Summary-data-based Mendelian randomisation and prediction of gene targets

Zhihong Zhu, Ph.D

Senior Researcher, NCRR, Aarhus University

Visitor, PCTG, University of Queensland

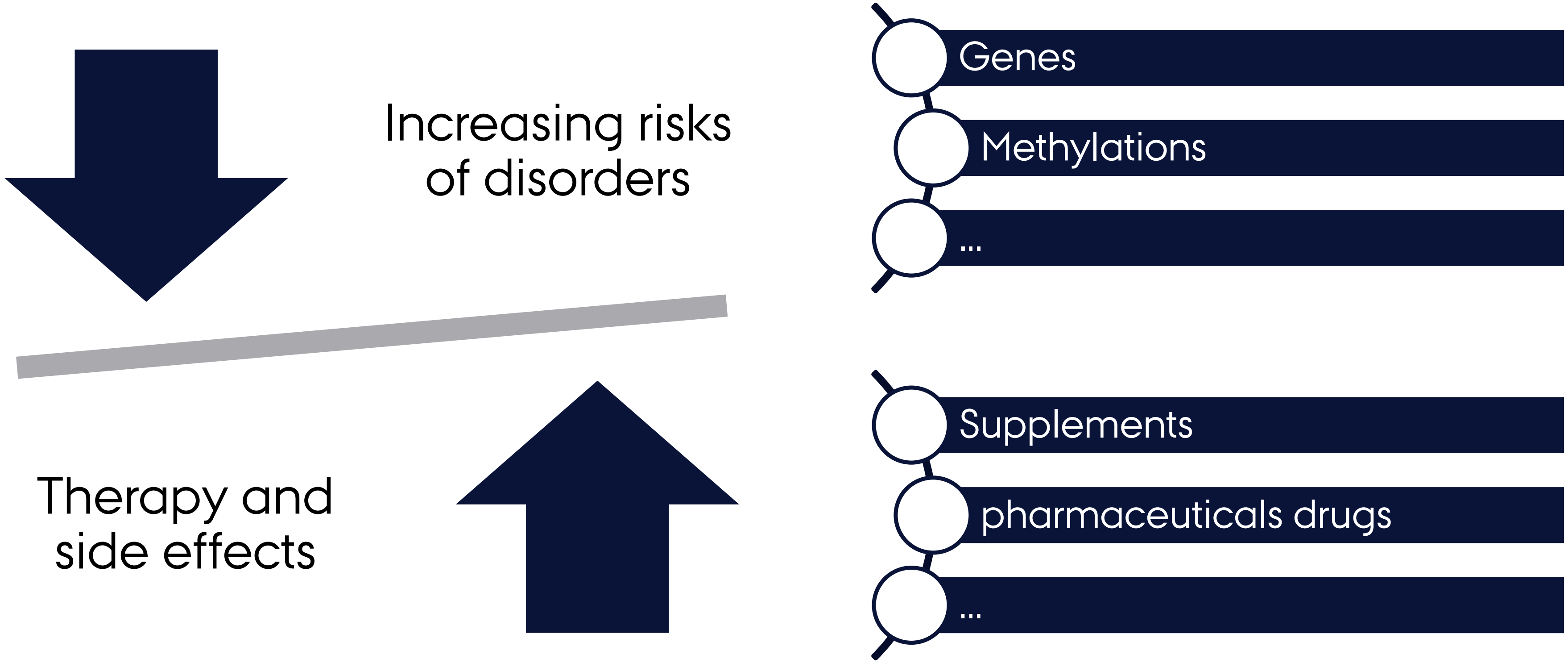
z.zhu@ncrr.au.dk | z.zhu1@uq.edu.au

Outlines

Summary-data-based Mendelian randomisation (SMR)

- Purposes of SMR
- Concept of SMR method
- A real example of SMR test
- SMR software
- Practical

Causal inference



Increasing risks of disorders

Therapy and side effects

- Genes
- Methylations
- ...

- Supplements
- pharmaceuticals drugs
- ...

Risk gene – *CACNA2D4*

The *CACNA2D4* gene, one of voltage-dependent calcium-channel genes, is an important gene target of anti-hypertensive drugs. It is a risk gene for both bipolar disorder and schizophrenia.

CACNA2D4 | hypertensive disorder -> schizophrenia / bipolar disorder | hypertensive disorder

Given the independence of hypertensive disorder and schizophrenia / bipolar disorder
CACNA2D4 -> schizophrenia / bipolar disorder

Observational study

In observational study, regression model is used to test association,

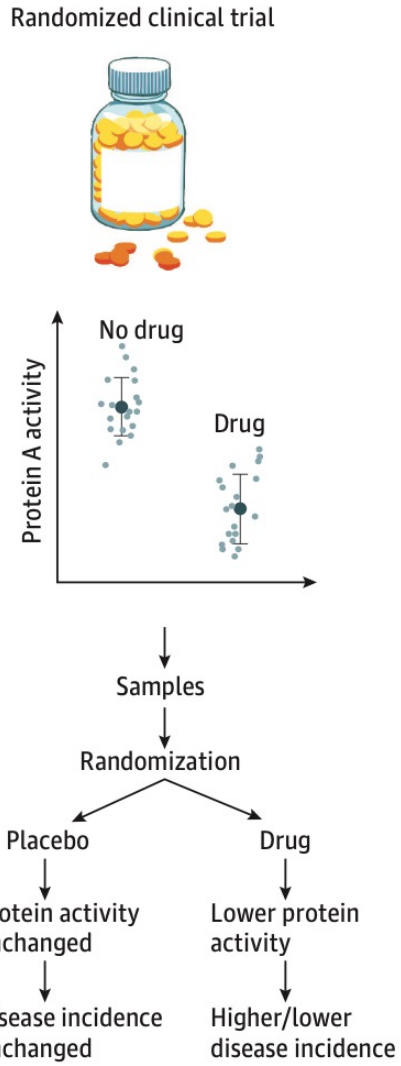
$$y_j = x_j\beta + e_j$$

The ordinary least square estimate,

$$\hat{\beta}_{OLS} = (x^T x)^{-1} x^T y = (x^T x)^{-1} x^T (x\beta + e) = \beta + (x^T x)^{-1} x^T e$$

If there is confounding factor, then $\hat{\beta}_{OLS}$ is biased.

Randomised controlled trail


Randomized clinical trial
Protein A activity
No drug
Drug
Samples
Randomization
Placebo
Drug
Protein activity unchanged
Lower protein activity
Disease incidence unchanged
Higher/lower disease incidence

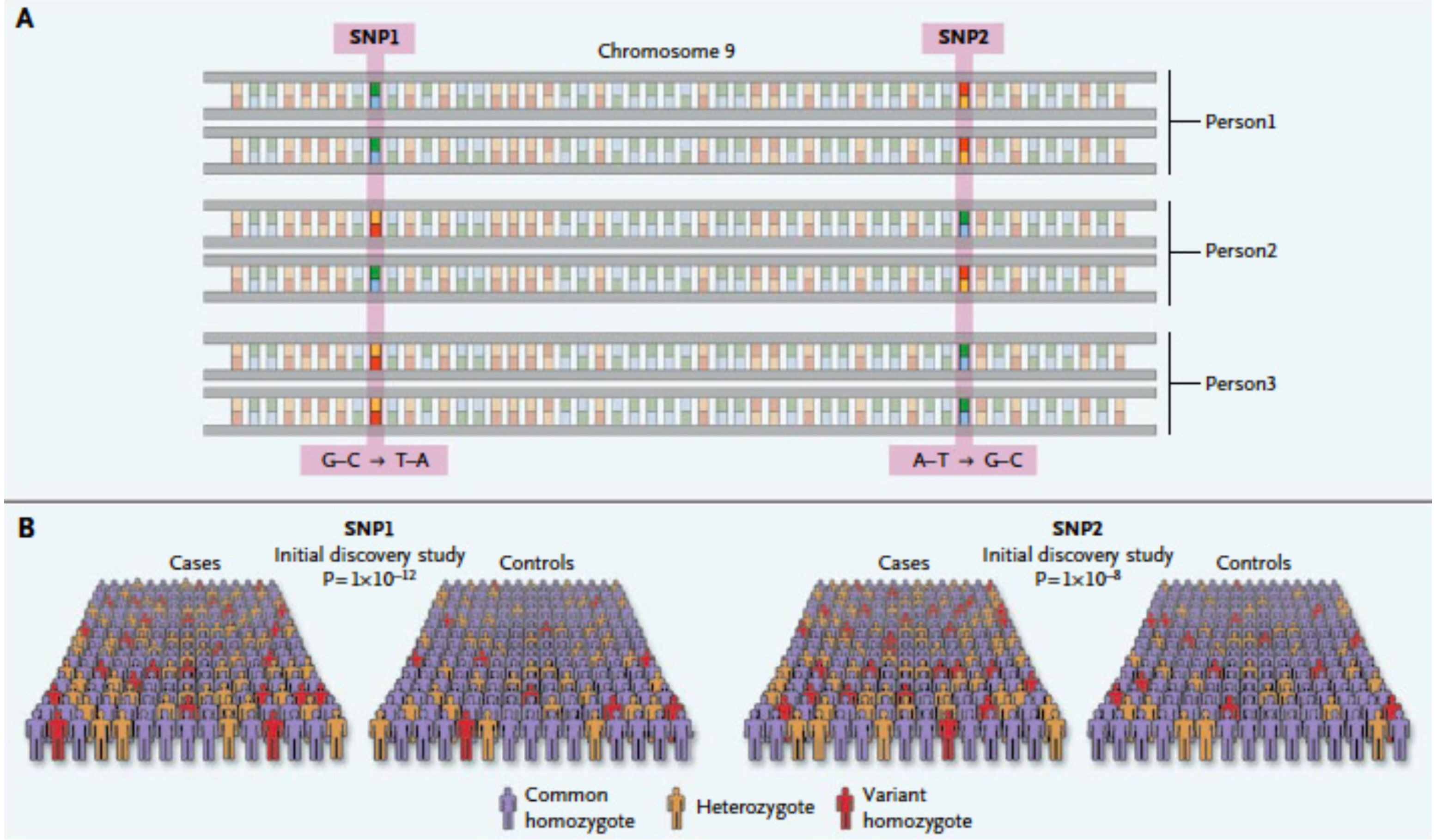
Assumptions	RCT
Two designed Groups	a) Treatment group b) Control group
Assignment	Randomly assigning subjects to treatment conditions
Confounder	Prior exposure and instrumentation do not threaten the internal validity.
Test	The difference must be driven by intervention.

Chauquet et al 2021 JAMA Psychiatry

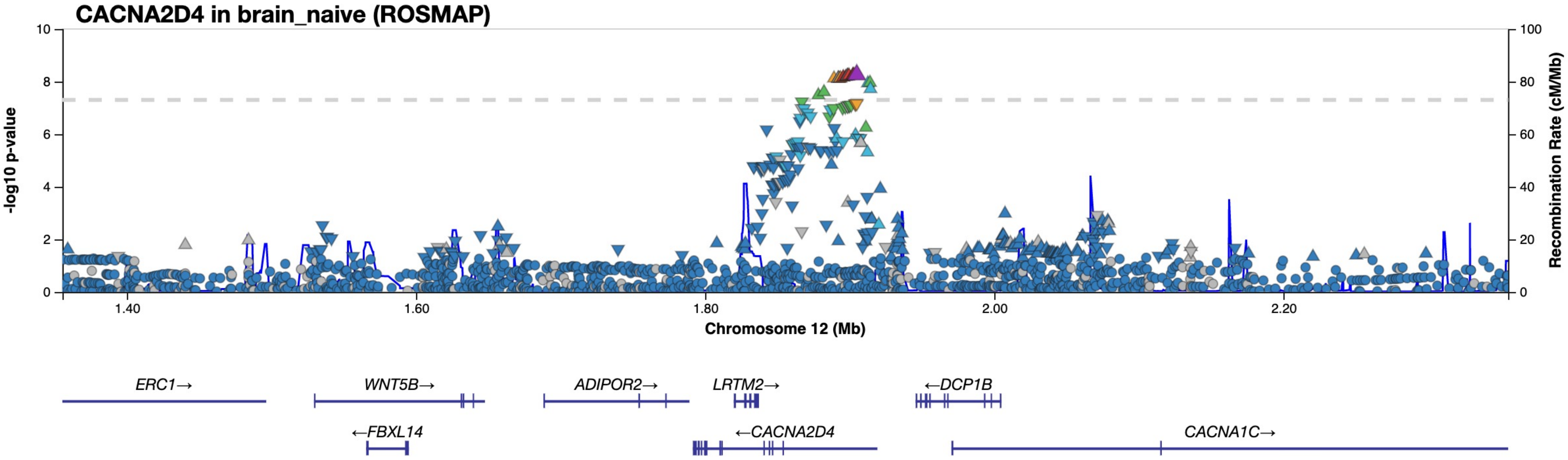
SNP (DNA variant)

Risk alleles

Increasing risks



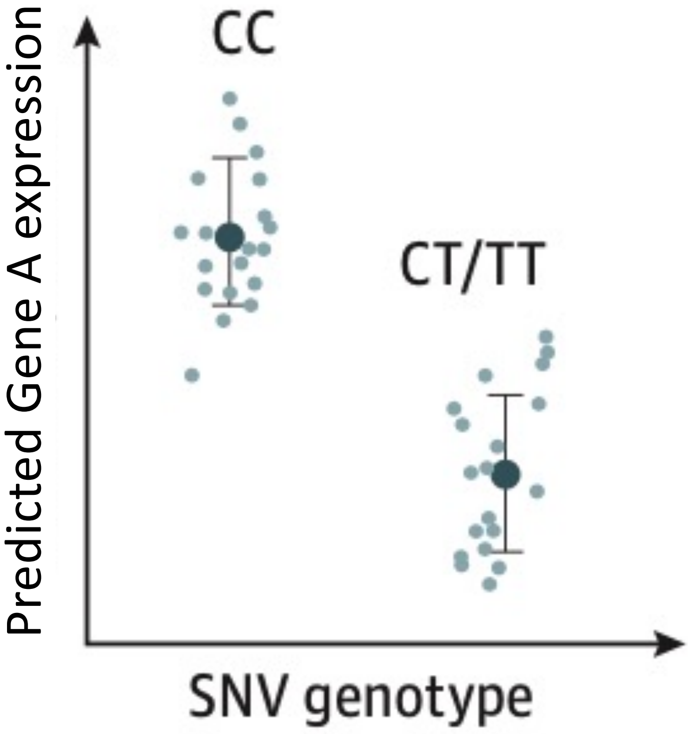
eQTL study



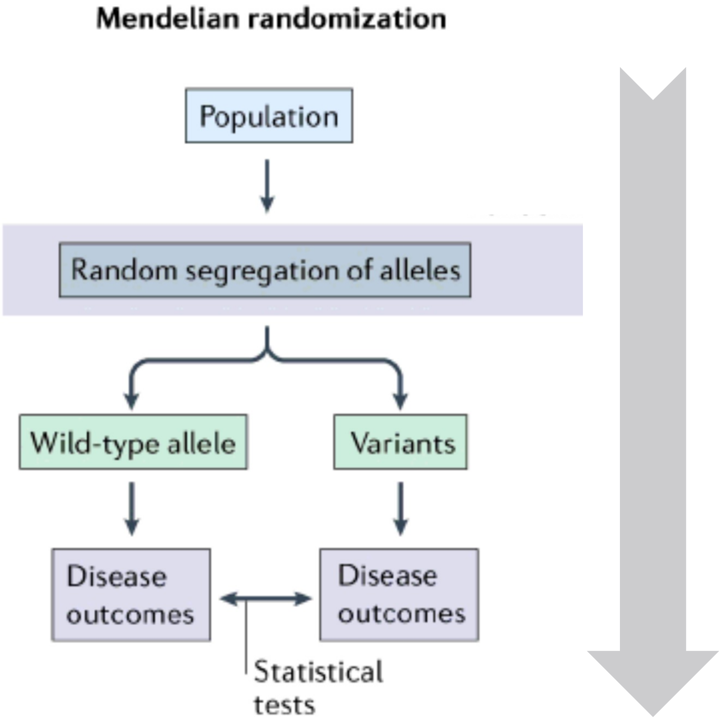
allele -> lower gene expression

Predicting heritable traits

Control Case

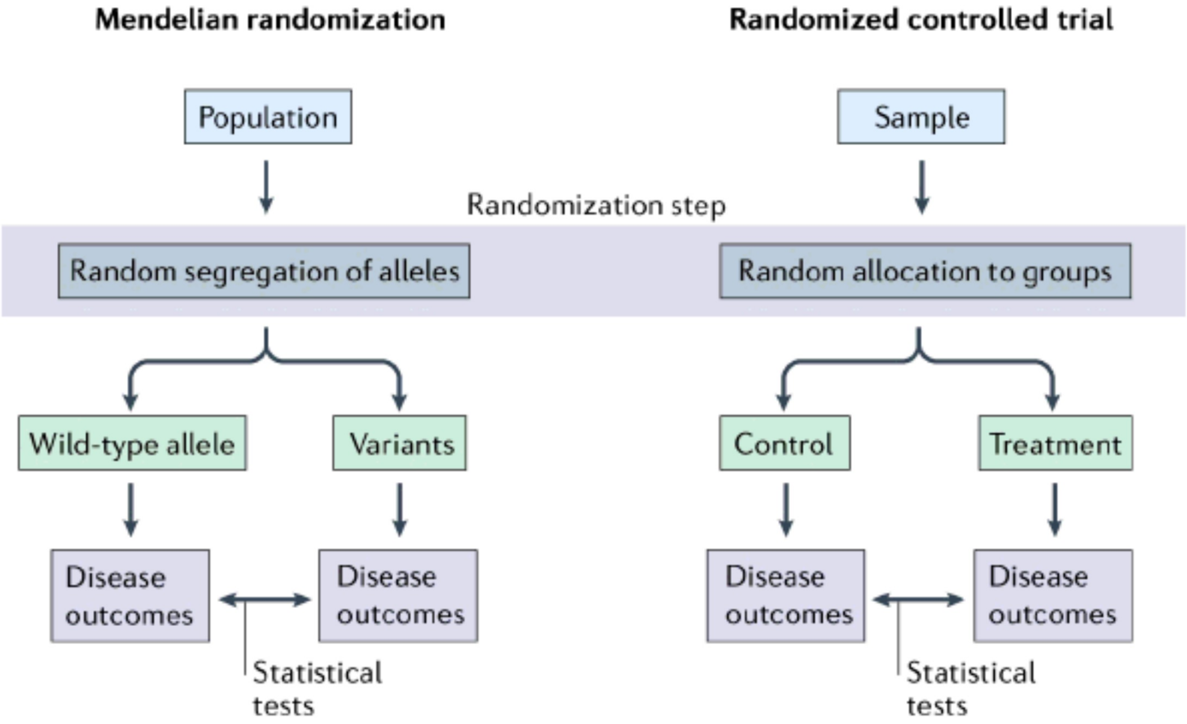


Mendelian randomisation



DNA variant	Non-risk allele	Risk allele
Risk factor	Normal	Deficiency
Outcome	Low risk	High risk

Similar concept

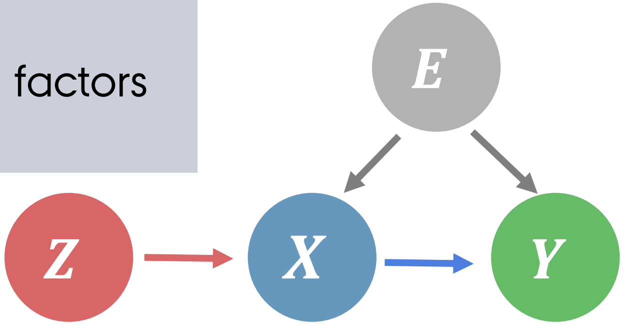


Wald ratio estimator

$$\beta = \frac{\mathrm{E}(\mathrm{Disorder}|A=1) - \mathrm{E}(\mathrm{Disorder}|A=0)}{\mathrm{E}(\mathrm{Risk\ factor}|A=1) - \mathrm{E}(\mathrm{Risk\ factor}|A=0)}$$

Strength of MR

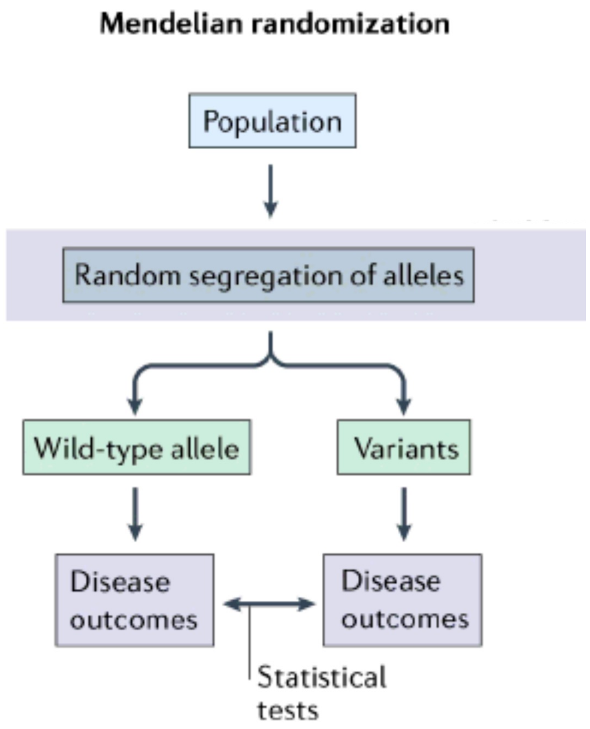
	RCT	MR
Ethics	Ethical issues, e.g . confidentiality, informed consent, etc.	Using SNPs (DNA variants) as instruments
Expense	Time-consuming and expensive	Many available genotyped populations and GWAS datasets
Confounder	Prior exposure and instrumentation do not threaten the internal validity	Free of environmental factors

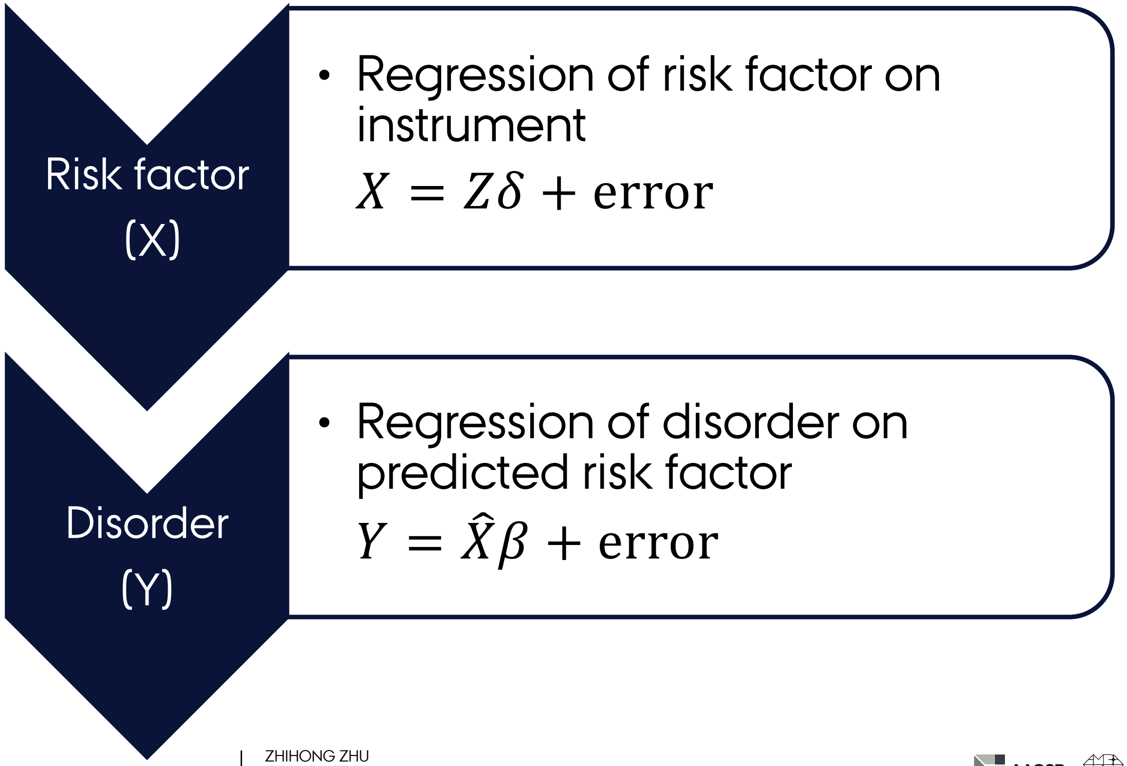


Two-stage least square estimate

$$\text{Disorder} = \text{Risk factor} + e$$

Instruments (Z)


Mendelian randomization



- Regression of risk factor on instrument

$$X = Z\delta + \text{error}$$

- Regression of disorder on predicted risk factor

$$Y = \hat{X}\beta + \text{error}$$

Two-stage least square estimate

$$\mathrm{E}\left(\hat{\beta}_{2LSL}\right) = (\hat{x}^T\hat{x})^{-1}\hat{x}^T y = \frac{x^T P_Z y}{x^T P_Z x} = \beta + \frac{x^T P_Z e}{x^T P_Z x}$$ where $P_Z = Z(Z^T Z)^{-1} Z^T$

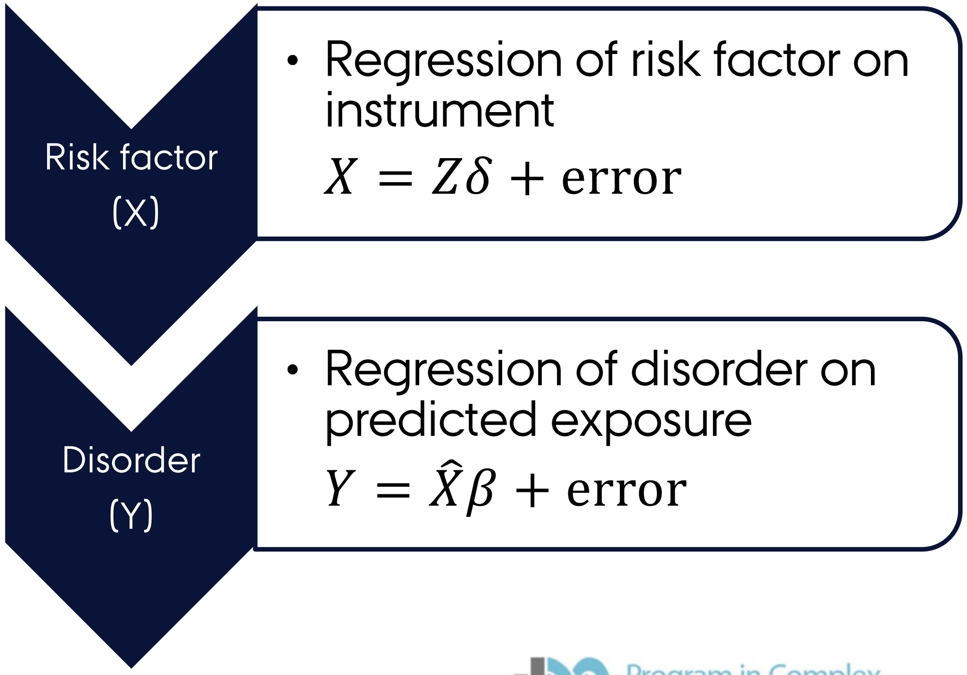
Note: Z should be associated with x, 1) $P_Z x \neq 0$, 2) attenuated effect

SNP instruments are independent of environmental factors, $Z^T e = 0$

$$\mathrm{E}\left(\hat{\beta}_{2LSL}\right) = \beta$$

MR using summary statistics

Individual-level data



Risk factor (X)

- Regression of risk factor on instrument

  $X = Z\delta + \text{error}$

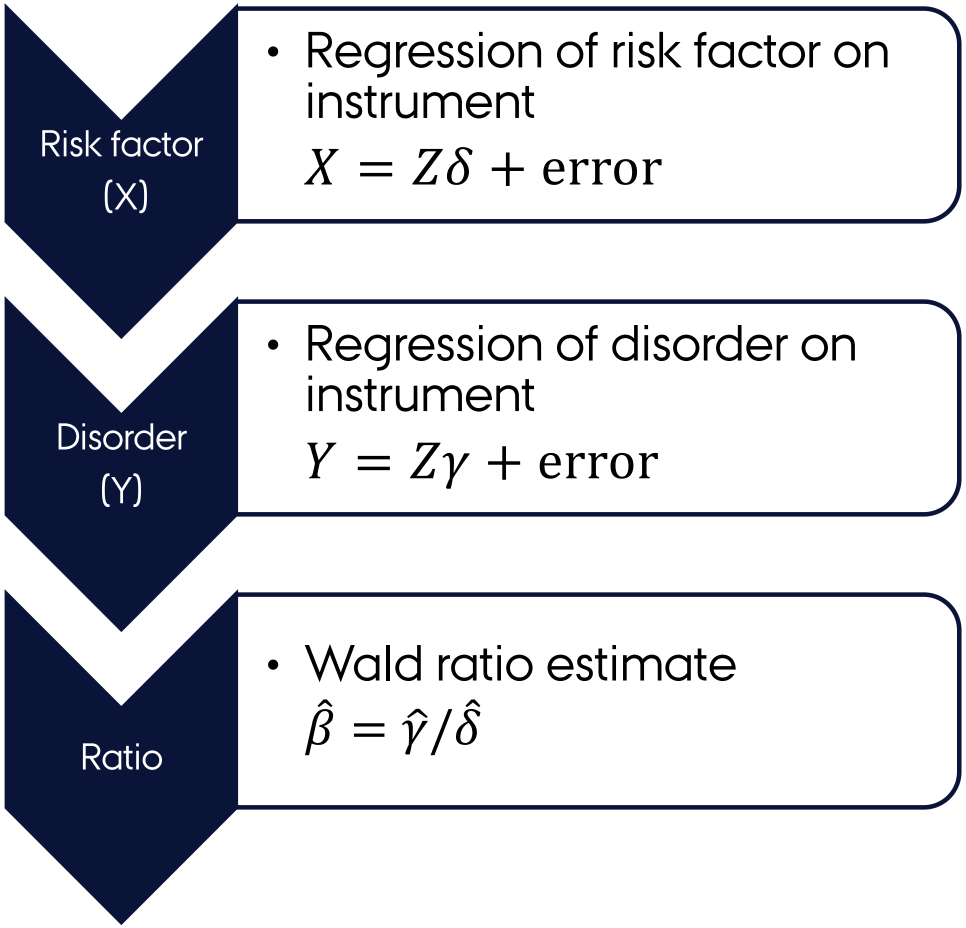
Disorder (Y)

- Regression of disorder on predicted exposure

  $Y = \hat{X}\beta + \text{error}$

=

Summary-level data



Risk factor (X)

- Regression of risk factor on instrument

  $X = Z\delta + \text{error}$

Disorder (Y)

- Regression of disorder on instrument

  $Y = Z\gamma + \text{error}$

Ratio

- Wald ratio estimate

  $\hat{\beta} = \hat{\gamma}/\hat{\delta}$

Program in Complex Trait Genomics

Summary-data based method

$$\mathrm{E}\left(\hat{\beta}_{2LSL}\right) = (\hat{x}^T\hat{x})^{-1}\hat{x}^T y = \frac{x^T P_Z y}{x^T P_Z x} = (\hat{x}^T\hat{x})^{-1}\hat{x}^T\hat{y} = \hat{\gamma}/\hat{\delta}$$

For a single SNP instrument

$\hat{\delta}$ from mQTL, eQTL, sQTL, etc.

$\hat{\gamma}$ from GWAS etc.

Summary-data-based MR

	2LSL – single instrument	Summary-data-based MR
Data	Individual-level data	Summary-level data
Availability	May not be available	eQTL, GWAS, etc.

Risk gene - *CACNA2D4*

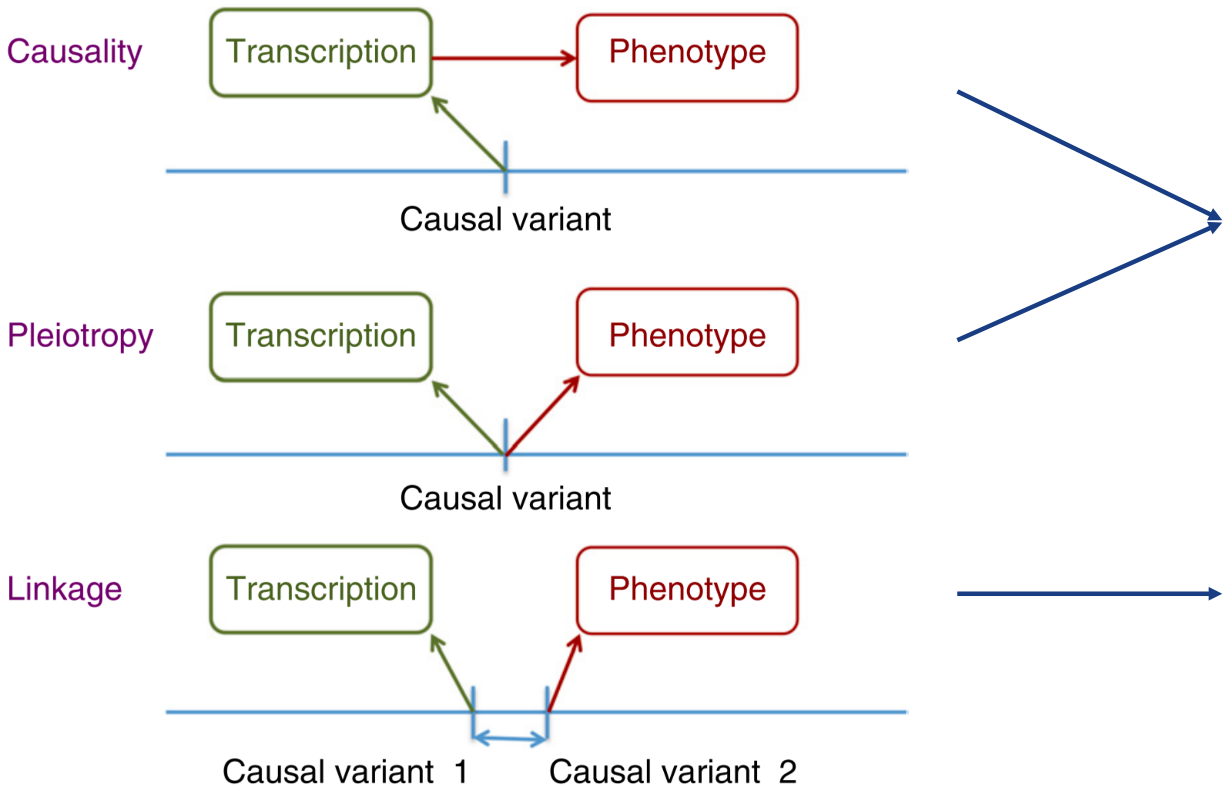
Gene	SNP	A1 / A2	Data	*b*	SE	*P*-value
CACNA2D4	rs1044825	G / T	eQTL (blood)	0.447	0.0186	4.1E-128
			GWAS (schizophrenia)	-0.0377	0.0087	1.3E-5

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

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

→ *P*-value = 2.0E-5

Linkage model

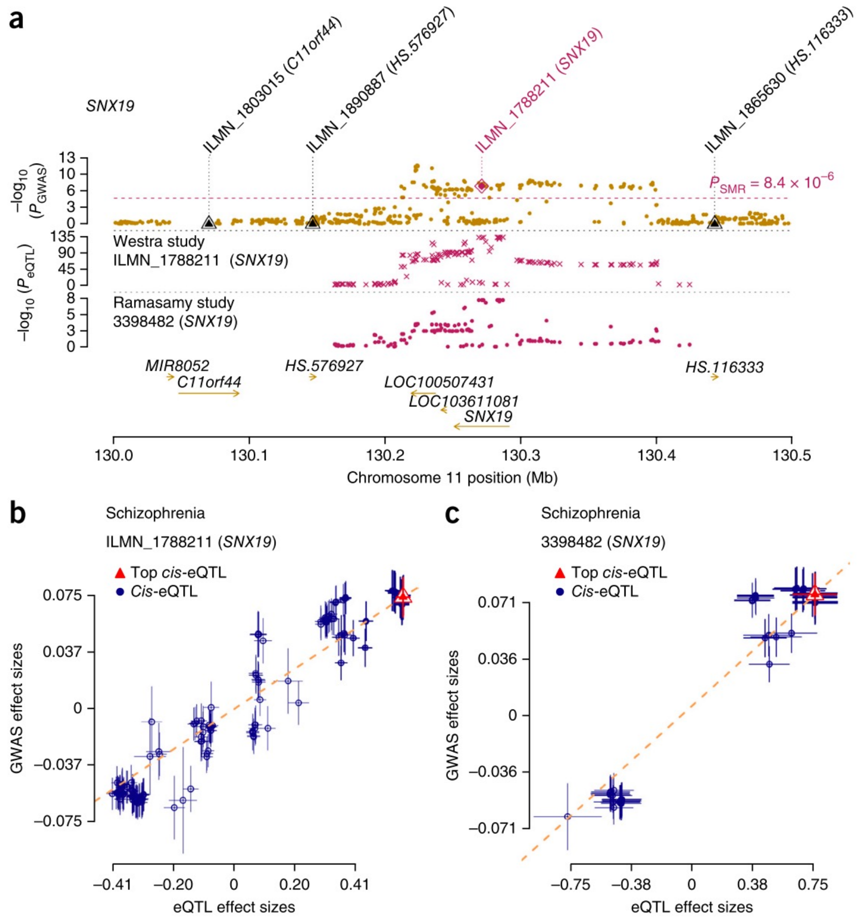


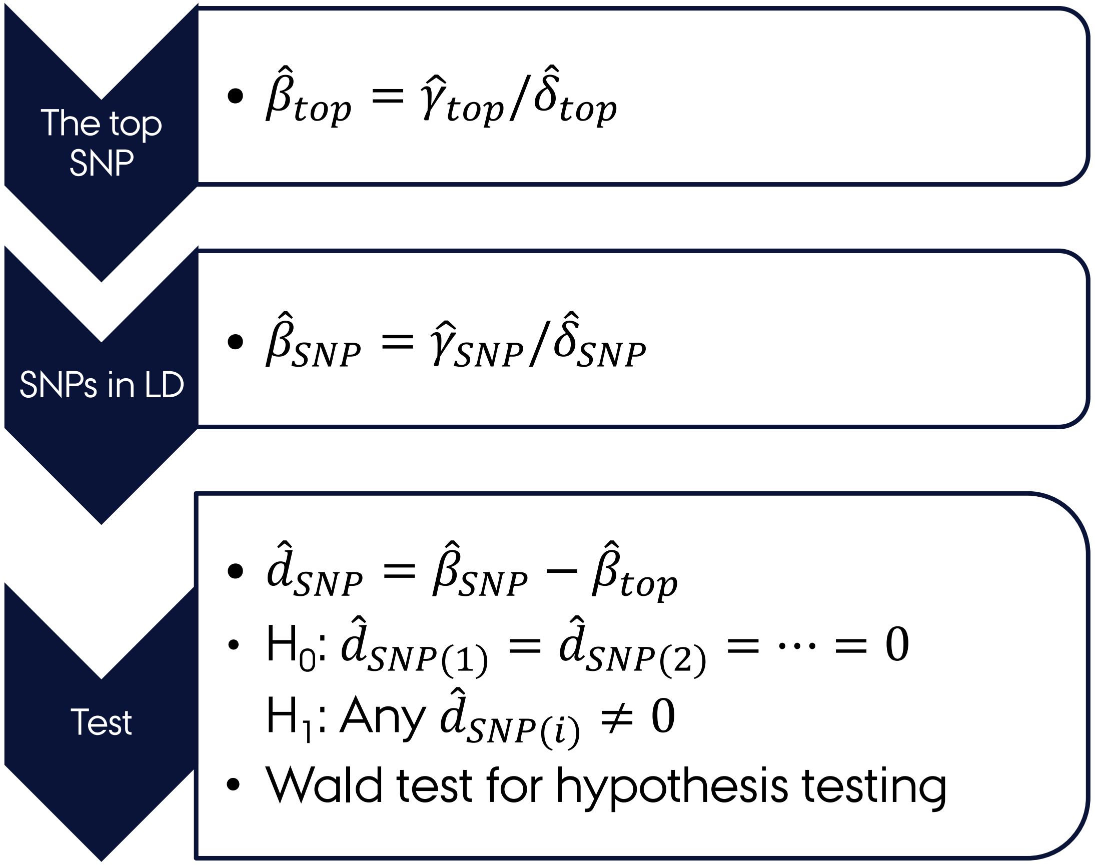
Unable to distinguish pleiotropy from causality

HEIDI (HEterogeneity In Dependent Instruments).

Using SNPs in LD with top *cis*-SNP

HEIDI





The top SNP

- $\hat{\beta}_{top} = \hat{\gamma}_{top}/\hat{\delta}_{top}$

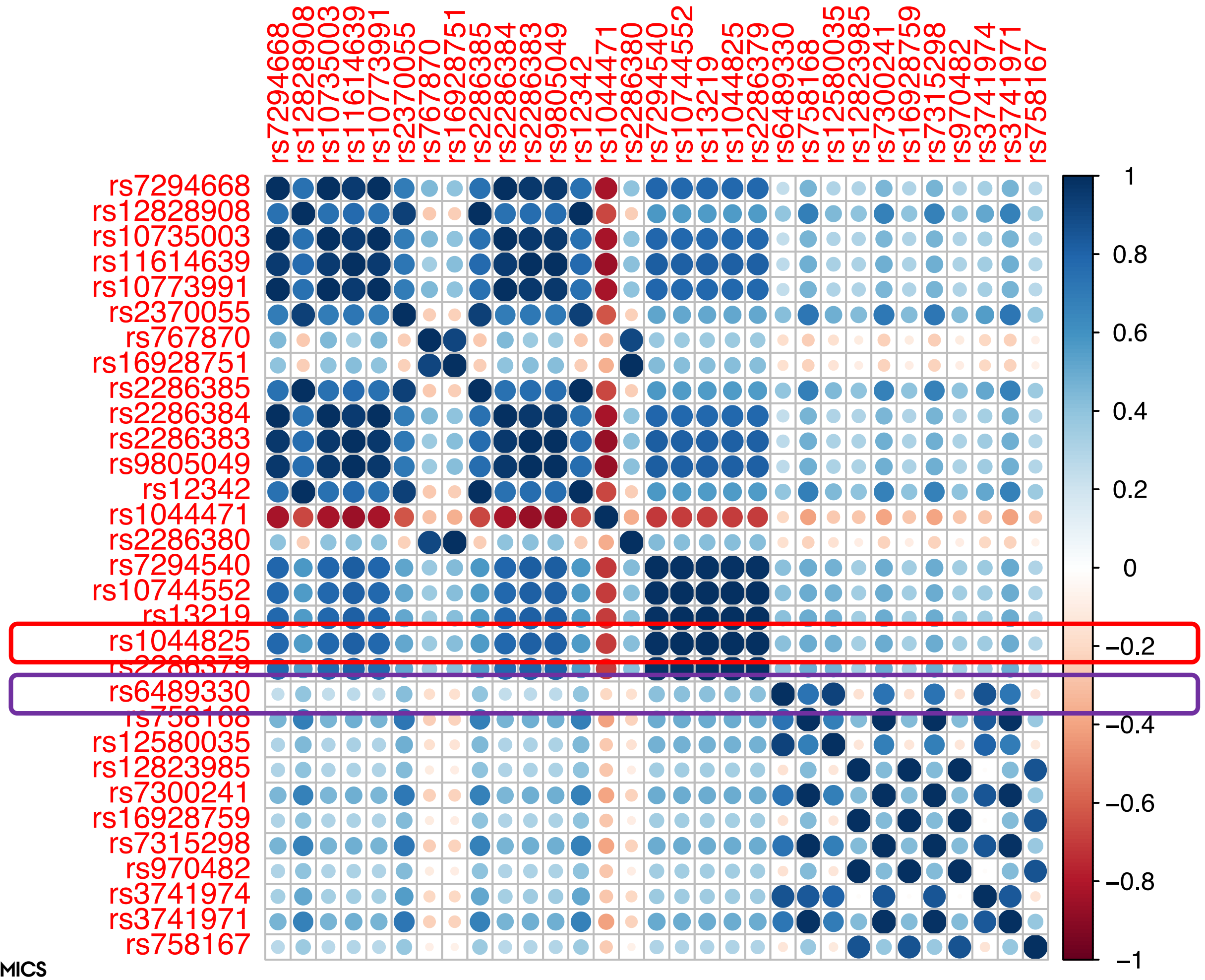
SNPs in LD

- $\hat{\beta}_{SNP} = \hat{\gamma}_{SNP}/\hat{\delta}_{SNP}$

Test

- $\hat{d}_{SNP} = \hat{\beta}_{SNP} - \hat{\beta}_{top}$
- H_0: $\hat{d}_{SNP(1)} = \hat{d}_{SNP(2)} = \cdots = 0$
  H_1: Any $\hat{d}_{SNP(i)} \neq 0$
- Wald test for hypothesis testing

Risk gene - *CACNA2D4*



The top-associated SNP

The SNP to test difference

Risk gene – *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, $\hat{\beta}_1 = -0.084$, $\mathrm{SE}\left(\hat{\beta}_1\right) \approx 0.020$

rs6489330, $\hat{\beta}_2 = -0.179$, $\mathrm{SE}\left(\hat{\beta}_2\right) \approx 0.055$

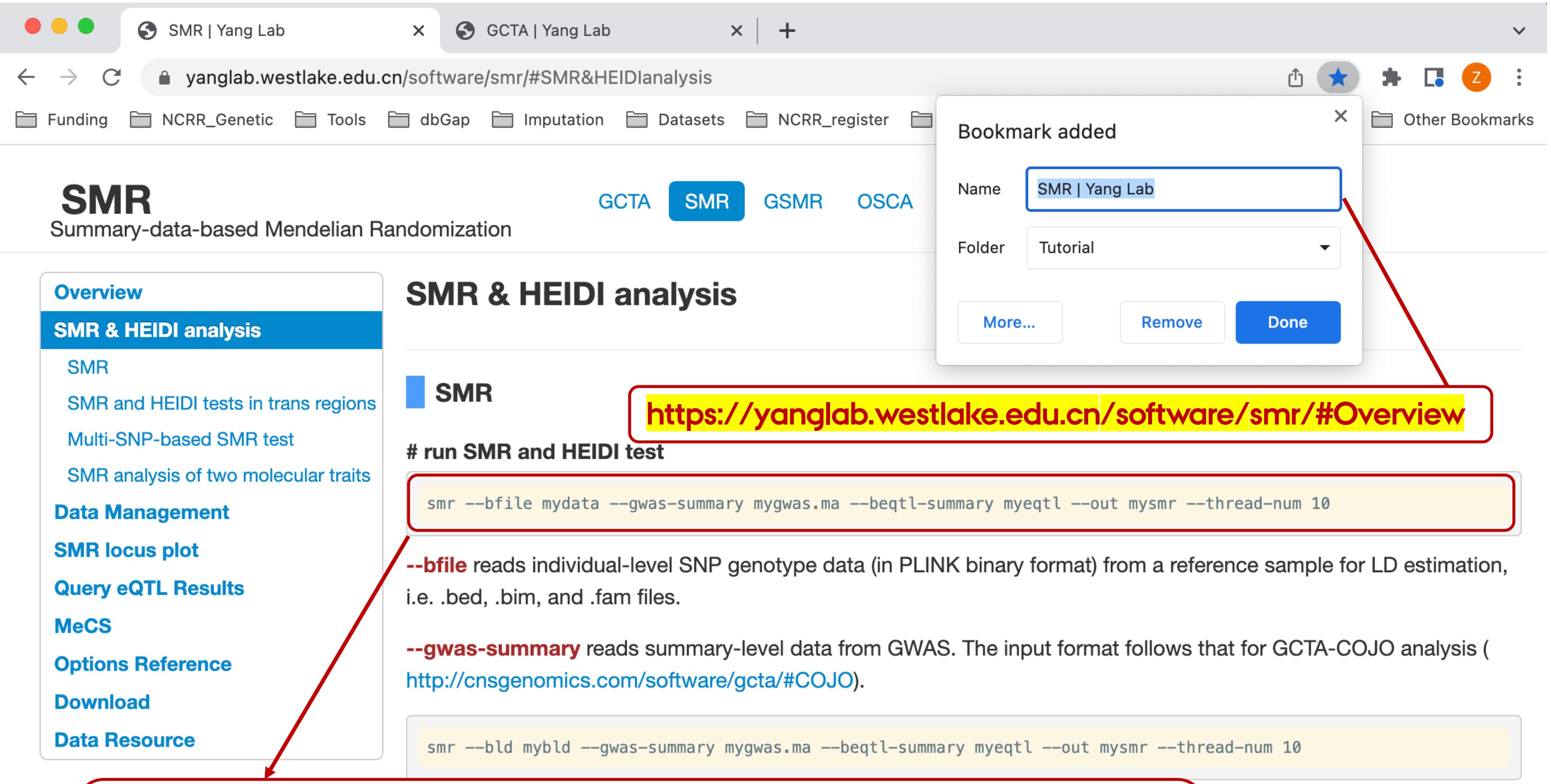
Difference, $\hat{d} = \hat{\beta}_2 - \hat{\beta}_1 = -0.179 + 0.084 = -0.095$

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

→ *P*-value = 0.06

Software

- SMR



SMR

Summary-data-based Mendelian Randomization

SMR & HEIDI analysis

SMR

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
```

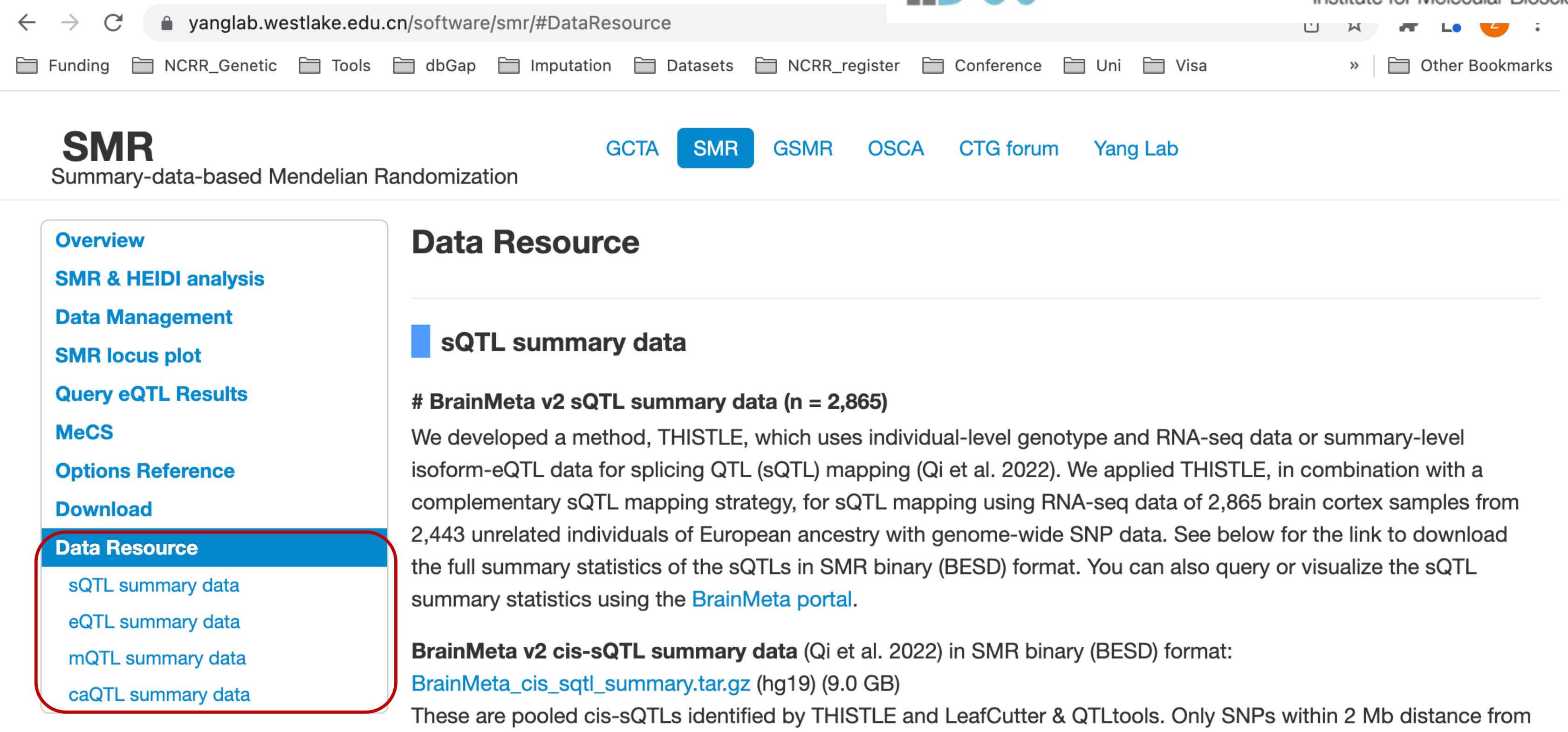
--bld reads LD information from a binary file in BLD format

https://yanglab.westlake.edu.cn/software/smr/#Overview

Command line:

**smr --bfile mydata --gwas-summary mygwas.ma --beqtl-summary myeqtl **
--out mysmr

SMR - Resources



Data Resource

sQTL summary data

# BrainMeta v2 sQTL summary data (n = 2,865)

We developed a method, THISTLE, which uses individual-level genotype and RNA-seq data or summary-level isoform-eQTL data for splicing QTL (sQTL) mapping (Qi et al. 2022). We applied THISTLE, in combination with a complementary sQTL mapping strategy, for sQTL mapping using RNA-seq data of 2,865 brain cortex samples from 2,443 unrelated individuals of European ancestry with genome-wide SNP data. See below for the link to download the full summary statistics of the sQTLs in SMR binary (BESD) format. You can also query or visualize the sQTL summary statistics using the BrainMeta portal.

BrainMeta v2 cis-sQTL summary data (Qi et al. 2022) in SMR binary (BESD) format:
BrainMeta_cis_sqtl_summary.tar.gz (hg19) (9.0 GB)
These are pooled cis-sQTLs identified by THISTLE and LeafCutter & QTLtools. Only SNPs within 2 Mb distance from

sQTL – Summary statistics of splicing QTLs

eQTL – Summary statistics from associations of gene expression

mQTL – Summary statistics from associations of methylation

…

Misuse of MR

- Assuming that study is performed in a population
  - Time-frame (youths vs adults)
  - Sex (males vs females)
  - Environment (e.g. low altitude vs high altitude)

- Tissue
  - Blood – the largest sample size, shared effects with other tissues
  - Mental disorders - brain
  - BMI – adipose
  - …

Summary

- Regression – bias due to environmental confounding factor
- Mendelian randomisation - similar concept to randomised controlled trial
  - RCT is the gold-standard approach
  - using genetic variant (e.g. SNP) as instrument
  - instrument should be strongly associated with exposure
  - 2SLS – individual-level data
  - Summary-data-based method – summary-level data
- Genetic architecture
  - Large genetic variation at a single SNP, large LD blocks
    CACNA2D4 -> schizophrenia
- SMR method
  - SMR – using a single SNP instrument
  - HEIDI – distinguishing linkage model from pleiotropy model
  - Misuse of SMR

Data agreement

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations.

Please send an email to pctgadmin@imb.uq.edu.au with your name and the below statement to confirm that you agree with the following:

"I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing account. "

Practical

- Software
  - SMR V1.3.1
- Data
  - eQTL dataset - the Westra eQTL data, Westra et al. 2013 Nature Genetics
  - GWAS dataset – GWAS of schizophrenia, Trubetskoy et al. 2022 Nature
  - LD reference cohort

eQTL dataset

- SMR format
  - .besd - summary statistics of eQTL dataset
  - .epi – probes

1	ILMN_1653466	0	934380	HES4	-
1	ILMN_2349633	0	1140818	TNFRSF18	-
1	ILMN_2112256	0	1146750	TNFRSF4	-
......					

  - .esi - SNPs

1	rs3131968	0	754192	A	G
1	rs2905035	0	775659	A	G
1	rs2980319	0	777122	A	T
......					

GWAS dataset

- COJO format

SNP	A1	A2	FREQ	BETA	SE	P	N
rs62513865	C	T	0.927	0.0119977384336167	0.0171	0.4847	58749.13
rs79643588	G	A	0.906	-0.00859684722551828	0.0148	0.5605	58749.13
rs17396518	T	G	0.566	-0.0021022080918702	0.0087	0.8145	58749.13
…							

Command

- LD reference cohort (PLINK format)

- Command

  CACNA2D4 -> schizophrenia

```
smr \
    --bfile ld_reference \
    --gwas-summary sz_2022.ma \
    --beqtl-summary westra \
    --out smr_westra_sz
```

Thank you!

Zhihong Zhu, Ph.D

z.zhu@econ.au.dk | z.zhu1@uq.edu.au

AARHUS BSS
DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

Program in Complex
Trait Genomics

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
IMB
Institute for Molecular Bioscience

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

EQUIS ACCREDITED

AARHUS
BSS

DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY