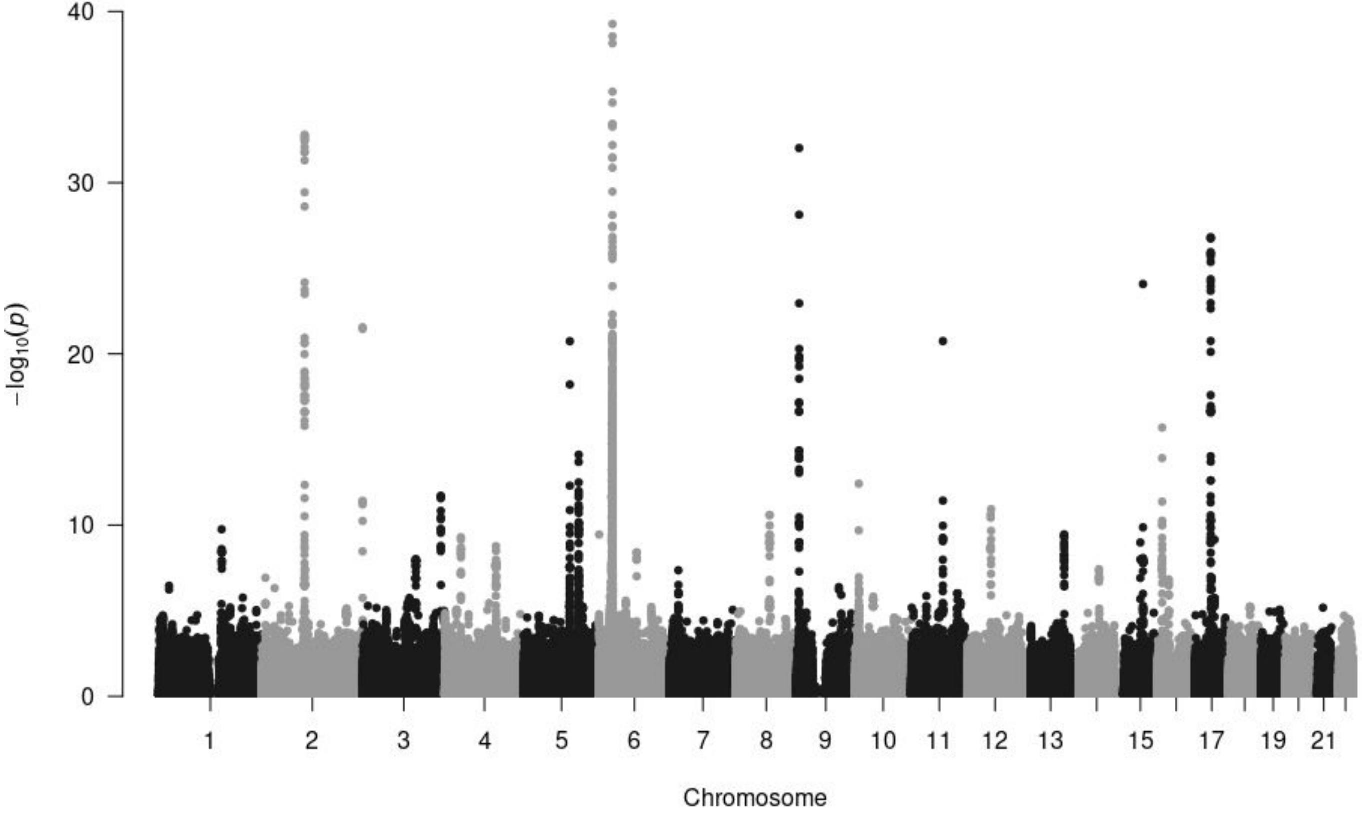
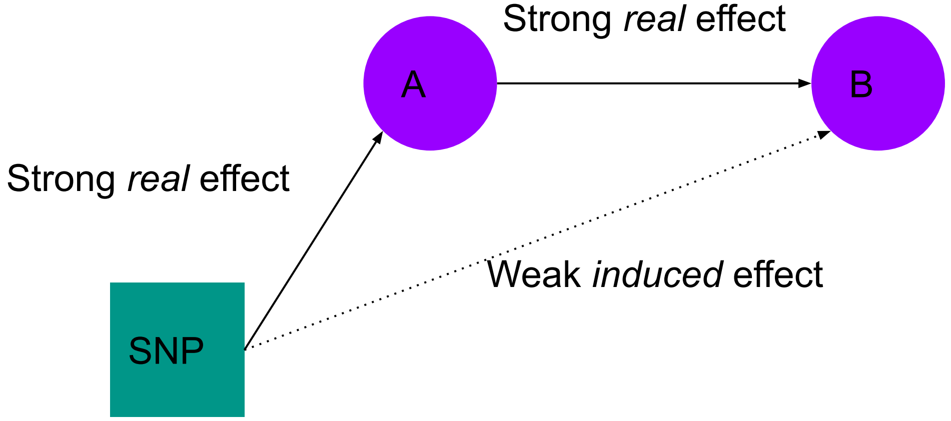
Phenome-scale directed network discovery with bi-directional mediated Mendelian randomization

Brielin C. Brown
Data Science Institute Fellow, Columbia University
Knowles and Lappalainen labs, New York Genome Center

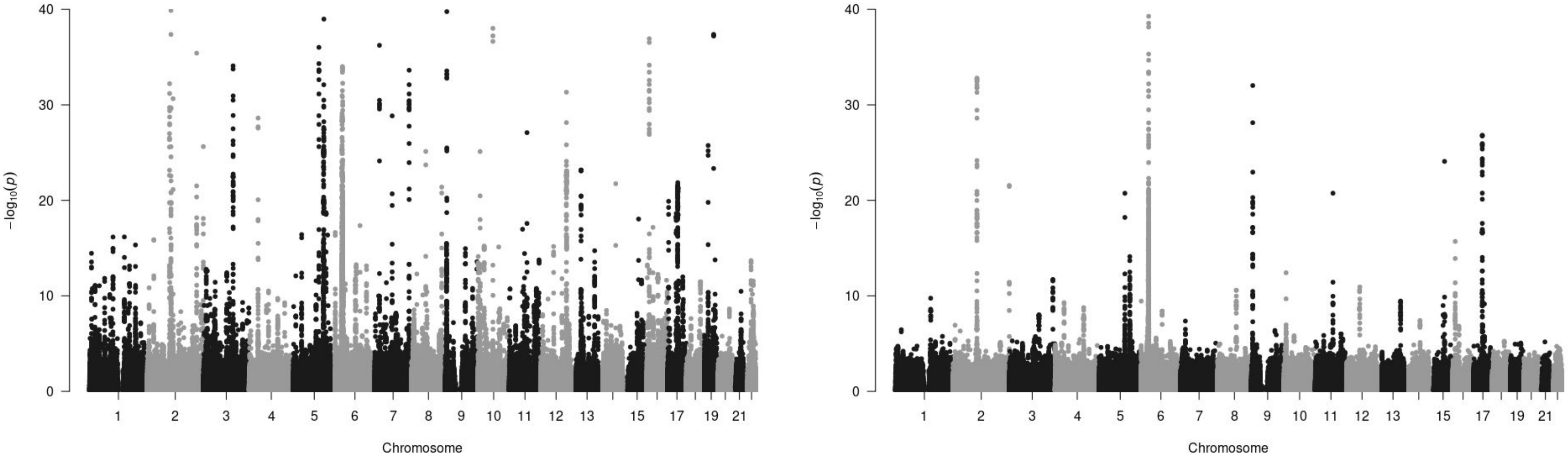
A GWAS

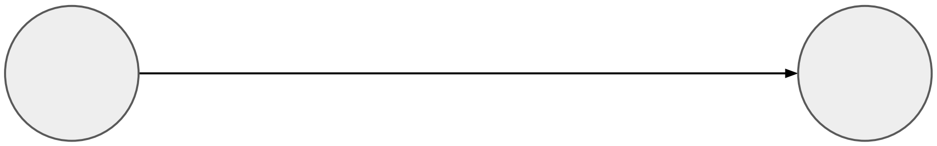

40
30
20
10
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
Chromosome

Some of the polygenic background comes from effects of other traits



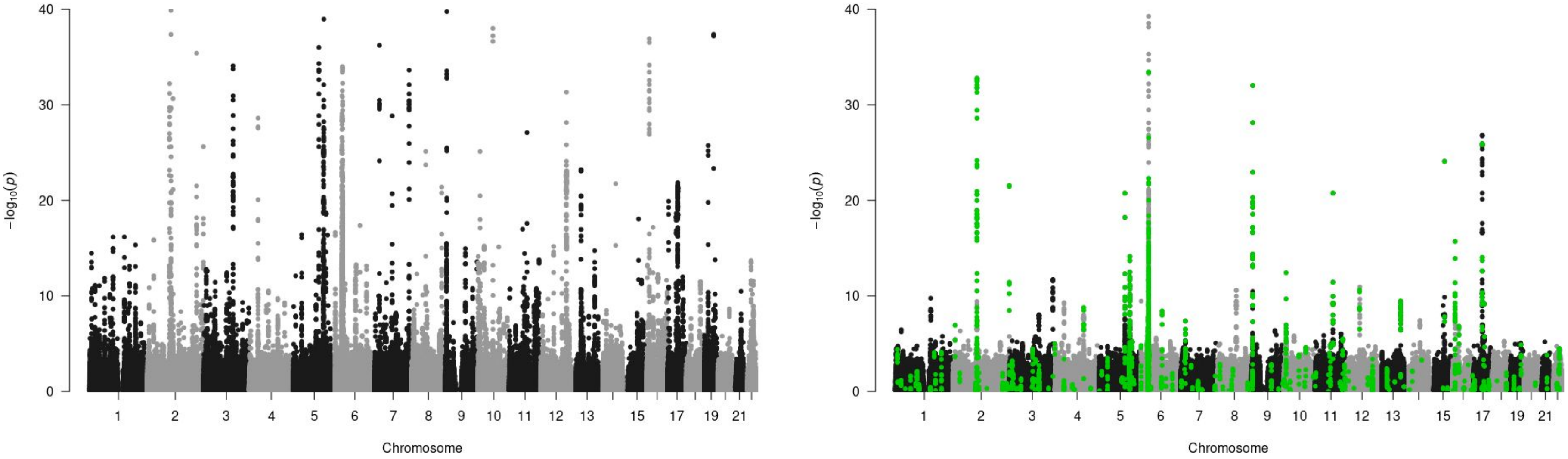
Y-axis clipped at 40 for comparison*

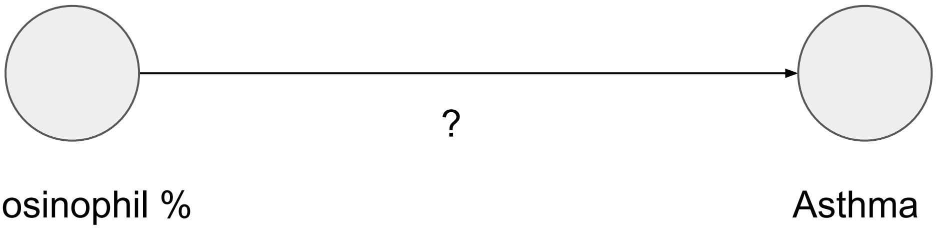




Asthma

Y-axis clipped at 40 for comparison*

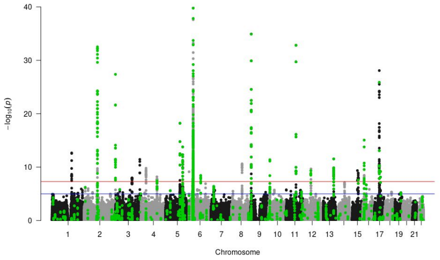


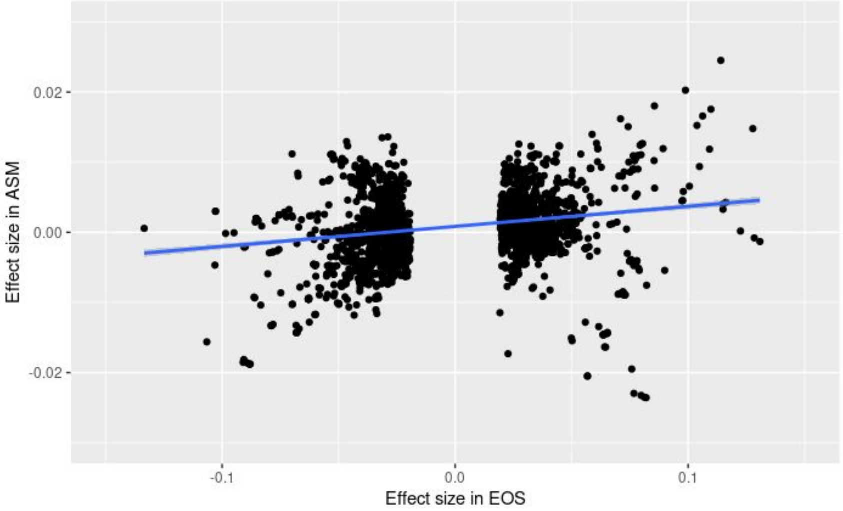

?

Eosinophil %

Asthma

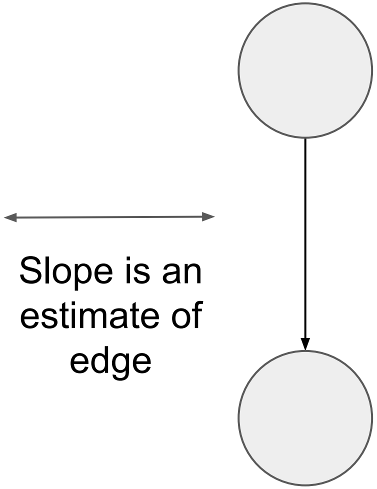
Mendelian Randomization is used to estimate the effect



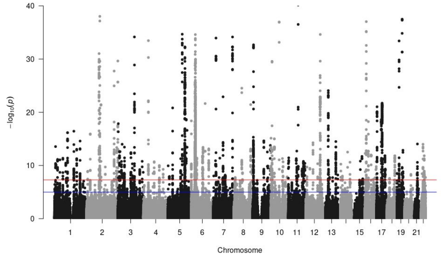


Slope is an estimate of edge

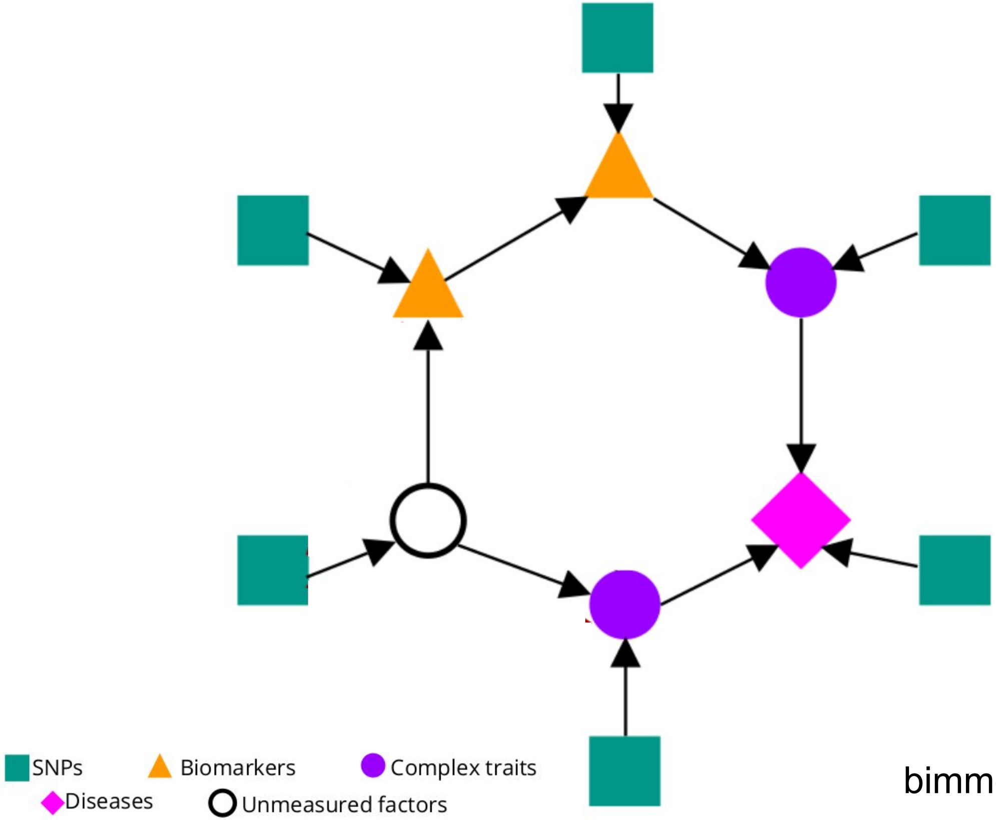
Eosinophil %



Asthma



Phenotypes live in directed networks



biobank uk

Trait Class	**Example**
Blood biomarker	Cholesterol
Blood composition	White blood cell count
Blood trait	Mean sphered cell volume
Immune-related disease	Psoriasis
Heart-related	Illness of father: heart disease
Other disease	Basal cell carcinoma
Morphological	Hip circumference
Dietary	Alcohol intake frequency
Behavioral	Getting up in morning
Neurological	Anxious feelings
Eye-related	3mm weak meridian
Other	Average income before tax

bimmer = estimation of directed networks in biobanks

A simple model can capture network and genetic effects

$$Y_j = \sum_{i \neq j} Y_i G_{ij} + \sum_m X_m \beta_{mj} + \gamma$$

Phenotype

Effects of other phenotypes

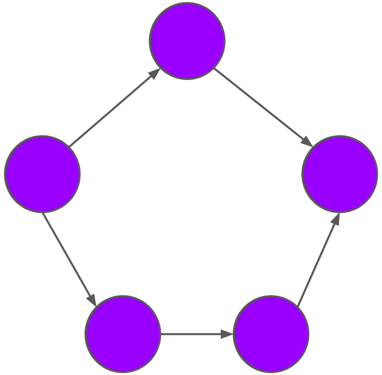
Genetic effects

Unmeasured factors

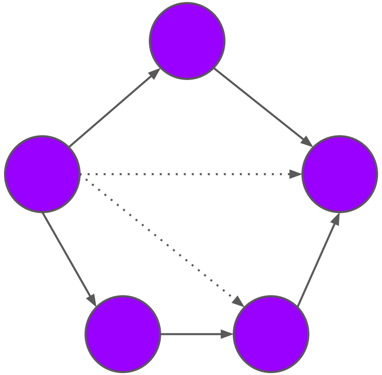
$$Y = YG + X\beta + \gamma$$

The graph can be determined from the total effects

G:



R:



$$Y = YG + X\beta + \gamma$$

$$\rightarrow G = I - R^{-1} D[1/R^{-1}]$$

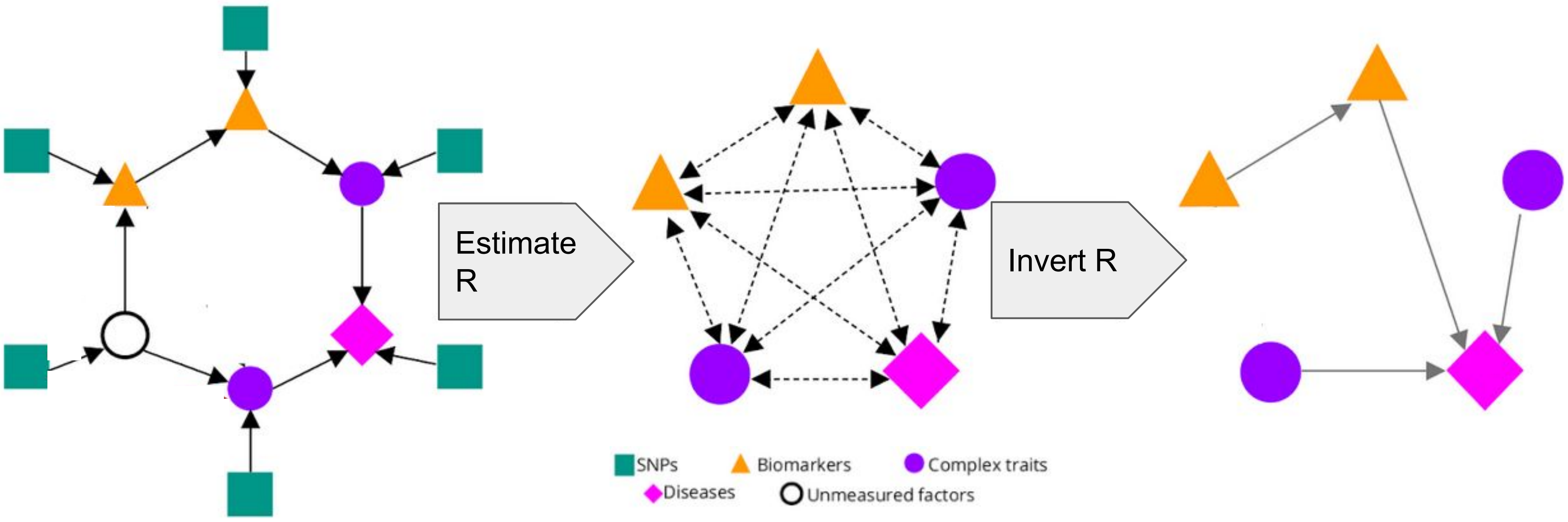
R: Total effects matrix (important!)
R_{ij} = Effect of trait i on j including all paths
Estimated through MR

D[A]: Diagonal of A

Gaussian graphical models: $\Omega \cong \Sigma^{-1}$

This result implies a two-step strategy for estimating G

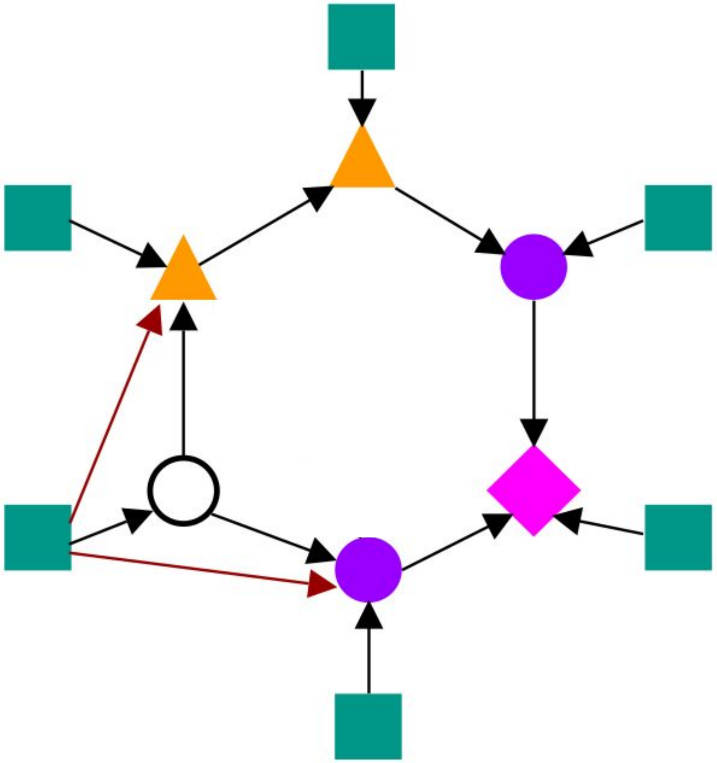
$$G = I - R^{-1} D[1/R^{-1}]$$

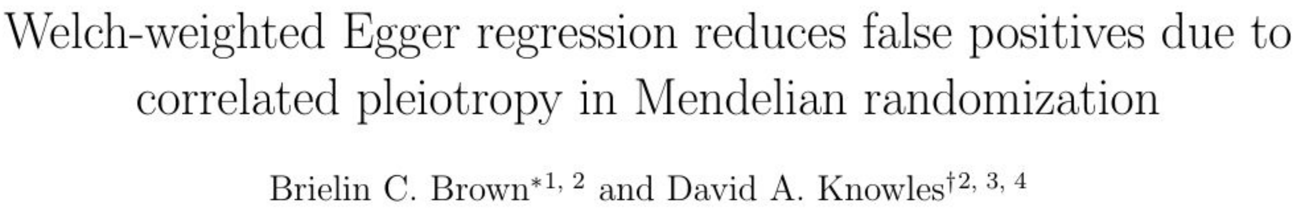


Estimate R

Invert R

Problem #1: not all traits are observed

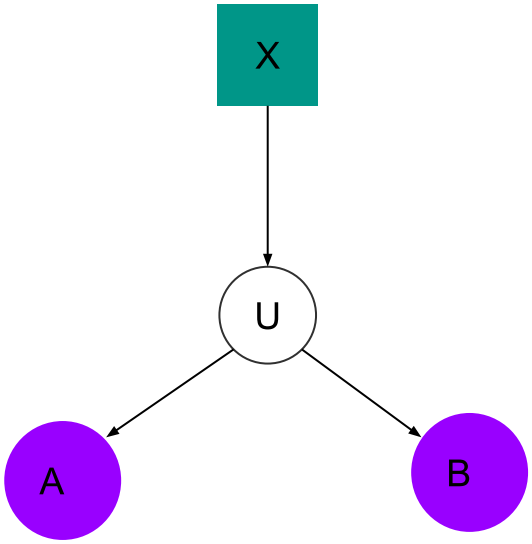


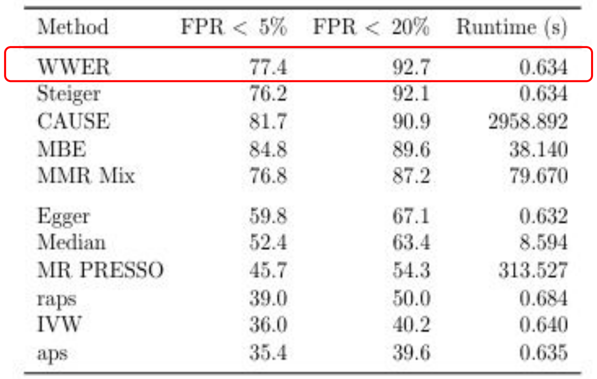


Welch-weighted Egger regression reduces false positives due to correlated pleiotropy in Mendelian randomization

Brielin C. Brown*[1, 2] and David A. Knowles†[2, 3, 4]

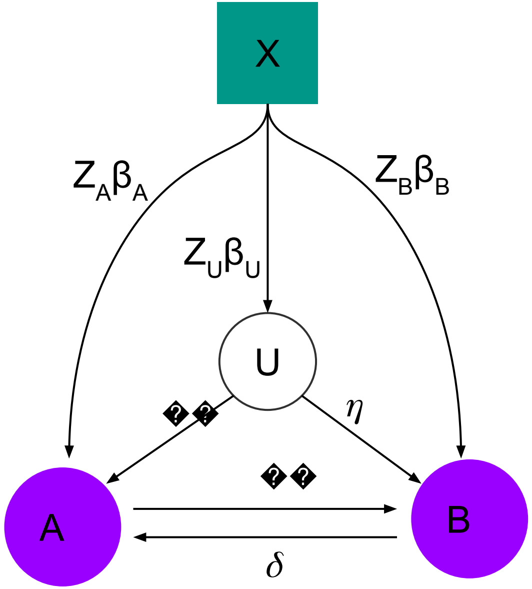
On bioRxiv now!

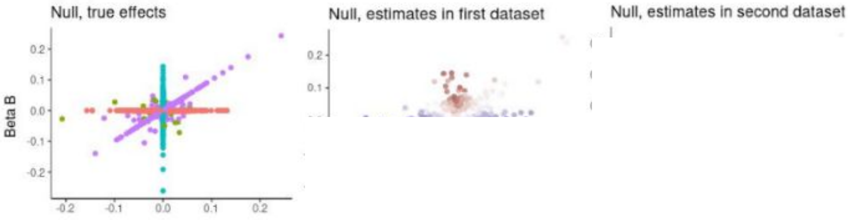


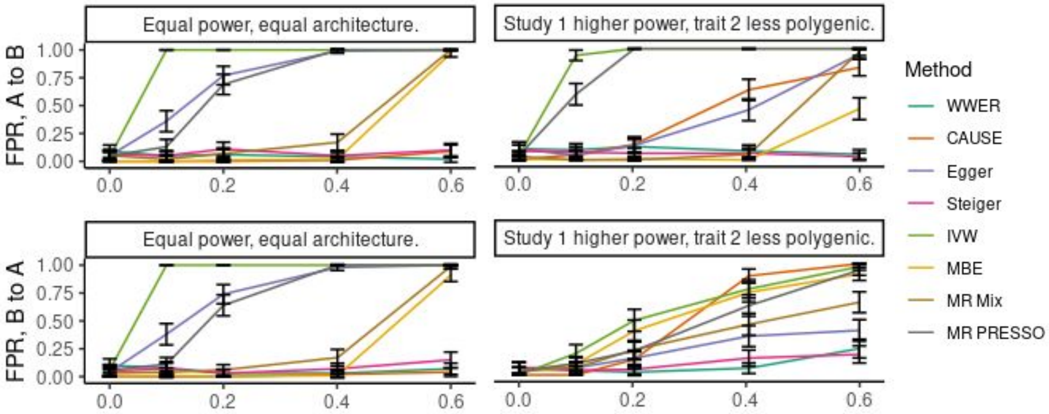


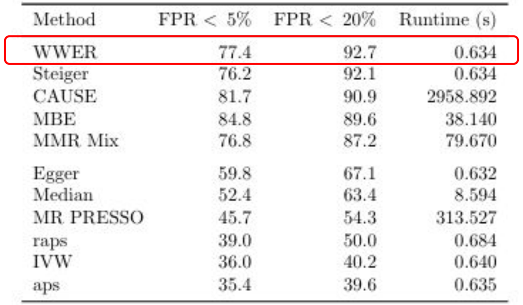
Method	FPR < 5%	FPR < 20%	Runtime (s)
WWER	77.4	92.7	0.634
Steiger	76.2	92.1	0.634
CAUSE	81.7	90.9	2958.892
MBE	84.8	89.6	38.140
MMR Mix	76.8	87.2	79.670
Egger	59.8	67.1	0.632
Median	52.4	63.4	8.594
MR PRESSO	45.7	54.3	313.527
raps	39.0	50.0	0.684
IVW	36.0	40.2	0.640
aps	35.4	39.6	0.635

WWER efficiently handles correlated pleiotropy









Method	FPR < 5%	FPR < 20%	Runtime (s)
WWER	77.4	92.7	0.634
Steiger	76.2	92.1	0.634
CAUSE	81.7	90.9	2958.892
MBE	84.8	89.6	38.140
MMR Mix	76.8	87.2	79.670
Egger	59.8	67.1	0.632
Median	52.4	63.4	8.594
MR PRESSO	45.7	54.3	313.527
raps	39.0	50.0	0.684
IVW	36.0	40.2	0.640
aps	35.4	39.6	0.635

Problem #2: we only have an estimate of R

$$G = I - R^{-1} D[1/R^{-1}]$$

versus

$$\hat{G} = I - \hat{R}^{-1} D[1/\hat{R}^{-1}]$$

Recast matrix inversion as a constrained optimization problem

Find matrices *U*, *V* such that *UV=I* that minimize the loss:

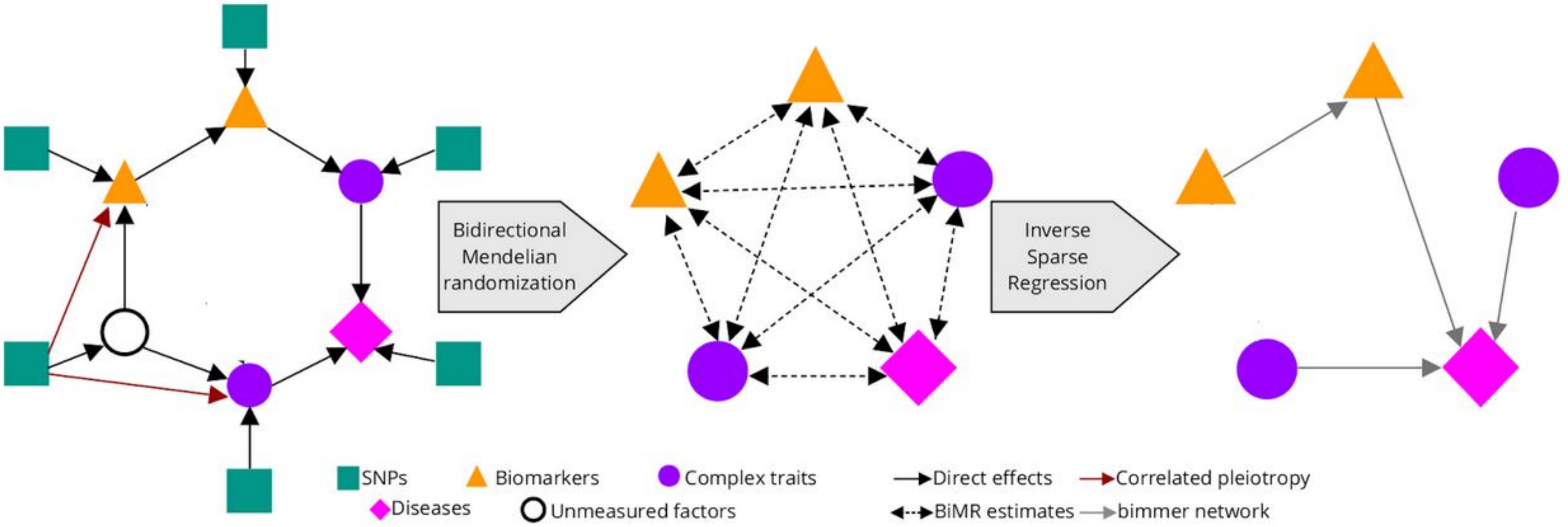
$$\left|\left|W \circ \left(\hat{R} - U\right)\right|\right|_F^2 + \lambda \sum_{i \neq j} |V_{ij}|$$

Fit with *alternating direction method of multipliers (ADMM)*

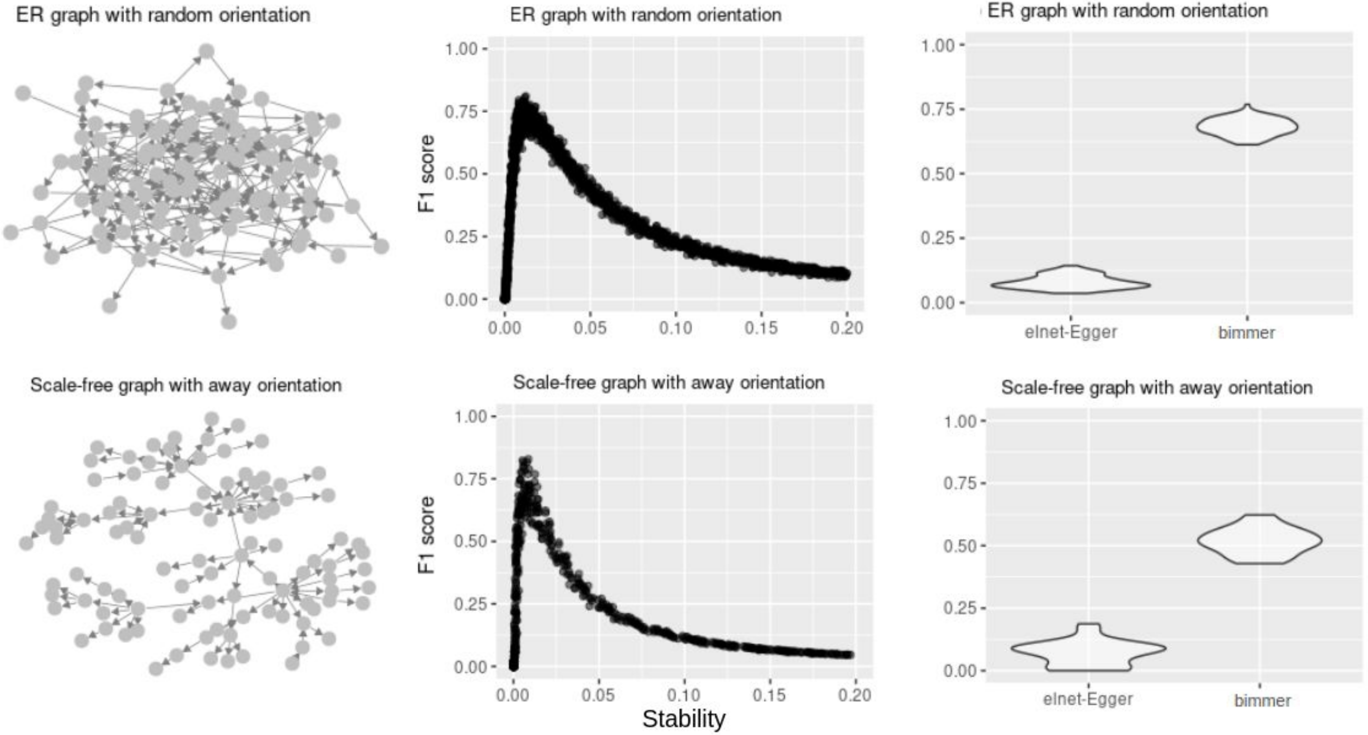
Select λ with adaptation of *stability* criteria in GLASSO

Inverse Sparse Regression (inspre)

bimmer = bi-directed MR + sparse mediation analysis



bimmer performs well in simulation



biobank uk

A wonderful resource for exploratory data analysis

>9000 phenotypes filtered to 411 based on h2 and rg

149 with >5 independent GWAS SNPs

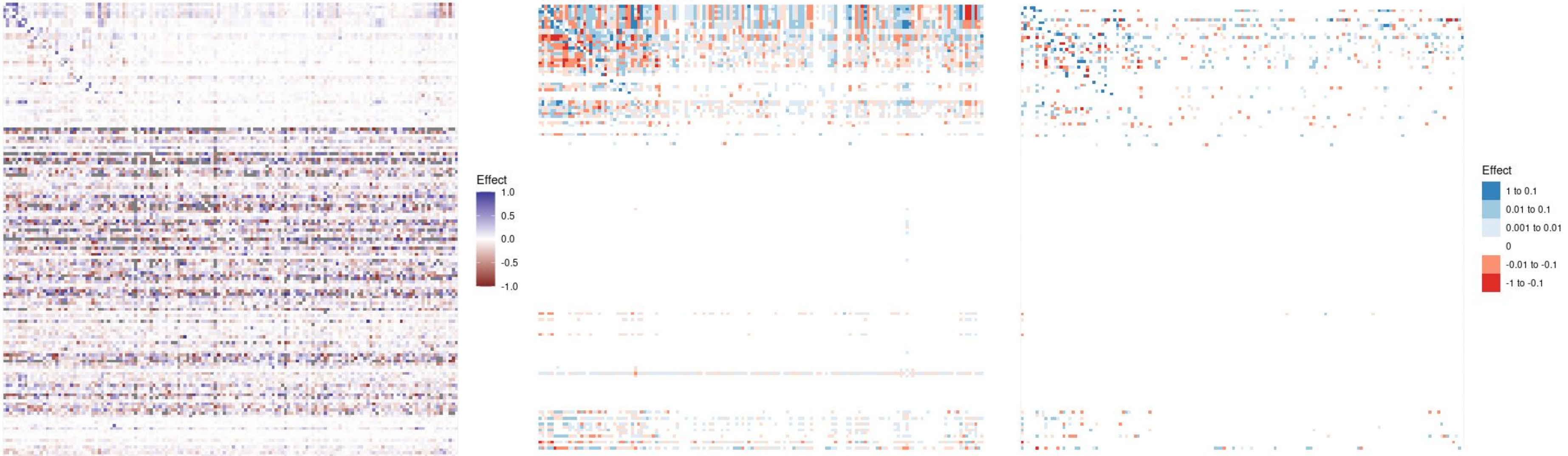
Trait Class	Example
Blood biomarker	Cholesterol
Blood composition	White blood cell count
Blood trait	Mean sphered cell volume
Immune-related disease	Psoriasis
Heart-related	Illness of father: heart disease
Other disease	Basal cell carcinoma
Morphological	Hip circumference
Dietary	Alcohol intake frequency
Behavioral	Getting up in morning
Neurological	Anxious feelings
Eye-related	3mm weak meridian
Other	Average income before tax

Directed graph on 149 UK Biobank phenotypes

WWER total effect estimates

Shrunk total effect estimate (U)

Directed graph estimate (G)



$\hat{R}$

$U \approx \hat{R}$

$UV \approx I$

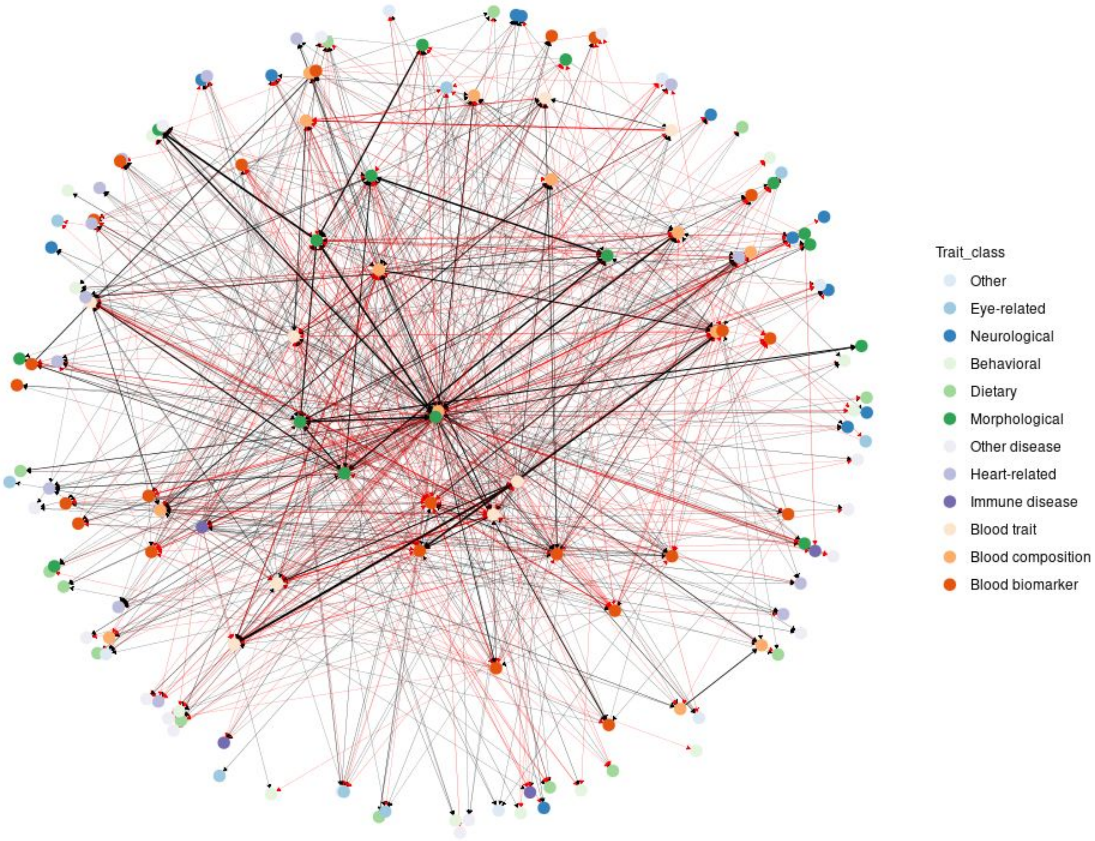
$\hat{G} = I - VD[1/V]$

2702 significant q <0.05

1658 entries > |0.01|
Ubiquitous smaller effects

843 non-zeros

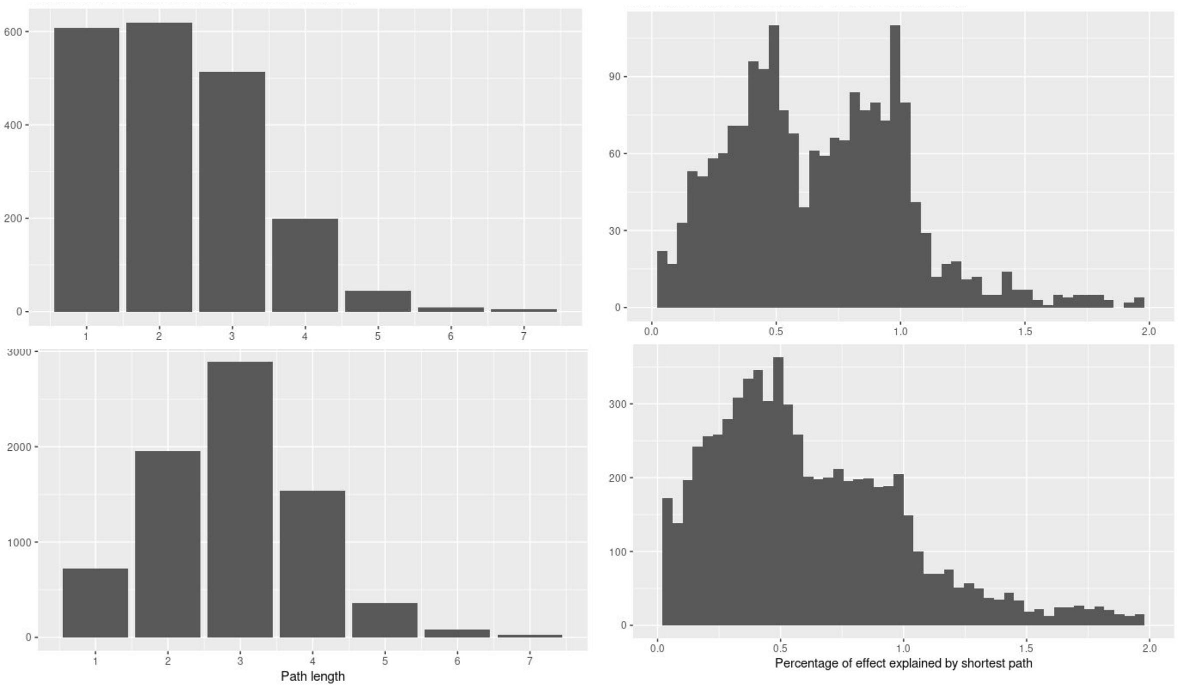
Directed graph on 149 UK Biobank phenotypes

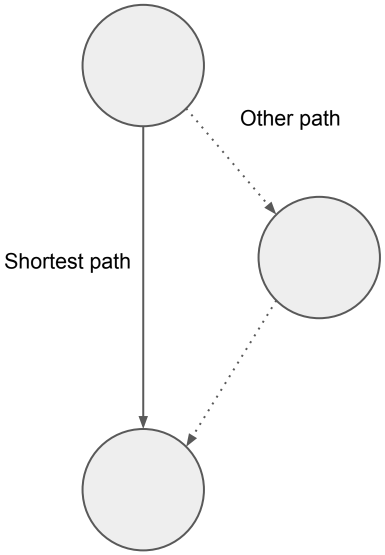


The shortest path often explains only some of the effect

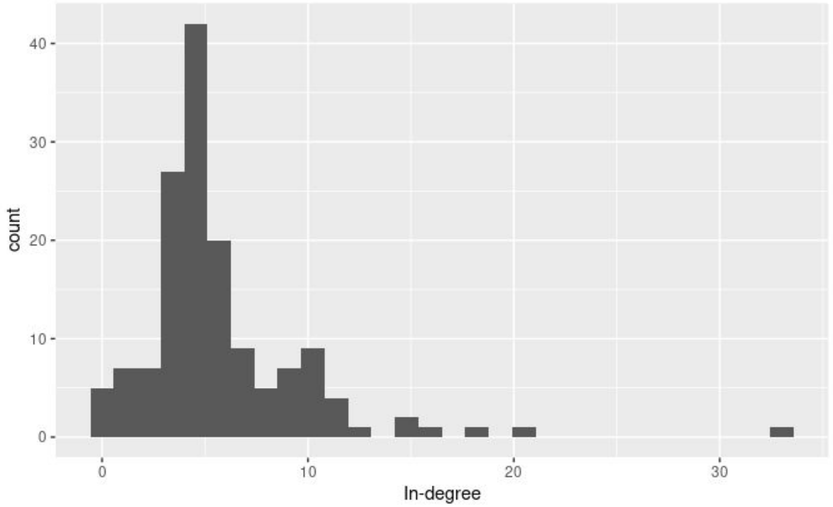
FDR 5% as total effects

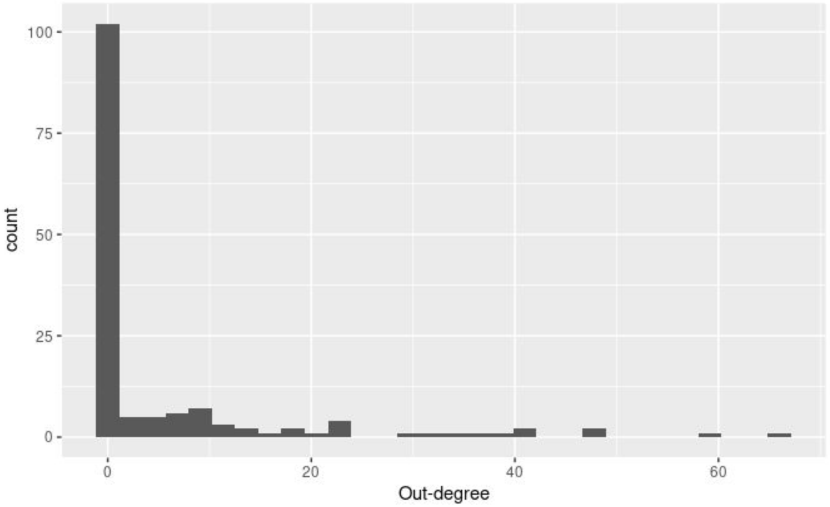
All connections

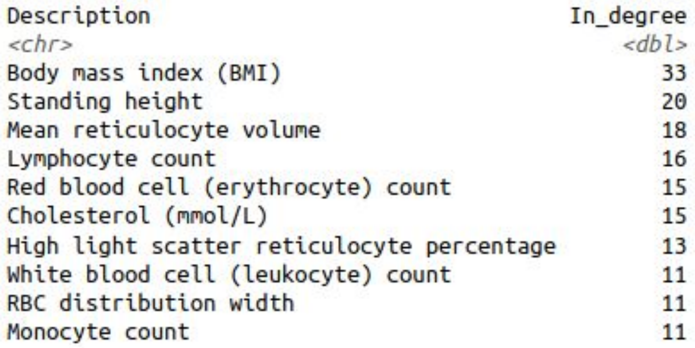




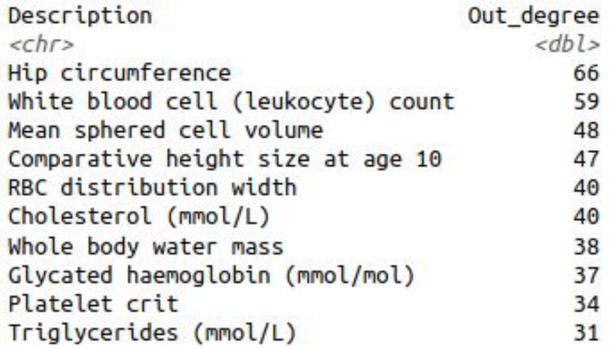
In and out-degree distributions are exponential







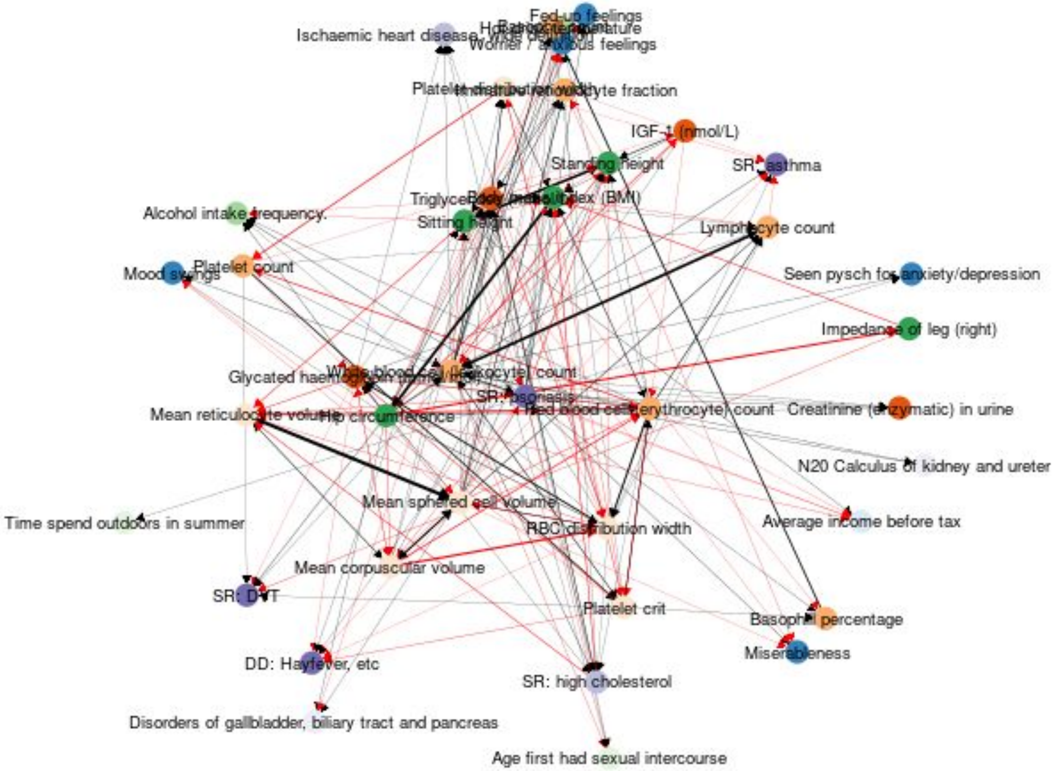
```
Description                                 In_degree
<chr>                                           <dbl>
Body mass index (BMI)                              33
Standing height                                    20
Mean reticulocyte volume                           18
Lymphocyte count                                   16
Red blood cell (erythrocyte) count                 15
Cholesterol (mmol/L)                               15
High light scatter reticulocyte percentage         13
White blood cell (leukocyte) count                 11
RBC distribution width                             11
Monocyte count                                     11
```



```
Description                             Out_degree
<chr>                                        <dbl>
Hip circumference                               66
White blood cell (leukocyte) count              59
Mean sphered cell volume                        48
Comparative height size at age 10               47
RBC distribution width                          40
Cholesterol (mmol/L)                            40
Whole body water mass                           38
Glycated haemoglobin (mmol/mol)                 37
Platelet crit                                   34
Triglycerides (mmol/L)                          31
```

Causes and consequences of IMID

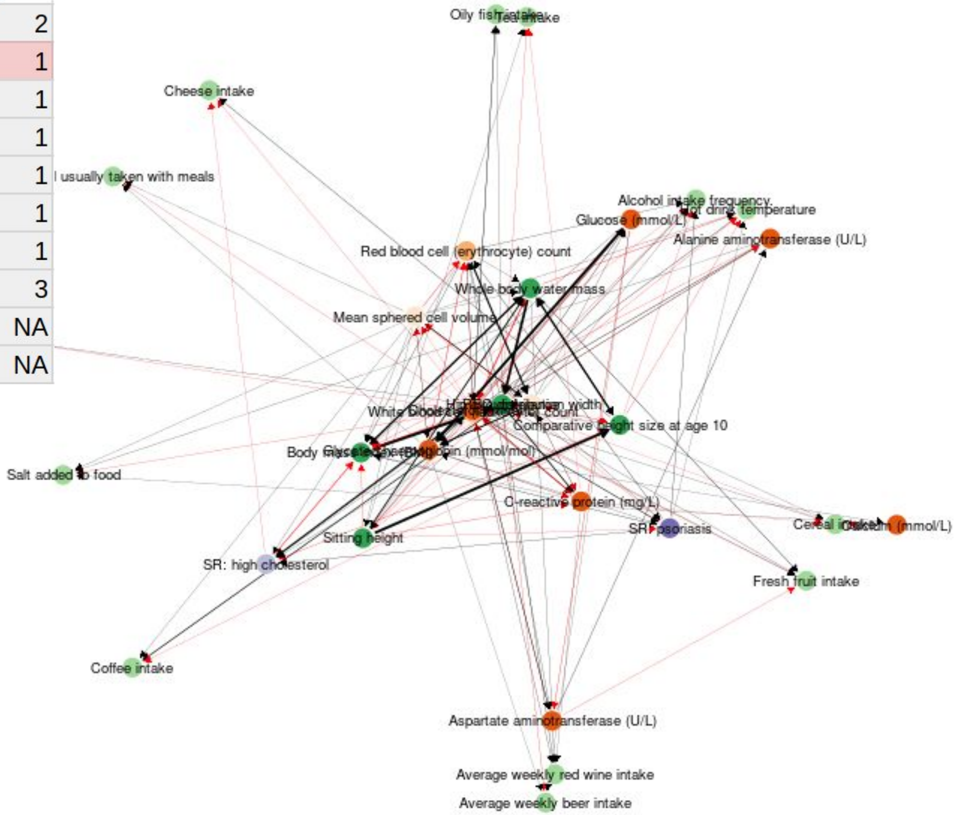
Exposure	Outcome	Path Length
SR: psoriasis	Alcohol intake frequency.	1
Impedance of leg (right)	SR: DVT	1
White blood cell (leukocyte) count	SR: psoriasis	1
Platelet crit	DD: Hayfever, etc	1
Platelet distribution width	SR: asthma	1
SR: psoriasis	Miserableness	1
Lymphocyte count	SR: psoriasis	2
SR: psoriasis	Mood swings	1
SR: psoriasis	SR: high cholesterol	1
SR: psoriasis	Fed-up feelings	1



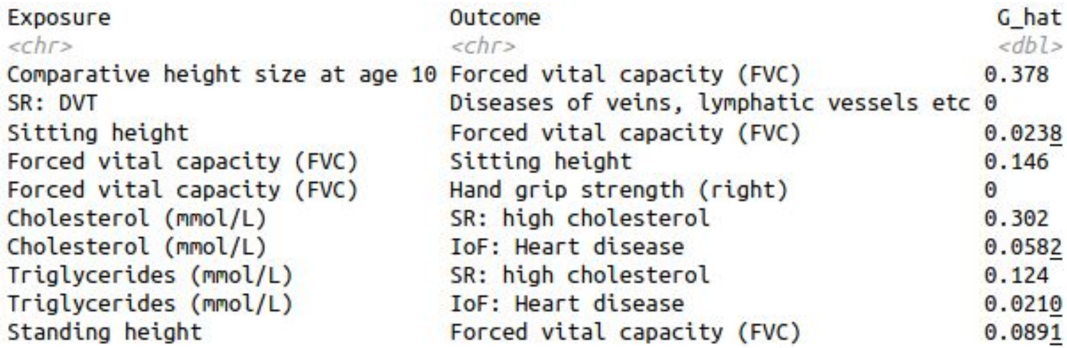
Fed-up feelings
Ischaemic heart disease, wide definition
Hot drink temperature
Worrier / anxious feelings
Platelet distribution width
Immature reticulocyte fraction
IGF-1 (nmol/L)
Standing height
SR: asthma
Triglycerides
Body mass index (BMI)
Sitting height
Alcohol intake frequency.
Lymphocyte count
Mood swings
Platelet count
Seen pysch for anxiety/depression
Impedance of leg (right)
Glycated haemoglobin
White blood cell (leukocyte) count
SR: psoriasis
Mean reticulocyte volume
Hip circumference
Red blood cell (erythrocyte) count
Creatinine (enzymatic) in urine
N20 Calculus of kidney and ureter
Time spend outdoors in summer
Mean sphered cell volume
RBC distribution width
Average income before tax
Mean corpuscular volume
SR: DVT
Platelet crit
Basophil percentage
Miserableness
DD: Hayfever, etc
SR: high cholesterol
Disorders of gallbladder, biliary tract and pancreas
Age first had sexual intercourse

Causes and consequences of diet

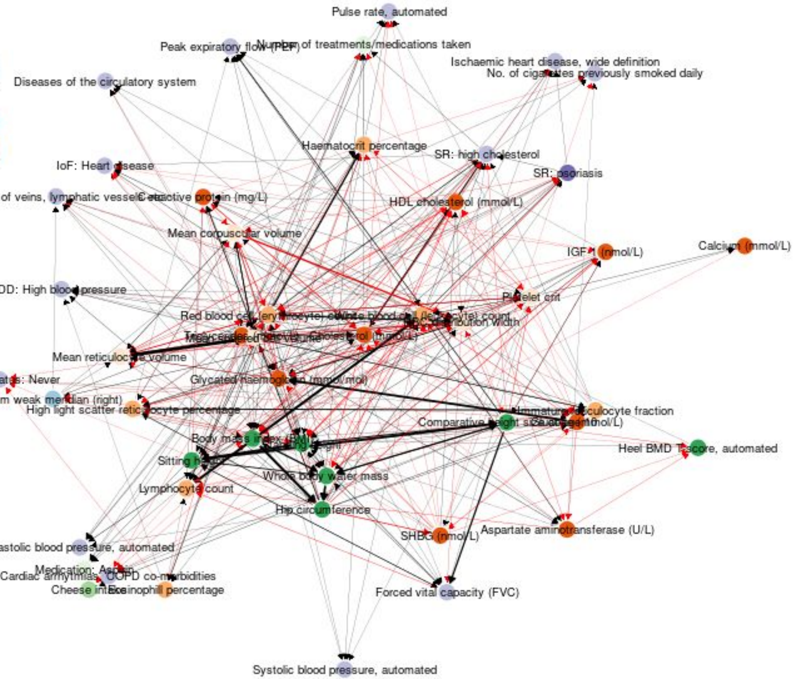
Exposure	Outcome	Path Length
Body mass index (BMI)	Coffee intake	2
Cholesterol (mmol/L)	Salt added to food	1
SR: psoriasis	Alcohol intake frequency.	1
Hip circumference	Coffee intake	1
Cholesterol (mmol/L)	Cereal intake	1
Cholesterol (mmol/L)	Oily fish intake	1
Cholesterol (mmol/L)	Fresh fruit intake	1
Comparative body size at age 10	Coffee intake	3
Tea intake	Creatinine (umol/L)	NA
Coffee intake	Creatinine (enzymatic) in urir	NA



Causes and consequences of heart-related traits



```
Exposure                          Outcome                                   G_hat
<chr>                             <chr>                                     <dbl>
Comparative height size at age 10 Forced vital capacity (FVC)              0.378
SR: DVT                           Diseases of veins, lymphatic vessels etc 0
Sitting height                    Forced vital capacity (FVC)              0.0238
Forced vital capacity (FVC)       Sitting height                           0.146
Forced vital capacity (FVC)       Hand grip strength (right)               0
Cholesterol (mmol/L)              SR: high cholesterol                     0.302
Cholesterol (mmol/L)              IoF: Heart disease                       0.0582
Triglycerides (mmol/L)            SR: high cholesterol                     0.124
Triglycerides (mmol/L)            IoF: Heart disease                       0.0210
Standing height                   Forced vital capacity (FVC)              0.0891
```



Pulse rate, automated
Peak expiratory flow (PEF)
Number of treatments/medications taken
Ischaemic heart disease, wide definition
No. of cigarettes previously smoked daily
Diseases of the circulatory system
Haematocrit percentage
SR: high cholesterol
SR: psoriasis
IoF: Heart disease
Diseases of veins, lymphatic vessels etc
C-reactive protein (mg/L)
HDL cholesterol (mmol/L)
Mean corpuscular volume
IGF-1 (nmol/L)
Calcium (mmol/L)
DD: High blood pressure
Platelet crit
Red blood cell (erythrocyte) count
White blood cell (leukocyte) count
Red blood cell (erythrocyte) distribution width
Triglycerides (mmol/L)
Mean sphered cell volume
Cholesterol (mmol/L)
Mean reticulocyte volume
Glycated haemoglobin (mmol/mol)
Smoking status: Never
3mm weak meridian (right)
High light scatter reticulocyte percentage
Immature reticulocyte fraction
Comparative height size at age 10
Body mass index (BMI)
Standing height
Heel BMD T-score, automated
Sitting height
Whole body water mass
Lymphocyte count
Hip circumference
SHBG (nmol/L)
Aspartate aminotransferase (U/L)
Diastolic blood pressure, automated
Medication: Aspirin
Cardiac arrhythmias
COPD co-morbidities
Cheese intake
Eosinophill percentage
Forced vital capacity (FVC)
Systolic blood pressure, automated

To conclude

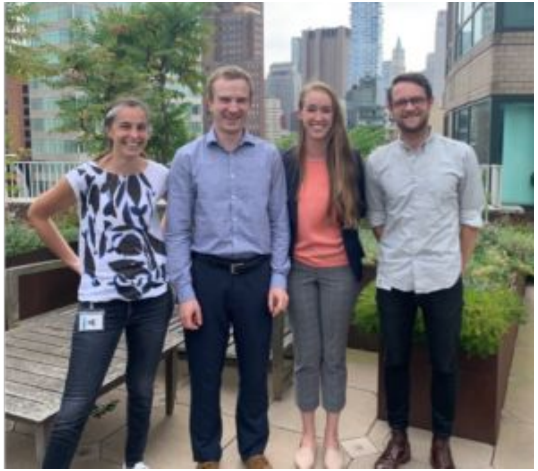
- Biobanks hold tremendous promise for exploratory data analysis
  - Can be challenging to find creative analyses that leverage their multifactoral nature
- *bimmer* combines genetic instruments with sparse graph methods to generate *putatively causal* directed graphs from biobank-style data
  - I discourage making formal causal claims from large-scale analyses
- Our results suggest a large amount of the polygenic background for complex traits is explained by small, long-range effects of other phenotypes
  - This is related to the omnigenic concept, but says nothing about core genes
- Preprints and code are available!
  - bimmer: https://bit.ly/3dY1Rl3, https://github.com/brielin/bimmer/
  - WWER: https://bit.ly/2PYrlXi, https://github.com/brielin/WWER/
  - bb2991@columbia.edu, Twitter: @brielinb

Thank you for listening!

Thanks to

Lappalainen lab

Knowles lab



Also: Andy Dahl, Harold Pimentel, Lior Pachter, Nick Bray

Funding: Data Science Institute at Columbia