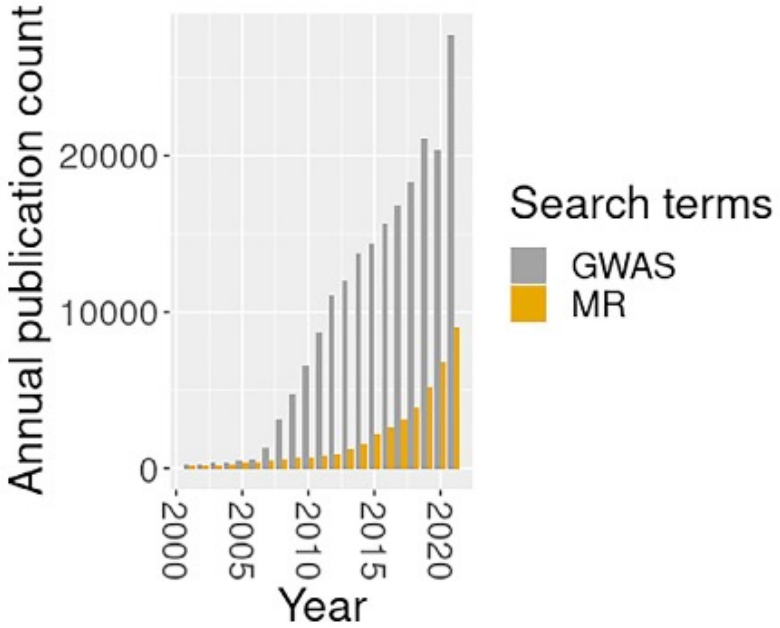
Leveraging Summary-data-based Mendelian Randomisation for Gene Target Discovery

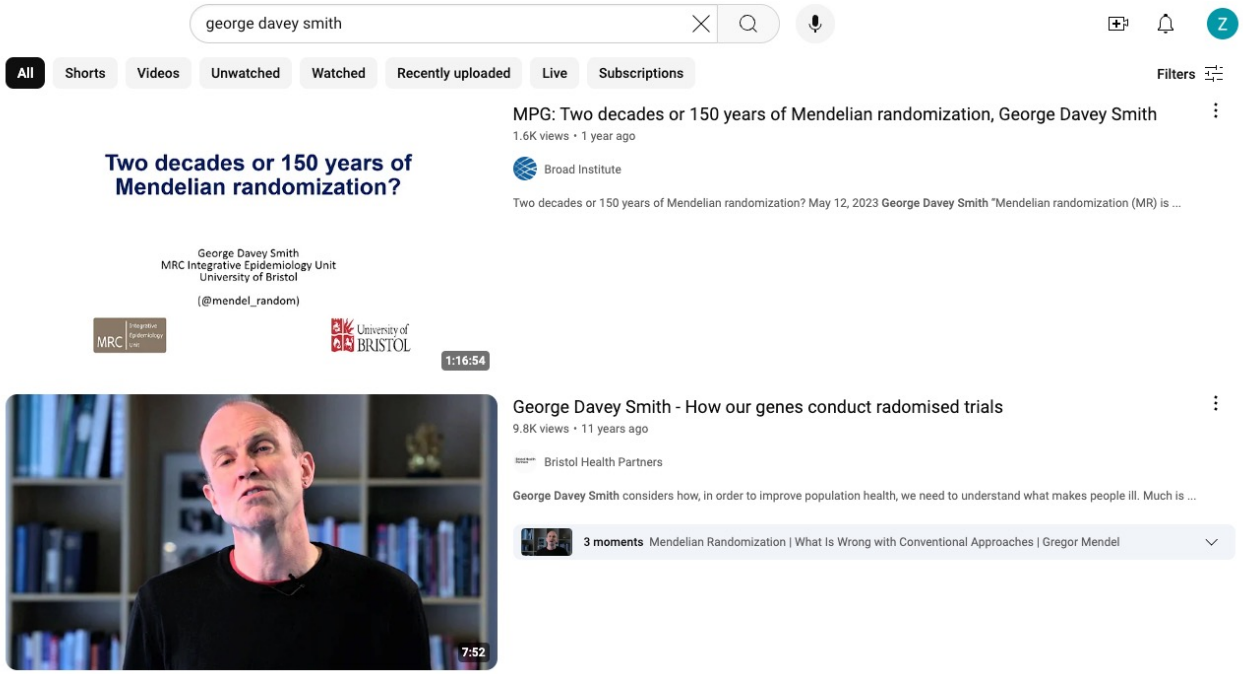
Zhihong Zhu, Ph.D

Senior Researcher, NCRR, Aarhus University
Scientific Visitor, PCTG, University of Queensland
z.zhu.ncrr@au.dk

Why Is Mendelian Randomisation?



Boehm et al. 2022



From YouTube

What Is Mendelian Randomisation?

- Flexible

  Fewer ethical restrictions

- Accessible

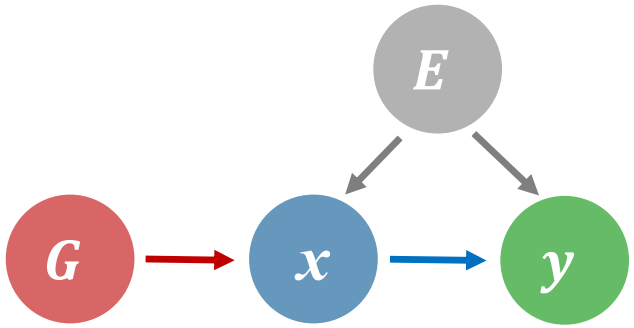
  GWAS summary data

  LD reference data

- Robust

  Less susceptible to environmental confounders and reverse causation

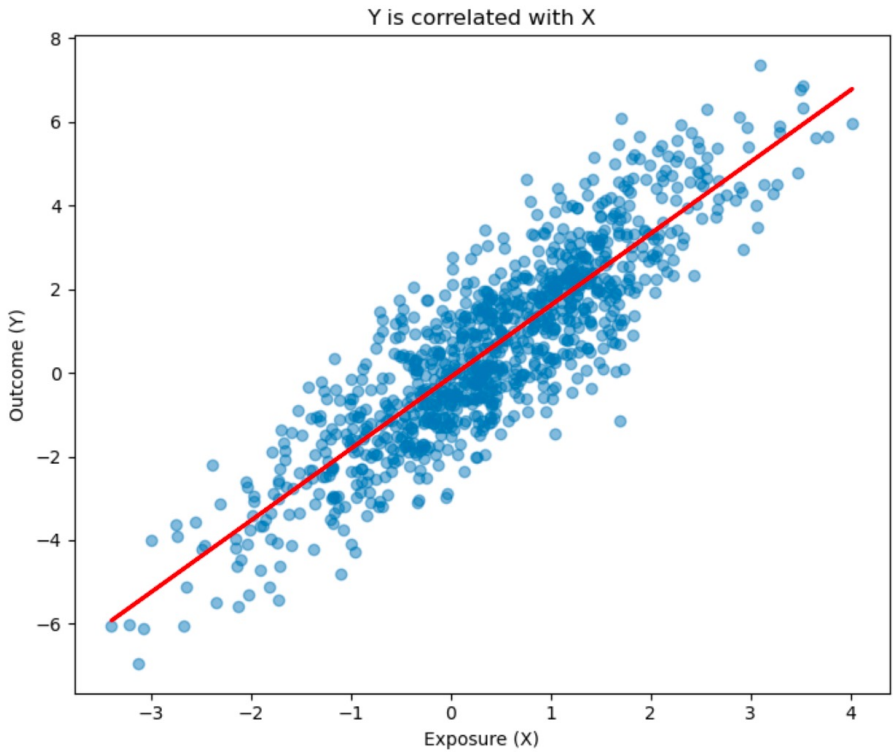
What Is Mendelian Randomisation?



GWAS data (G -> Y)

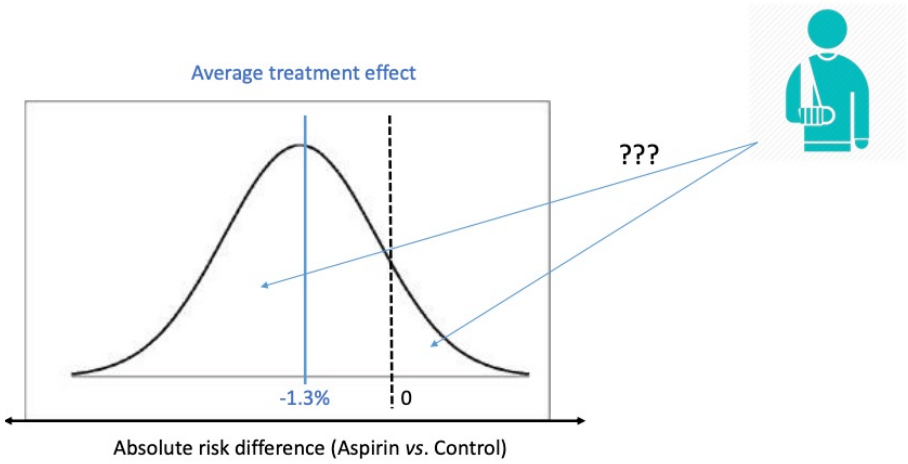
eQTL data (G -> X)

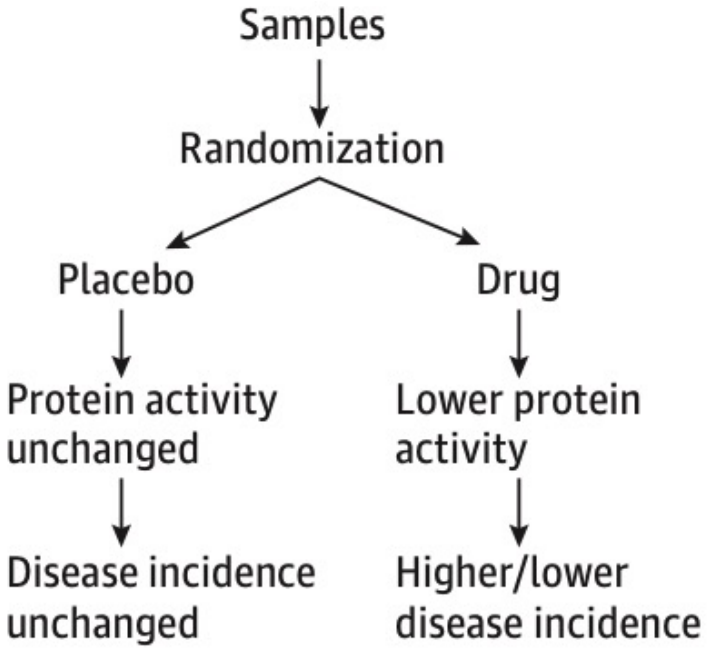
LD reference data



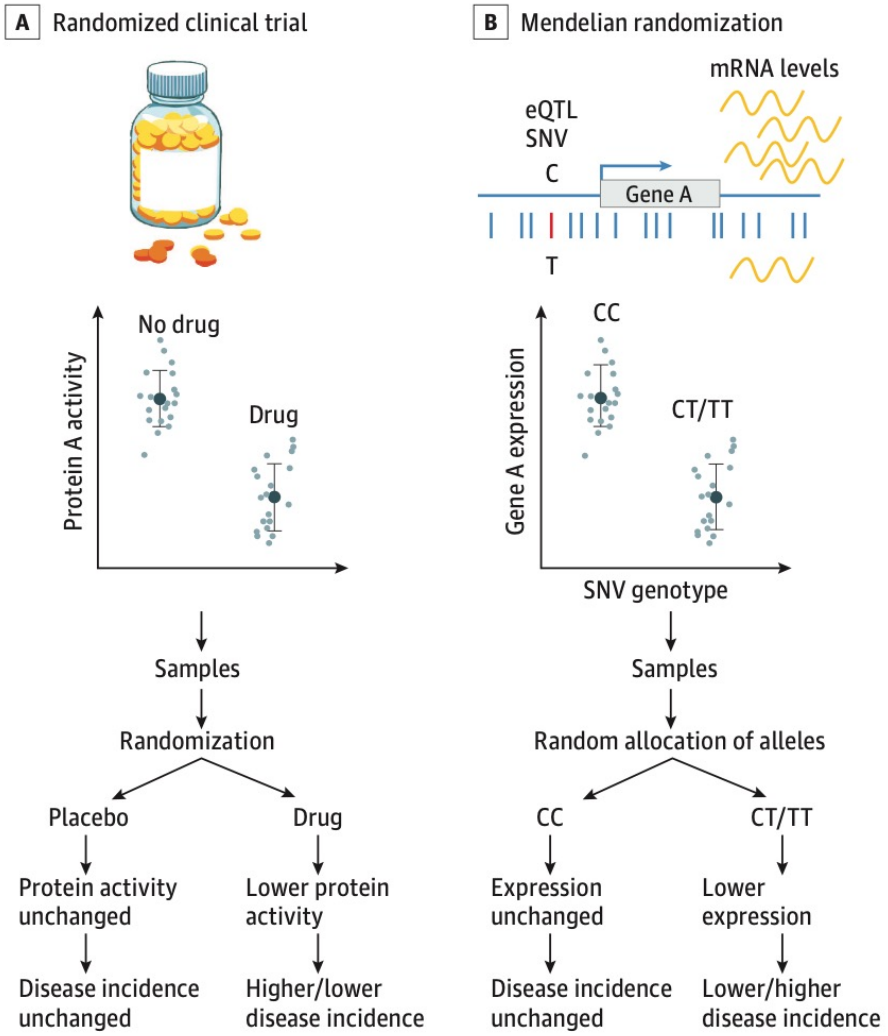
How Does Mendelian Randomisation Work?

Causal Inference





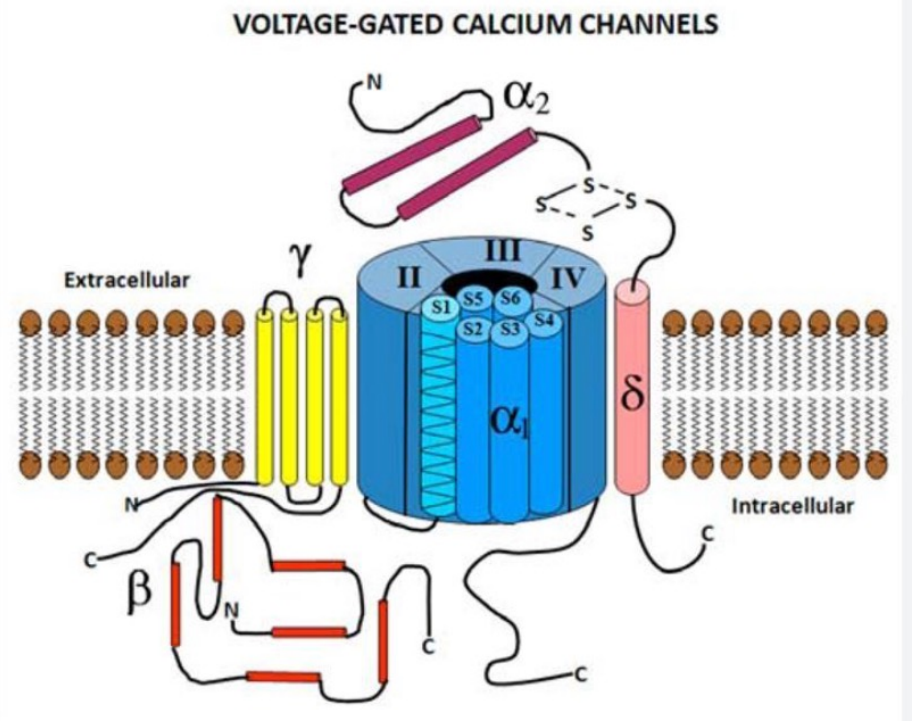
How Does Mendelian Randomisation Work?



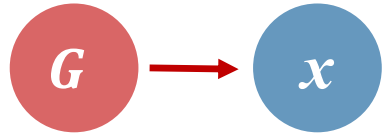
NATIONAL CENTRE FOR REGISTER-BASED RESEARCH

Voltage-Dependent Calcium Channel Complex

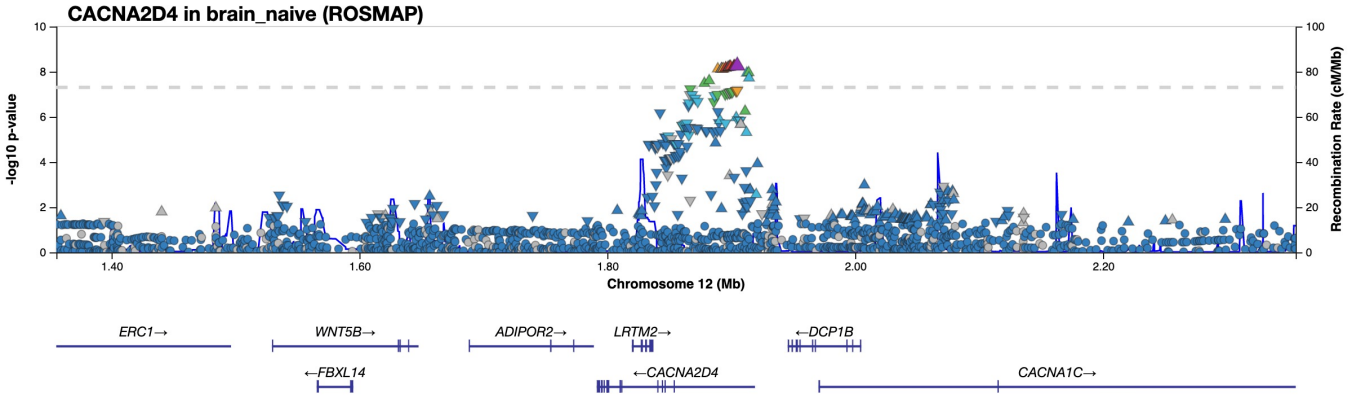
- Regulates calcium ion entry into cells
- Drug targets for hypertensive disease, cardiovascular diseases, psychiatric diseases, neurologic diseases, retinal diseases etc.
- Coding gene of CACNA2D4, part of the voltage-dependent calcium channel complex



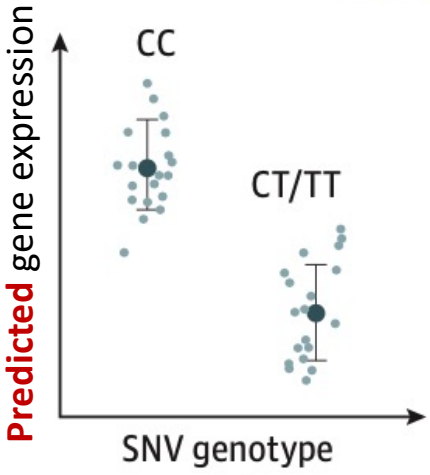
CACNA2D4, Predicting Gene Expression



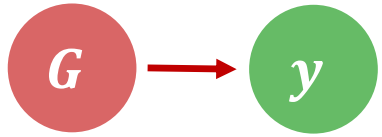
- *cis*-eQTL data
  - ❑ Association between SNP and exposure



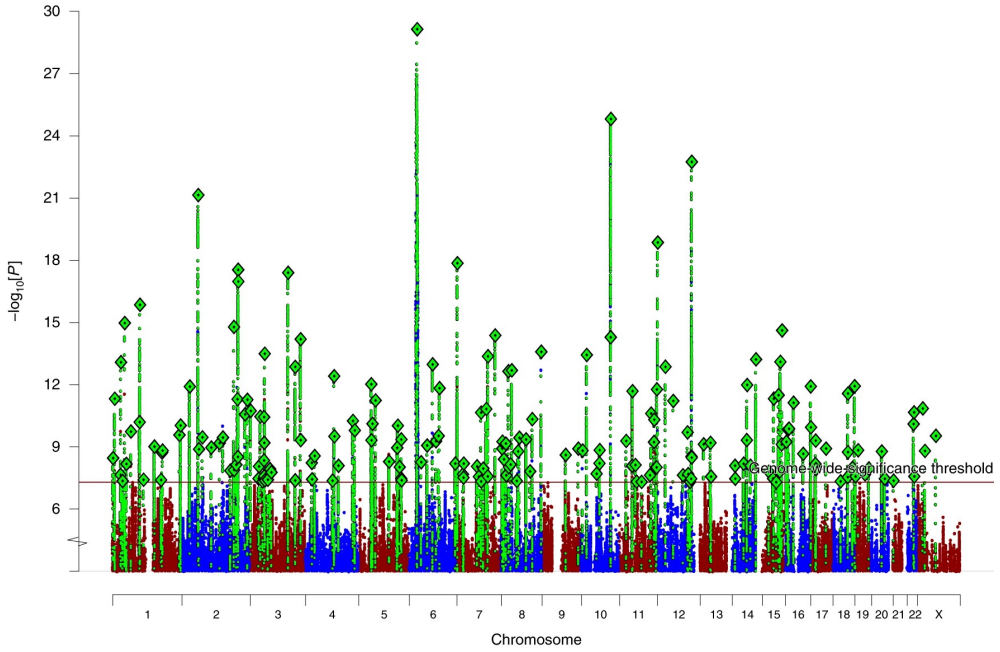
- Assumption
  - ❑ Strong association between SNP and exposure
  - ❑ Linear relationship between SNP and exposure
- Genetically predicted exposure
  - ❑ $\hat{x}_j = z_j\hat{\delta} + e_{x(j)}$



CACNA2D4, Predicting Disease Risk



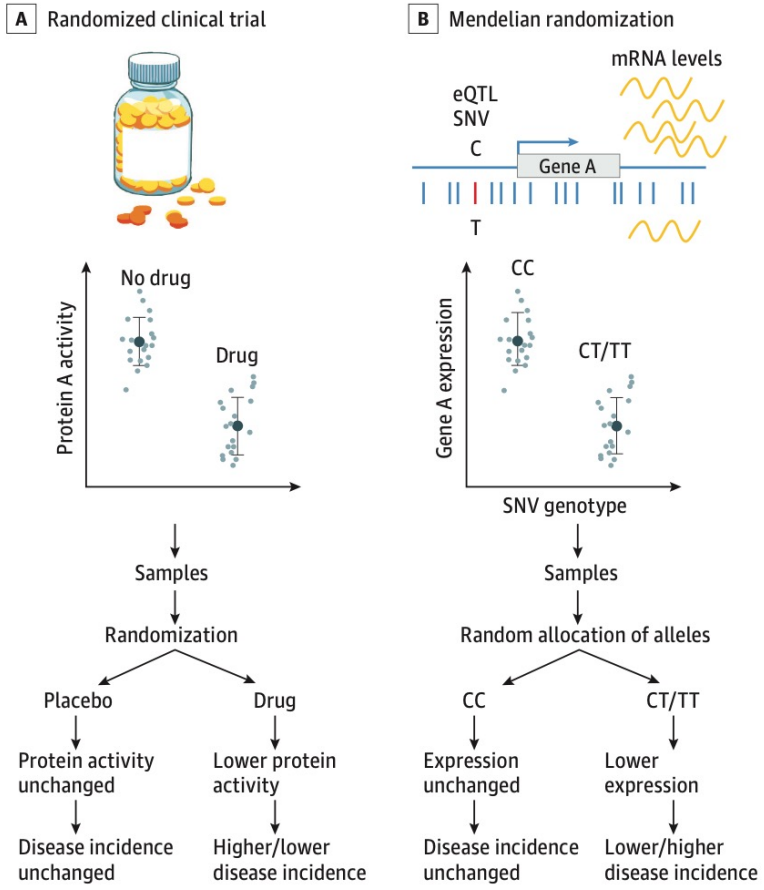
- GWAS data
  - ❑ Association between SNP and outcome
- Assumption
  - ❑ Linear relationship between SNP and outcome
- Genetically predicted outcome
  - ❑ $\hat{y}_j = z_j\hat{\gamma} + e_{y(j)}$



Lam et al 2019 Nature Genetics

Summary-data-based MR

- SMR estimate
  - ❑ Assumption: Single causal variant, linear relationship between exposure and outcome
  - ❑ Estimate: $\hat{\beta}_{\mathrm{SMR}} = \hat{\gamma}/\hat{\delta}$
  - ❑ Interpretation: a x mg/L (1SD unit) higher exposure increase/decrease y% genetic risk of disease outcome
  - ❑ Equivalent to estimate from regression if individual-level data are available.



Proof of MR Estimate

- $\mathrm{E}(\hat{\beta}_{\mathrm{SMR}}) = \hat{\gamma}/\hat{\delta} = (\hat{\gamma}\times\hat{\delta})/(\hat{\delta}\times\hat{\delta})$
- SNP-exposure association: $\hat{\delta} = (z^T z)^{-1} z^T x$
  SNP-outcome association: $\hat{\gamma} = (z^T z)^{-1} z^T y$
- $\mathrm{E}(\hat{\beta}_{\mathrm{SMR}}) = \hat{\gamma}/\hat{\delta} = (\hat{\gamma}\times\hat{\delta})/(\hat{\delta}\times\hat{\delta}) = \frac{x^T P_z y}{x^T P_z x} = \beta + \frac{x^T P_z e}{x^T P_z x}$, where $P_z = z(z^T z)^{-1} z^T$
- DNA variants are independent of environmental factors, $z^T e = 0$

  $\mathrm{E}(\hat{\beta}_{\mathrm{SMR}}) = \beta$

CACNA2D4 – SMR Calculation

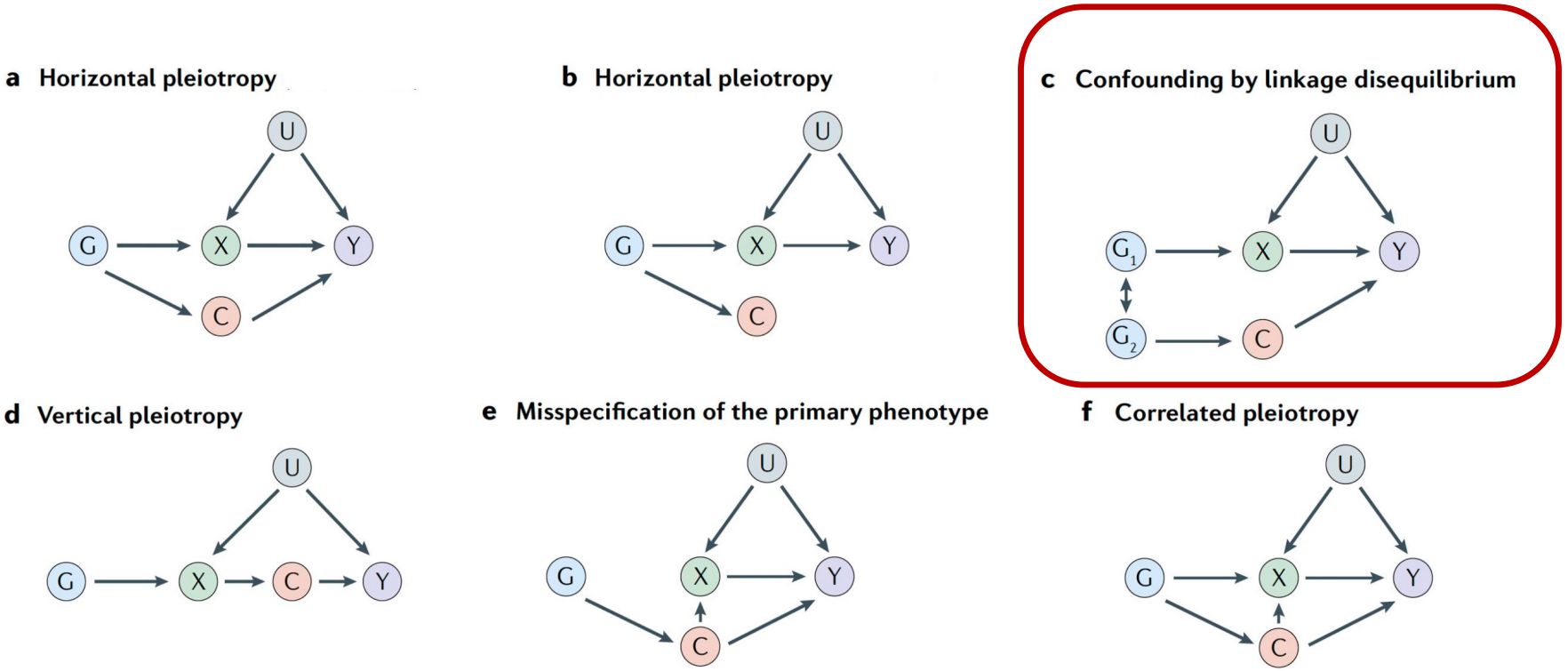
Gene	SNP	A1 / A2	Data	*b*	SE	*P*-value
CACNA2D4	rs1044825	G / T	eQTL (blood)	0.447	0.0186	4.1E-128
			GWAS (schizophrenia)	-0.0377	0.0087	1.3E-5

$$\hat{\beta} \approx \frac{\hat{\gamma}}{\hat{\delta}} = -\frac{0.0377}{0.447} = -0.084$$

→ *P*-value = 2.0E-5

$$\mathrm{SE}(\hat{\beta}) \approx \sqrt{\left(\frac{\gamma}{\delta}\right)^2 \left[\frac{var(\delta)}{\delta^2} + \frac{var(\gamma)}{\gamma^2}\right]} = \sqrt{\left(\frac{0.45}{-0.04}\right)^2 \left[\frac{0.02^2}{0.45^2} + \frac{0.01^2}{(-0.04)^2}\right]} = 0.020$$

Bias Due to Pleiotropy



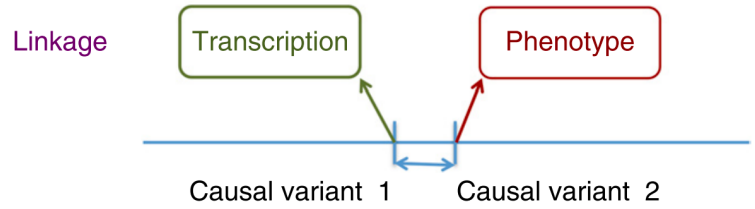
Sanderson et al. Nature Reviews 2022

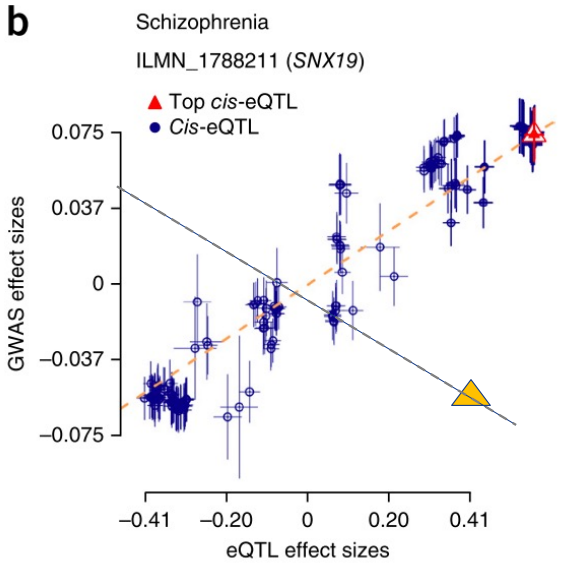
How to Examine the LD Confounding

- HEIDI (Heterogeneity in Dependent Instruments)

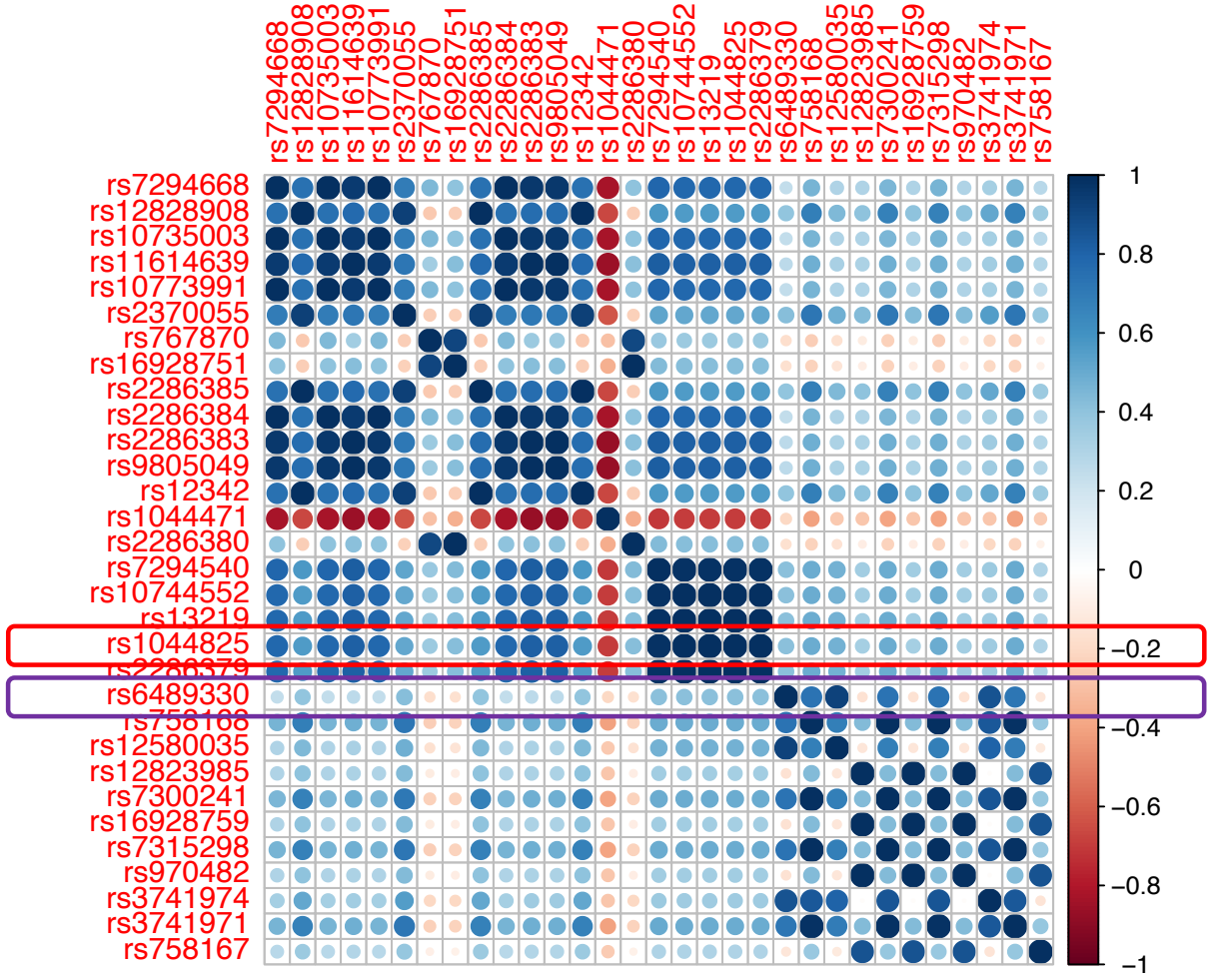
  H0: No difference

  H1: Significant difference at two correlated SNPs





SMR/HEIDI – *CACNA2D4*



The top-associated SNP

The SNP to test difference

NATIONAL CENTRE FOR REGISTER-BASED RESEARCH

SMR/HEIDI – *CACNA2D4*

SNP	A1 / A2	Data	*b*	SE	*P*-value
rs1044825	G / T	eQTL (blood)	0.447	0.0186	4.1E-128
		GWAS (schizophrenia)	-0.0377	0.0087	1.3E-5
rs6489330	A / G	eQTL (blood)	0.211	0.02384	9.5E-19
LD $r = 0.413$		GWAS (schizophrenia)	-0.0378	0.0108	4.7E-4

rs1044825, $\hat{\beta}_1 = -0.084$, $\mathrm{SE}(\hat{\beta}_1) \approx 0.020$ rs6489330, $\hat{\beta}_2 = -0.179$, $\mathrm{SE}(\hat{\beta}_2) \approx 0.055$

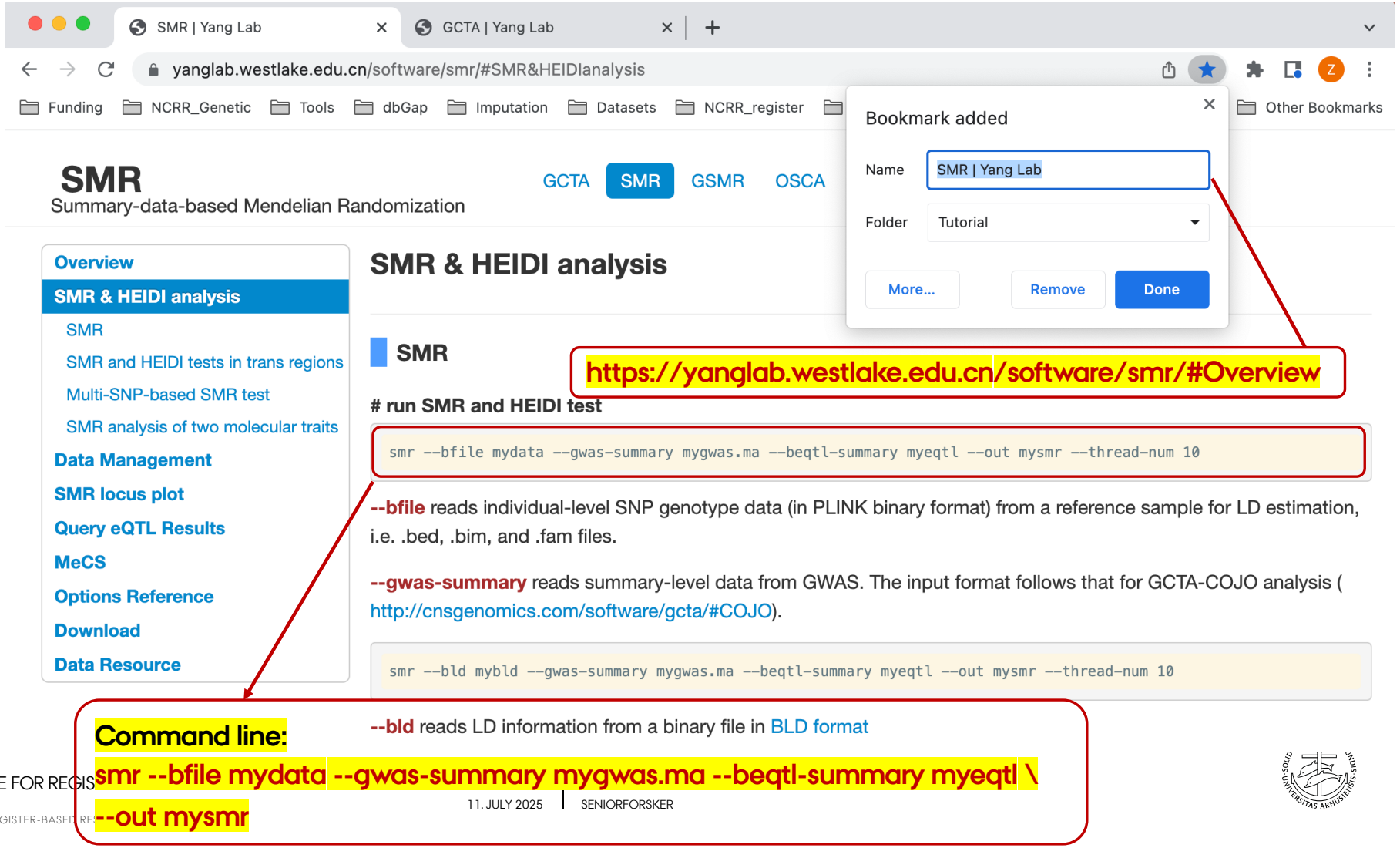
Difference, $\hat{d} = \hat{\beta}_2 - \hat{\beta}_1 = -0.179 + 0.084 = -0.095$

$$\mathrm{SE}(\hat{d}) = \sqrt{\mathrm{var}(\hat{\beta}_2 - \hat{\beta}_1)} = \sqrt{var(\hat{\beta}_2) + var(\hat{\beta}_1) - 2 \times cov(\hat{\beta}_1, \hat{\beta}_2)} = 0.050$$

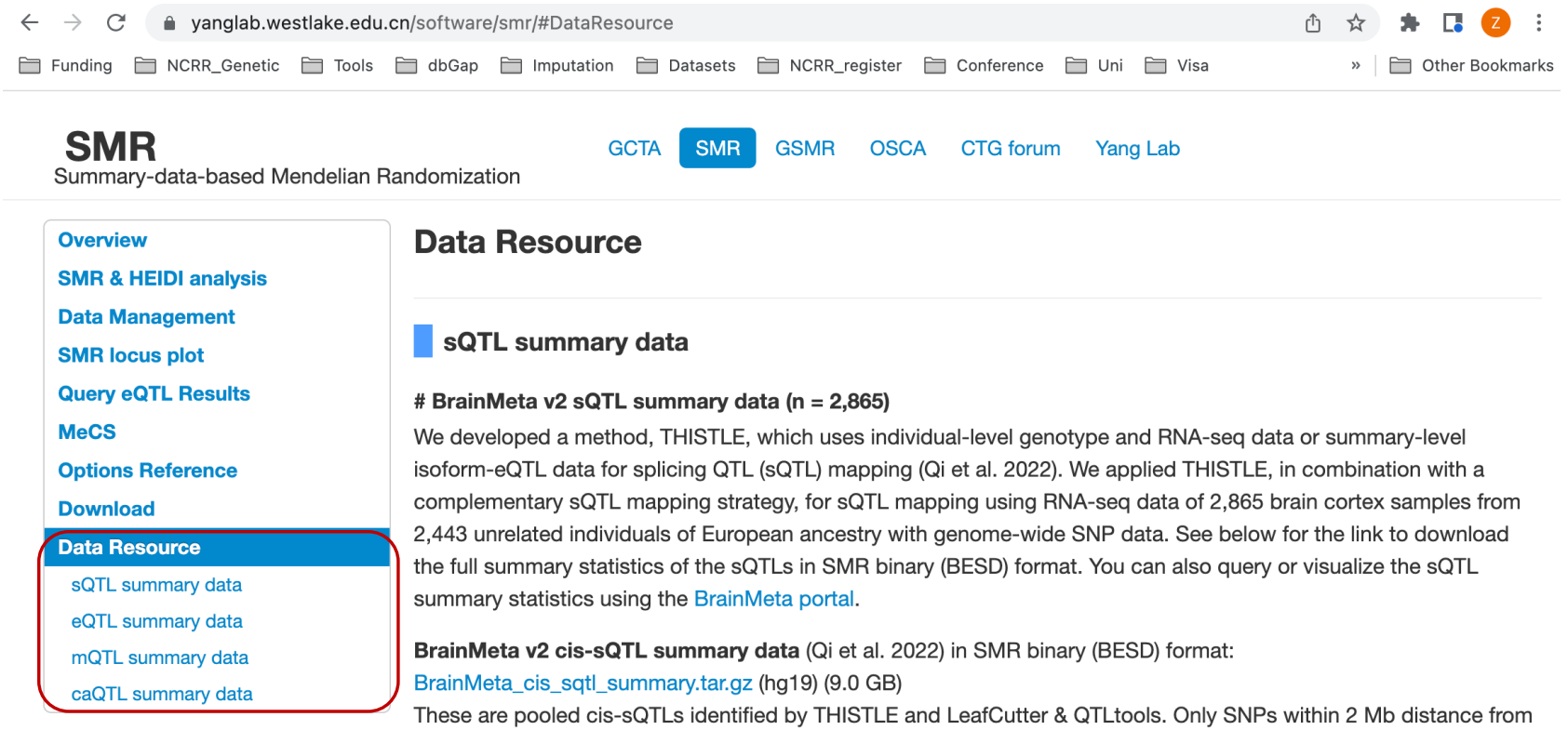
→ *P*-value = 0.06

NATIONAL CENTRE FOR REGISTER-BASED RESEARCH

Software


SMR | Yang Lab
GCTA | Yang Lab
yanglab.westlake.edu.cn/software/smr/#SMR&HEIDIanalysis
Funding
NCRR_Genetic
Tools
dbGap
Imputation
Datasets
NCRR_register
Other Bookmarks
Bookmark added
Name
SMR | Yang Lab
Folder
Tutorial
More...
Remove
Done
SMR
Summary-data-based Mendelian Randomization
GCTA
SMR
GSMR
OSCA
Overview
SMR & HEIDI analysis
SMR
SMR and HEIDI tests in trans regions
Multi-SNP-based SMR test
SMR analysis of two molecular traits
Data Management
SMR locus plot
Query eQTL Results
MeCS
Options Reference
Download
Data Resource
SMR & HEIDI analysis
SMR
https://yanglab.westlake.edu.cn/software/smr/#Overview
run SMR and HEIDI test
smr --bfile mydata --gwas-summary mygwas.ma --beqtl-summary myeqtl --out mysmr --thread-num 10
--bfile reads individual-level SNP genotype data (in PLINK binary format) from a reference sample for LD estimation, i.e. .bed, .bim, and .fam files.
--gwas-summary reads summary-level data from GWAS. The input format follows that for GCTA-COJO analysis (http://cnsgenomics.com/software/gcta/#COJO).
smr --bld mybld --gwas-summary mygwas.ma --beqtl-summary myeqtl --out mysmr --thread-num 10
--bld reads LD information from a binary file in BLD format
Command line:
smr --bfile mydata --gwas-summary mygwas.ma --beqtl-summary myeqtl \
--out mysmr

Data Resource



sQTL – Summary statistics of splicing QTLs

eQTL – Summary statistics from associations of gene expression

mQTL – Summary statistics from associations of methylation

Finished? - No

- Interpretation is crucial
- Checklist

  Experimental design

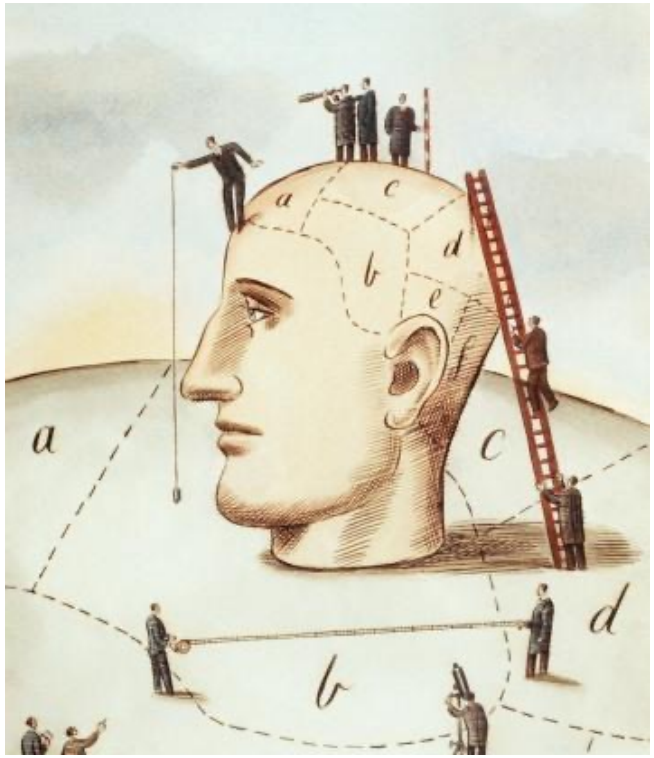
  Cohorts used in GWAS studies

  Biological context, such as age, sex, tissues, cell types etc.

  Selection of SNP instruments

  Sample sizes

  …



Validating Assumptions

Table S13 Correlations between effect sizes of SNPs on brain *C4* gene expressions and neonatal circulating C4 protein concentration, related to the STAR Methods

Tissue	Brain *C4A* expression		Brain *C4B* expression	
	r_b (SE)	*P*-value	r_b (SE)	*P*-value
Brain - Amygdala	0.73 (0.02)	<1.0E-100	0.14 (0.11)	0.20
Brain - Anterior cingulate cortex BA24	0.79 (0.03)	<1.0E-100	-0.99 (0.22)	7.2E-06
Brain - Caudate basal ganglia	0.67 (0.01)	<1.0E-100	-1.01 (0.09)	7.9E-27
Brain - Cerebellar hemisphere	0.59 (0.01)	<1.0E-100	-0.78 (0.04)	5.3E-100
Brain - Cerebellum	0.55 (0.01)	<1.0E-100	-0.70 (0.02)	<1.0E-100
Brain - Cortex	0.64 (0.02)	<1.0E-100	-0.84 (0.04)	3.2E-85
Brain - Frontal cortex BA9	0.88 (0.02)	<1.0E-100	-0.88 (0.05)	5.8E-64
Brain - Hippocampus	0.70 (0.01)	<1.0E-100	0.77 (0.09)	2.0E-17
Brain - Hypothalamus	0.85 (0.02)	<1.0E-100	-0.44 (0.05)	2.1E-20
Brain - Nucleus accumbens basal ganglia	0.77 (0.02)	<1.0E-100	-0.91 (0.08)	5.6E-33
Brain - Putamen basal ganglia	0.72 (0.02)	<1.0E-100	-0.78 (0.04)	3.2E-71
Brain - Spinal cord cervical c-1	1.01 (0.04)	<1.0E-100	-0.96 (0.12)	1.0E-16
Brain - Substantia nigra	0.75 (0.03)	<1.0E-100	0.12 (0.02)	3.2E-11
Nerve - Tibial	0.58 (0.01)	<1.0E-100	0.36 (0.02)	5.1E-65
Pituitary	0.70 (0.01)	<1.0E-100	-0.01 (0.03)	0.63
Average	0.73		0.46	

Note: the Bonferroni-corrected threshold is 0.003 (= 0.05/15).

Borbye-Lorenzen et al.
2023 Cell Genomics

STROBE MR Checklist

STROBE-MR checklist of recommended items to address in reports of Mendelian randomization studies[1 2]

Item No.	Section	Checklist item	Page No.	Relevant text from manuscript
1	**TITLE and ABSTRACT**	Indicate Mendelian randomization (MR) as the study's design in the title and/or the abstract if that is a main purpose of the study	1	Title page
	INTRODUCTION			
2	**Background**	Explain the scientific background and rationale for the reported study. What is the exposure? Is a potential causal relationship between exposure and outcome plausible? Justify why MR is a helpful method to address the study question	4-5	Introduction section
3	**Objectives**	State specific objectives clearly, including pre-specified causal hypotheses (if any). State that MR is a method that, under specific assumptions, intends to estimate causal effects	5	Final paragraph
	METHODS			
4	**Study design and data sources**	Present key elements of the study design early in the article. Consider including a table listing sources of data for all phases of the study. For each data source contributing to the analysis, describe the following:	6	Study design
	a)	Setting: Describe the study design and the underlying population, if possible. Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection, when available.	6-7	Data sources section
	b)	Participants: Give the eligibility criteria, and the sources and methods of selection of participants. Report the sample size, and whether any power or sample size calculations were carried out prior to the main analysis	6-7	Data sources section
	c)	Describe measurement, quality control and selection of genetic variants	6-7	Data sources section
	d)	For each exposure, outcome, and other relevant variables, describe methods of assessment and diagnostic criteria for diseases	7	Instrument selection section
	e)	Provide details of ethics committee approval and participant informed consent, if relevant	NA	NA

Links

- SMR: https://yanglab.westlake.edu.cn/software/smr/
- SMR portal: https://yanglab.westlake.edu.cn/smr-portal/
- STROBE-MR checklist

https://www.bmj.com/sites/default/files/attachments/bmj-article/pre-pub-history/strobe-mr-checklist-fillable_r2.pdf


NATIONAL CENTRE FOR REG