Discovering Mechanistic Heterogeneity using Mendelian Randomization

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

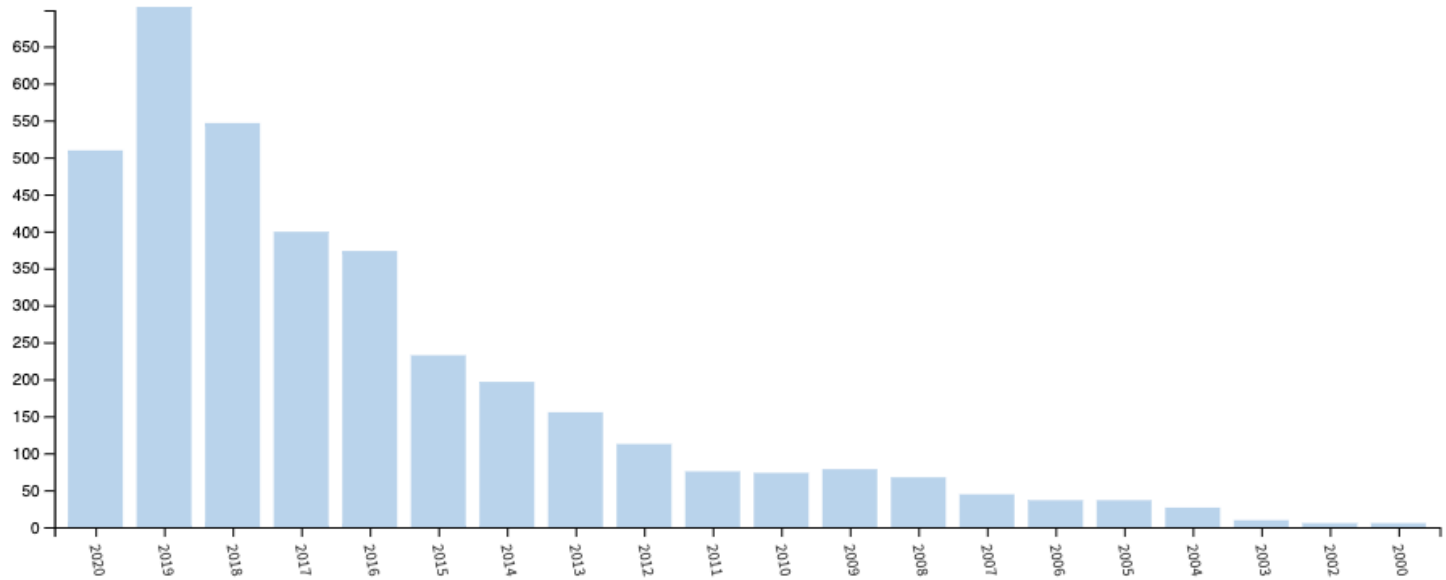
Joint work with Daniel Iong (who made most of the slides) and Yang Chen

September 26, 2020 @ PCIC

Outline

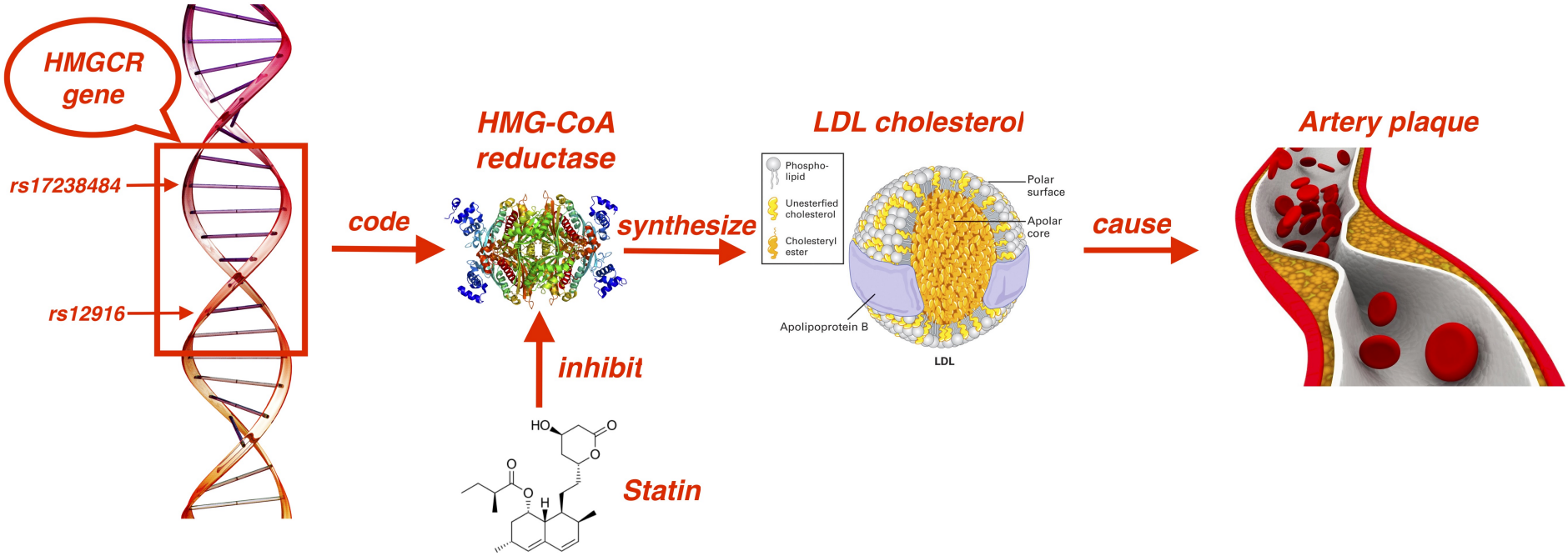
Mendelian randomization (MR)

- MR = Using genetic variation as instrumental varibles.
- Surging interest in epidemiology and genetics.



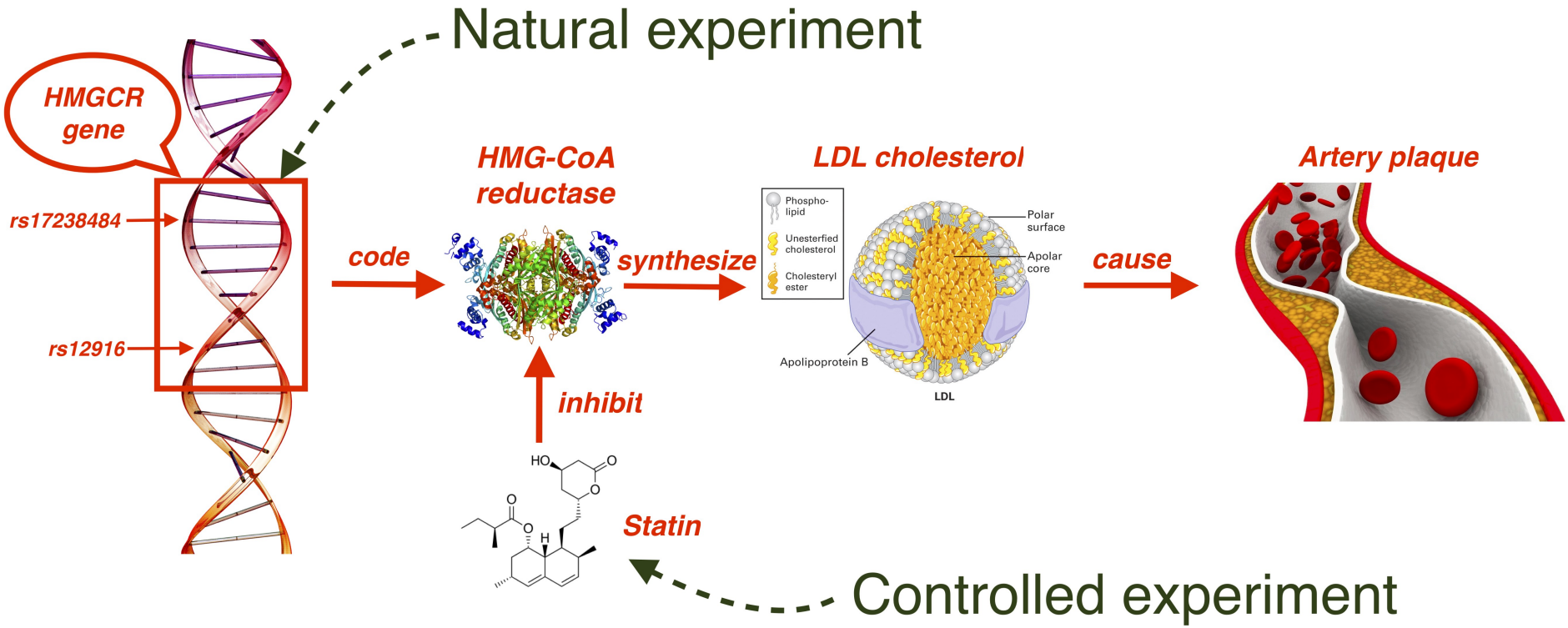
Number of publications in MR by year (Source: Web of Science).

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.

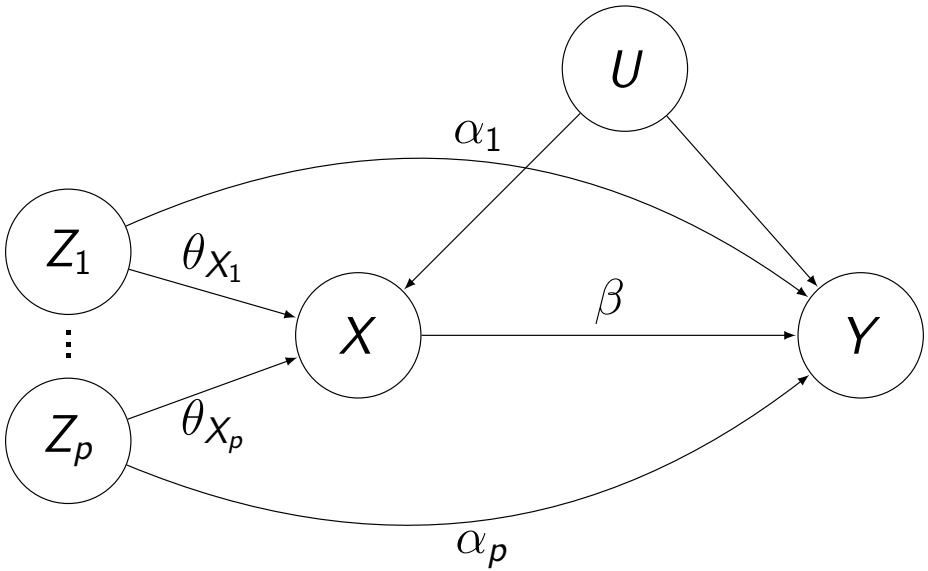
Motivation for this work

- **Exclusion restriction**: Instruments (genetic variants) can only affect the outcome through the risk exposure.
  - In MR, this assumption may be violated due to **pleiotropy**.
  - Many pleiotropy-robust MR methods (e.g. MR-RAPS) have been developed.
- Most robust MR methods rely on the **"effect homogeneity" assumption**: the risk exposure has the same causal effect for every individual.

Our contributions

1. A novel concept—**Mechanistic heterogeneity**.
2. A transparent mixture model—**MR-PATH**.

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

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

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

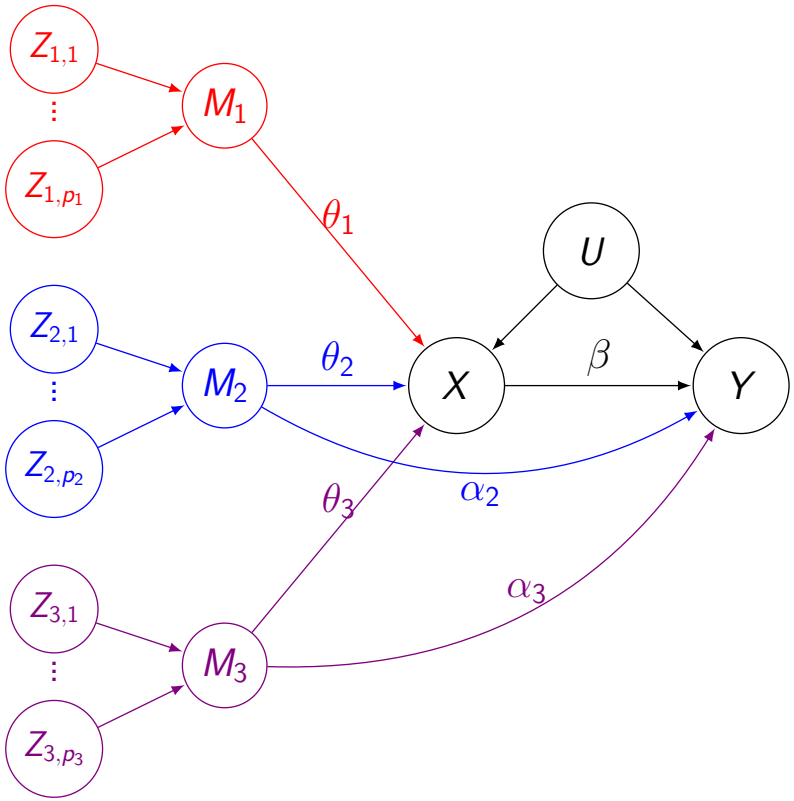
Review: Linear structural equation model for MR

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

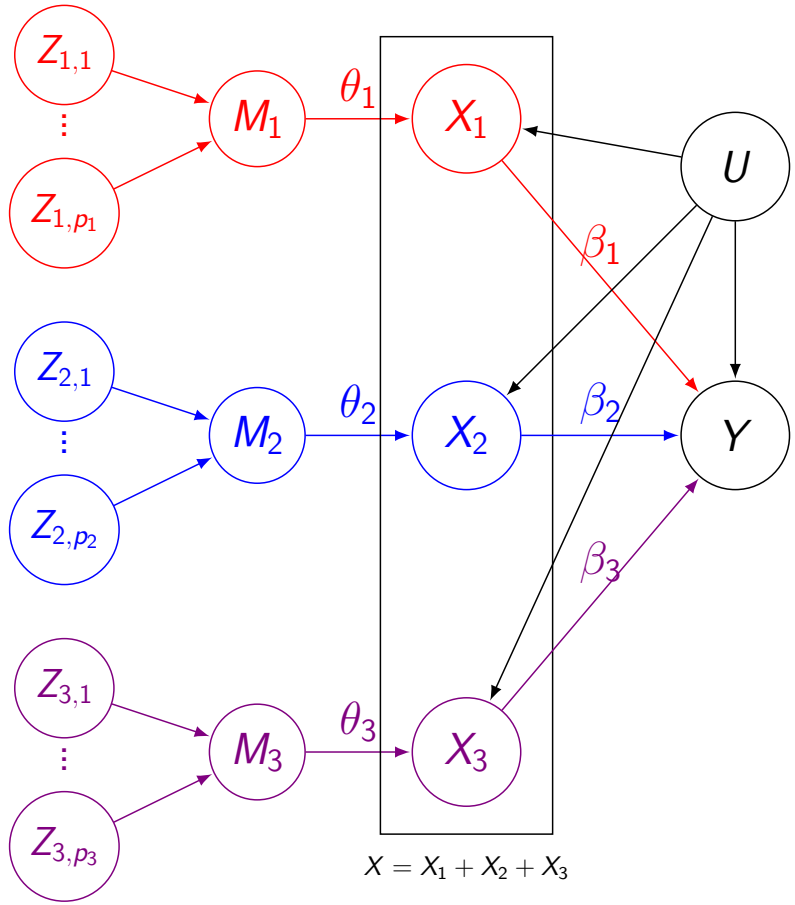
$$Y = \beta X + \sum_{i=1}^{p} \alpha_i Z_i + \eta_Y U + E_Y$$

- 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, it is often the case that $\alpha_i \neq 0$ due to pleiotropy and multiple causal pathways.
- 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.

Two scenarios of mechanistic heterogeneity



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



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

Two scenarios of mechanistic heterogeneity

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).

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 et al., *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.

MR-PATH: Model Assumptions

Assumption (Error-in-variables regression)

The observed SNP-exposure and SNP-outcome associations are distributed as

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

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

Assumption (Mixture model for mechanistic heterogeneity)

$$Z_i \sim Categorical\,(\pi_1, \ldots, \pi_K),$$
$$\beta_i | Z_i = k \sim N(\mu_k, \sigma^2_k), \quad k = 1, \ldots, K.$$

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

Data (Three-sample MR design)

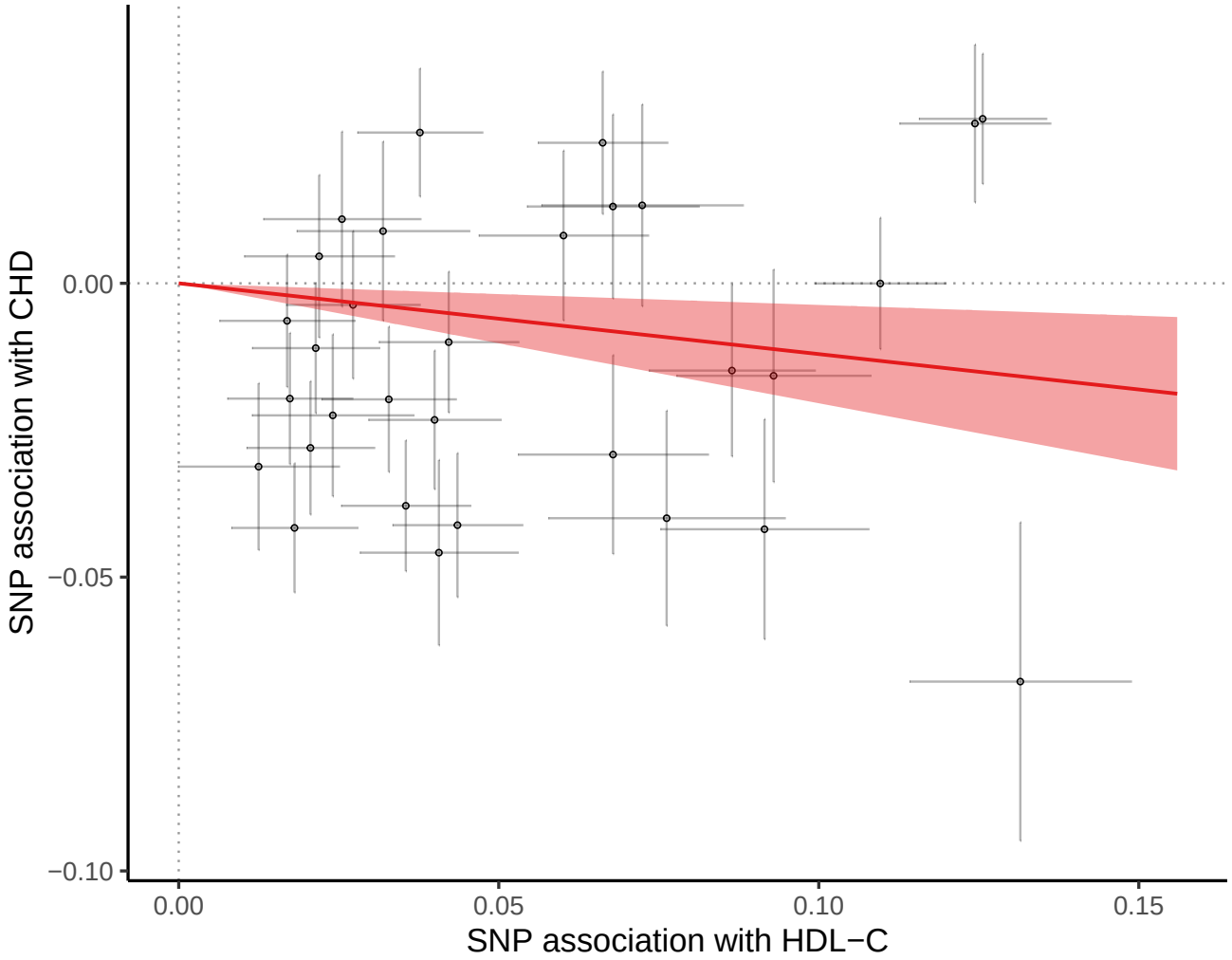
- **Selection dataset**: Teslovich et al. 2010[1]
- **Exposure dataset**: Kettunen et al. 2016[2]
- **Outcome dataset**: Nikpay et al. 2015[3]

[1] Tanya M Teslovich et al. "Biological, clinical and population relevance of 95 loci for blood lipids". In: *Nature* 466.7307 (2010), pp. 707–713.

[2] Johannes Kettunen et al. "Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA". In: *Nature communications* 7.1 (2016), pp. 1–9.

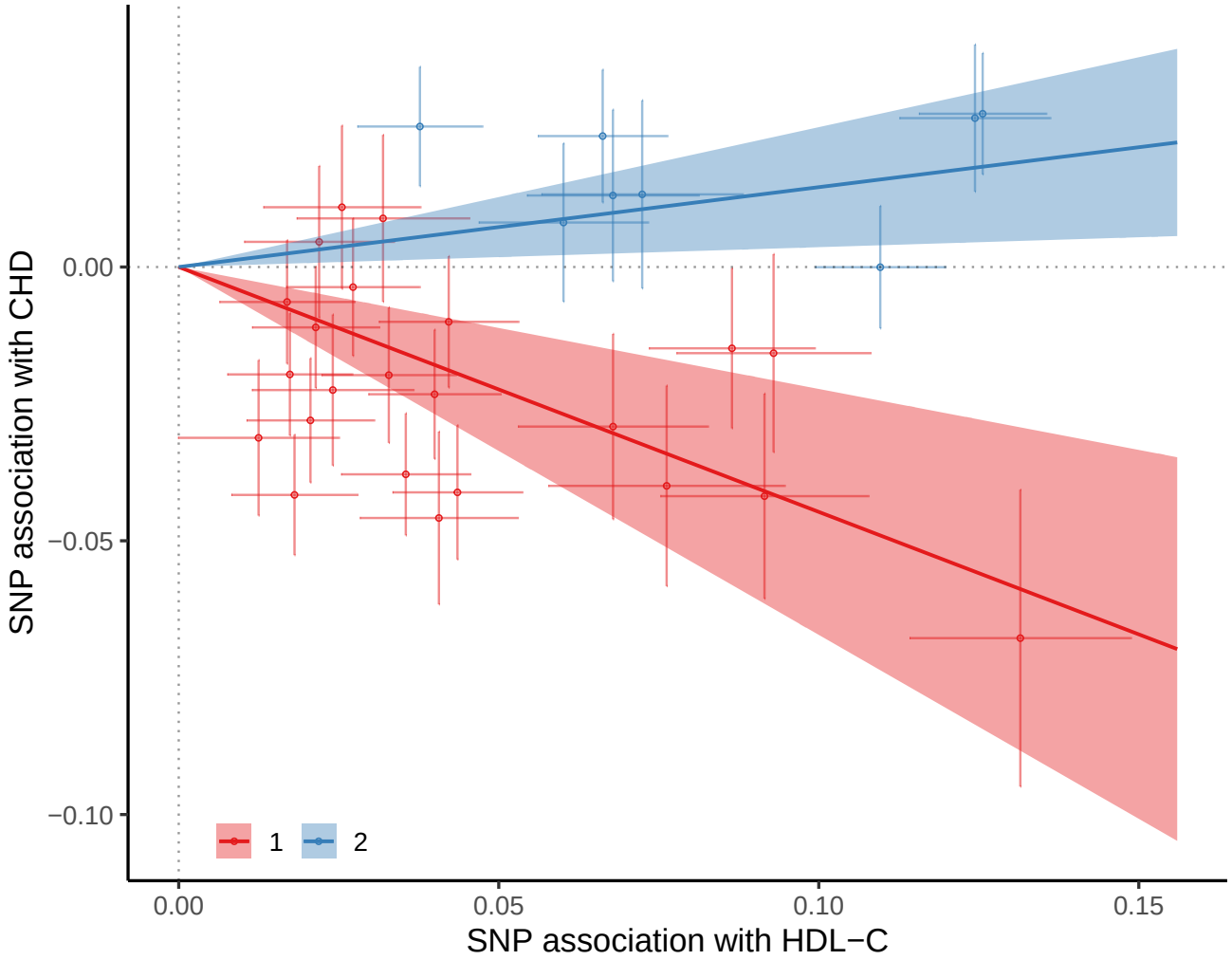
[3] Majid Nikpay et al. "A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease". In: *Nature Genetics* 47.10 (2015), p. 1121.

Example: HDL-CHD



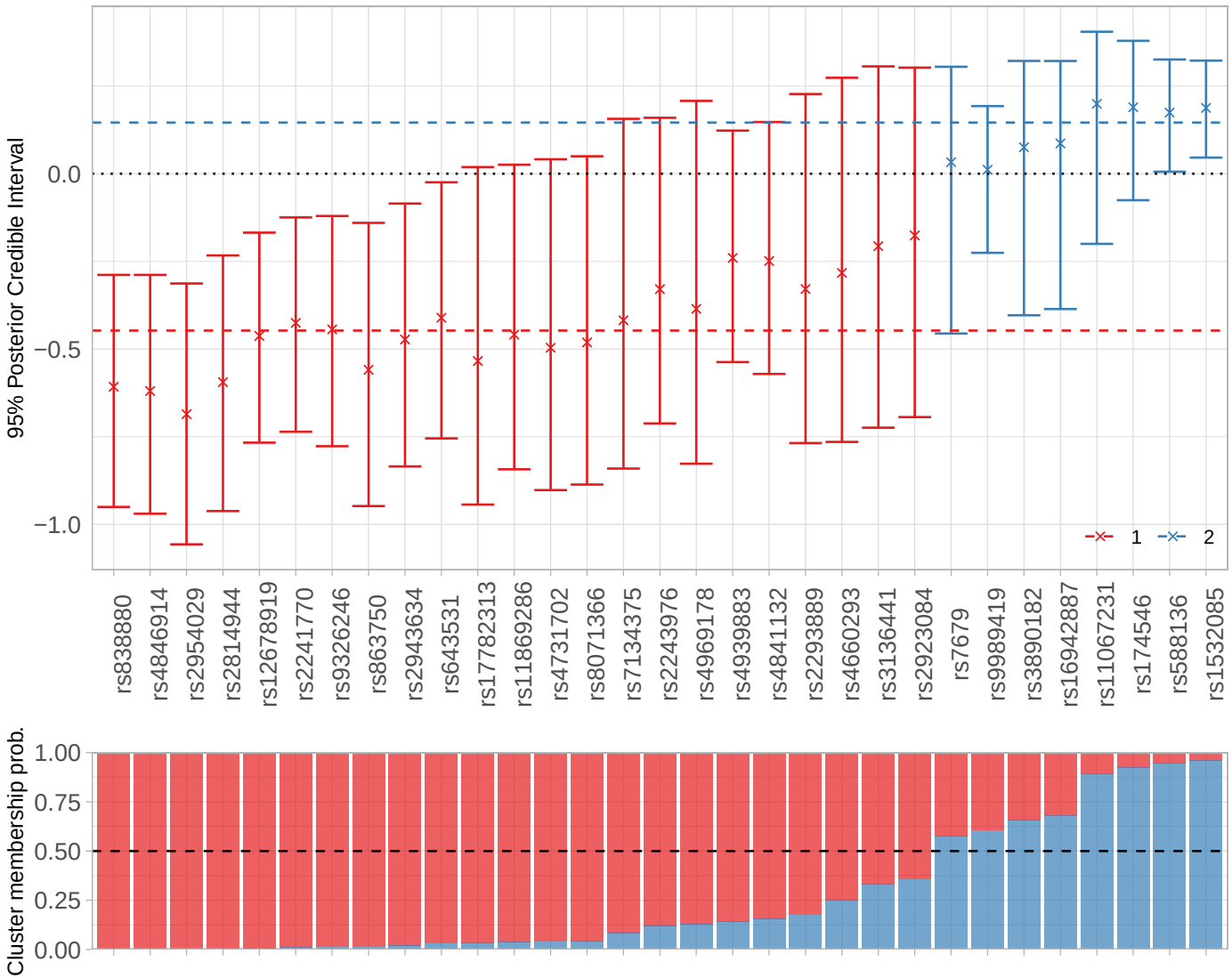
Results of MR-RAPS.

Example: HDL-CHD

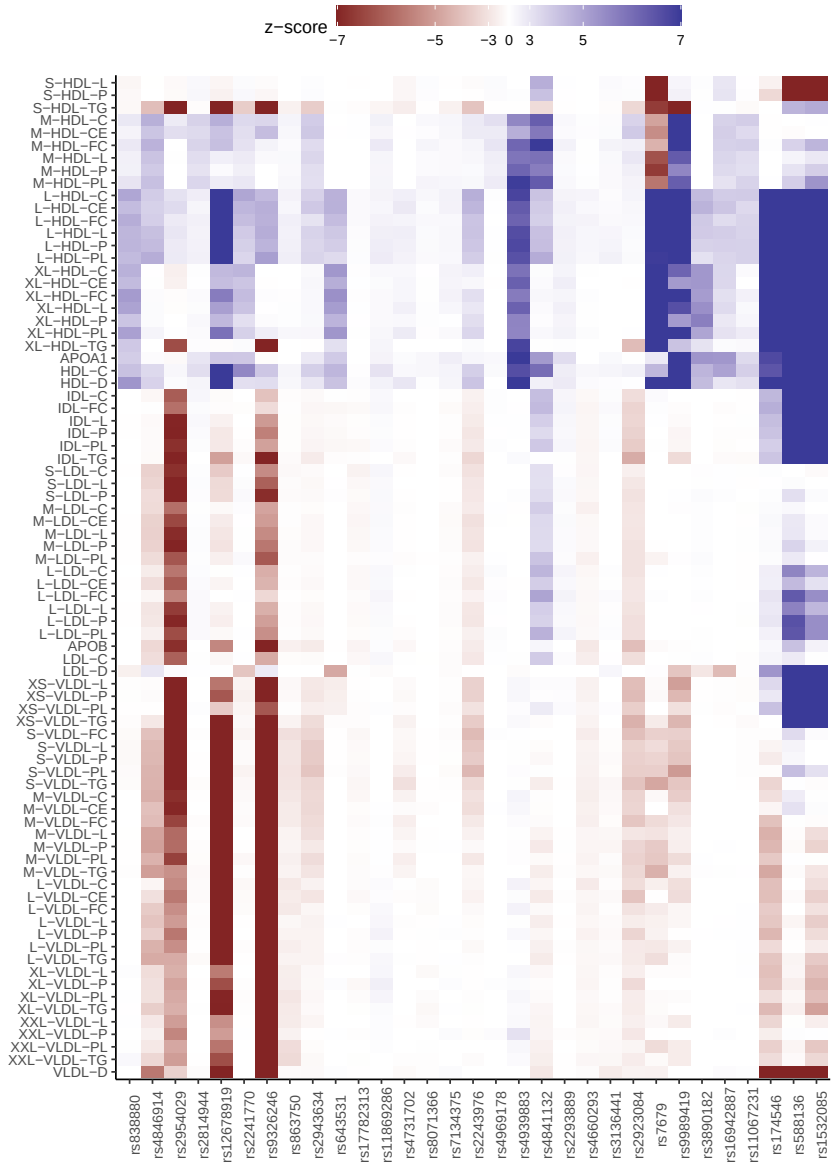


Results of MR-PATH (`http://danieliong.me/mr-path/`.)

Example: HDL-CHD



Example: HDL-CHD



Example: BMI-T2D

Data (Three-sample MR design)

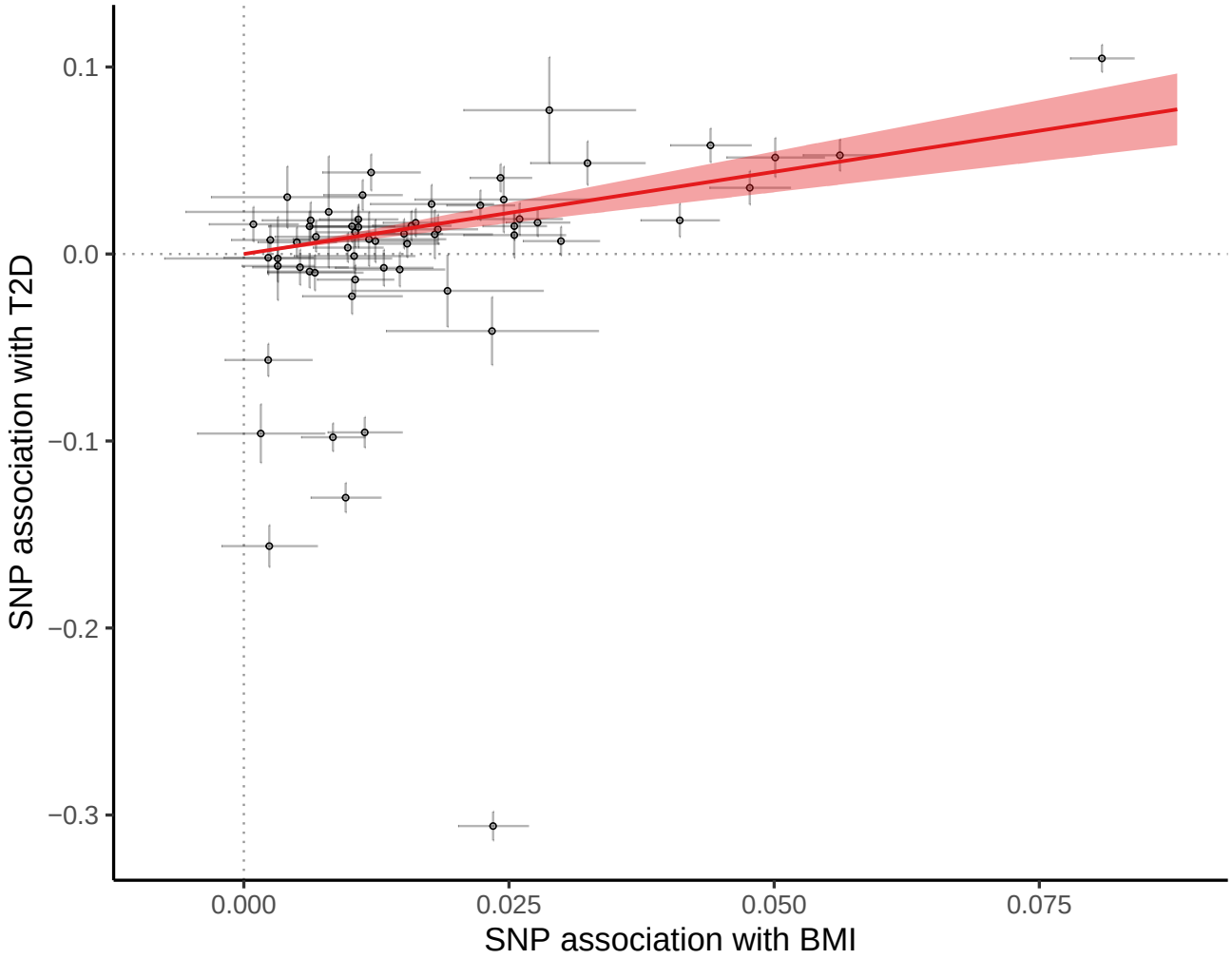
- **Selection dataset**: Akiyama et al. 2017[1]
- **Exposure dataset**: Locke et al. 2015[2]
- **Outcome dataset**: Mahajan et al. 2018[3]

[1] Masato Akiyama et al. "Genome-wide association study identifies 112 new loci for body mass index in the Japanese population". In: *Nature Genetics* 49.10 (2017), p. 1458.

[2] Adam E Locke et al. "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.7538 (2015), pp. 197–206.

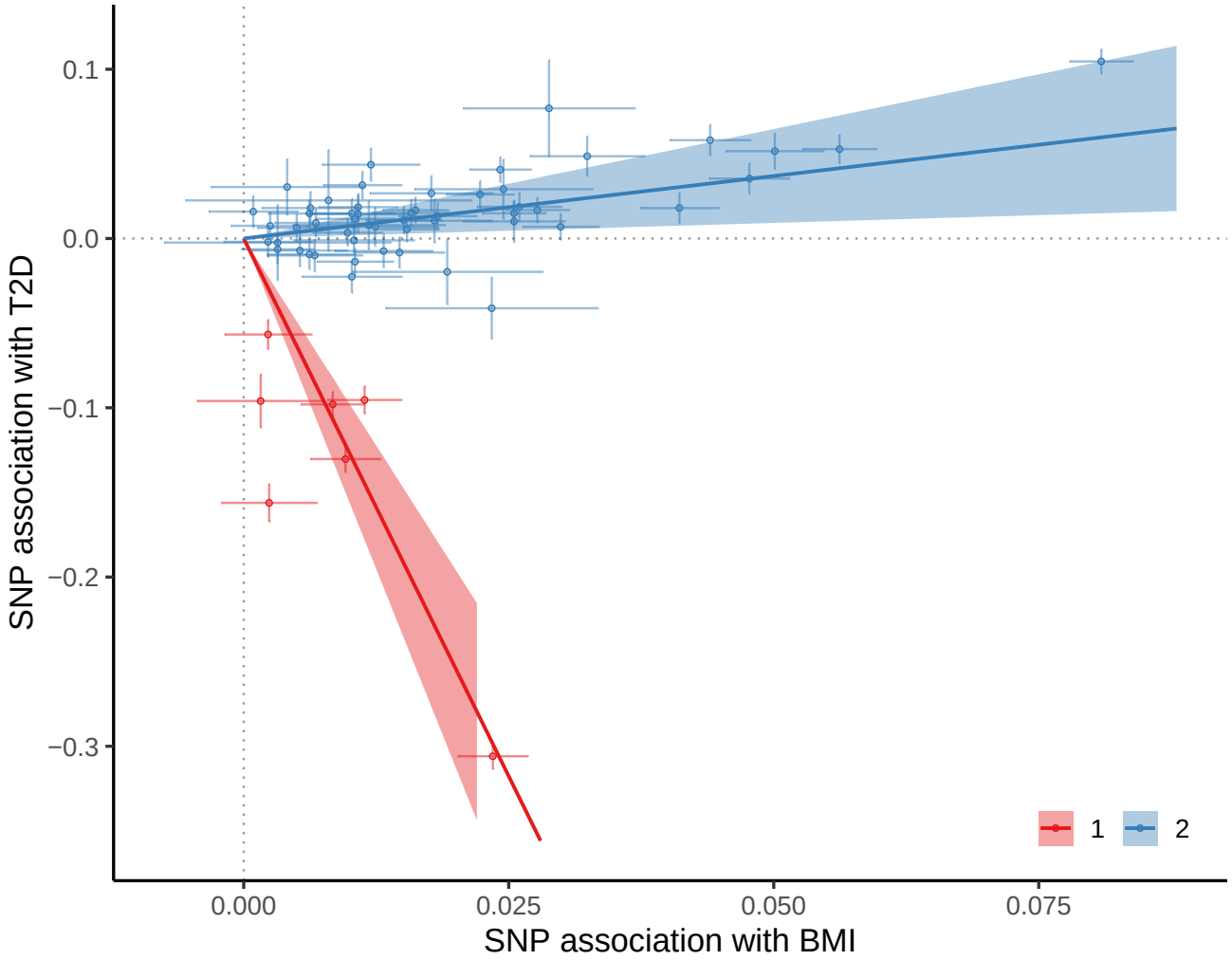
[3] Anubha Mahajan et al. "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps". In: *Nature genetics* 50.11 (2018), pp. 1505–1513.

Example: BMI-T2D



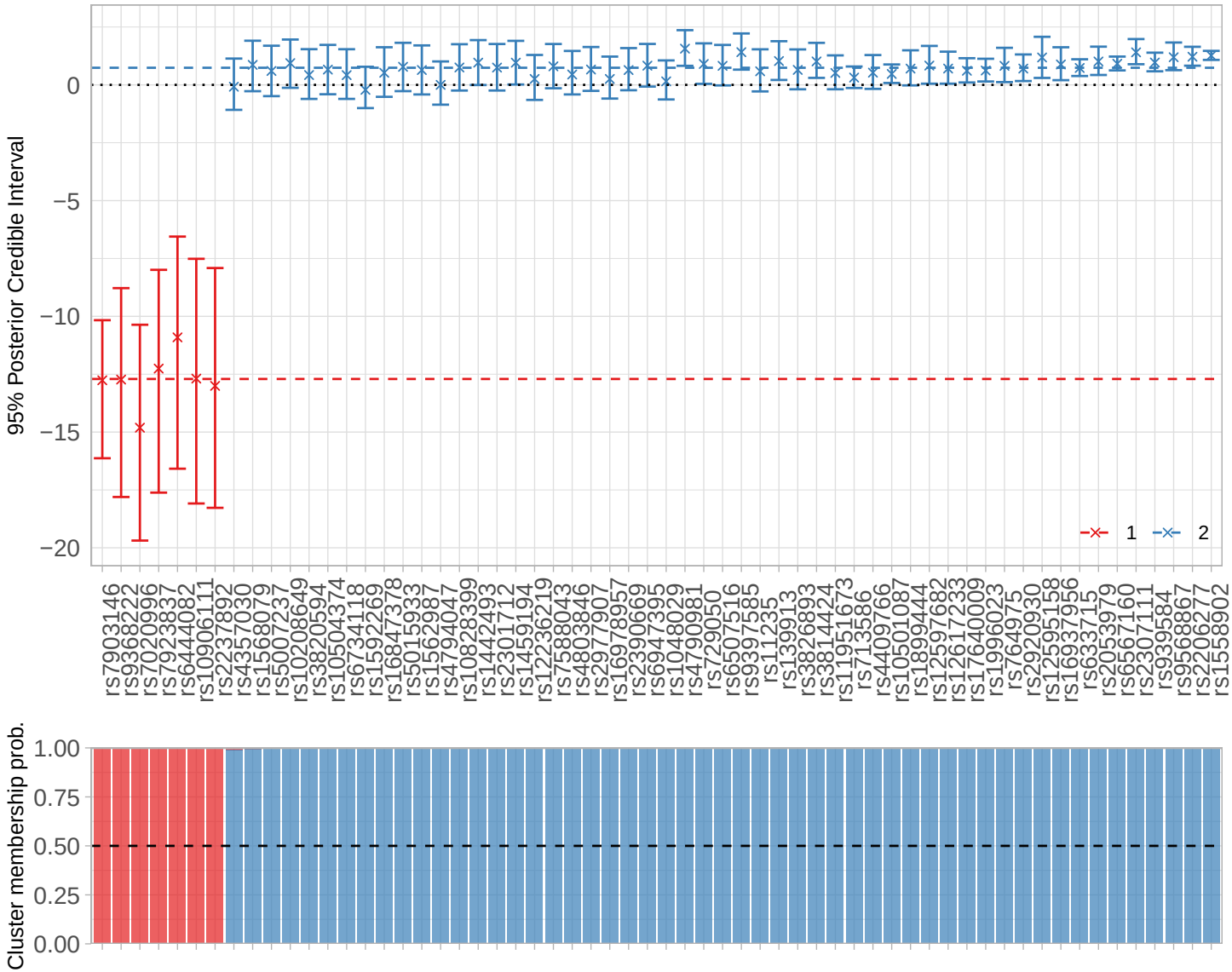
Results of MR-RAPS.

Example: BMI-T2D

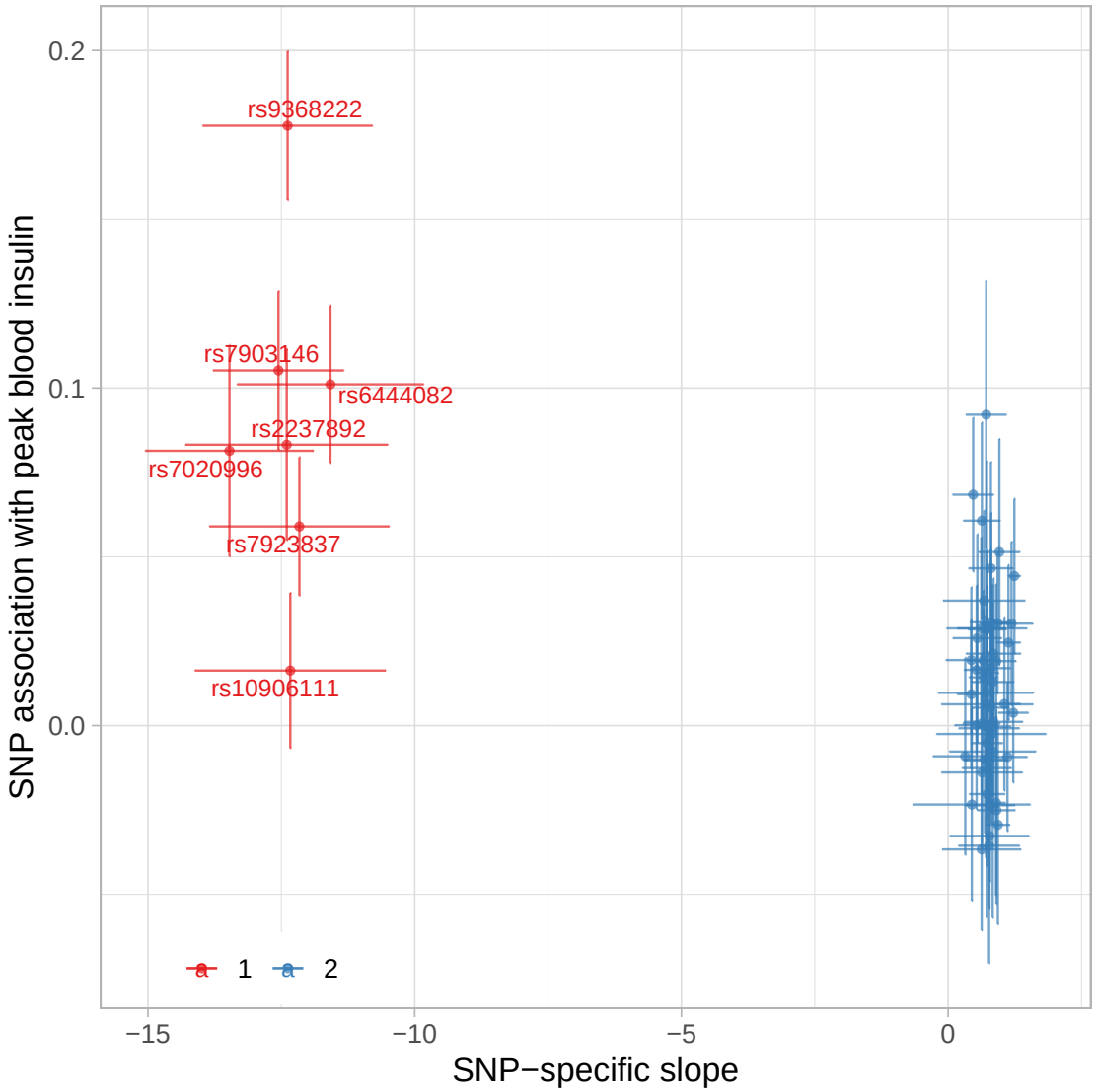


Results of MR-PATH.

Example: BMI-T2D



Example: BMI-T2D



Concluding remarks

- A few other related methods:
  - MR-Clust: Constructs mixture model based on SNP-specific Wald estimators.
  - GRAPPLE: A visualization tool that does not attempt to model different mechanisms explicitly.
  - BESIDE-MR: A Bayesian model averaging approach extends the profile likelihood used in MR RAPS.
- Advantages of MR-PATH:
  - Does not require individually strong instruments.
  - Accounts for measurement error in the summary data.
  - An interpretable generative model for multiple causal mechanisms.
  - Potential extensions to multivariable MR with correlated SNPs.
- Further information: `http://danieliong.me/mr-path/`.