

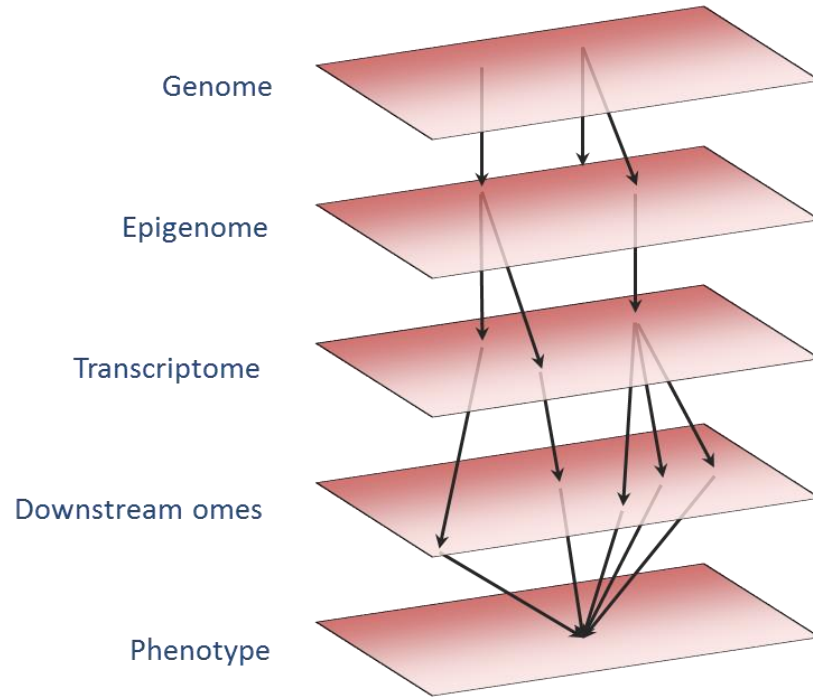
Analyzing multiple omics levels - Mendelian randomization

Bas Heijmans
Molecular Epidemiology
Dept. of Biomedical Data Sciences
Leiden University Medical Center
The Netherlands
bas.heijmans@lumc.nl



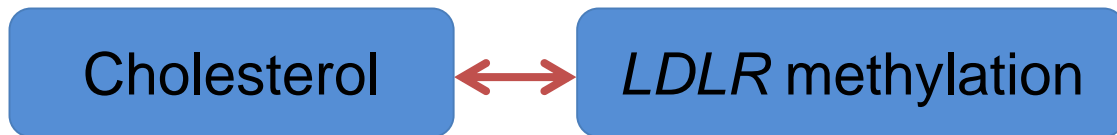
Koen Dekkers

The single most important distinction in study designs

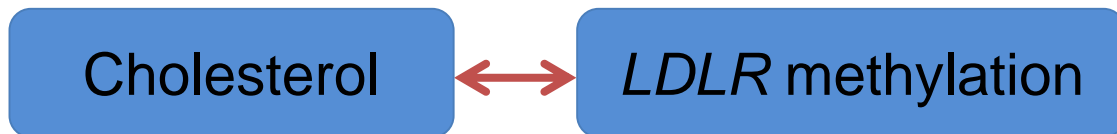


Epigenome-wide Association Study

1. Data: methylation at 450 thousand CpGs + lipids levels in 2000 individuals
2. Test per CpG: $\text{DNAm} \sim \text{cholesterol} + \text{sex} + \text{age} + \text{cell counts} + \text{batches}$



What's next?



- Can we make conclusions stronger?
- What are the main limitations in observational epidemiology?

What's next?

Cholesterol



LDLR methylation

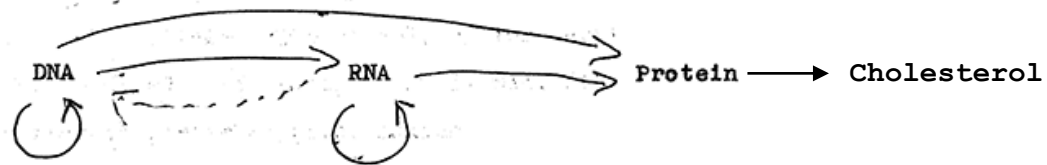
Alternative for experiment

Cholesterol

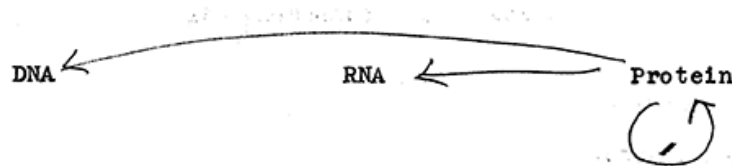


LDLR methylation

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have



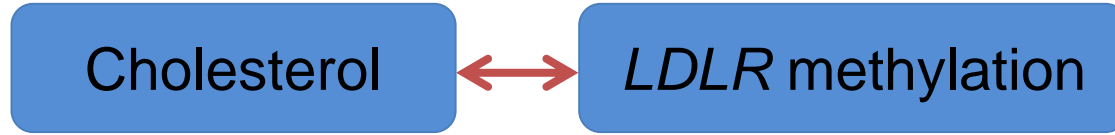
but never



where the arrows show the transfer of information.

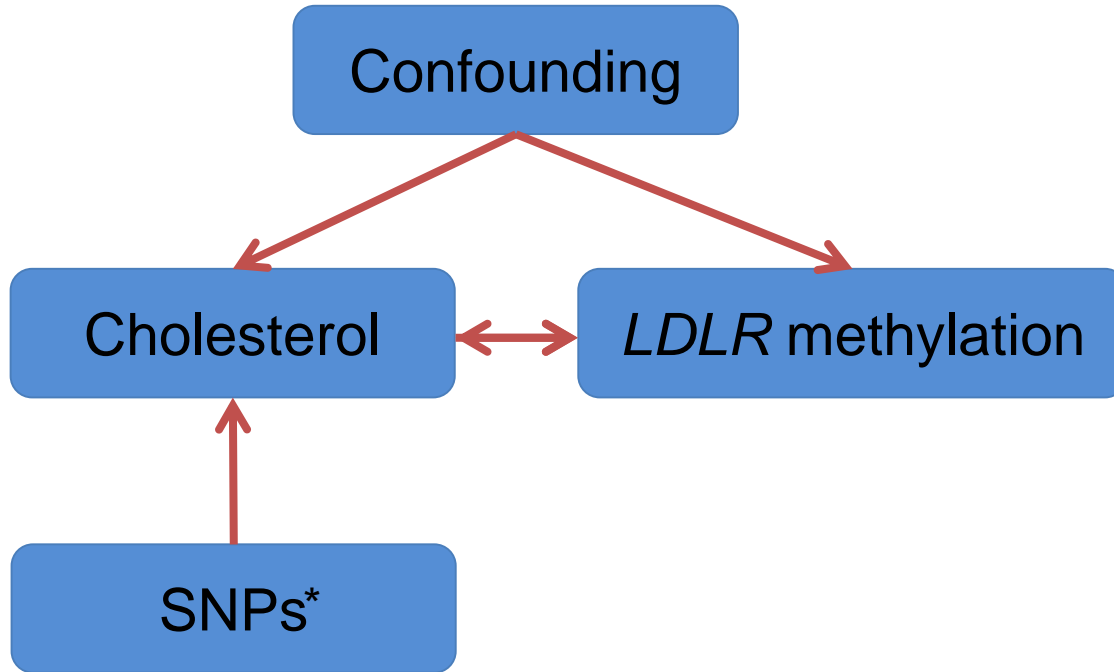
Crick, 1956

Alternative for experiment



- An experiment of nature using genetic variation as causal anchor
- ‘Mendelian randomization’: a natural trial with exposures to genetic variations randomized according to Mendel’s law and with the exposed blinded towards exposure.
- Uses genetic variant(s) as ‘instrumental variable’ instead of measured variable itself.

Alternative for experiment



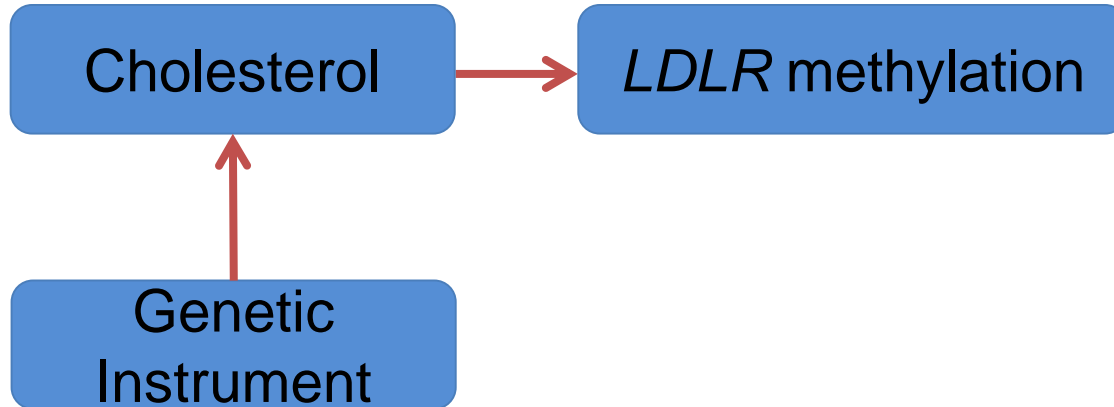
**Which we know from GWASs*

Mendelian randomization

1. Identify (sufficiently strong) genetic instrument.
Here: SNP associated with cholesterol from GWAS (see last week's practical)
2. Predict cholesterol level for every individual on basis of one's genotype.
Will explain only small proportion of variation
3. Test whether predicted (genetic) level is associated with methylation.
No confounding (unbiased effect estimate) & only one possible direction of causality

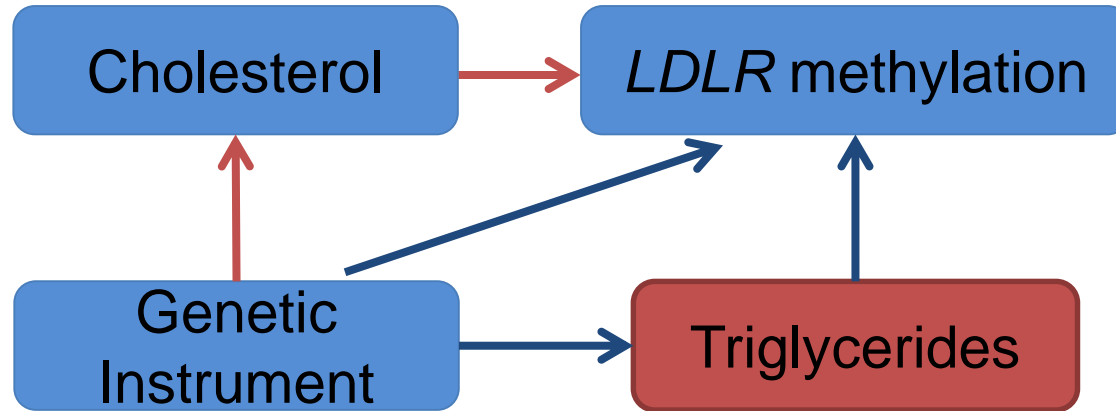
Two-stage least squares model

1. $\text{Pred(Chol)} = \gamma_0 + \gamma_1 \times \text{genotype} + \gamma_2 \times \text{age} + \gamma_3 \times \text{batch} + \dots$
2. $\text{DNAm} = \beta_0 + \beta_1 \times \text{pred(Chol)} + \beta_2 \times \text{age} + \beta_3 \times \text{batch} + \dots$



Beware of assumptions

Pleiotropy



Example: good and bad cholesterol

HDL

LDL

- What is the evidence for being good and bad?
- Do you know medication targeting good or bad cholesterol?

MR – validation PolyGenic Scores

Unrestricted

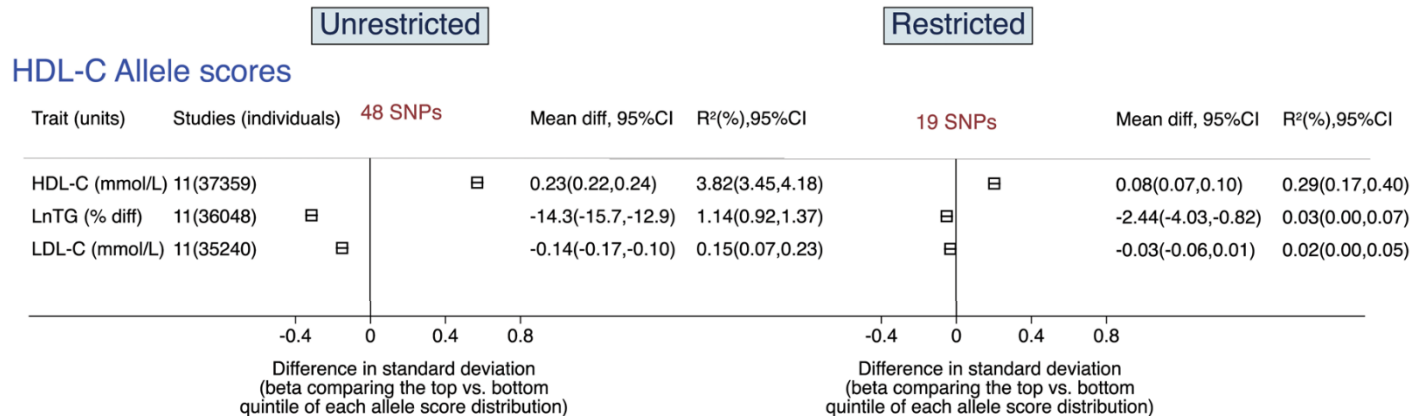
HDL-C Allele scores

Trait (units)	Studies (individuals)	48 SNPs	Mean diff, 95%CI	R ² (%),95%CI
HDL-C (mmol/L)	11(37359)	⊖	0.23(0.22,0.24)	3.82(3.45,4.18)
LnTG (% diff)	11(36048)	⊖	-14.3(-15.7,-12.9)	1.14(0.92,1.37)
LDL-C (mmol/L)	11(35240)	⊖	-0.14(-0.17,-0.10)	0.15(0.07,0.23)

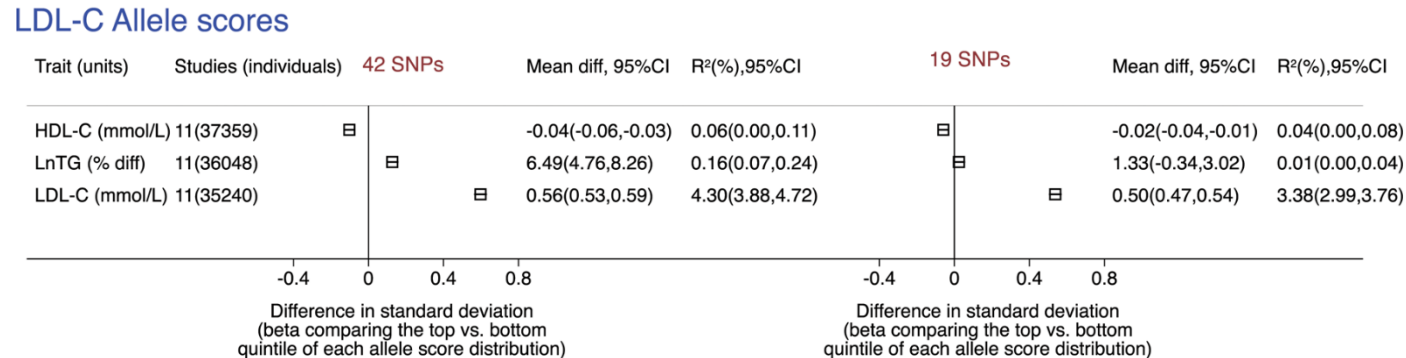
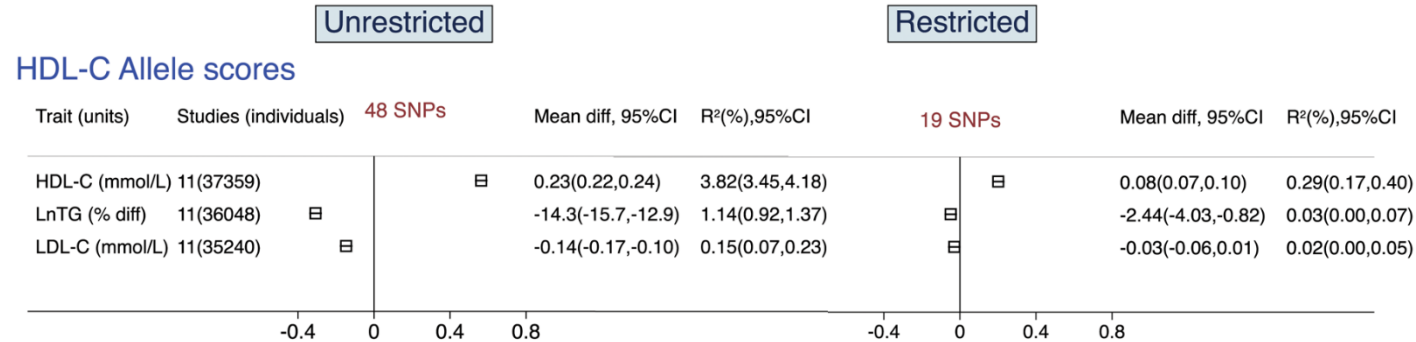
-0.4 0 0.4 0.8

Difference in standard deviation
(beta comparing the top vs. bottom
quintile of each allele score distribution)

MR – validation PolyGenic Scores



MR – validation PolyGenic Scores



$$1. \text{Pred(Lipid)} = Y_0 + Y_1 \times \text{PGS} + \dots$$

PolyGenic Scores (PGSs)

- Find a GWAS publication with SNPs and effect sizes.
- Have genotypes for your own study.
- Count the number of minor alleles an individual has.
- Multiply this number by the effect size of the allele.
- The result is the genetically predicted level for an individual.

Constructing PGS

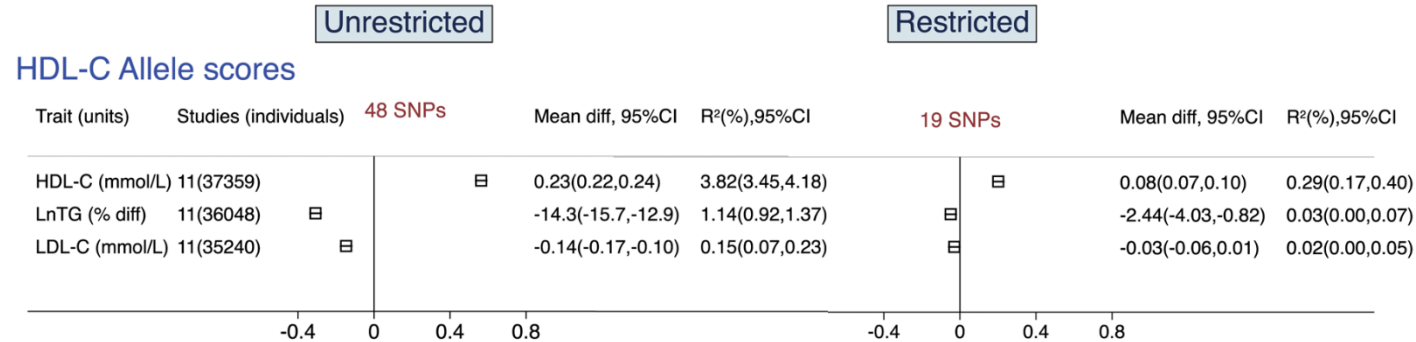
Nearest gene	MarkerName	Chr	hg19 Position (Mb)	Primary trait, Secondary trait(s)	MAF	Alleles minor/major	Minor allele effect size	Joint N (in 1000s)	Joint P-value
							↓ Effect		
Loci Primarily Associated with HDL Cholesterol									
LCAT	rs16942887	16	67.93	HDL	.14	A/G	.083	186	8x10 ⁻⁵⁴
CMIP	rs2925979	16	81.53	HDL	.31	T/C	-.035	186	1x10 ⁻¹⁹
STARD3	rs11869286	17	37.81	HDL	.35	G/C	-.032	178	3x10 ⁻¹⁷

Let genotypes for individual 'Harry' be:

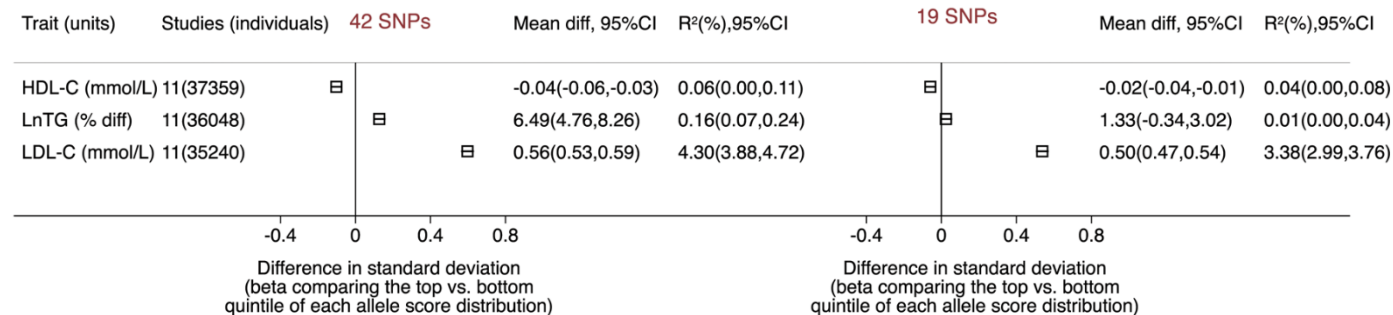
GG (*LCAT*), TC (*CMIP*), GG (*STARD3*)

- What is his PGS (or genetically predicted HDL level)?

MR – validation PolyGenic Scores



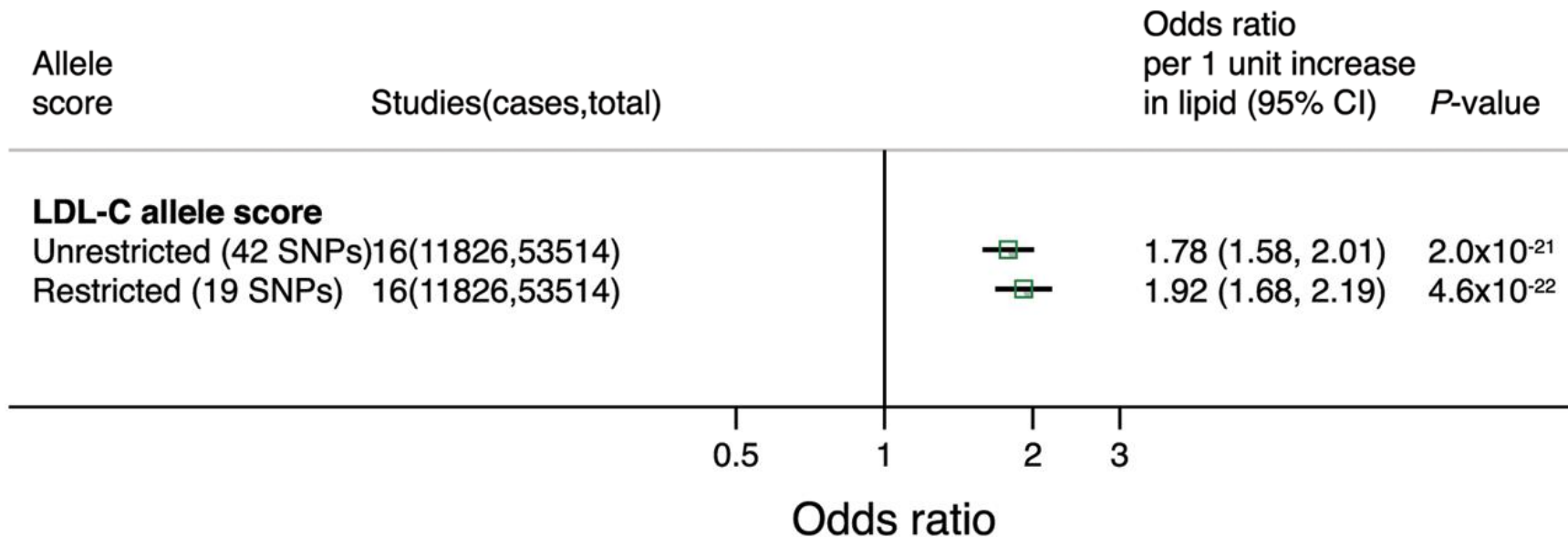
LDL-C Allele scores



$$1. \text{Pred(Lipid)} = Y_0 + Y_1 \times \text{PGS} + \dots$$

MR – causal inference

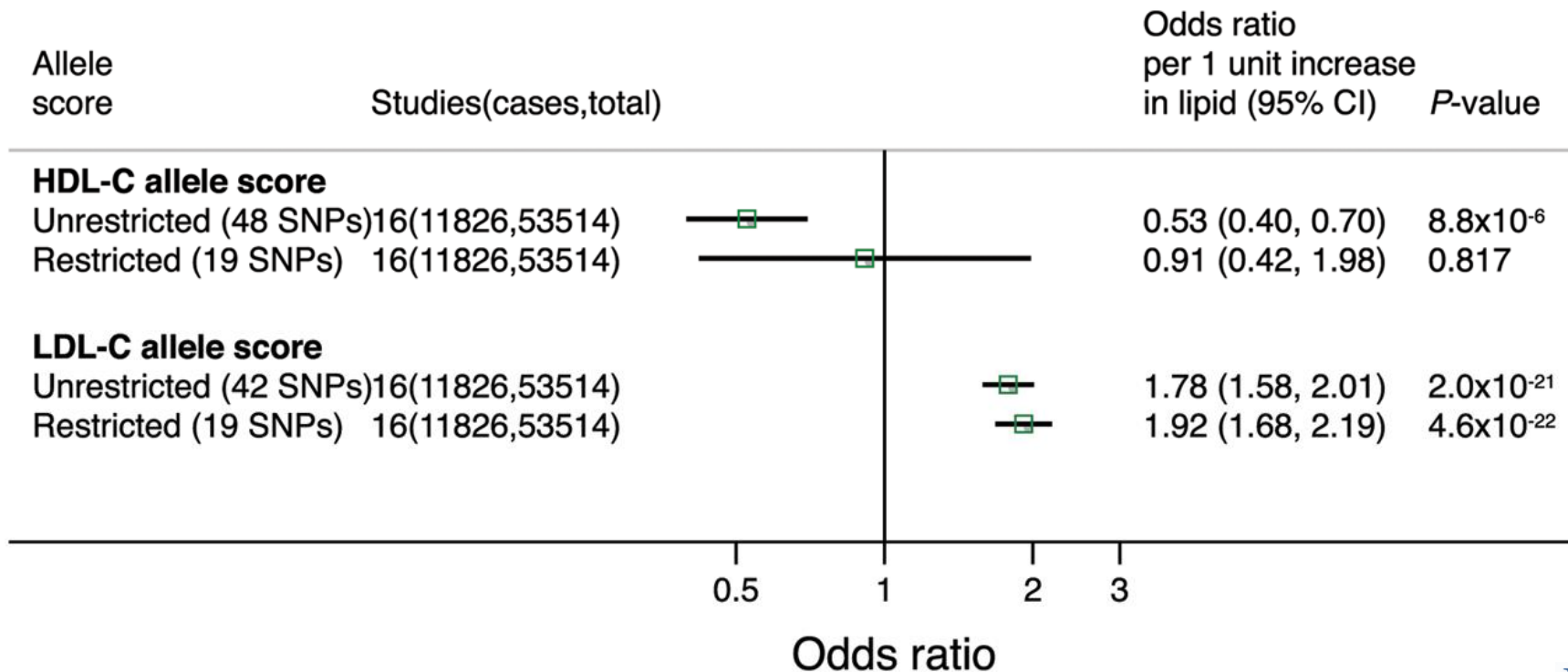
CHD (incident/prevalent)



$$2. \text{CHD} = \beta_0 + \beta_1 \times \text{pred}(\text{Lipid}) + \dots$$

MR – causal inference

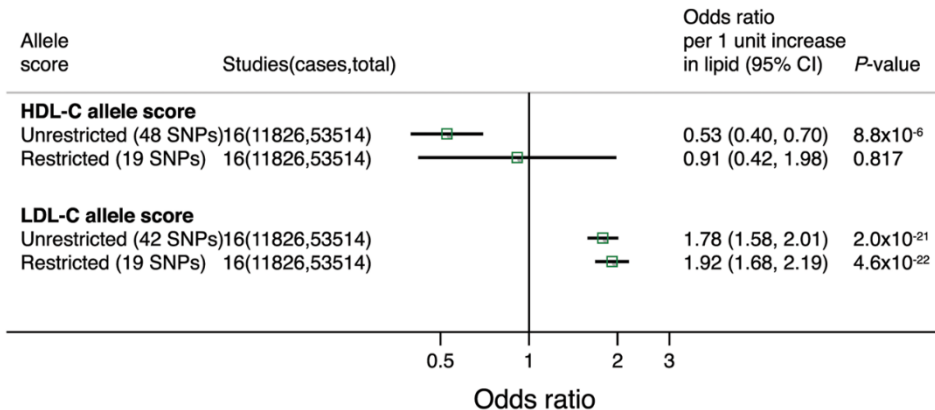
CHD (incident/prevalent)



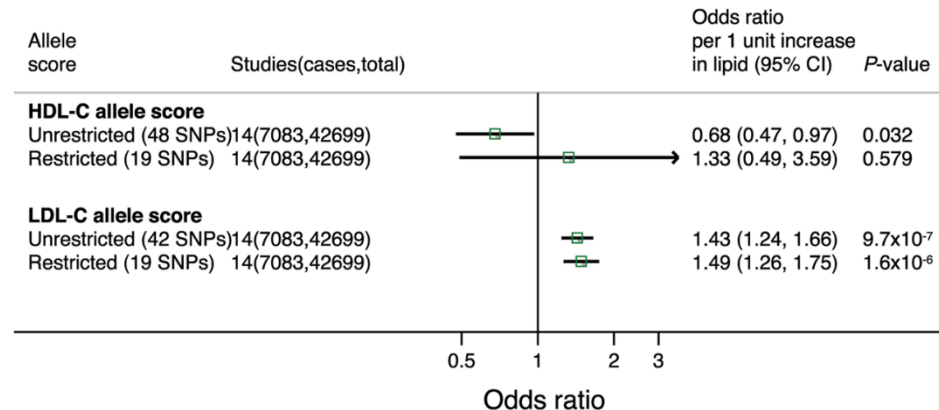
2. CHD = $\beta_0 + \beta_1 \times \text{pred}(\text{Lipid}) + \dots$

Mendelian randomization

CHD (incident/prevalent)



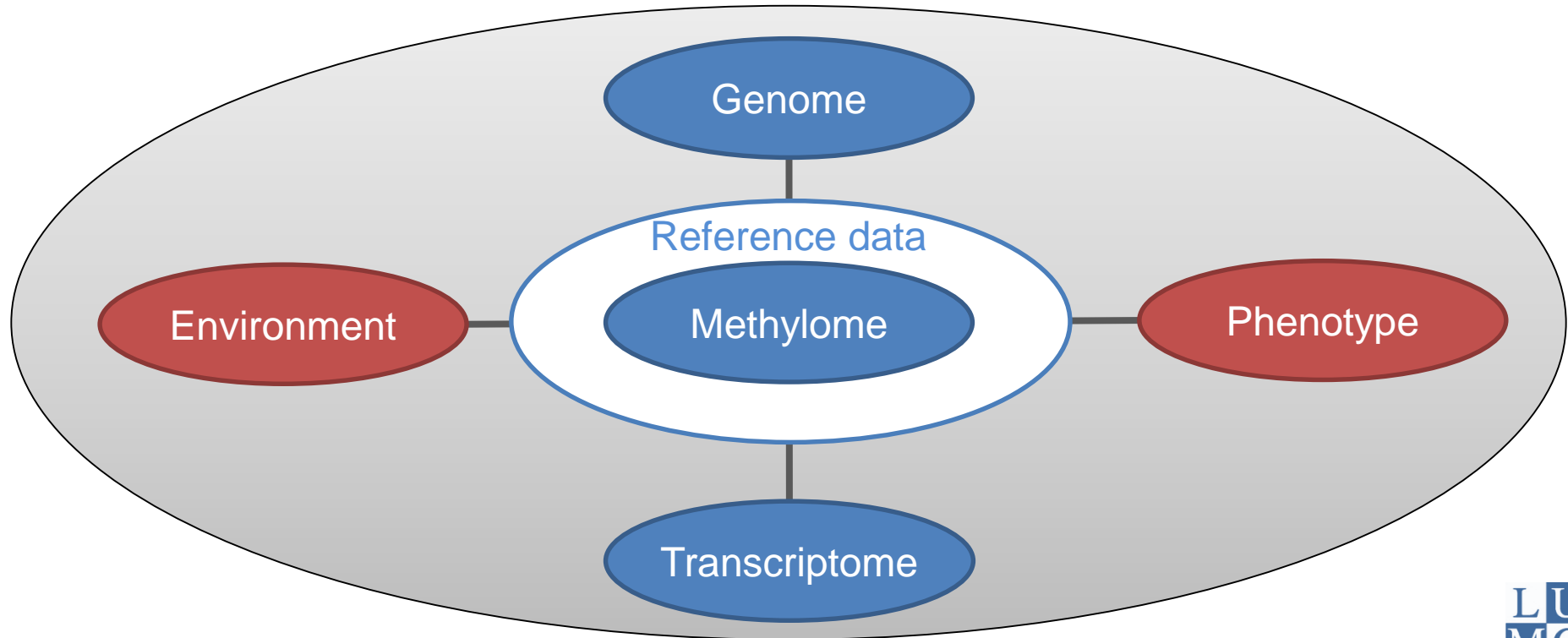
CHD (incident only)



HDL and LDL in human studies

- Observational: cross-sectional & prospective case/control studies
- Experimental: drugs in clinical trial
- 'Natural experiment': Mendelian randomization

Integrative Genomics



Molecular Data Science in populations

The human as 'model organism':

Exploiting natural variation in large-scale population studies

- Genome biology
- Disease mechanisms
- Biomarkers

