight, and acceleration, stratified by origin of car.

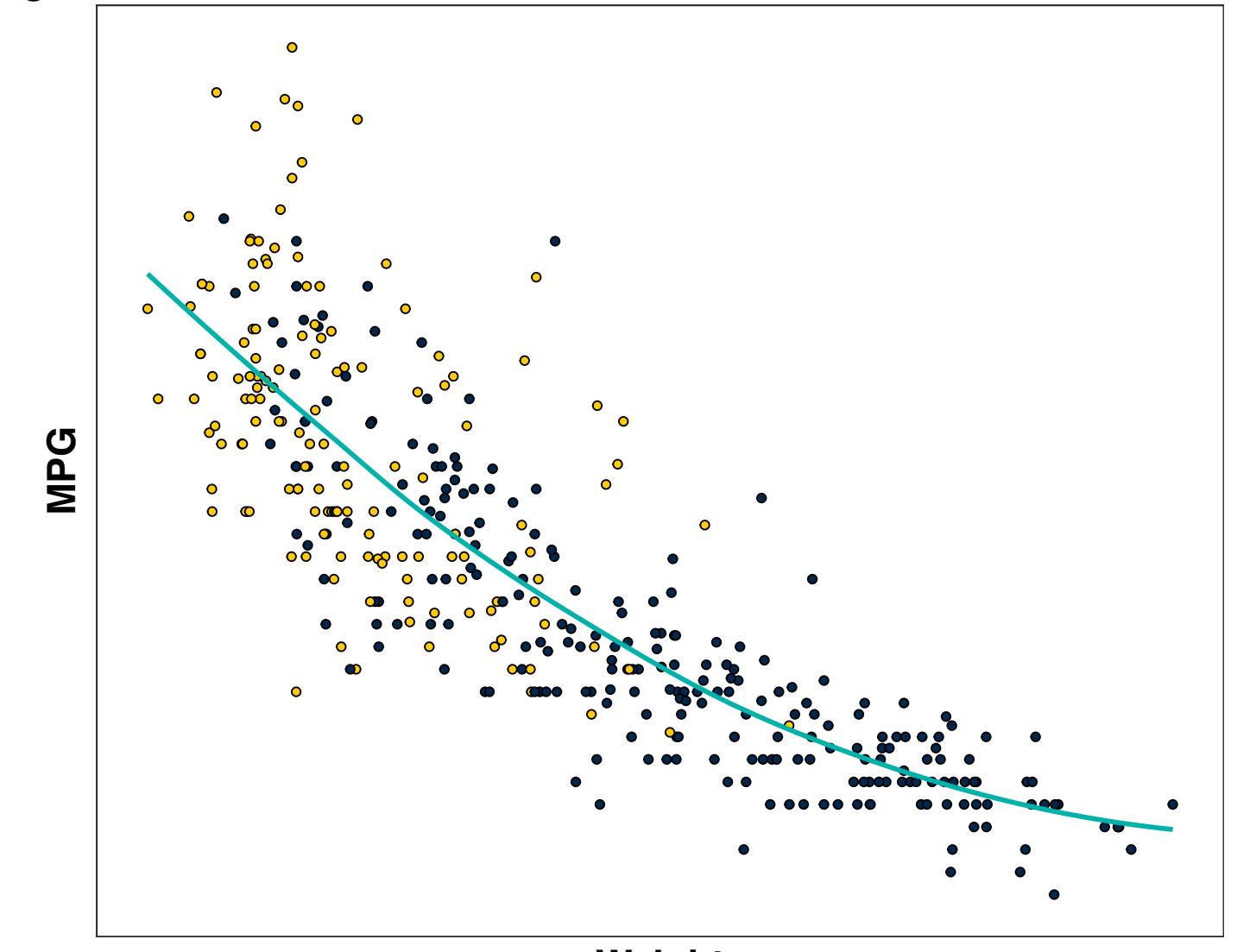
A



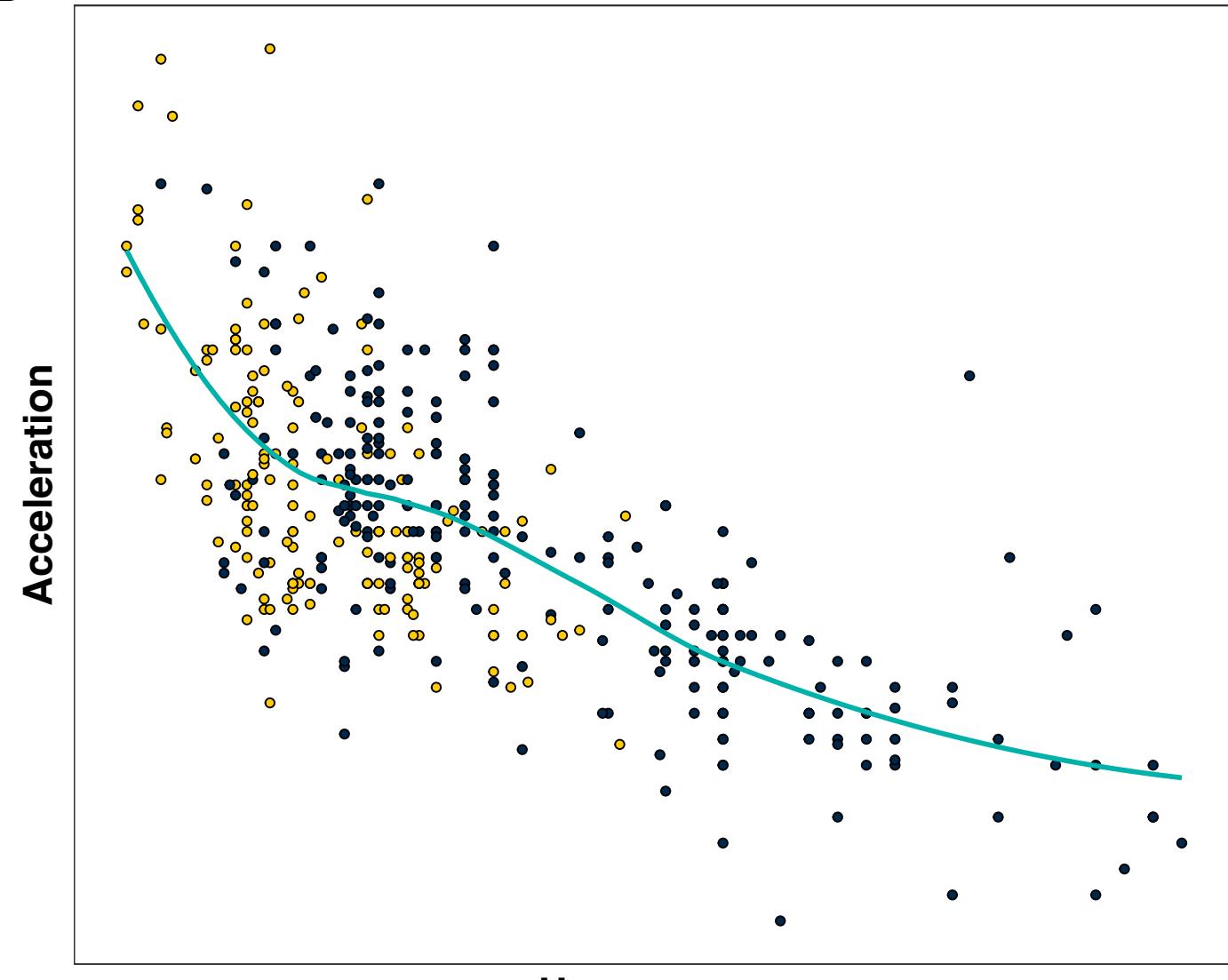
B



C



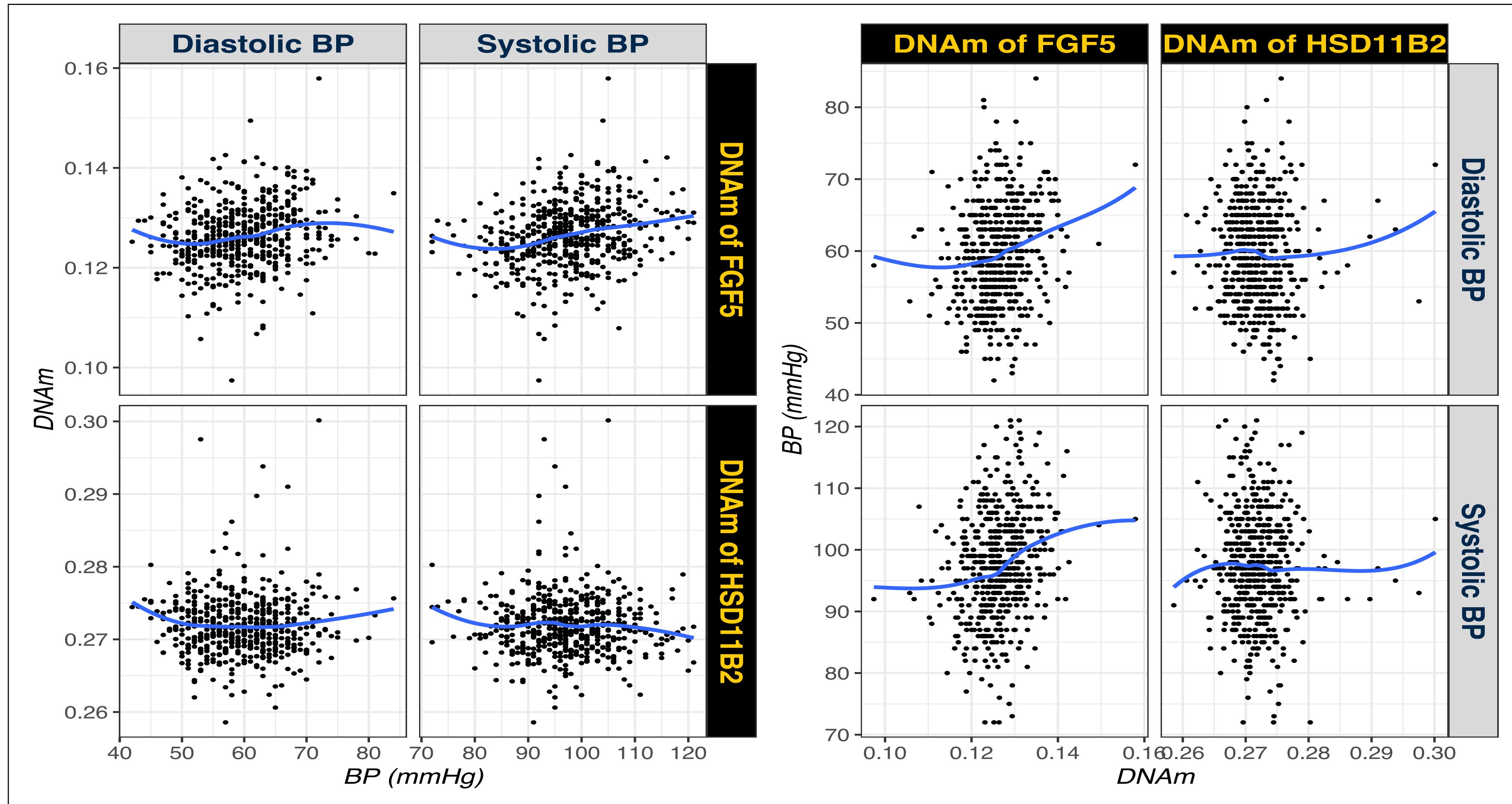
D



Origin • North America • Rest of the world

Appendix V

Epigenetics and blood pressure Association and direction?



Appendix VI

Fourier transform-based copula density estimation

(1) Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

(2) O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., & O'Brien, J. P. (2016). A fast and objective multidimensional kernel density estimation method: *fastKDE*. *Computational Statistics & Data Analysis*, 101, 148-160.

$\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$. Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

Fourier transform-based copula density estimation

(1) Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

(2) O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., & O'Brien, J. P. (2016). A fast and objective multidimensional kernel density estimation method: fastKDE. *Computational Statistics & Data Analysis*, 101, 148-160.

$\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$. Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

Fourier transform $\hat{\phi}_{\mathbf{Z}}$ of $\hat{c}_{\mathbf{Z}}$ depends on empirical characteristic function.

$$\hat{\mathcal{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^n \exp(it' \mathbf{Z}_j)$$

Fourier transform-based copula density estimation

(1) Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3), 407-422.

(2) O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., & O'Brien, J. P. (2016). A fast and objective multidimensional kernel density estimation method: fastKDE. *Computational Statistics & Data Analysis*, 101, 148-160.

$\mathbf{Z} = (X, Y)'$ with copula density $c_{\mathbf{Z}}$. Using $\mathbf{Z}_1, \dots, \mathbf{Z}_n \stackrel{\text{i.i.d}}{\sim} c_{\mathbf{Z}}$, obtain $\hat{c}_{\mathbf{Z}}$

Fourier transform $\hat{\phi}_{\mathbf{Z}}$ of $\hat{c}_{\mathbf{Z}}$ depends on empirical characteristic function.

$$\hat{\mathcal{C}}(\mathbf{t}) := n^{-1} \sum_{j=1}^n \exp(i\mathbf{t}'\mathbf{Z}_j)$$

$$\text{fastMI} = n^{-1} \sum_{j=1}^n \log \left\{ \hat{c}_{\mathbf{Z}}(\mathbf{z}_i) \right\}$$