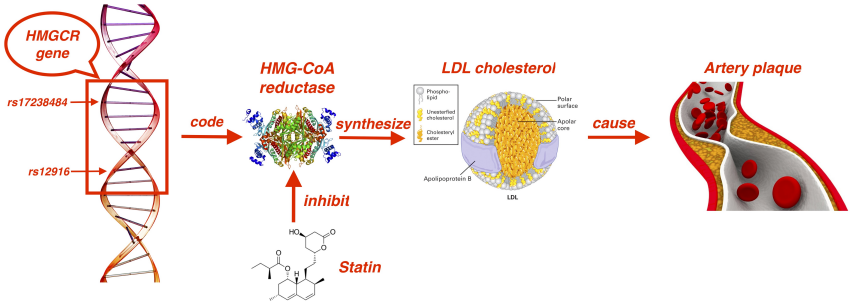
Using Mendelian randomization to discover biological mechanisms

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

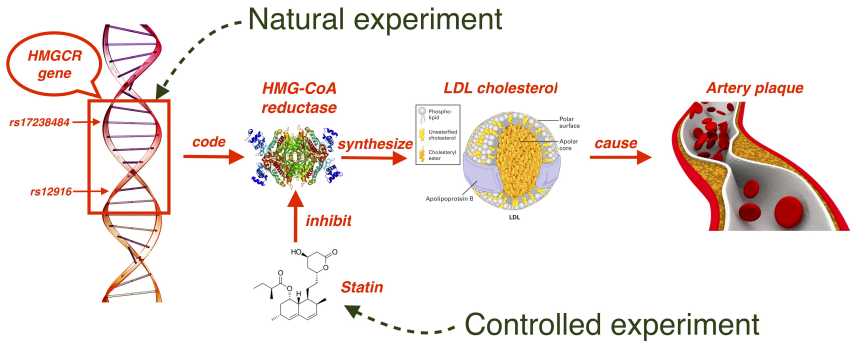
August 10, 2023 @ JSM

Example: Causal effect of the LDL-cholesterol



Basic idea: People who inherited certain alleles of *rs17238484* and *rs12916* have **naturally** higher concentration of LDL cholesterol.

Example: Causal effect of the LDL-cholesterol



Basic idea: People who inherited certain alleles of *rs17238484* and *rs12916* have **naturally** higher concentration of LDL cholesterol.

Different perspectives

- Statistical: much of the previous methodological work focuses on violations of **exclusion restriction** due to **pleiotropy**.
- Biological: the fundamental assumption of MR is **gene-environment equivalence**.
- Drug development: the most useful is to discover relevant pathways.

This talk

How can we use MR to discover biological mechanisms?

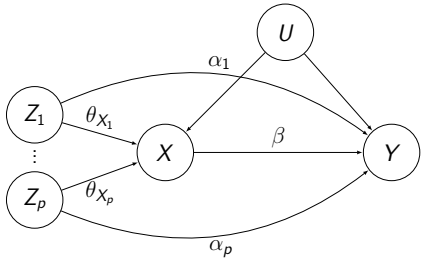
MR-PATH Daniel Iong, Qingyuan Zhao, and Yang Chen. "A Latent Mixture Model for Heterogeneous Causal Mechanisms in Mendelian Randomization". In: (2020). arXiv: 2007.06476 [stat.AP].

PathGPS Zijun Gao, Trevor Hastie, and Qingyuan Zhao. *PathGPS: Discover shared genetic architecture using biobank data*. 2022. bioRxiv: 10.1101/2022.05.01.490230.

Outline

Outline

Review: Linear structural equation model for MR



For exposure X, outcome Y, unobserved confounding variables U, and SNPs $Z_1, \ldots, Z_p$, the commonly assumed linear structural equation model is given by

$$X = \sum_{i=1}^{p} \theta_{X_i} Z_i + \eta_X U + E_X, \quad Y = \beta X + \sum_{i=1}^{p} \alpha_i Z_i + \eta_Y U + E_Y.$$

Review: Linear structural equation model for MR

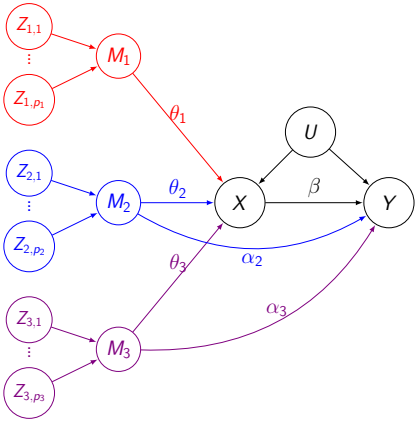
$$X = \sum_{i=1}^{p} \theta_{X_i} Z_i + \eta_X U + E_X,$$

$$Y = \beta X + \sum_{i=1}^{p} \alpha_i Z_i + \eta_Y U + E_Y.$$

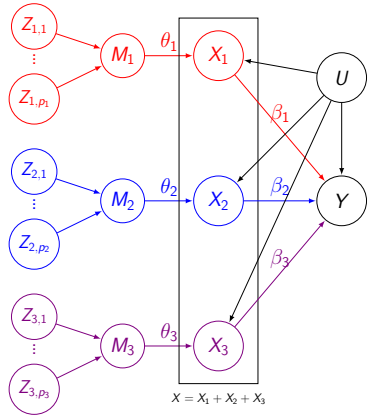
$$\Rightarrow Y = \sum_{i=1}^{p} (\theta_{X_i}\beta + \alpha_i) Z_i + (\beta\eta_X + \eta_Y) U + E_Y.$$

- If Z_i is a valid instrument, $\theta_{X_i} \neq 0$, $Z_i \perp\!\!\!\perp \{U, E_X, E_Y\}$, and $\alpha_i = 0$.
- However, often $\alpha_i \neq 0$ due to pleiotropy and mechanistic heterogeneity.
- If $\alpha_i \neq 0$ for some SNPs, then the causal effect β cannot be estimated consistently without further assumptions on α_i.
  - e.g. $\alpha_i \sim N(0, \tau^2)$ for most SNPs.

Mechanistic heterogeneity: Two scenarios



(a) Scenario 1: Multiple pathways of correlated pleiotropy.



(b) Scenario 2: Multiple mechanisms for the exposure X.

Mechanistic heterogeneity: Two scenarios

If we interpret the diagrams in the previous slide as linear structural equations as before, we can derive the Wald estimands for each pathway.

Instruments Z	**Pathway** M	**Effect of** M **on** X	**Effect of** M **on** Y	**Wald estimand**
		Scenario 1		
$Z_{1,1}, \ldots, Z_{1,p_1}$	M_1	θ_1	$\theta_1\beta$	β
$Z_{2,1}, \ldots, Z_{2,p_2}$	M_2	θ_2	$\theta_2\beta + \alpha_2$	$\beta + \alpha_2/\theta_2$
$Z_{3,1}, \ldots, Z_{3,p_3}$	M_3	θ_3	$\theta_3\beta + \alpha_3$	$\beta + \alpha_3/\theta_3$
		Scenario 2		
$Z_{1,1}, \ldots, Z_{1,p_1}$	M_1	θ_1	$\theta_1\beta_1$	β_1
$Z_{2,1}, \ldots, Z_{2,p_2}$	M_2	θ_2	$\theta_2\beta_2$	β_2
$Z_{3,1}, \ldots, Z_{3,p_3}$	M_3	θ_3	$\theta_3\beta_3$	β_3

- SNPs on the same pathway have the same Wald estimand, while SNPs across different pathways generally have different estimands.
- Mechanistic heterogeneity can arise even when all SNPs are valid instruments (Scenario 2).

Nonparametric perspective: Mechanism-specific causal effect

The same clustering phenomenon also occurs in nonlinear models.

- It is well known that assuming **monotonicity**, IV nonparametrically estimates the **complier average treatment effect** (Angrist, Imbens, Rubin, *JASA*, 1996).
- By assuming **monotonicity** and Pearl's nonparametric structural equation model with independent errors (**NPSEM-IE**), our paper showed that (if X, Z, M are all binary variables)

$$\begin{aligned}&\mathbb{E}[Y(X=1)-Y(X=0)\mid X(Z_{kj}=1)>X(Z_{kj}=0)]\\ =&\mathbb{E}[Y(X=1)-Y(X=0)\mid X(M_k=1)>X(M_k=0)],\end{aligned}$$

where k indexes the mechanism and j indexes the gene within.

Outline

MR-PATH: Model Assumptions

We make the following assumptions on GWAS summary data.

Assumption (Error-in-variables regression)

The observed (marginal) SNP-exposure and SNP-outcome associations are distributed as

$$\begin{pmatrix} \hat{\theta}_{X_i} \\ \hat{\theta}_{Y_i} \end{pmatrix} \overset{indep.}{\sim} N\Big(\begin{pmatrix} \theta_{X_i} \\ \beta_i \theta_{X_i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{X_i} & 0 \\ 0 & \sigma^2_{Y_i} \end{pmatrix} \Big), \quad i = 1, \ldots, p,$$

where σ_{X_i}, σ_{Y_i} are (fixed) measurement errors.

Assumption (Mixture model for mechanistic heterogeneity)

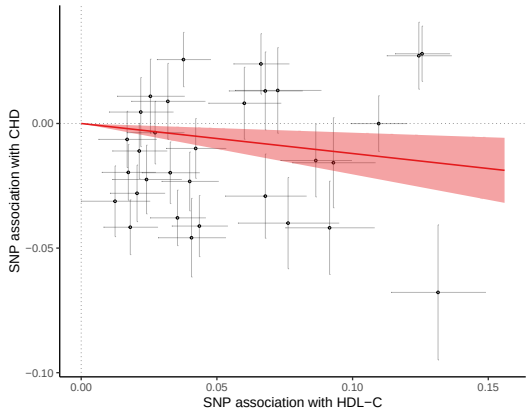
$$\begin{aligned} Z_i &\sim \mathit{Multinomial}(\pi_1, \ldots, \pi_K), \\ \beta_i | Z_i = k &\sim N(\mu_k, \sigma^2_k), \quad k = 1, \ldots, K. \end{aligned}$$

MR-PATH: Statistical Inference

1. **Monte-Carlo EM algorithm** for obtaining model parameter estimates
2. **Approximate confidence intervals** for quantifying uncertainty of the estimates
3. **Modified Bayesian Information criterion (BIC)** for selecting number of clusters

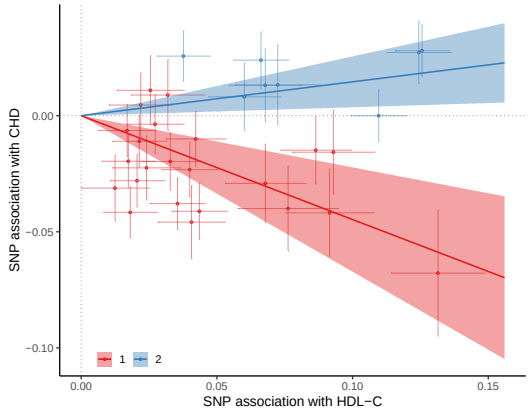
- We perform simulation studies to verify the efficacy of these inference procedures.
- See paper for implementation details.

Example: HDL-CHD



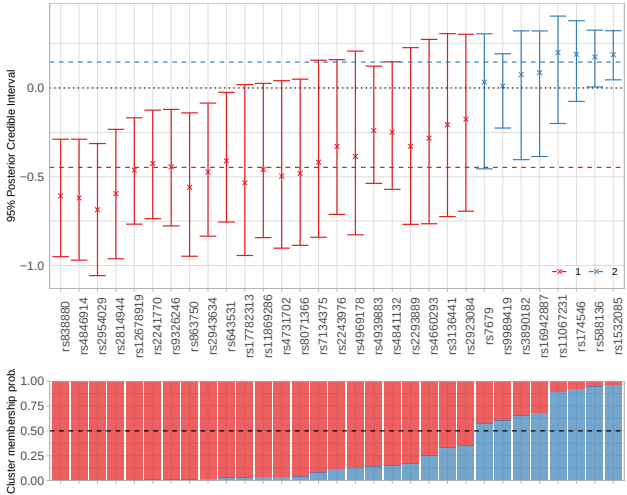
Results of MR-RAPS.

Example: HDL-CHD

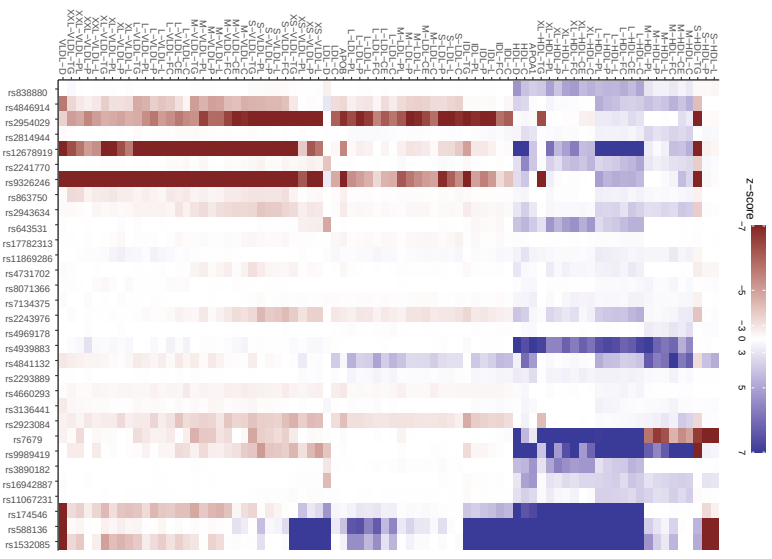


Results of MR-PATH (`http://danieliong.me/mr-path/`.)

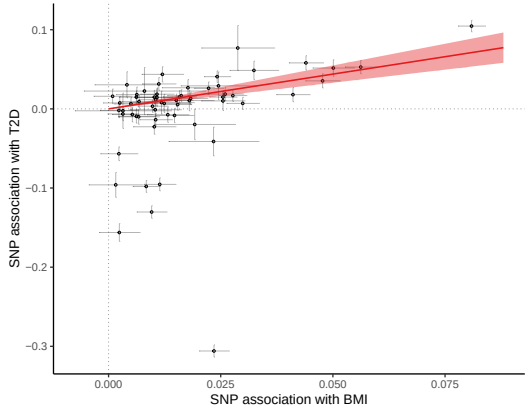
Example: HDL-CHD (individual β_i)



Example: HDL-CHD (heatmap against metabolites)

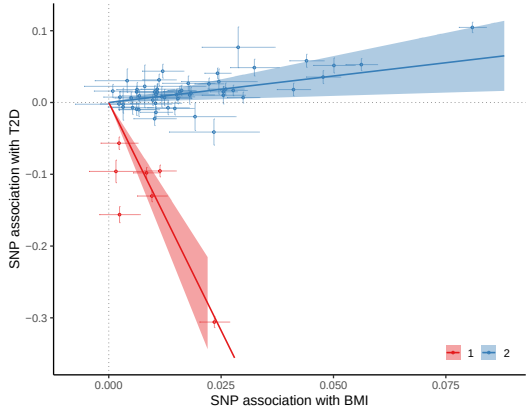


Example: BMI-T2D



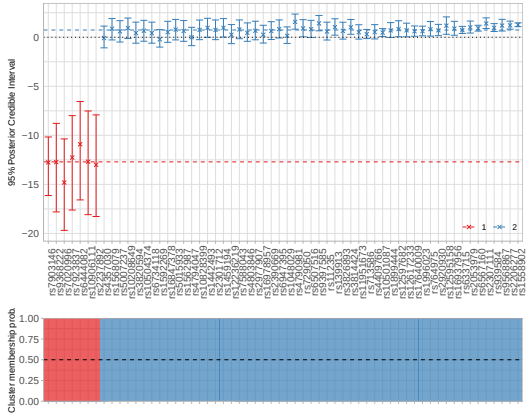
Results of MR-RAPS.

Example: BMI-T2D

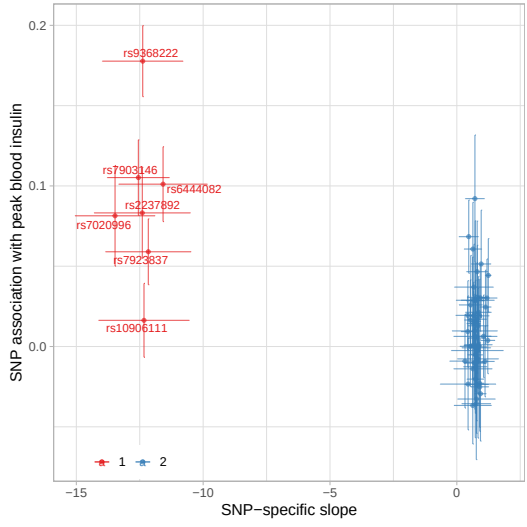


Results of MR-PATH.

Example: BMI-T2D (individual β_i)



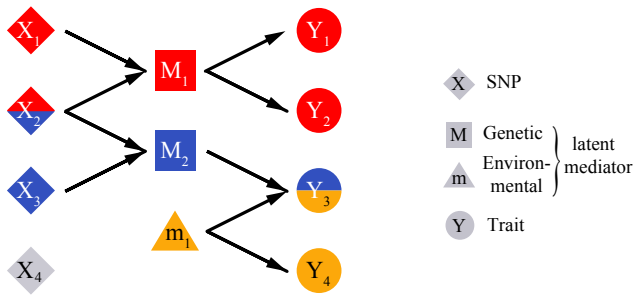
Example: BMI-T2D (possible correlated pleiotropy: insulin)



Outline

PathGPS: PATHway discovery through Genome-Phenome Summary data

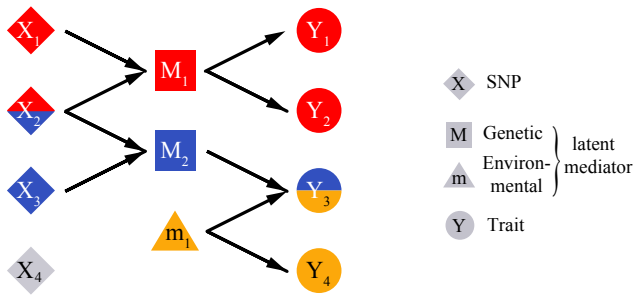
Goal: Discover SNP-trait clusters corresponding to the same genetic pathway.



Genetic and environmental pathways.

PathGPS: PATHway discovery through Genome-Phenome Summary data

Goal: Discover SNP-trait clusters corresponding to the same genetic pathway.

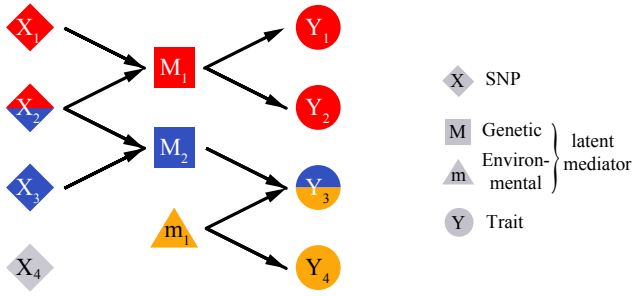


Two SNP-trait clusters: (1) X_1, X_2, Y_1, Y_2; (2) X_2, X_3, Y_3.

Statistical model: Linear SEM

- Observed data: $\boldsymbol{X}$ genotypes/SNPs ($n \times p$), $\boldsymbol{Y}$ phenotypes/traits ($n \times q$);
- Latent mediators/confounders: $\boldsymbol{M}$ genetic ($n \times r$), $\boldsymbol{m}$ environmental ($n \times s$).

$$\underbrace{\boldsymbol{M}}_{n\times r} = \underbrace{\boldsymbol{X}}_{n\times p}\underbrace{\boldsymbol{U}}_{p\times r} + \underbrace{\varepsilon_M}_{n\times r}, \quad \underbrace{\boldsymbol{Y}}_{n\times q} = \underbrace{\boldsymbol{M}}_{n\times r}\underbrace{\boldsymbol{V}^\top}_{r\times q} + \underbrace{\boldsymbol{m}}_{n\times s}\underbrace{\boldsymbol{W}^\top}_{s\times q} + \underbrace{\varepsilon_Y}_{n\times q}.$$



Statistical model: Linear SEM

- Observed data: $\boldsymbol{X}$ genotypes/SNPs ($n \times p$), $\boldsymbol{Y}$ phenotypes/traits ($n \times q$);
- Latent mediators/confounders: $\boldsymbol{M}$ genetic ($n \times r$), $\boldsymbol{m}$ environmental ($n \times s$).

$$\underbrace{\boldsymbol{M}}_{n\times r} = \underbrace{\boldsymbol{X}}_{n\times p}\underbrace{\boldsymbol{U}}_{p\times r} + \underbrace{\varepsilon_M}_{n\times r}, \quad \underbrace{\boldsymbol{Y}}_{n\times q} = \underbrace{\boldsymbol{M}}_{n\times r}\underbrace{\boldsymbol{V}^\top}_{r\times q} + \underbrace{\boldsymbol{m}}_{n\times s}\underbrace{\boldsymbol{W}^\top}_{s\times q} + \underbrace{\varepsilon_Y}_{n\times q}.$$

$$\Longrightarrow \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{m}\boldsymbol{W}^\top + \varepsilon, \quad \varepsilon = \varepsilon_M \boldsymbol{V}^\top + \varepsilon_Y.$$

- Non-zero elements in $\boldsymbol{U}_{\cdot k}$, $\boldsymbol{V}_{\cdot k} \rightarrow k$-th SNP-trait cluster.

Statistical model: Linear SEM

- Observed data: $\boldsymbol{X}$ genotypes/SNPs ($n \times p$), $\boldsymbol{Y}$ phenotypes/traits ($n \times q$);
- Latent mediators/confounders: $\boldsymbol{M}$ genetic ($n \times r$), $\boldsymbol{m}$ environmental ($n \times s$).

$$\underbrace{\boldsymbol{M}}_{n\times r} = \underbrace{\boldsymbol{X}}_{n\times p}\underbrace{\boldsymbol{U}}_{p\times r} + \underbrace{\varepsilon_M}_{n\times r}, \quad \underbrace{\boldsymbol{Y}}_{n\times q} = \underbrace{\boldsymbol{M}}_{n\times r}\underbrace{\boldsymbol{V}^\top}_{r\times q} + \underbrace{\boldsymbol{m}}_{n\times s}\underbrace{\boldsymbol{W}^\top}_{s\times q} + \underbrace{\varepsilon_Y}_{n\times q}.$$

$$\Longrightarrow \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{m}\boldsymbol{W}^\top + \varepsilon, \quad \varepsilon = \varepsilon_M \boldsymbol{V}^\top + \varepsilon_Y.$$

- Non-zero elements in $\boldsymbol{U}_{\cdot k}$, $\boldsymbol{V}_{\cdot k} \to k$-th SNP-trait cluster.

Assumptions

- Low-rank $\boldsymbol{U}\boldsymbol{V}^\top$: #genetic mediators $\ll$ #SNPs, #traits.
- Sparse $\boldsymbol{U}$, $\boldsymbol{V}$: a few influential SNPs/influenced traits.

Statistical method

Singular value decomposition (SVD) for $\hat{\boldsymbol{\beta}}$.

- Suppose $\mathrm{Var}(\boldsymbol{m}) = \sigma_m^2 \boldsymbol{I}_s$, $\mathrm{Var}(\boldsymbol{\varepsilon}_M) = \sigma_M^2 \boldsymbol{I}_r$, $\mathrm{Var}(\boldsymbol{\varepsilon}_Y) = \sigma_Y^2 \boldsymbol{I}_q$,

$$\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}} \approx \underbrace{\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{V}^\top}_{\text{gene}} + \frac{p}{n}\left(\underbrace{\sigma_m^2 \boldsymbol{W}\boldsymbol{W}^\top}_{\text{environment}} + \sigma_M^2 \boldsymbol{V}\boldsymbol{V}^\top + \sigma_Y^2 \boldsymbol{I}_q \right).$$

- Right singular vectors: genetic effect ($\boldsymbol{V}$) & environmental effect ($\boldsymbol{W}$).

Statistical method

Singular value decomposition (SVD) for $\hat{\boldsymbol{\beta}}$.

- Suppose $\mathrm{Var}(\boldsymbol{m}) = \sigma_m^2 \boldsymbol{I}_s$, $\mathrm{Var}(\boldsymbol{\varepsilon}_M) = \sigma_M^2 \boldsymbol{I}_r$, $\mathrm{Var}(\boldsymbol{\varepsilon}_Y) = \sigma_Y^2 \boldsymbol{I}_q$,

$$\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}} \approx \underbrace{\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{V}^\top}_{\text{gene}} + \frac{p}{n}\left(\underbrace{\sigma_m^2 \boldsymbol{W}\boldsymbol{W}^\top}_{\text{environment}} + \sigma_M^2 \boldsymbol{V}\boldsymbol{V}^\top + \sigma_Y^2 \boldsymbol{I}_q \right).$$

- Right singular vectors: genetic effect ($\boldsymbol{V}$) & environmental effect ($\boldsymbol{W}$).

Challenge 1: Genetics vs. Environment

Solution: Use negative control SNPs ($\boldsymbol{U}_{\text{nc}} = 0$) and apply SVD to the contrast.

$$\hat{\boldsymbol{\beta}}_{\text{nc}}^\top \hat{\boldsymbol{\beta}}_{\text{nc}} \approx \underbrace{\boldsymbol{0}}_{\text{gene}} + \frac{p_{\text{nc}}}{n}\left(\underbrace{\sigma_m^2 \boldsymbol{W}\boldsymbol{W}^\top}_{\text{environment}} + \sigma_M^2 \boldsymbol{V}\boldsymbol{V}^\top + \sigma_Y^2 \boldsymbol{I}_q \right)$$

Statistical method

- Next step: Cluster SNPs and traits by non-zero entries of $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{U}} = \hat{\boldsymbol{\beta}}\hat{\boldsymbol{V}}^{\top}$.

Challenge 2: Instability of unsupervised learning

- Many hyper-parameters: number of principal components and clusters, rotation and thresholding, clustering algorithm.

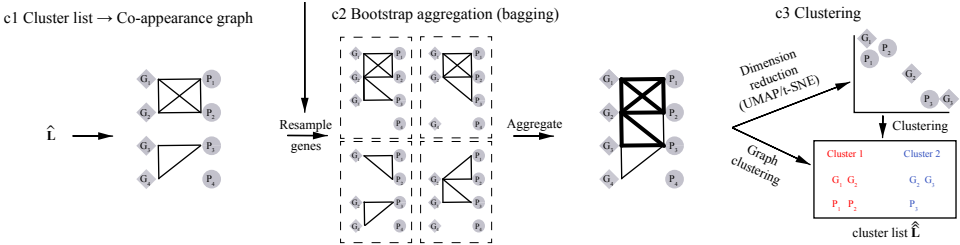
Statistical method

- Next step: Cluster SNPs and traits by non-zero entries of $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{U}} = \hat{\boldsymbol{\beta}}\hat{\boldsymbol{V}}^{\top}$.

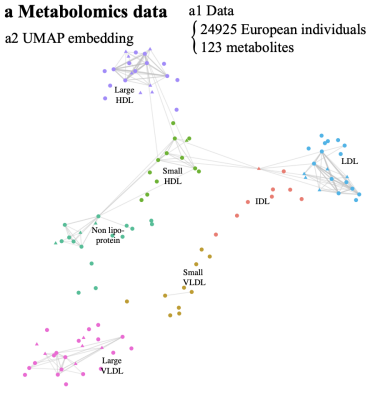
Challenge 2: Instability of unsupervised learning

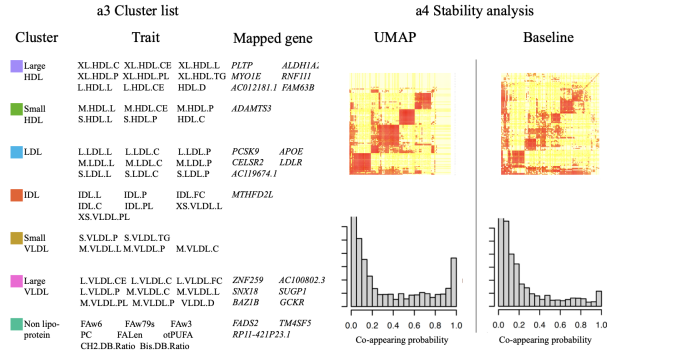
- Many hyper-parameters: number of principal components and clusters, rotation and thresholding, clustering algorithm.

Solution: Bootstrap aggregating (bagging)



Example: Metabolomics dataset





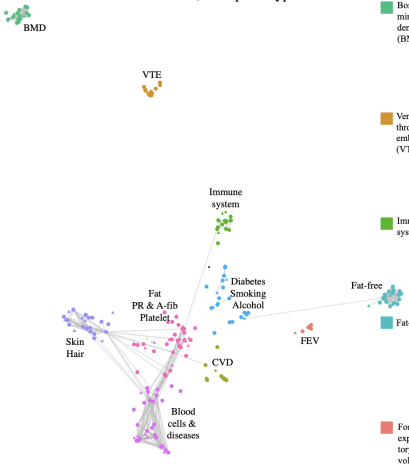
Example: UK Biobank

b UK BioBank data

b1 Data

{ 361194 British individuals
{ 4359 phenotypes

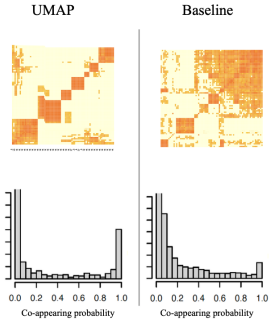
b2 UMAP embedding



b3 Cluster list

Cluster	Trait	Mapped gene
Bone mineral density (BMD)	Heel bone BMD Heel BMD T-score Heel quantitative ultrasound index Heel Broadband ultrasound attenuation Medication: atorvastatin	*AC000093.3* *WNT16* *RP11-389-E17.1* *RP11-556-E13.1*
Venous thrombo-embolism (VTE)	Pulmonary embolism Deep vein thrombosis Phlebitis and thrombophlebitis Blood clot Hypotension Medication: warfarin	*NME7*
Immune system	Leukocyte count Monocyte count Neutrophill count Lymphocyte count Asthma Diabetes Medication: metformin	*CYP8B1* *GFI1* *LYZ* *ITGA4* *PAN3* *FLT3*
Fat-free	Whole body/trunk/ leg/arm fat-free mass Whole body water mass Trunk/leg/arm predicted mass Standing/sitting height Basal metabolic rate	*ID4* *SOX6* *Metazoa_SRP* *LEPREL1*
Forced expira-tory volume (FEV)	Forced expiratory volume in 1 sec. Forced vital capacity Peak expiratory flow	*RP11-257-A22.1*
Cardio-vascular disease (CVD)	Major coronary heart disease event Myocardial infarction Ischaemic heart disease High cholesterol Hypertension Smoking	*SNORD112*
Skin Hair	Hair color Skin color Childhood sunburn occasions Ease of skin tanning Use of sun protection Malignant neoplasms of skin	*HERC2* *PIGU* *NRXN3* *RXFP3* *SLC45A2* *ZDHHC17* *DBNDD1*
Blood cells & diseases	Red blood cell count/ distribution width Reticulocyte count Mean reticulocyte volume Diseases of the circulatory system Hereditary/genetic haematological disorder Clotting disorder Excessive bleeding	*PELO* *KANK2* *SLC17A3* *HBS1L* *CCND3* *ANK1* *HCG17* *AC006552.1* *RP11-764E7.1*
Fat Pause rate (PR) Atrial fibrill-ation (A-fib) Platelet	Fat percentage Systolic blood pressure Heart rate Pulse rate Atrial fibrillation and flutter Wheeze or whistling in the chest Medication: simvas-tatin, bezafibrate, fenofibrate, digoxin	*BCL2* *DNM3* *TUBB1* *NCKAP5* *RNPEPL1* *CTC-454M9.1* *FAM101A* *RP11-512M8.5* *TMEM120B* *CTD-2384B11.2* *RP11-347I19.3*
Diabetes Smoking Alcohol	Diabetes Diabetes related eye disease Medication: insulin product, pioglitazone Ever smoked Smoking status Past tobacco smoking Frequency of drinking alcohol Amount of alcohol drunk	U6 AC090825.2 RP11-665G4.1 RP11-720L2.4

b4 Stability analysis



Outline

Concluding remarks

- Genetic variation can be used as instrumental variables for biological mechanisms.
- MR-PATH: many SNPs, 2 traits (exposure and outcome).
- PathGPS: many SNPs, many traits.
- Biobanks provide a rich data source for MR and pathway discovery.
- Many open problems: unified methodology? theoretical guarantees? more sophisticated/realistic models? integration with drug discovery?