Investigating shared genetics

Chris Wallace

chr1swallace chr1swallace.github.io

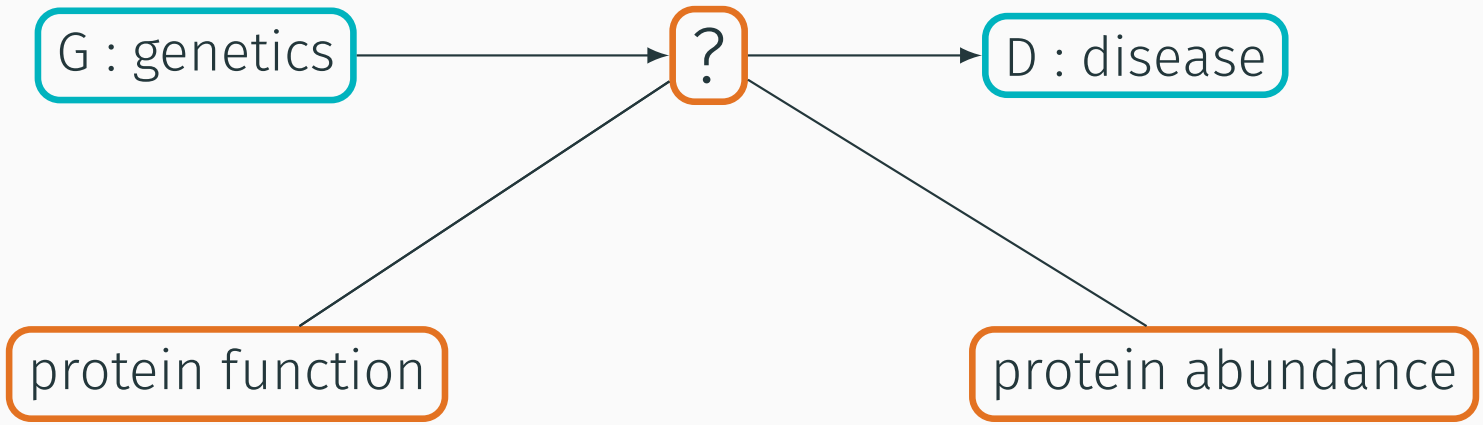
9th May 2019

University of Cambridge

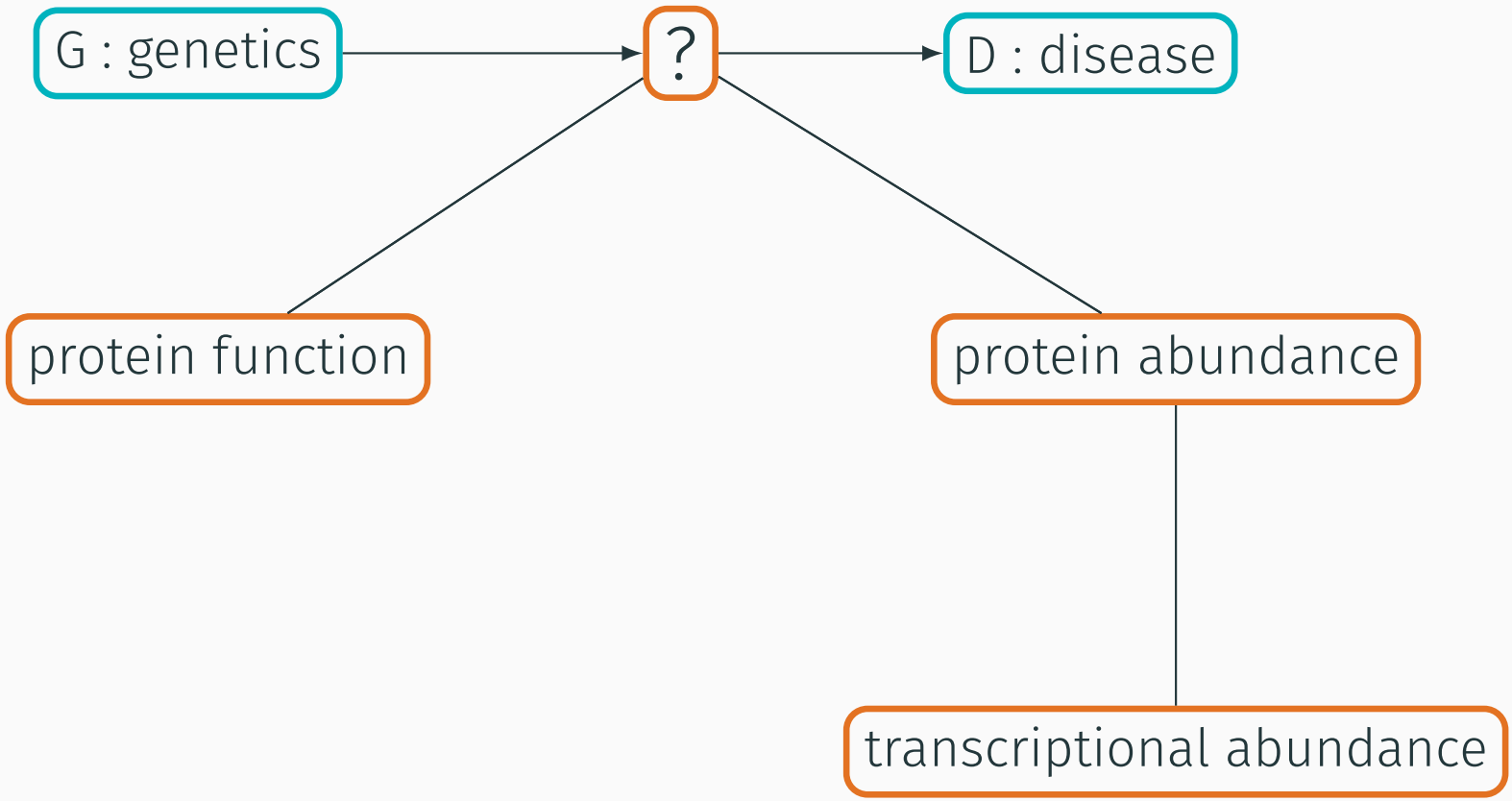
Identifying molecular traits that mediate a disease association

Identifying molecular traits that mediate a disease association

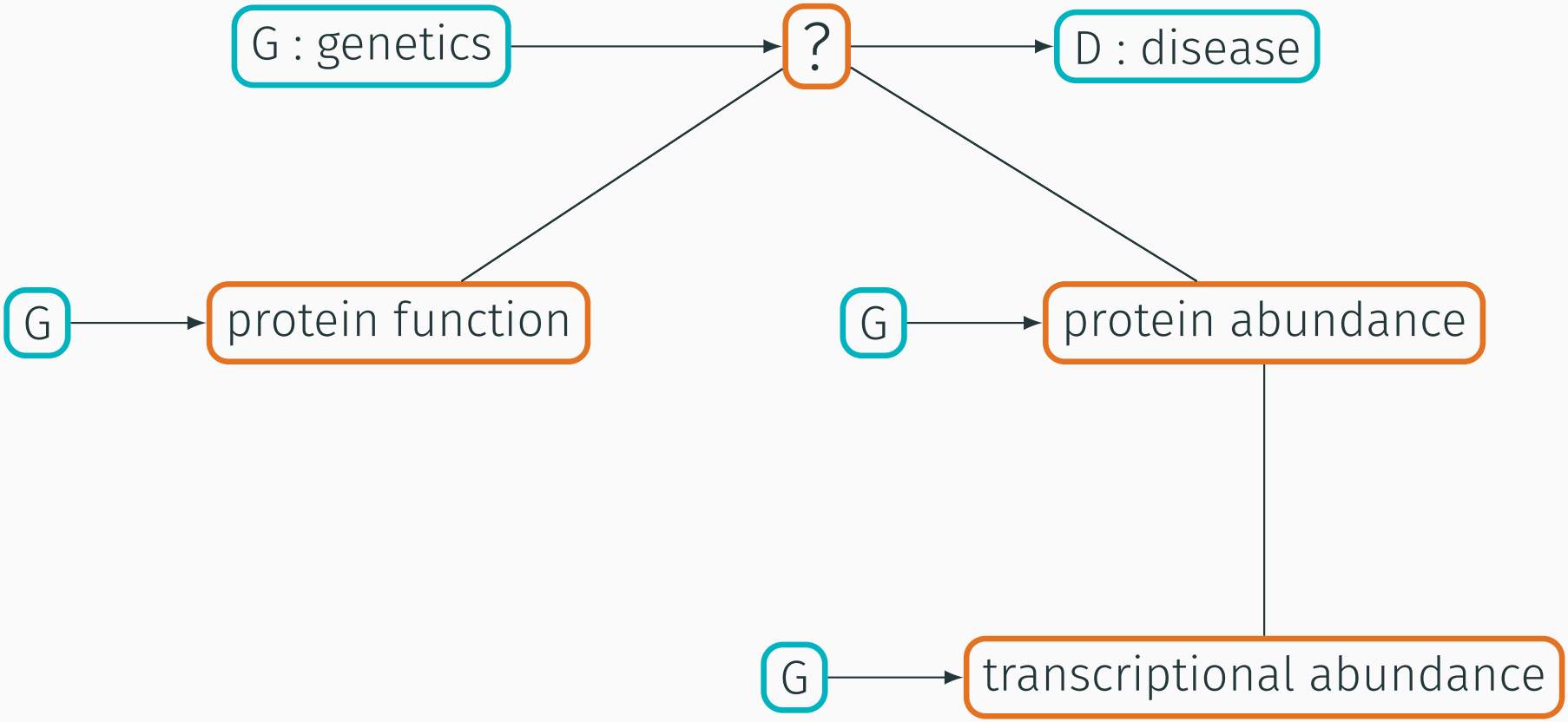
Identifying molecular traits that mediate a disease association



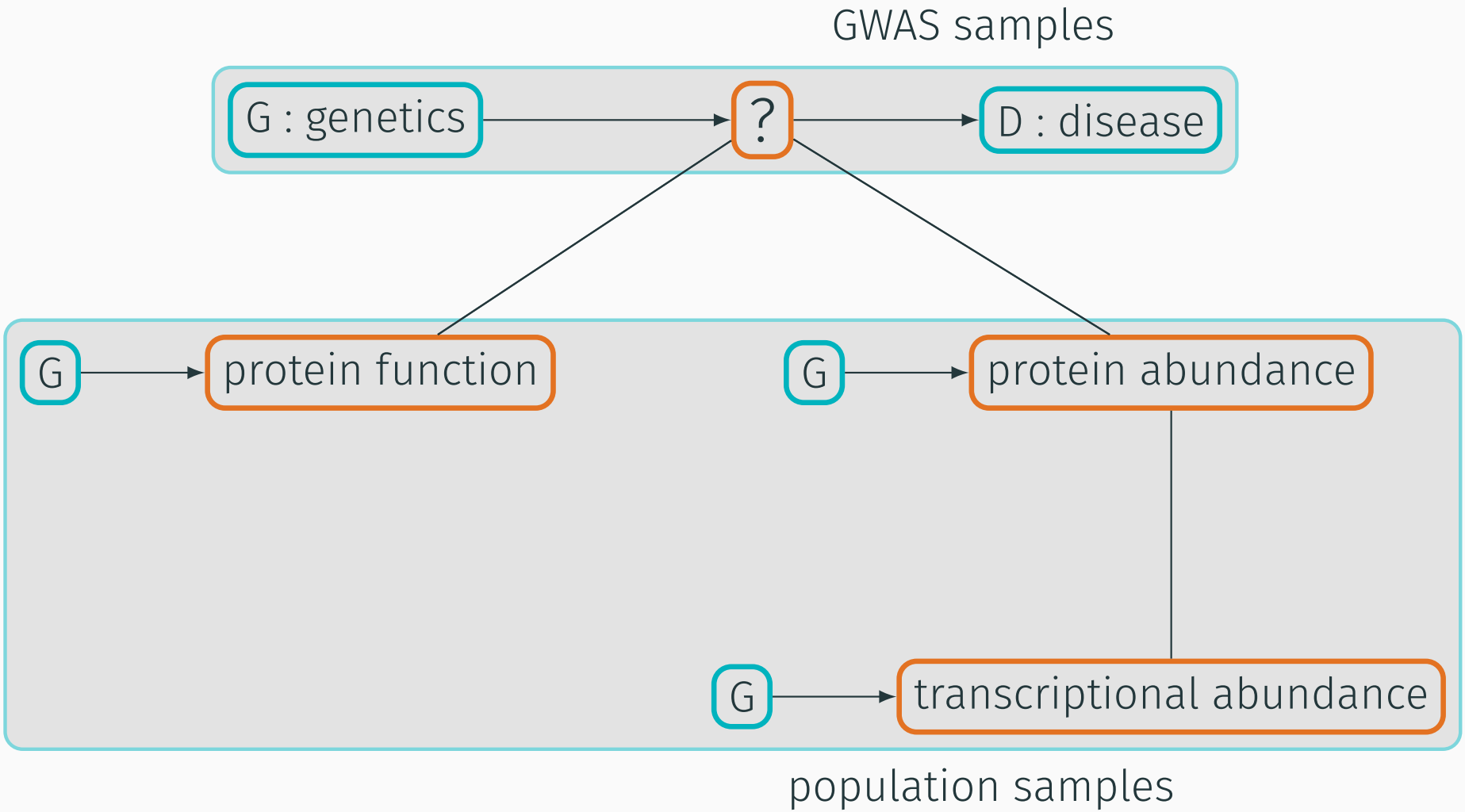
Identifying molecular traits that mediate a disease association



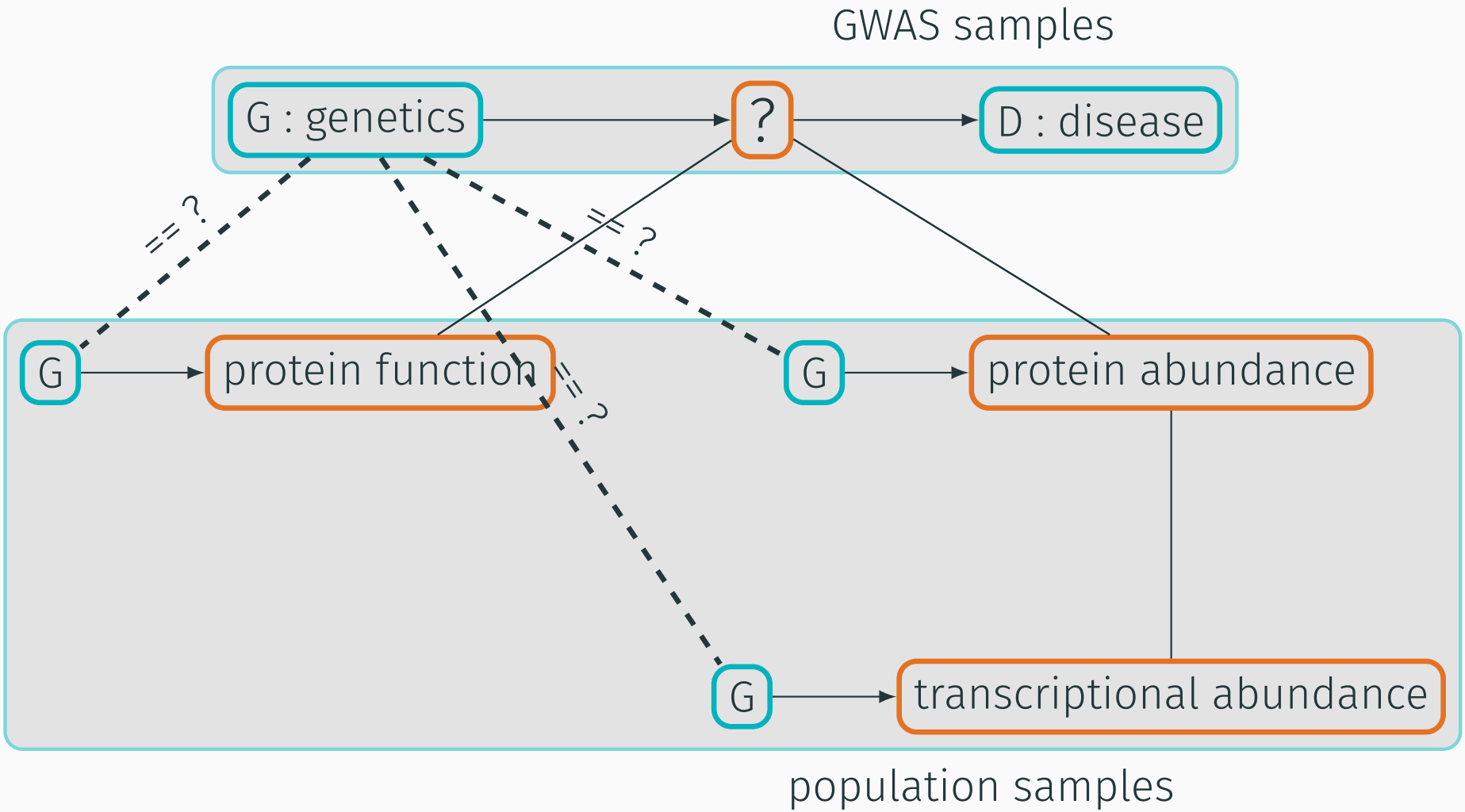
Identifying molecular traits that mediate a disease association



Identifying molecular traits that mediate a disease association



Identifying molecular traits that mediate a disease association

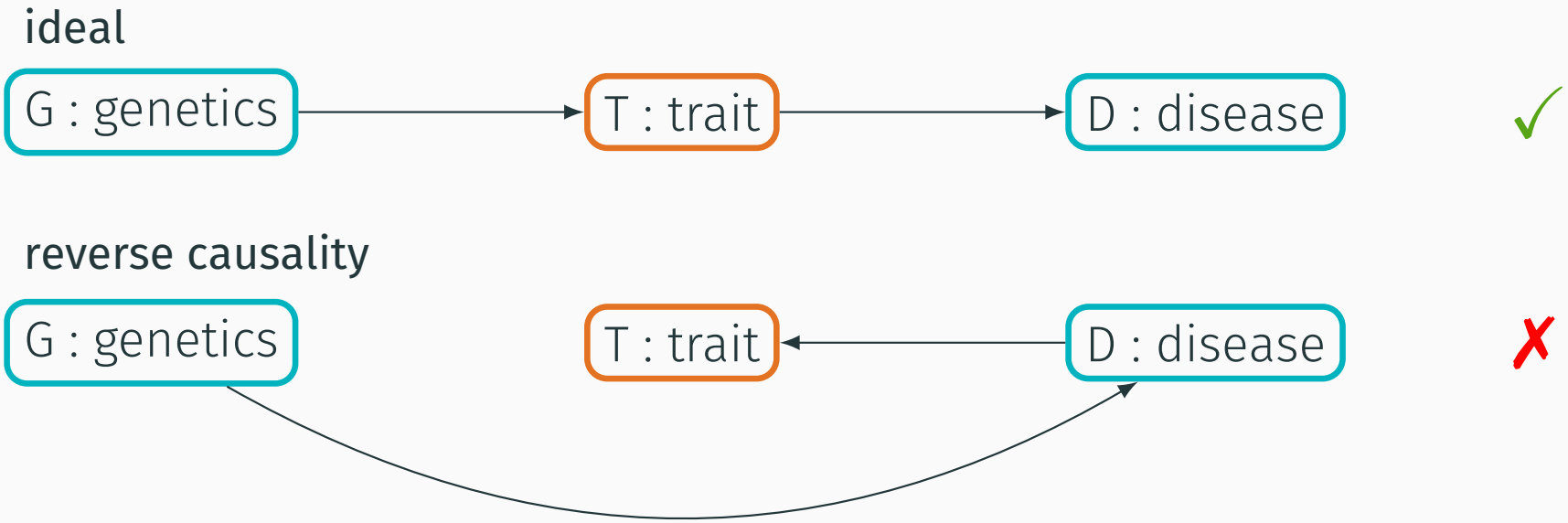


Identifying molecular traits that mediate a disease association

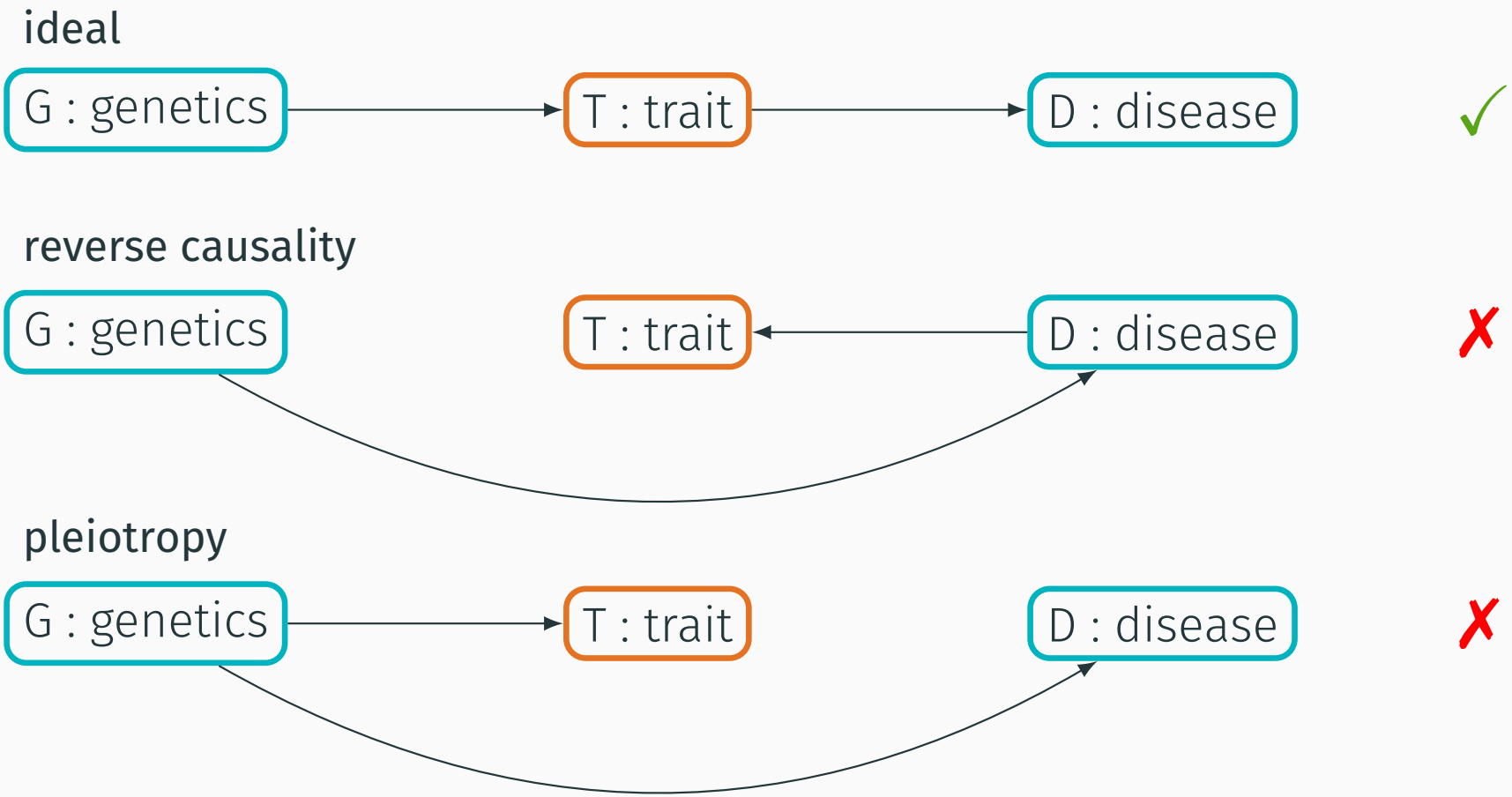
ideal

✓

Identifying molecular traits that mediate a disease association



Identifying molecular traits that mediate a disease association

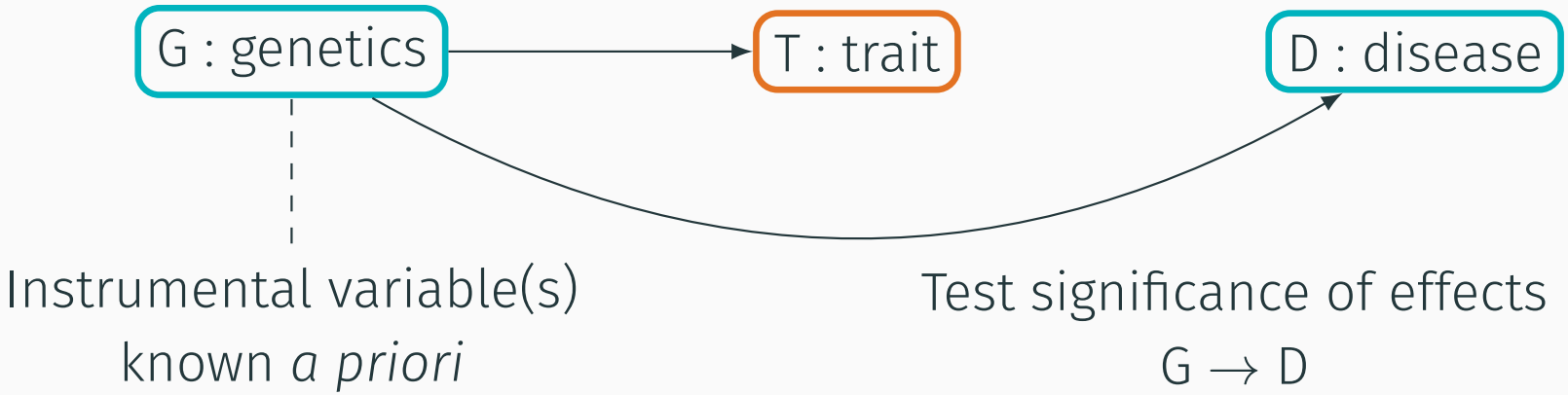


Identifying molecular traits that mediate a disease association

Mendelian Randomisation

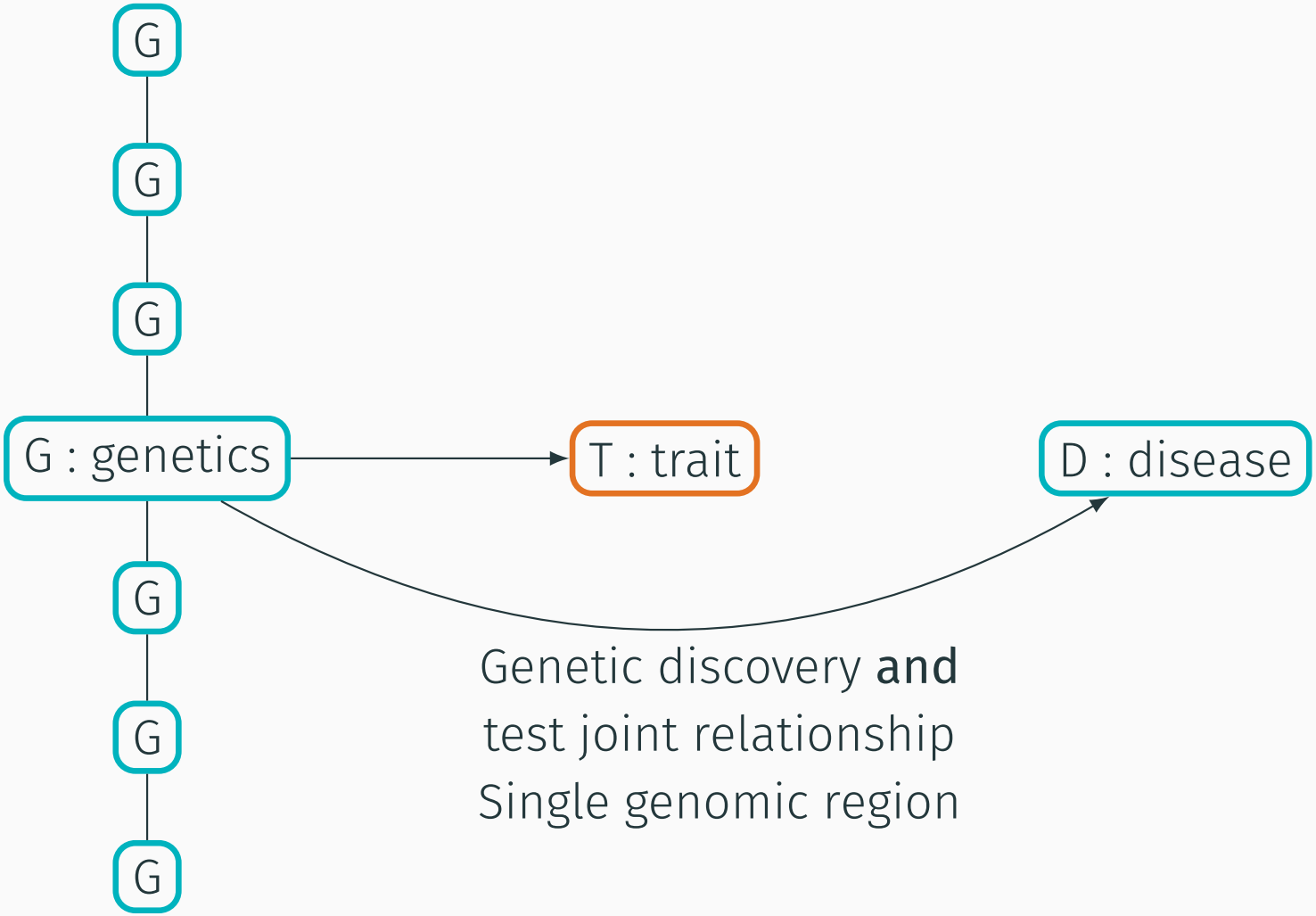
Identifying molecular traits that mediate a disease association

Mendelian Randomisation



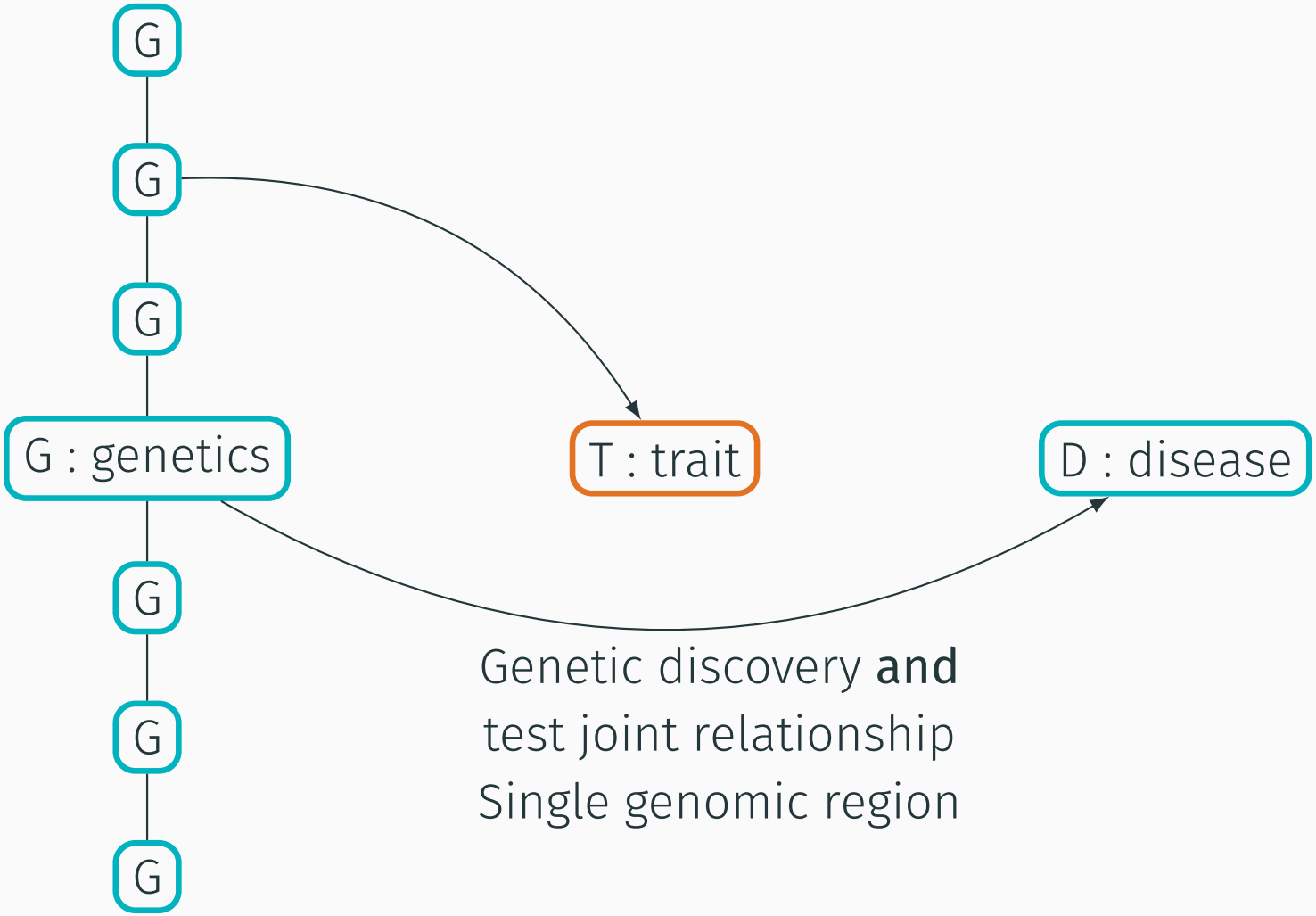
Identifying molecular traits that mediate a disease association

Colocalisation



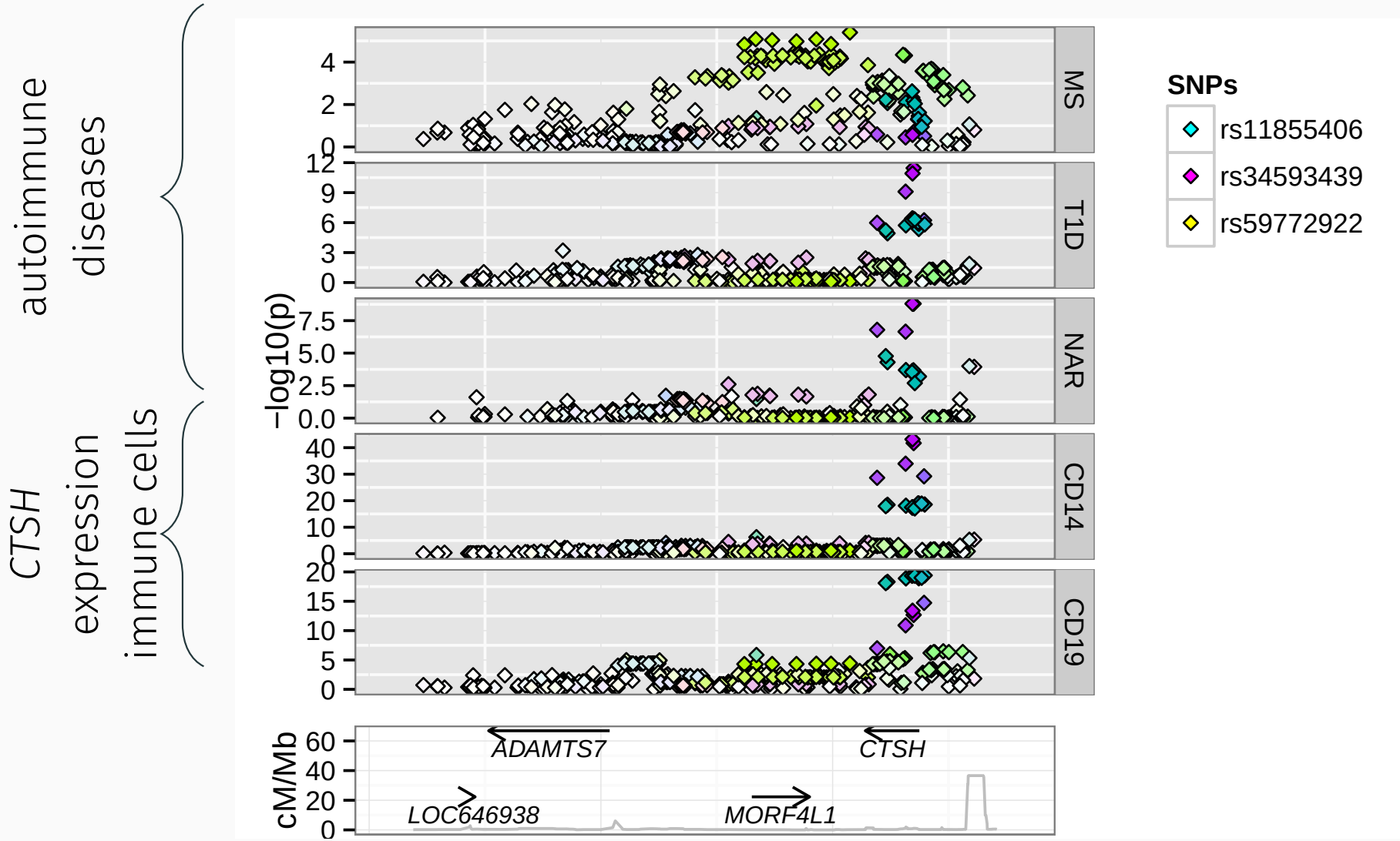
Identifying molecular traits that mediate a disease association

Colocalisation

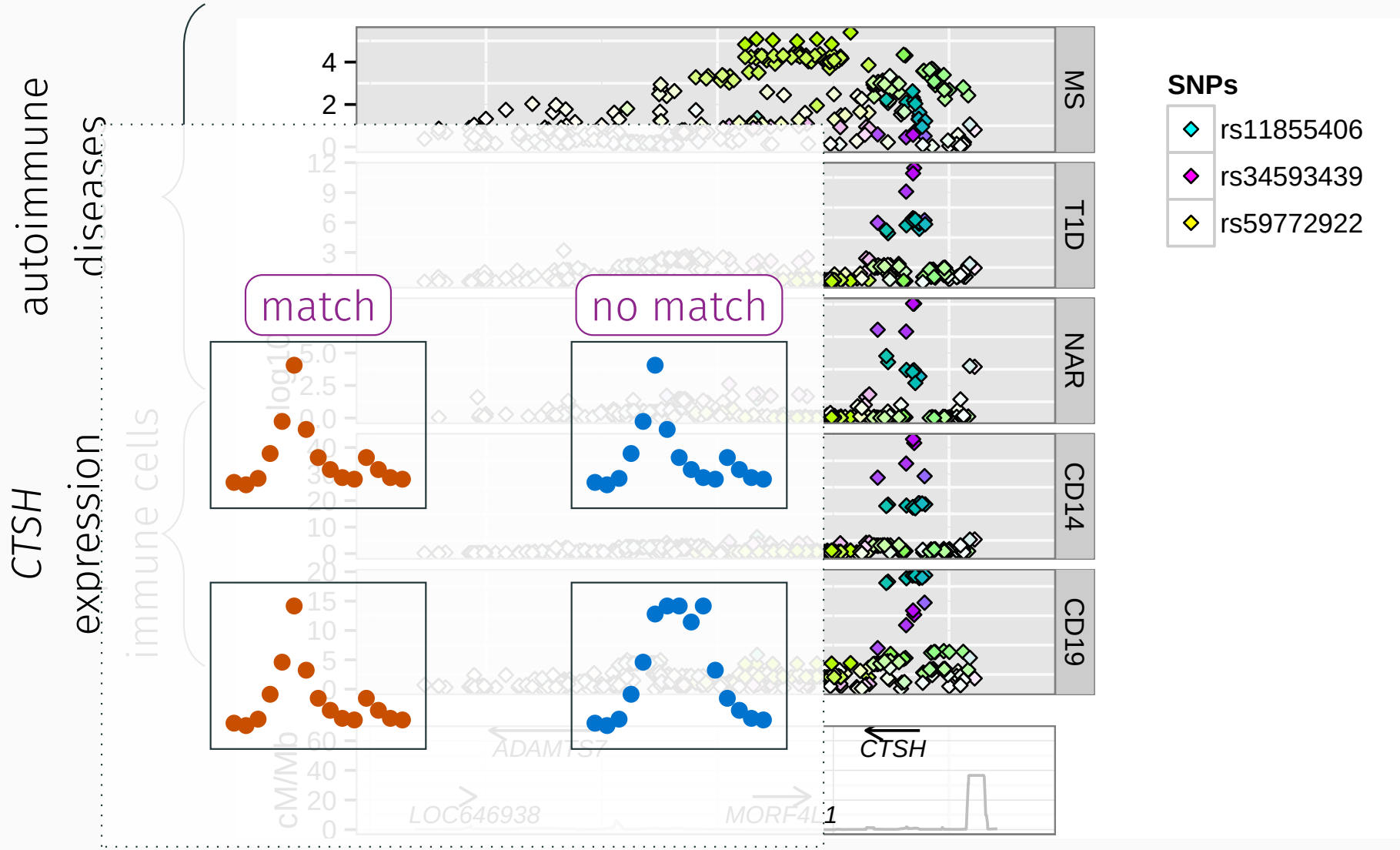


Colocalisation:
the coloc approach

Coloc: at essence tests whether patterns match



Coloc: at essence tests whether patterns match



Giambartolomei et al, 2014. Guo et al, 2016. chr1swallace/coloc

Coloc enumerates hypotheses

We have a pair of traits.

For a single, LD-defined, genomic region, assume at most one association per trait.

Then exactly 5 possibilities:

H_0 : no association

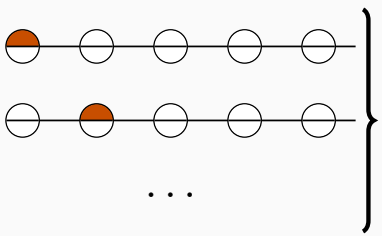
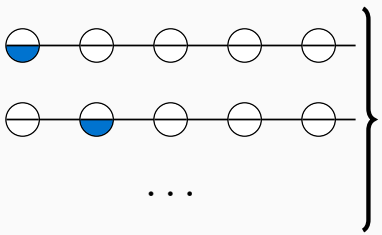
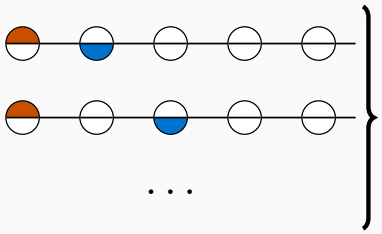
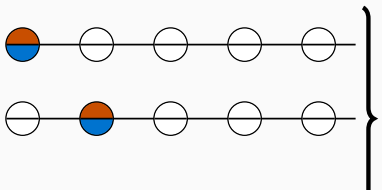
H_1 : association to trait 1 only

H_2 : association to trait 2 only

H_3 : association to both traits, distinct causal variants

H_4 : association to both traits, shared causal variant

Coloc enumerates hypotheses

hyp	configuration	num	prior
H_0		$\times 1$	
H_1		$\times n$	p_1 p_1
H_2		$\times n$	p_2 p_2
H_3		$\times n(n-1)$	p_1p_2 p_1p_2
H_4		$\times n$	p_{12} p_{12}

Values of prior parameters

Original context:

- GWAS of lipid traits (100,000 individuals)
- Liver eQTLs (1000 individuals)

i.e. two large studies, *a priori* biologically relevant to each other

$$p_1 = p_2 = 10^{-4} \qquad p_{12} = 10^{-5}$$

For a 1000 SNP region, this gives

$$P(H_1) = 0.1 \qquad P(H_2) = 0.1$$
$$P(H_3) = 0.01 \qquad P(H_4) = 0.01$$

Giambartolomei et al, PLoS Gen 2014

Review of current practice

Out of 25 papers which used coloc in 2018 ...

... 22 used software default priors

Review of current practice

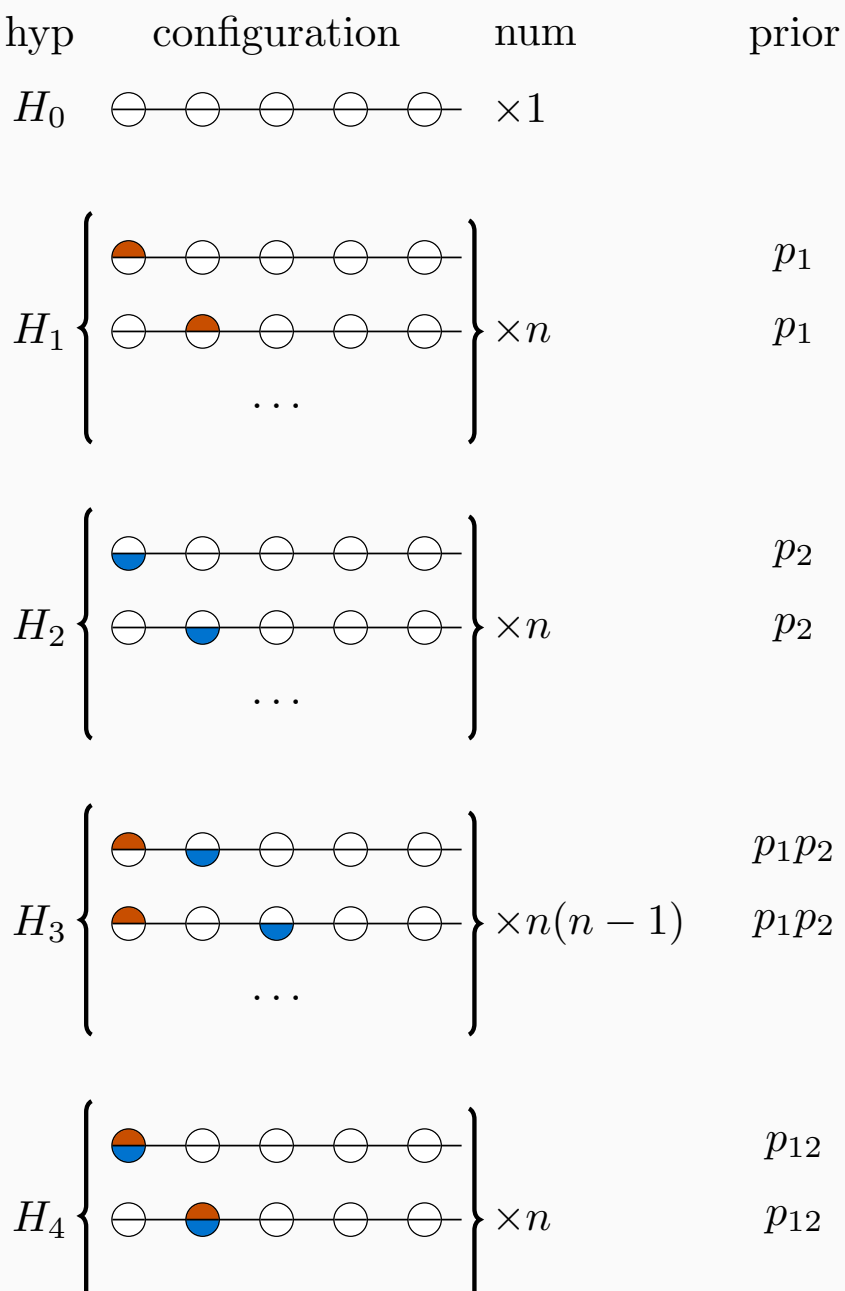
Out of 25 papers which used coloc in 2018 ...

... 22 used software default priors

Varied trait pairs:

eQTL-pQTL	GWAS-chromatinQTL	GWAS-eQTL	GWAS-pQTL
2	2	15	1
GWAS-GWAS	GWAS-meQTL	GWAS-molecular QTL	
1	3	1	

Alternative prior parameterisation



Prob. causal

$q_1 = p_1 + p_{12}$

$q_2 = p_2 + p_{12}$

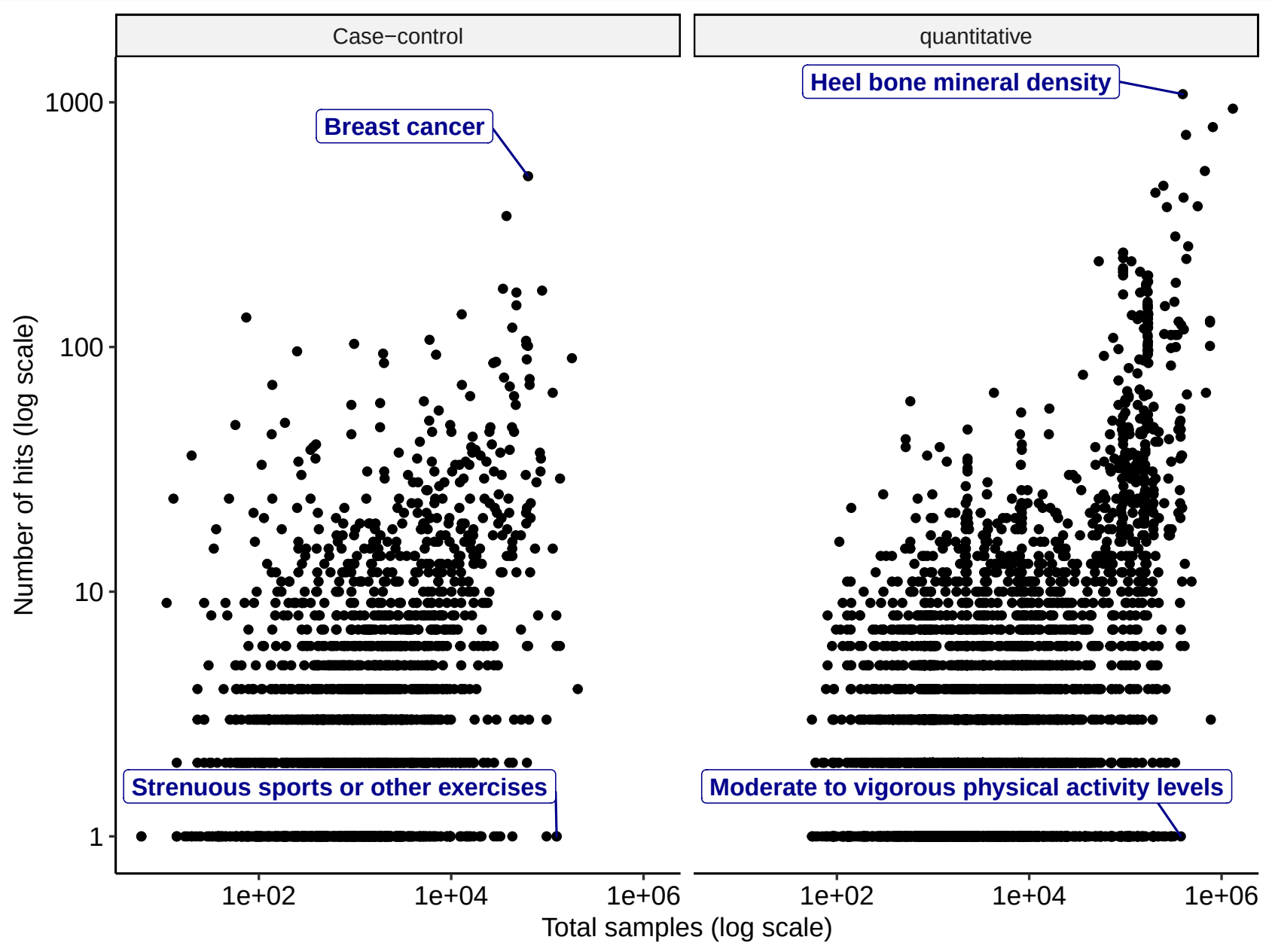
Cond. Prob. jointly causal

$q_{1|2} = p_{12}/q_1$

$q_{2|1} = p_{12}/q_2$

Marginal priors: q_1, q_2

Empirical prior for single trait, GWAS catalogue



Empirical prior for single trait, GWAS catalogue

Largest relevant GWAS disease studies:

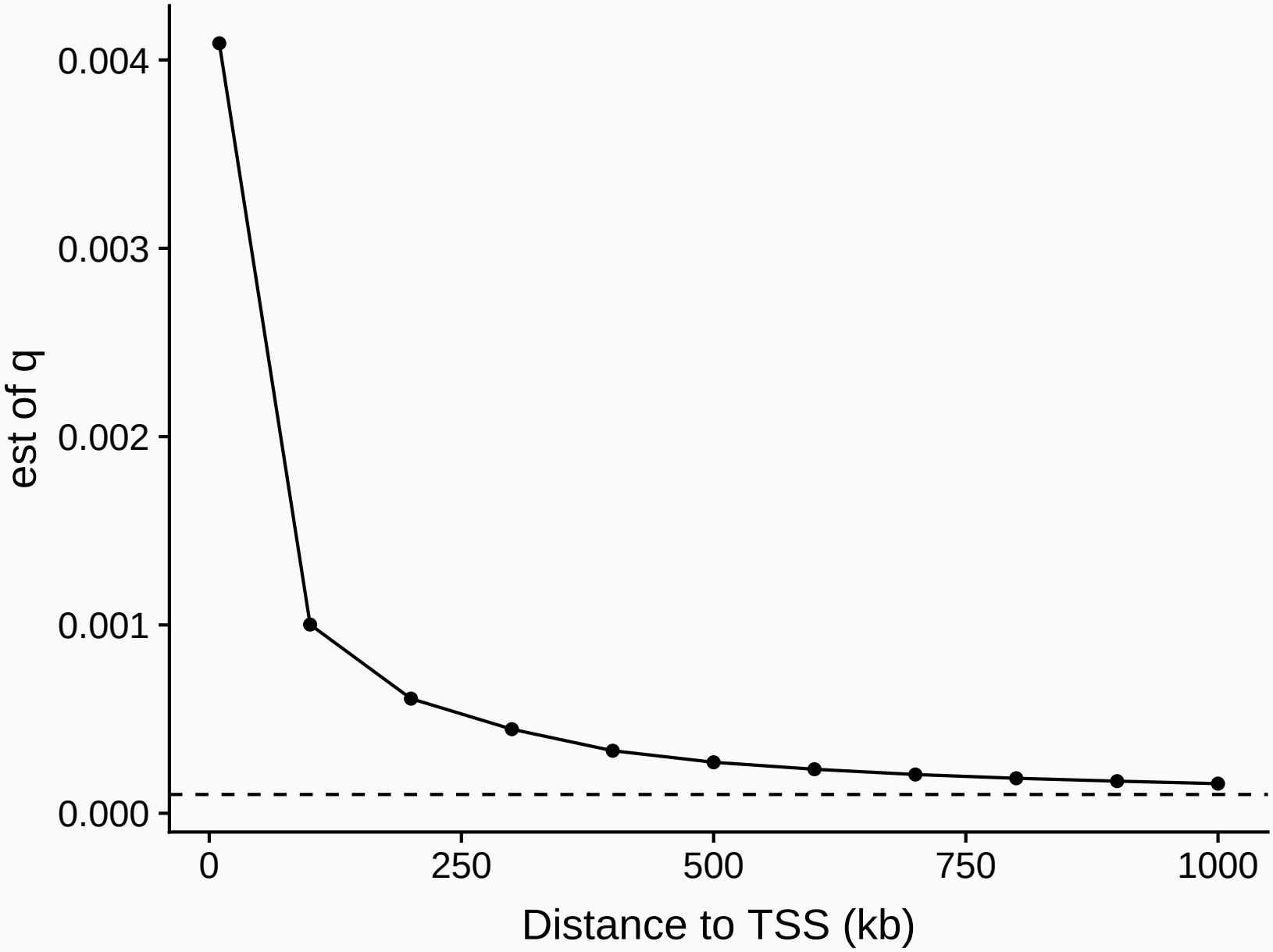
eg IBD, 60,000 subjects

$\sim$ 200 ”hits” from 2 million common SNPs $\rightarrow$

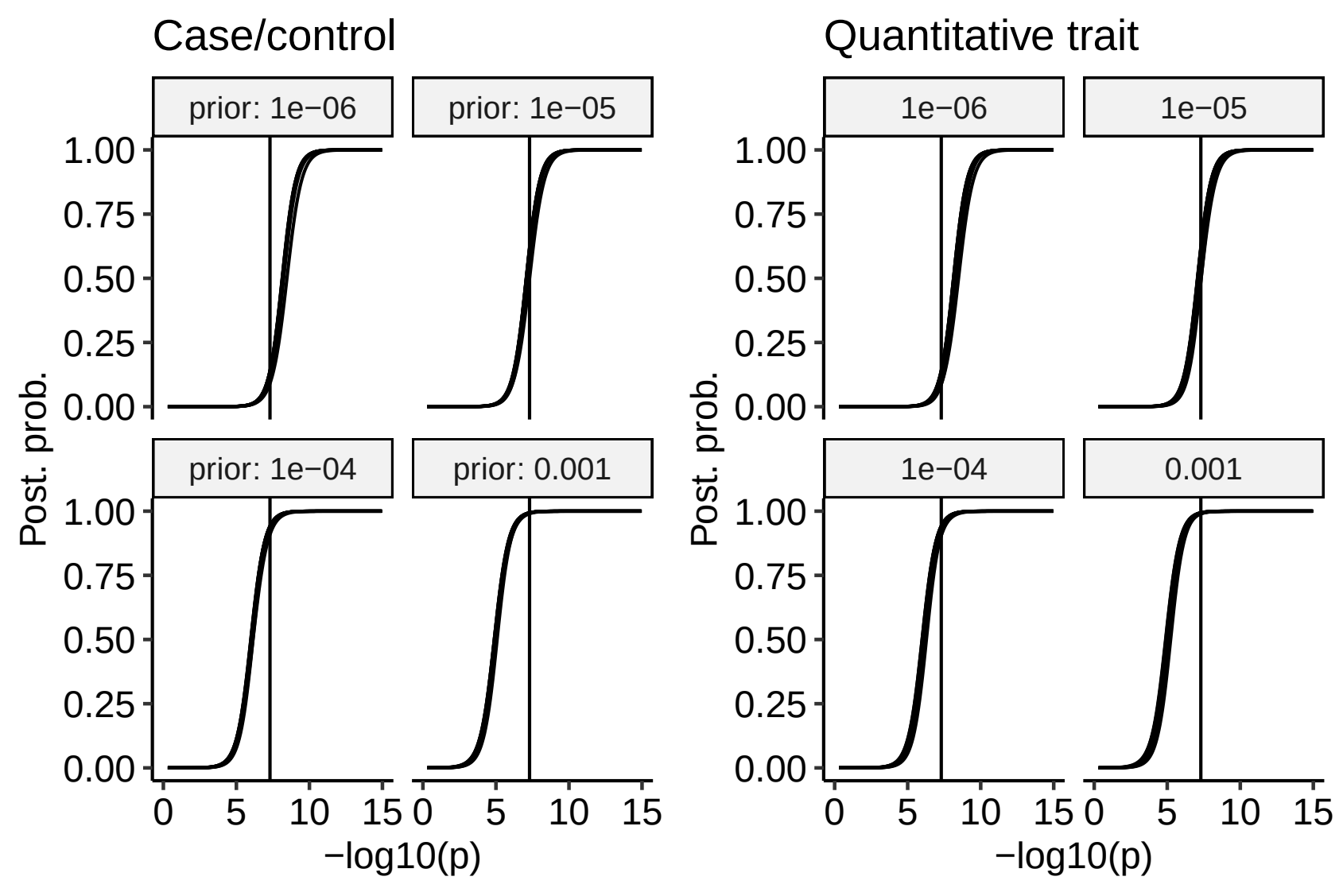
$$q_i = \frac{200}{2,000,000} = \frac{1}{10,000}$$

de Lange et al, Nat Genet 2017

Empirical prior for single trait, eQTL (GTeX whole blood)



Pragmatic prior for single trait



Conditional priors $q_{1|2}$, $q_{2|1}$

Empirical priors

?

Bounds on conditional prior for joint association

For a SNP to be causal, assume it has to be "functional" in some sense

$$P(\text{SNP is functional}) = f$$

Also, prob a SNP is causal for both traits $\leq$ prob it is causal for one.

Then

$$\frac{q_1 q_2}{f} \leq p_{12} \leq min(q_1, q_2)$$

Estimates of f range from 5%–80%[1]

$f = 0.05$ gives

$$2 \times 10^{-7} \leq p_{12} \leq 10^{-4}$$

[1] *Kellis et al, PNAS 2014*

Genetic correlation is a conservative estimate of $q_{1|2}$

Assume two traits Y_1, Y_2 can be modelled as

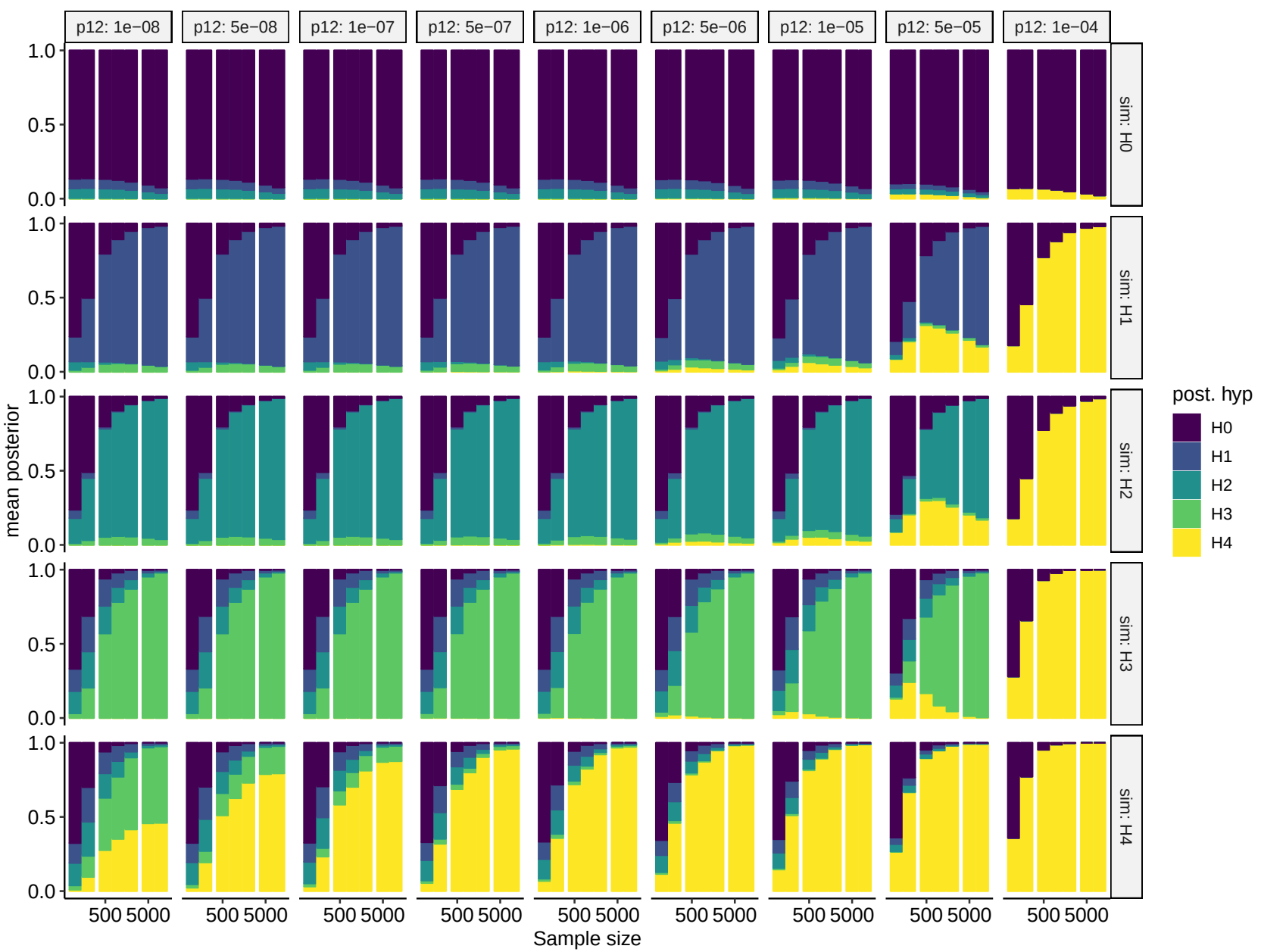
$$Y_1 = \sum_{i=1}^{n_{12}} \alpha_i G_i + \sum_{i=1}^{n_1} \beta_i H_i + E_1 \qquad Y_2 = \sum_{i=1}^{n_{12}} \alpha_i' G_i + \sum_{i=1}^{n_2} \gamma_i J_i + E_2$$

Key point: n_{12} variants shared, n_1, n_2 distinct to traits 1 and 2.

Then (under reasonable assumptions)

$$|r_g| \geq \frac{n_{12}}{\sqrt{(n_{12}+n_1)(n_{12}+n_2)}} = \frac{p_{12}}{\sqrt{q_1 q_2}}$$

Pragmatic p_{12}



Summary

Priors matter!

Default values for marginal priors seem reasonable

Always reasonable priors for joint causality probably don't exist

Questions:

1. How can I convince analysts to think about priors?
2. What information can we use to get better priors?
   - genetic correlation
   - relative overlap of open chromatin with GWAS trait between biologically relevant tissue and another tissue
3. Will it break your pipelines if I change (/remove?) the defaults?
4. Would sensitivity analysis be reportable?

Simplifying assumption of single causal variant

Single Causal Variant Assumption

At most one causal variant per trait in the studied region

Allows modelling the joint distribution via single SNP summary statistics

$$P(D|i \text{ causal}) = P(D_i|i \text{ causal}) \times P(D_{-i}|D_i, i \text{ causal})$$

✓ Simplicity of data: can use only p values if needed, no LD information or per-allele effect estimates

Maller et al, 2012

Single Causal Variant Assumption

At most one causal variant per trait in the studied region

Allows modelling the joint distribution via single SNP summary statistics

$$P(D|i \text{ causal}) = P(D_i|i \text{ causal}) \times P(D_{-i}|D_i, i \text{ causal})$$

✓ Simplicity of data: can use only p values if needed, no LD information or per-allele effect estimates

✗ Unrealistic to always assume single causal variant

Can be relaxed using **stepwise conditioning**, but then effect estimates **aligned to** reference LD matrix are needed

Maller et al, 2012

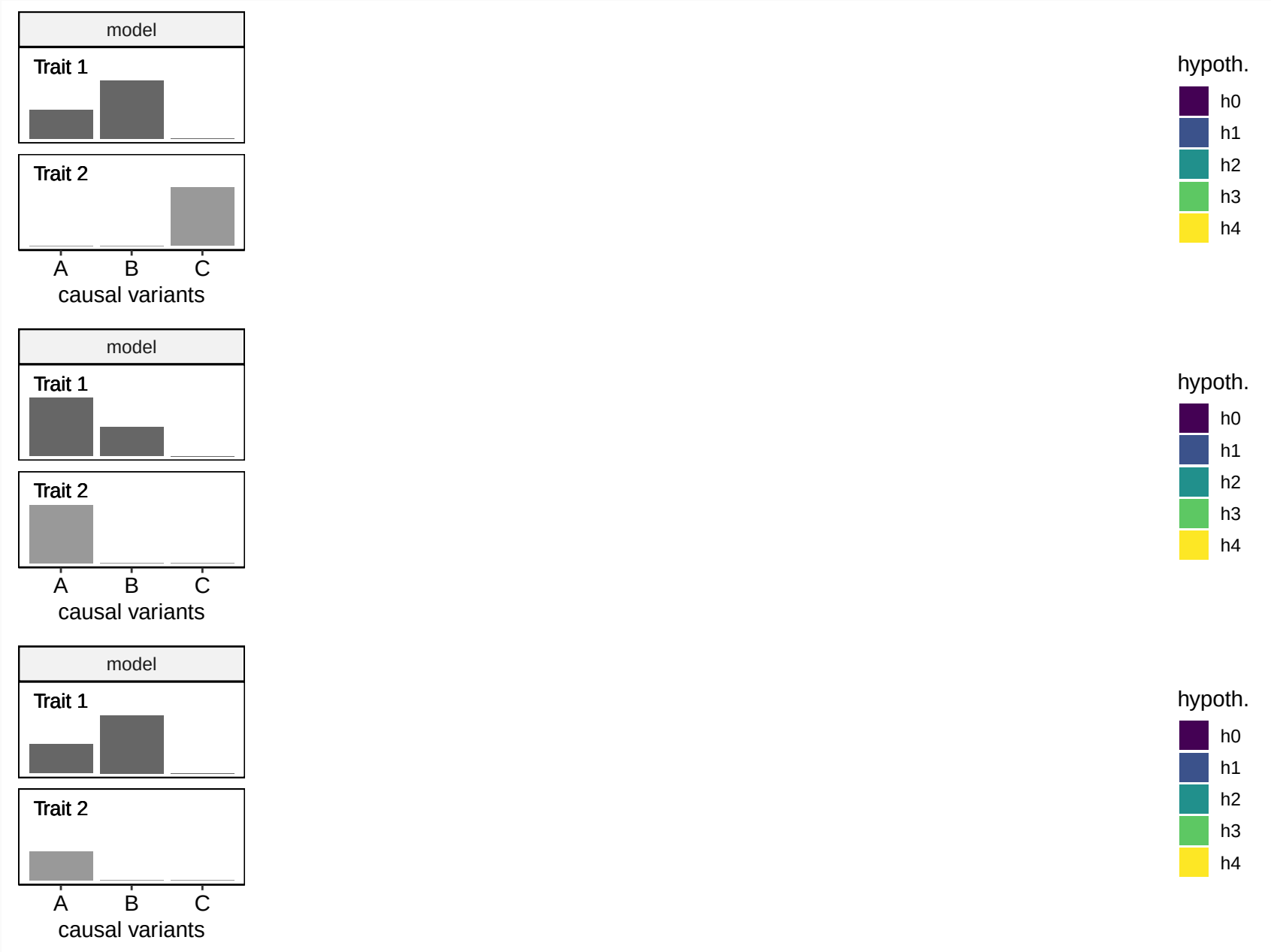
Review of current practice

Out of 25 papers which used coloc in 2018 ...

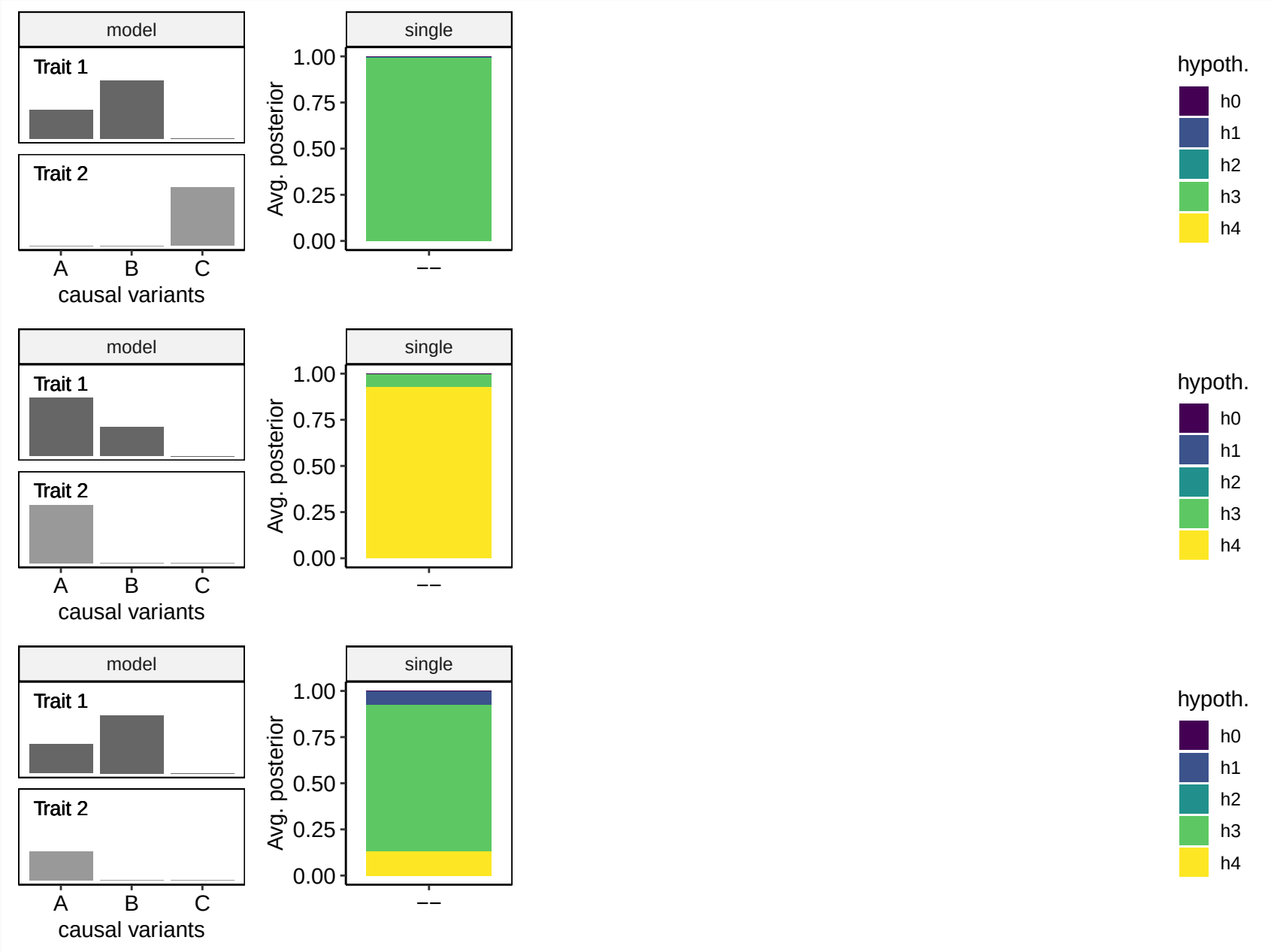
... 1 used conditioning*

*on one trait for which they had full data

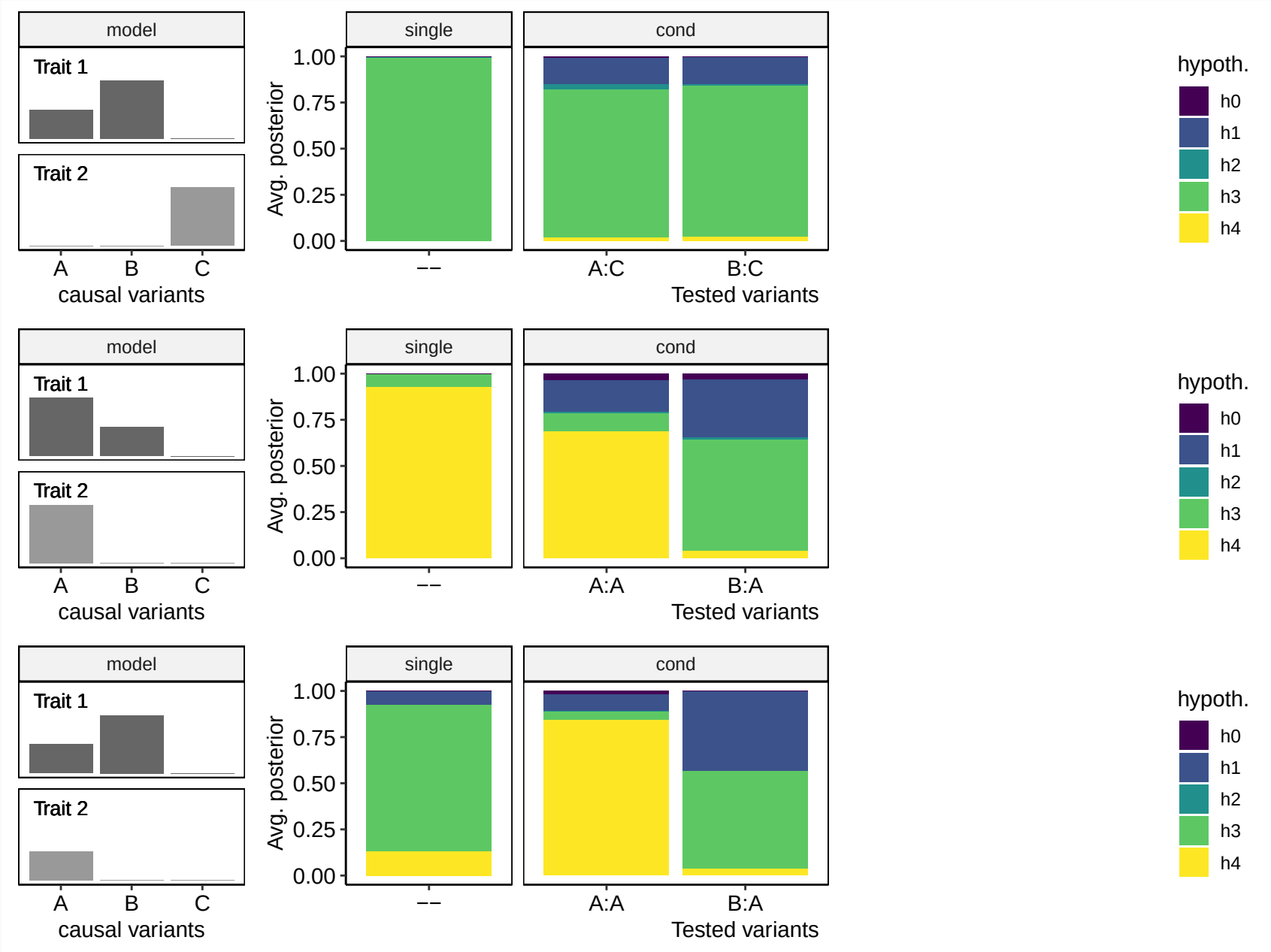
Violating the single causal variant assumption



Violating the single causal variant assumption



Violating the single causal variant assumption



Masking as an alternative to Conditioning

Algorithm 1: Conditioning

Result: S = a set of selected SNPs
Initialise S as an empty set;
while *TRUE* **do**
    Consider T = all SNPs not in S;
    Fit a model to each SNP + S and compare to a model containing only S;
    Choose the *best SNP* amongst these;
    if *the best SNP is significant* **then**
        add it to S;
    else
        break

Masking as an alternative to Conditioning

Algorithm 1: Conditioning

Result: S = a set of selected SNPs
Initialise S as an empty set;
while *TRUE* **do**
  Consider T = all SNPs not in S;
  Fit a model to each SNP + S and compare to a model containing only S;
  Choose the *best SNP* amongst these;
  if *the best SNP is significant* **then**
    add it to S;
  else
    break

Algorithm 2: Masking

Result: S = a set of selected SNPs
Initialise S as an empty set;
while *TRUE* **do**
  Consider T = all SNPs not in LD with any SNP in S;
  Fit a model to each SNP in T;
  Choose the *best SNP* among these;
  if *the best SNP is significant* **then**
    add it to S;
  else
    break

Masking as an alternative to Conditioning

Algorithm 1: Conditioning

Result: S = a set of selected SNPs
Initialise S as an empty set;
while *TRUE* **do**
    Consider T = all SNPs not in S;
    Fit a model to each SNP + S and compare to a model containing only S;
    Choose the *best SNP* amongst these;
    if *the best SNP is significant* **then**
        add it to S;
    else
        break

Algorithm 2: Masking

Result: S = a set of selected SNPs
Initialise S as an empty set;
while *TRUE* **do**
    Consider T = all SNPs not in LD with any SNP in S;
    Fit a model to each SNP in T;
    Choose the *best SNP* among these;
    if *the best SNP is significant* **then**
        add it to S;
    else
        break

Masking as an alternative to Conditioning

Algorithm 1: Conditioning

Result: S = a set of selected SNPs
Initialise S as an empty set;
while *TRUE* **do**
  Consider T = all SNPs not in S;
  Fit a model to each SNP + S and compare to a model containing only S;
  Choose the *best SNP* amongst these;
  if *the best SNP*
    add it to S;
  else
    break

Algorithm 2: Mask

Result: S = a set
Initialise S as an e
while *TRUE* **do**
  Consider T = all SNPs not in LD with any SNP in S;
  Fit a model to each SNP in T;
  Choose the *best SNP* among these;
  if *the best SNP is significant* **then**
    add it to S;
  else
    break

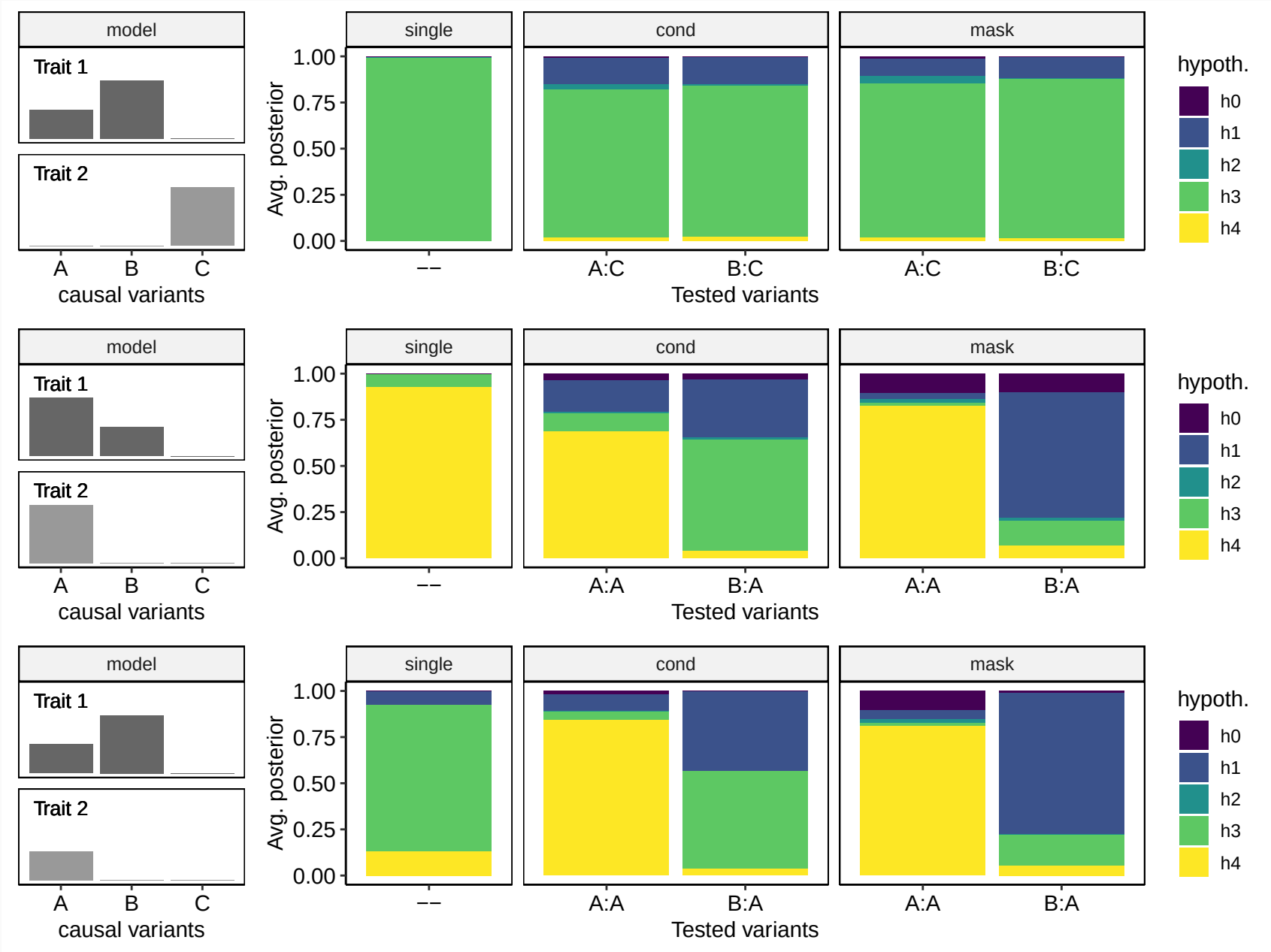
Masking...

✓Data requirements remain “simple”

✓can help with multi-ethnic studies

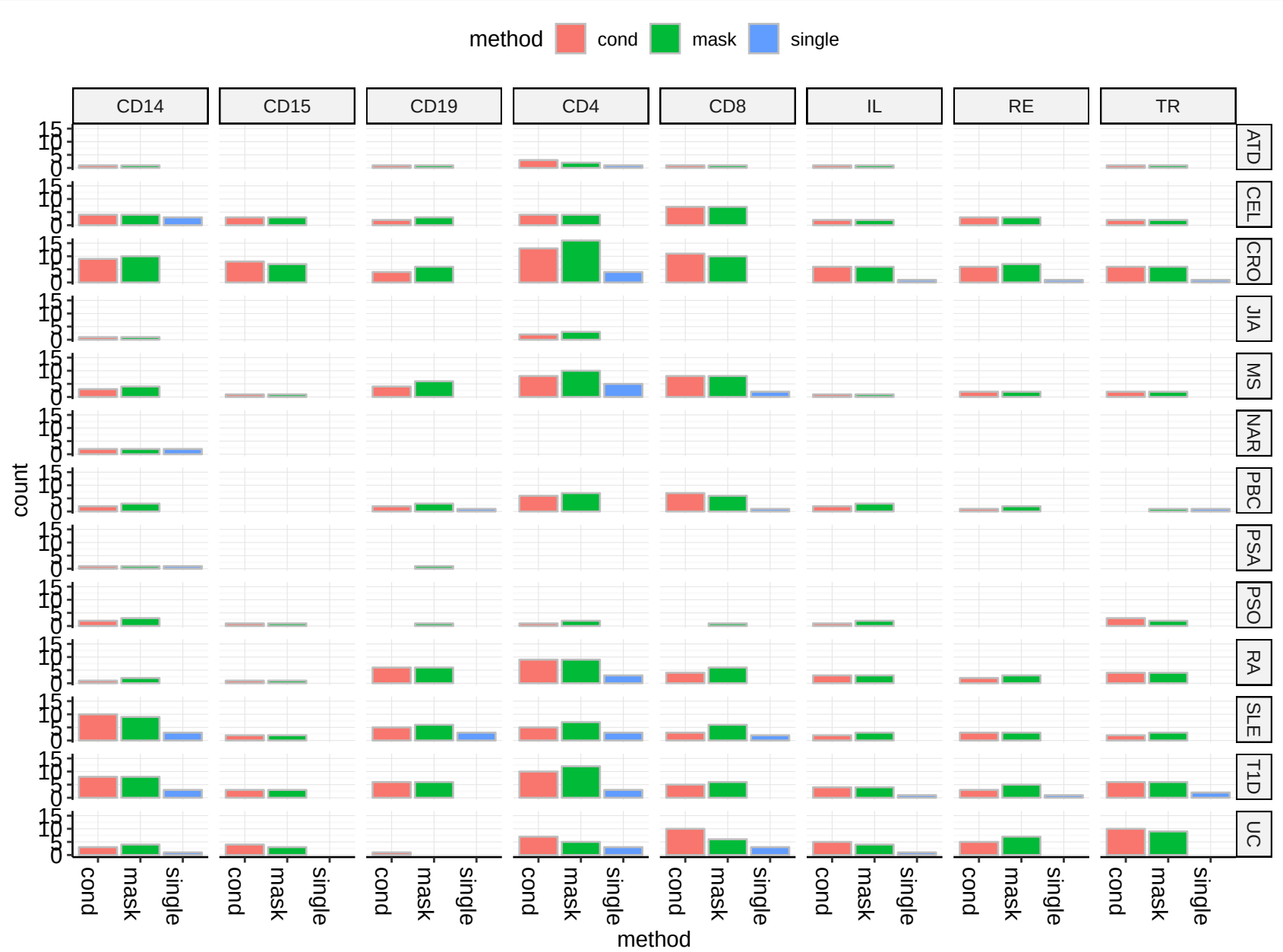
✗Difficult to balance maximising discovery/ avoiding false second signals

Conditioning and masking allow multiple comparisons

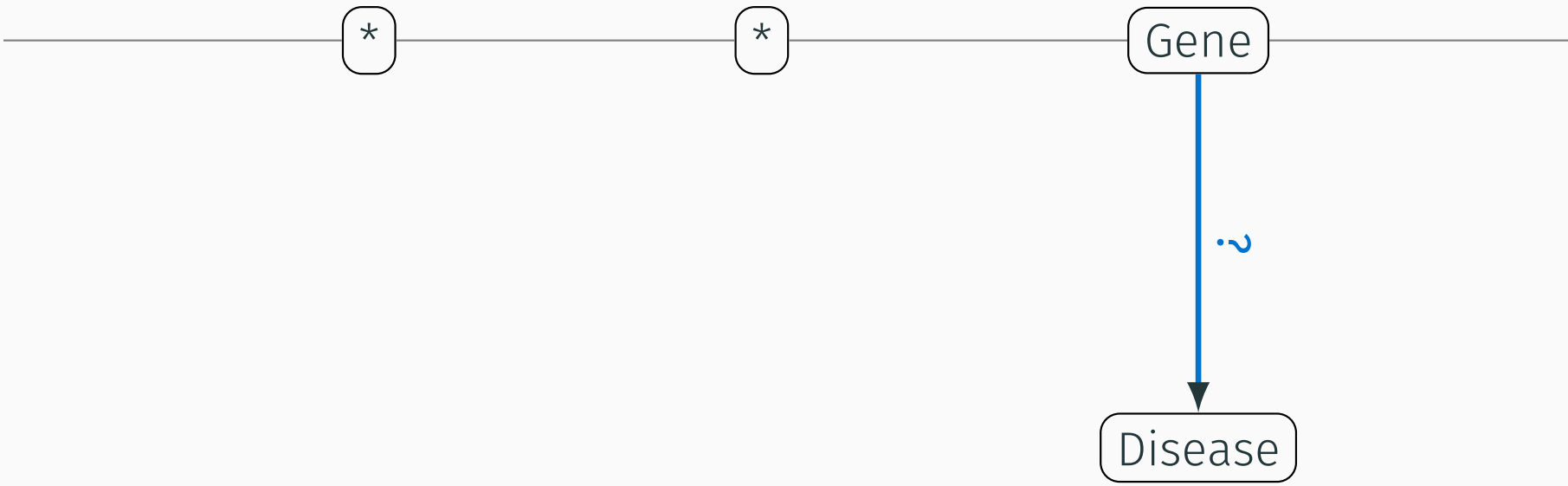


Impact on results

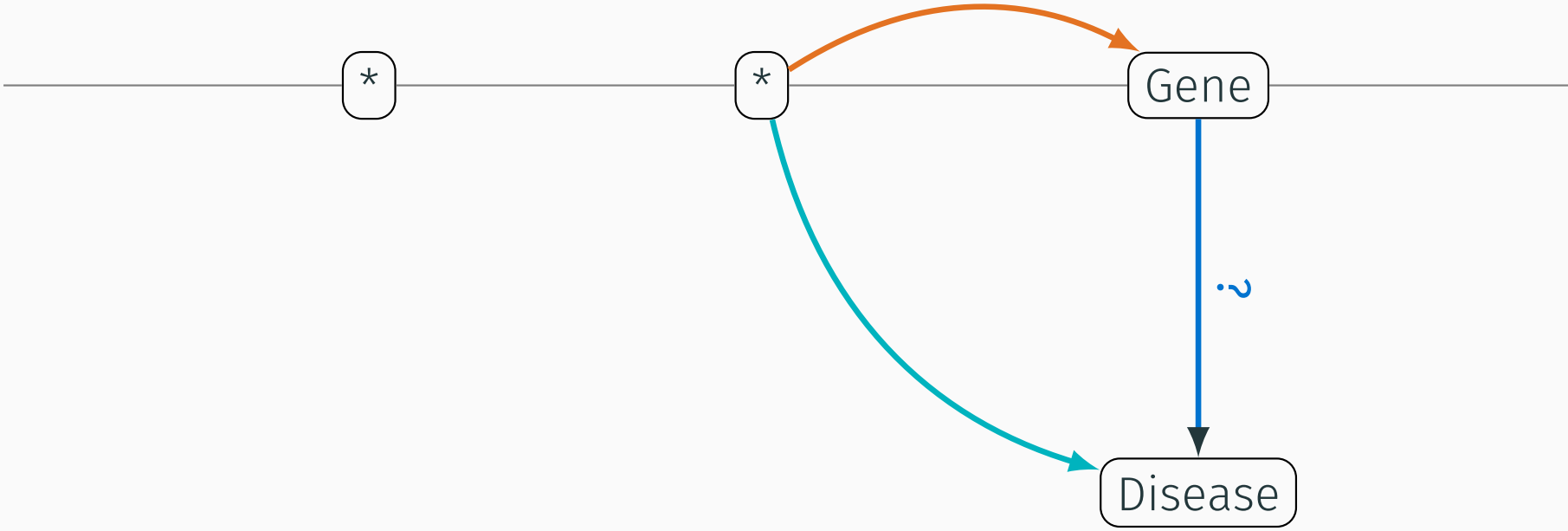
Colocalisation of eQTL (8 cell types) vs 13 diseases



Impact on interpretation

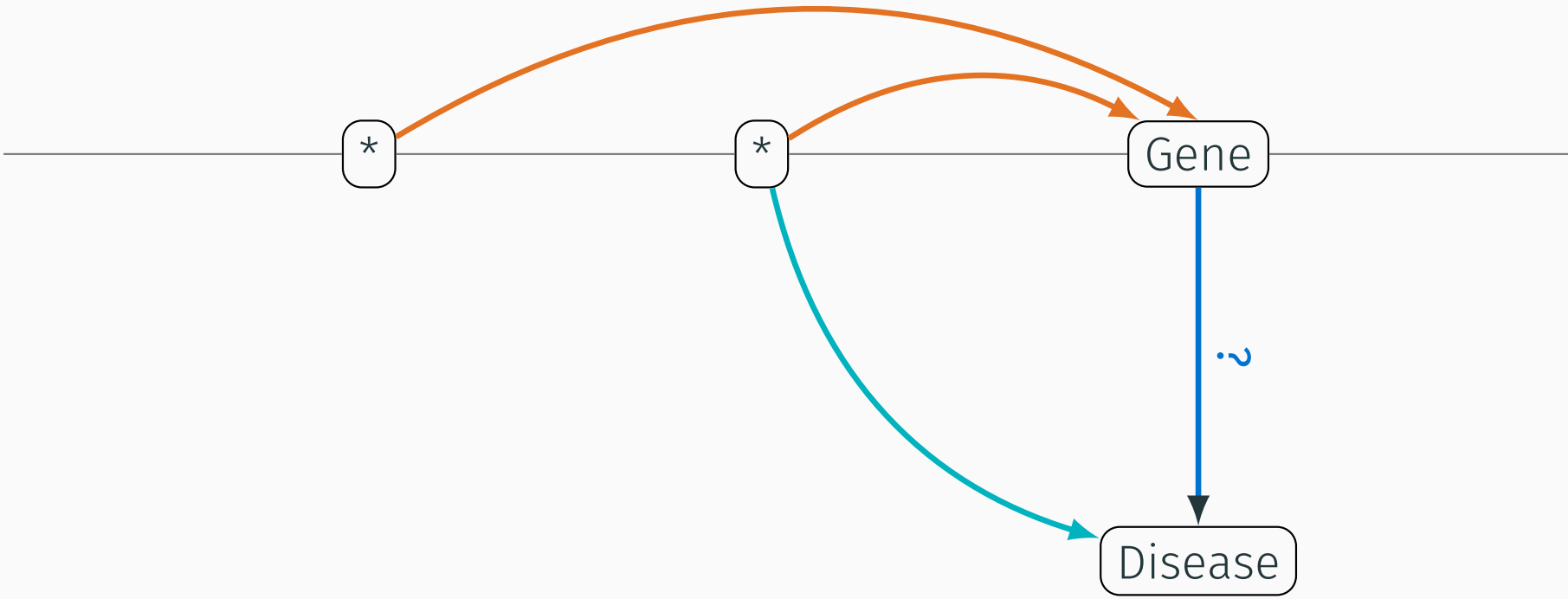


Impact on interpretation



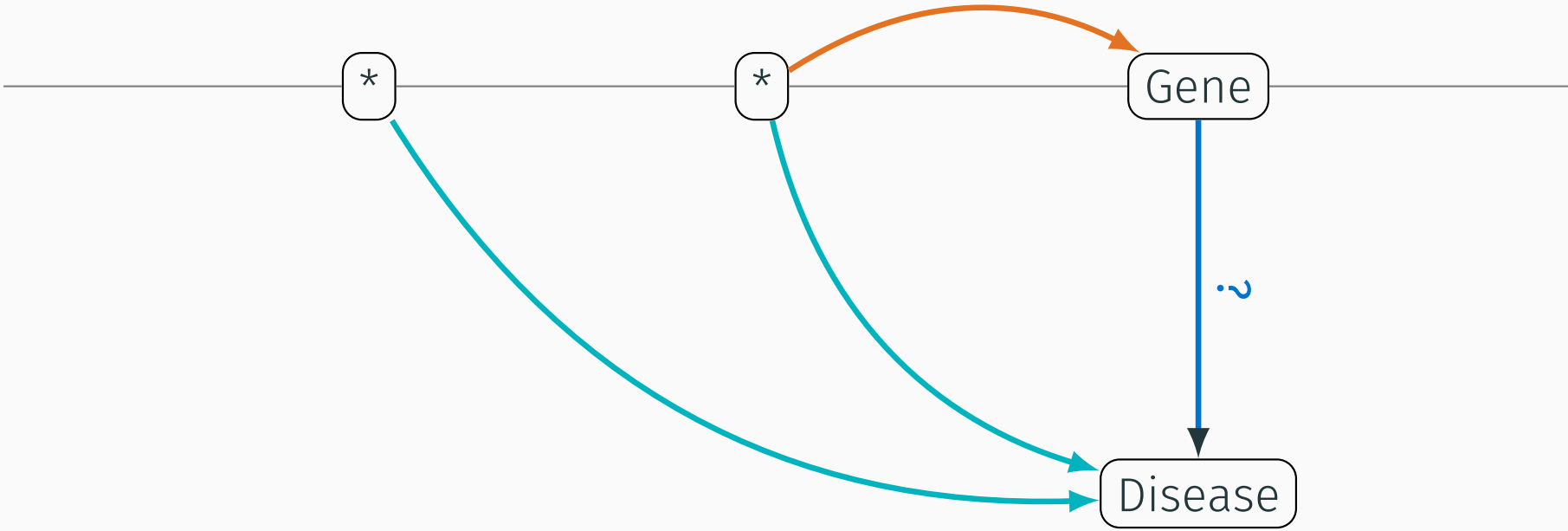
Strong evidencepleiotropy still possible

Impact on interpretation



Suggestive evidence
loss of power or ”wrong” cell type/state

Impact on interpretation



Medium-strong evidence
secondary effect can have different mechanism

Summary

Conditioning gives more detailed view, but need to be more nuanced in interpretation of partial colocalisation

Masking can approximate conditioning when aligned alleles not known

Methods still in development

Questions:

1. How important is the “no alignment” problem?
2. How important is multi-ethnic summary statistic analysis?

Thanks to...

Stasia Grinberg

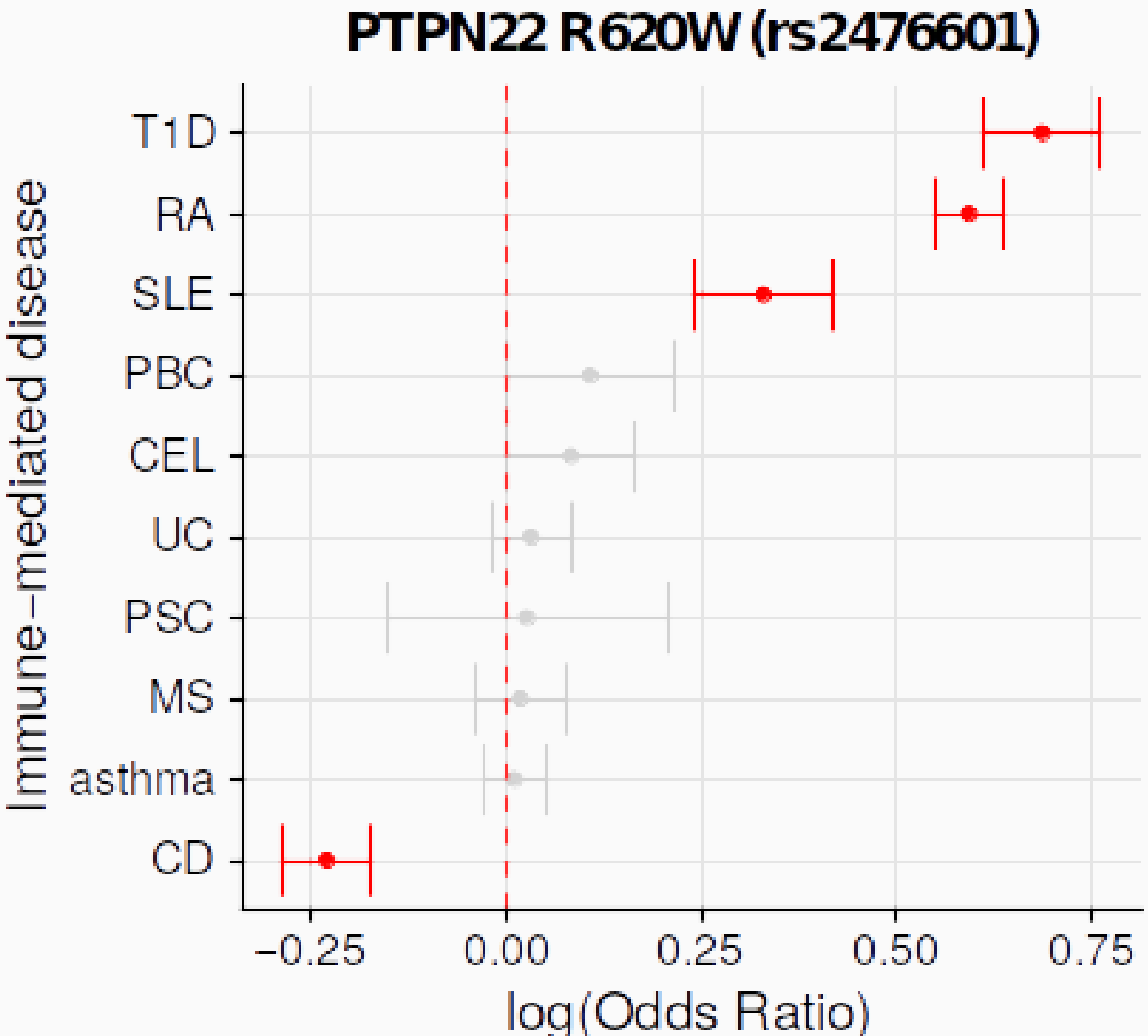
wellcome

UNIVERSITY OF CAMBRIDGE

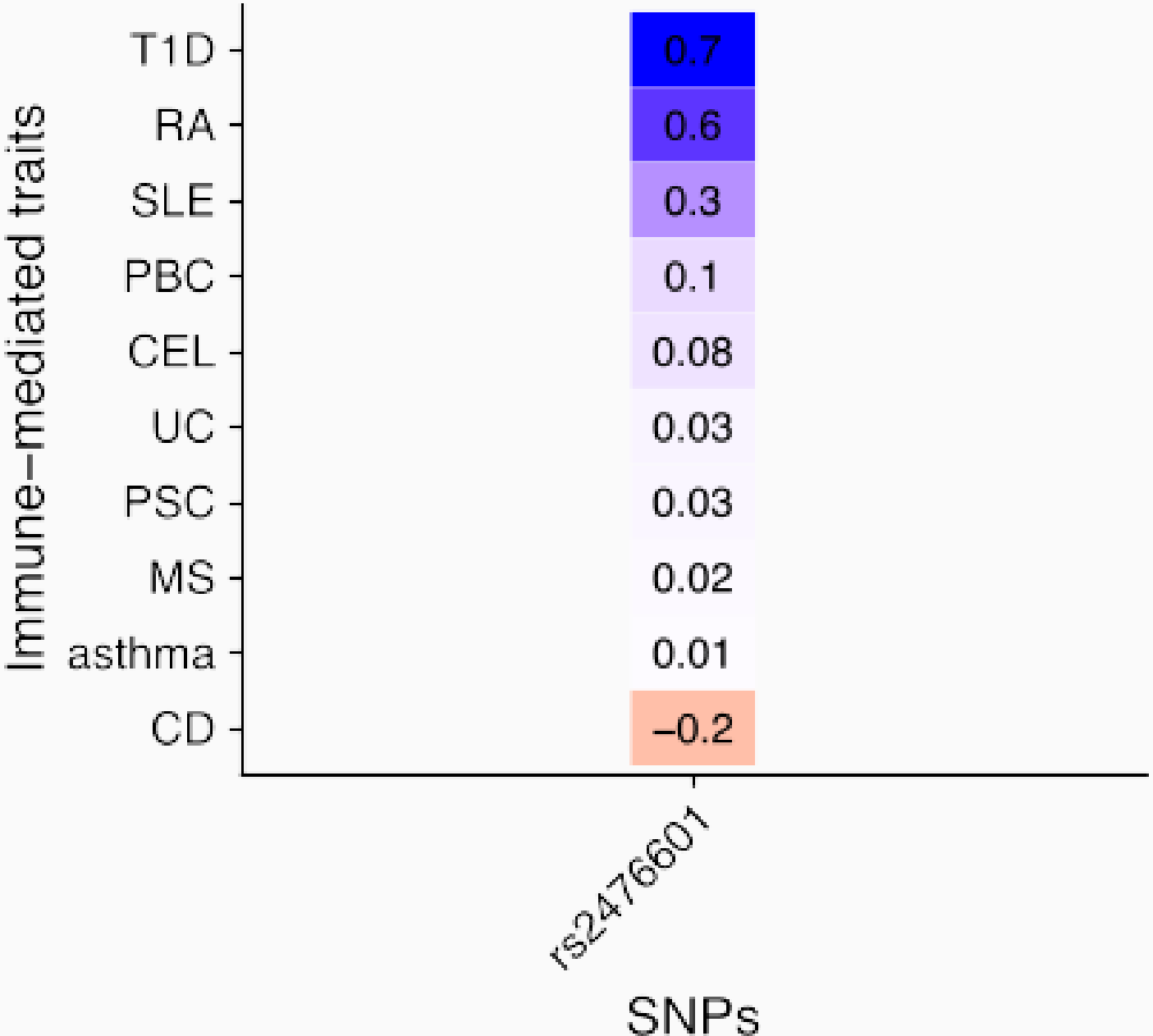
MRC | Biostatistics Unit

Informed dimension reduction

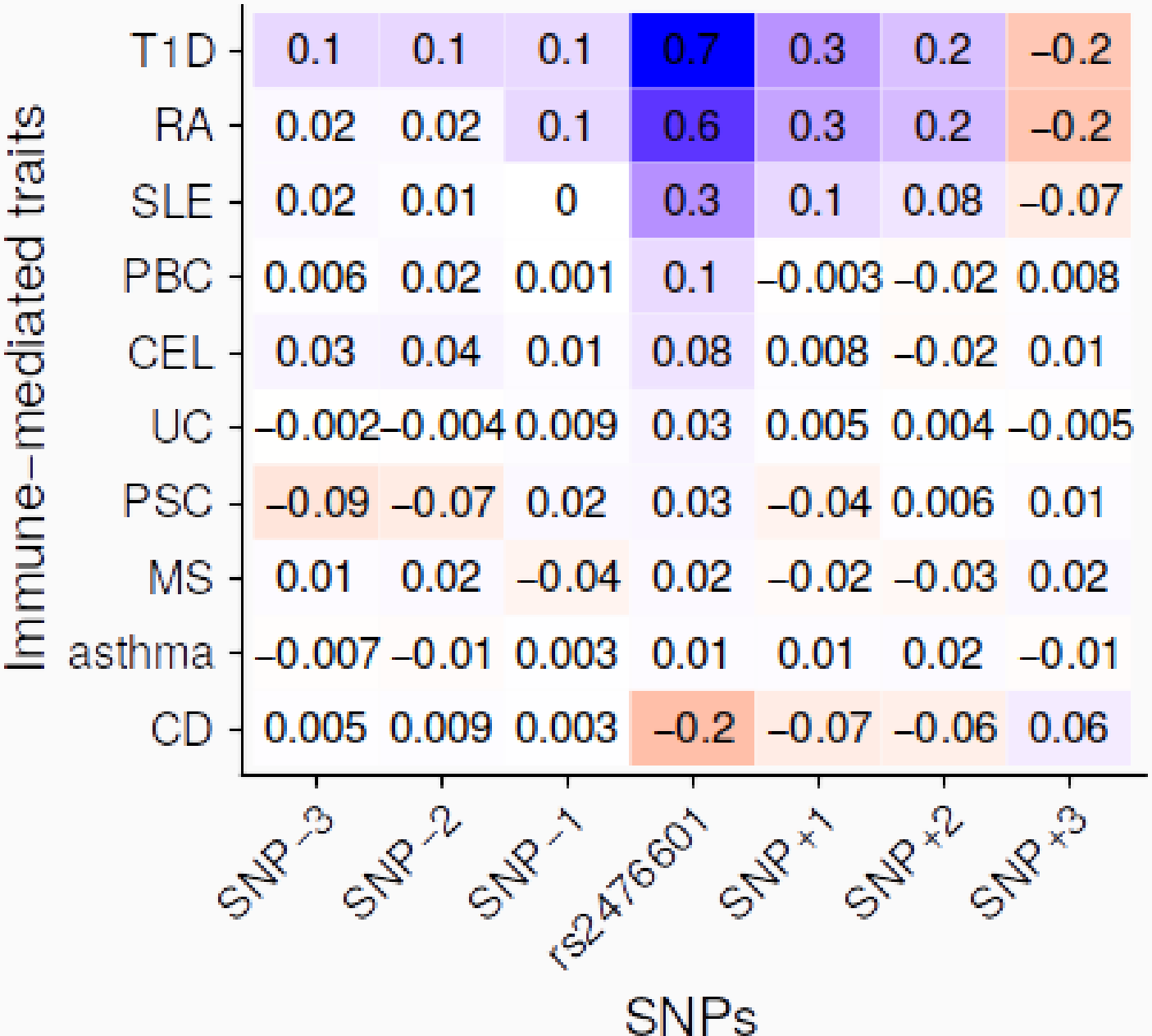
Heterogeneity in effects across diseases



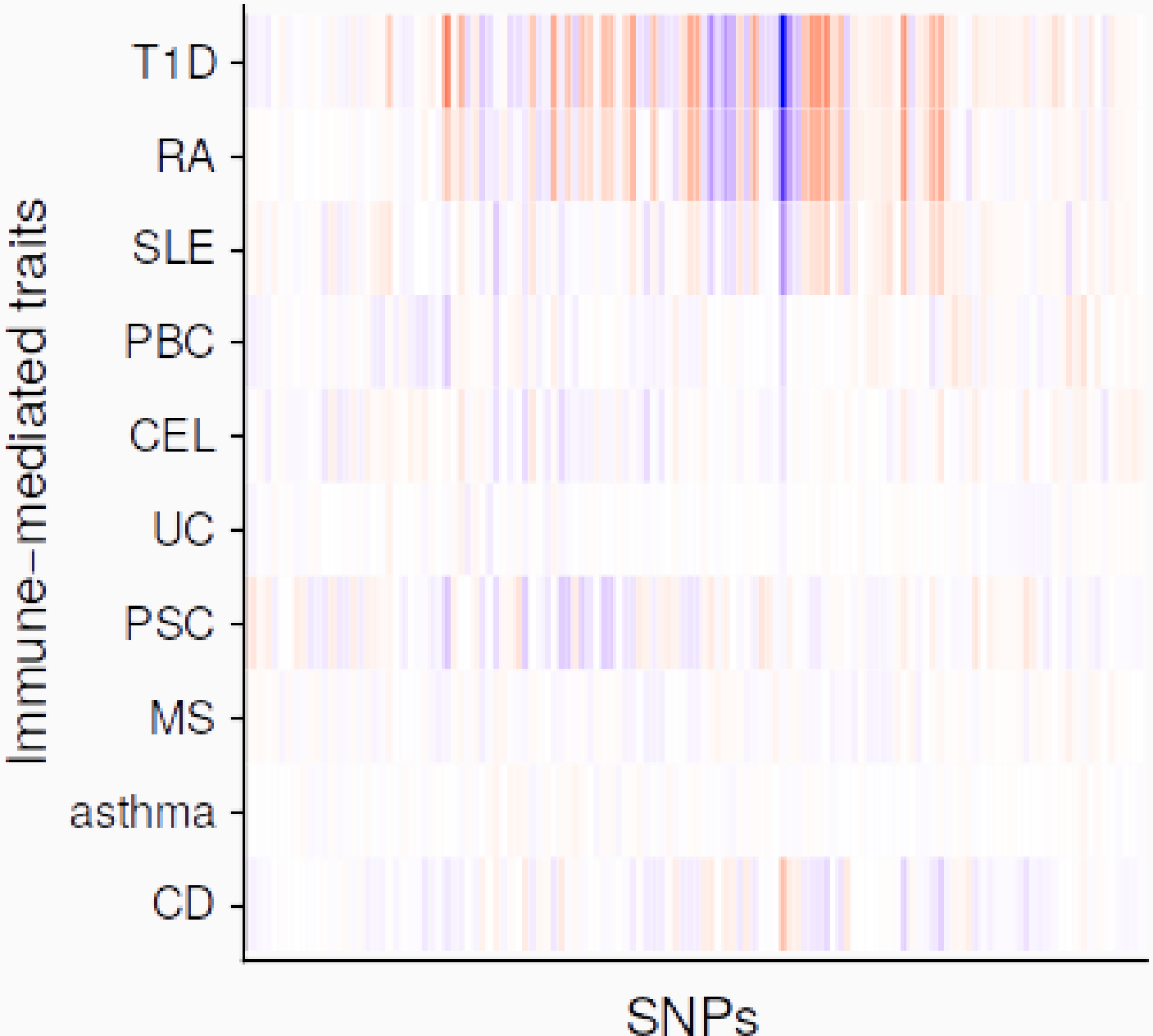
Heterogeneity in effects across diseases



Heterogeneity in effects across diseases

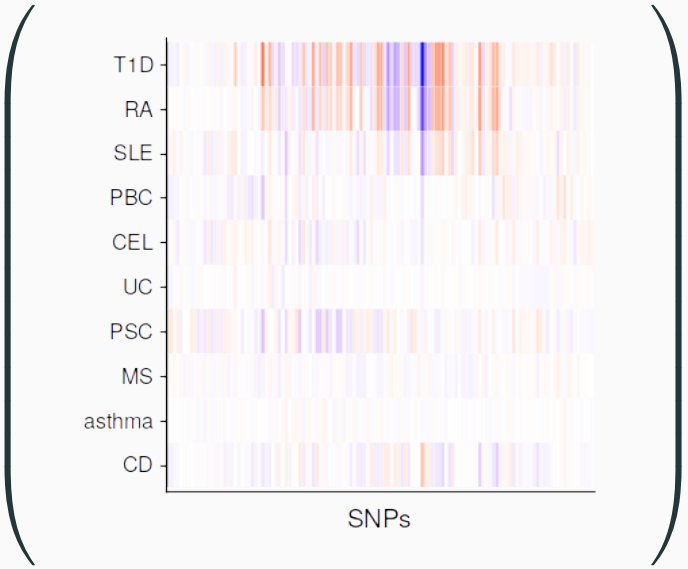


Heterogeneity in effects across diseases

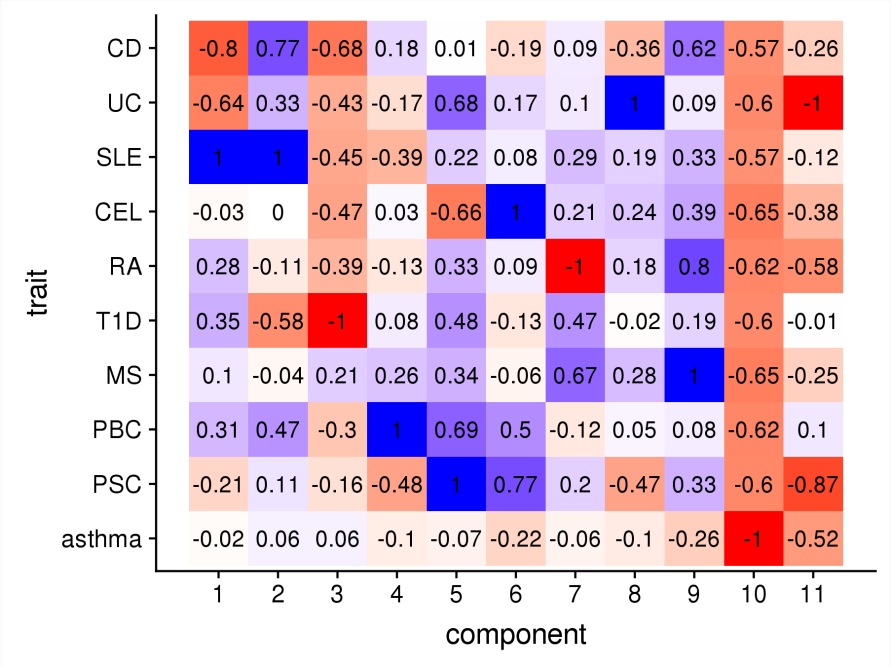


Principal components analysis to generate a new “basis”

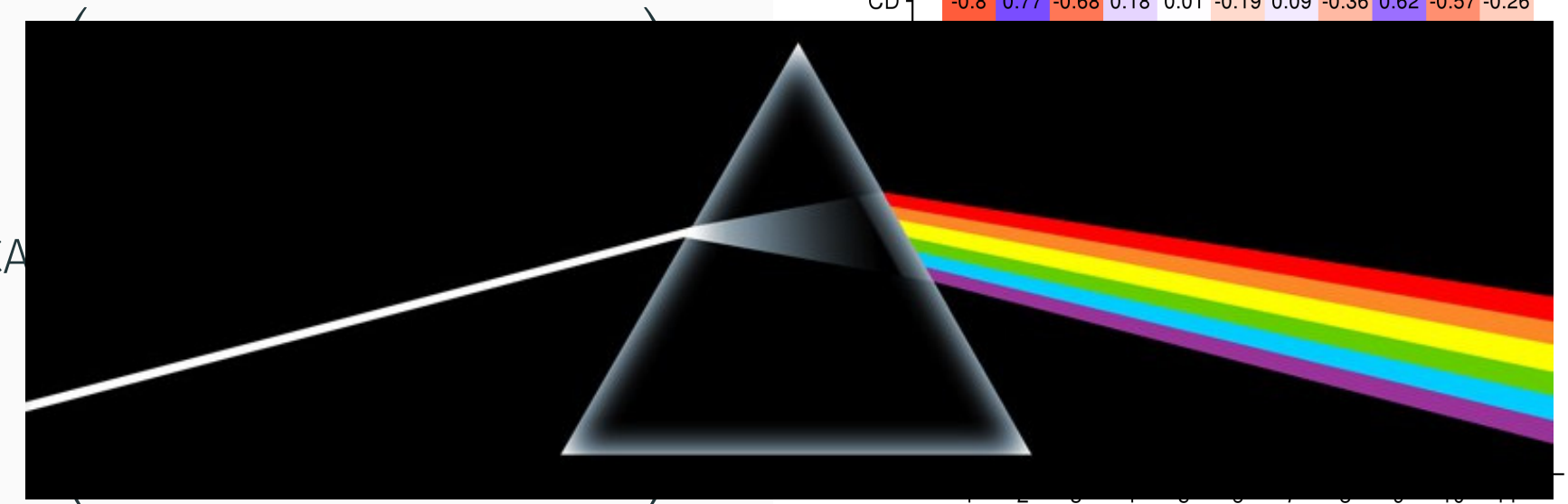
PCA



=



Principal components analysis to generate a new “basis”

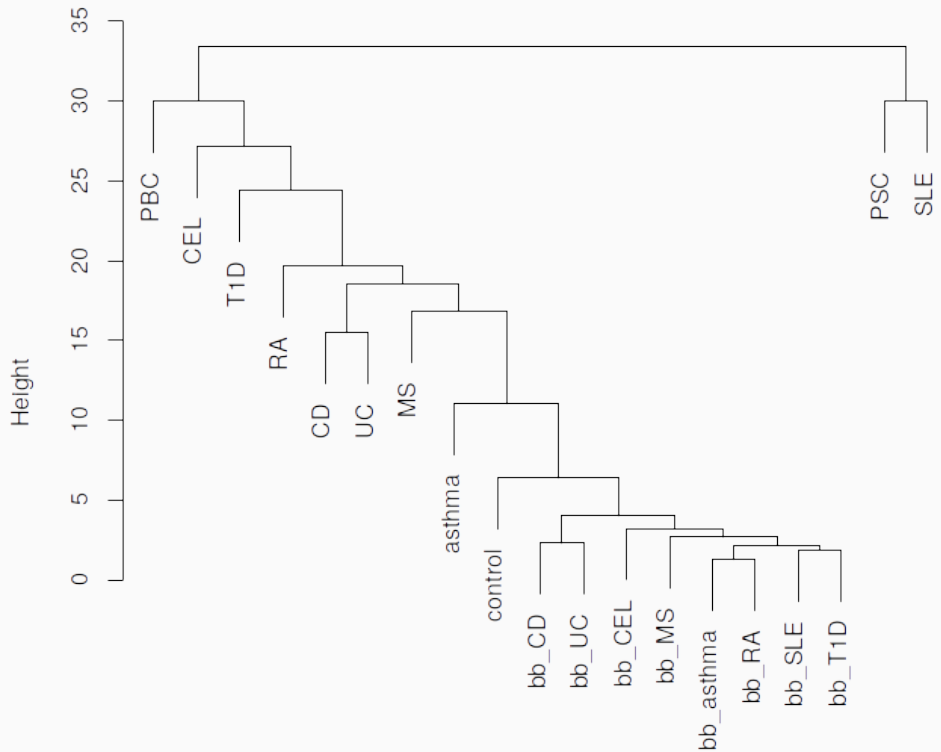


Basis diseases

Disease	Study	Cases	Controls
asthma	Demenais	19954	107715
Rheumatoid arthritis	Okada	14361	43923
Ulcerative colitis	de Lange	12366	33609
Crohn's disease	de Laange	12194	28072
Multiple sclerosis	IMSGC	9772	17376
Type 1 diabetes	Cooper	5913	8829
Primary sclerosing cholangitis	Ji	4796	19955
Coeliac disease	Dubois	4533	10750
Systemic lupus erythematosus	Bentham	4036	6959
Primary biliary cholangitis	Cordell	2764	10475

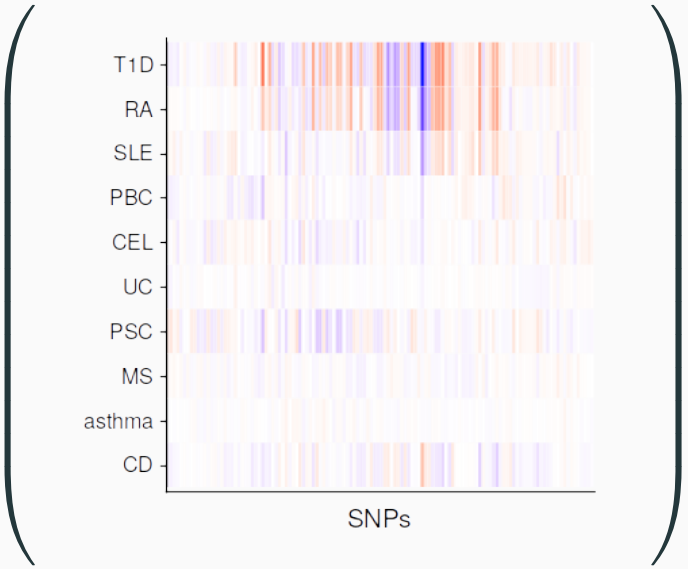
Naive basis captures dataset, not disease

UK Biobank	Cases
Type 1 diabetes	286
Systemic lupus erythematosus	366
Multiple sclerosis	1,228
Coeliac disease	1,452
Ulcerative colitis	1,795
Crohn's disease	1,032
Rheumatoid arthritis	3,730
Asthma	39,049

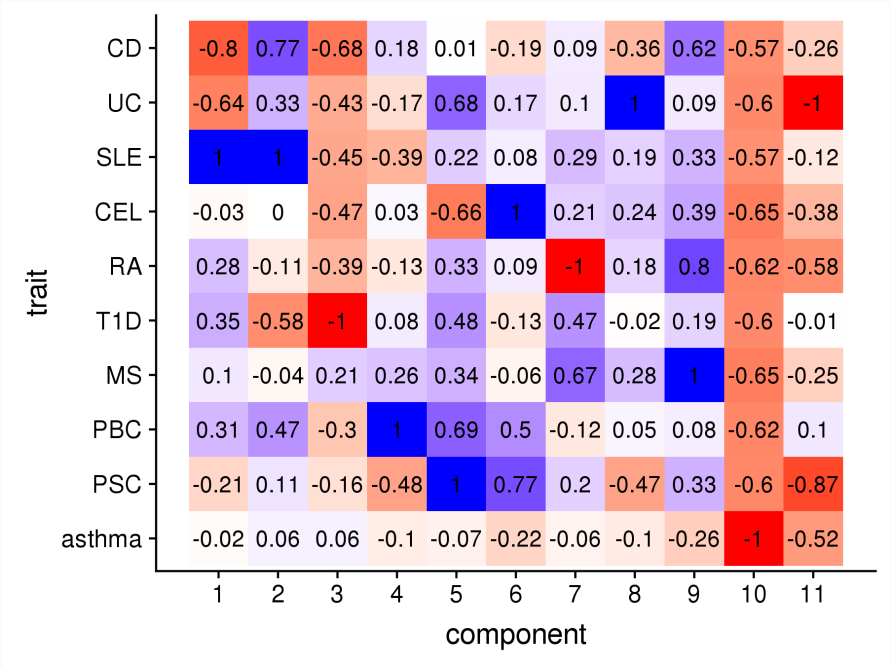


Solution: focus data before principal components analysis

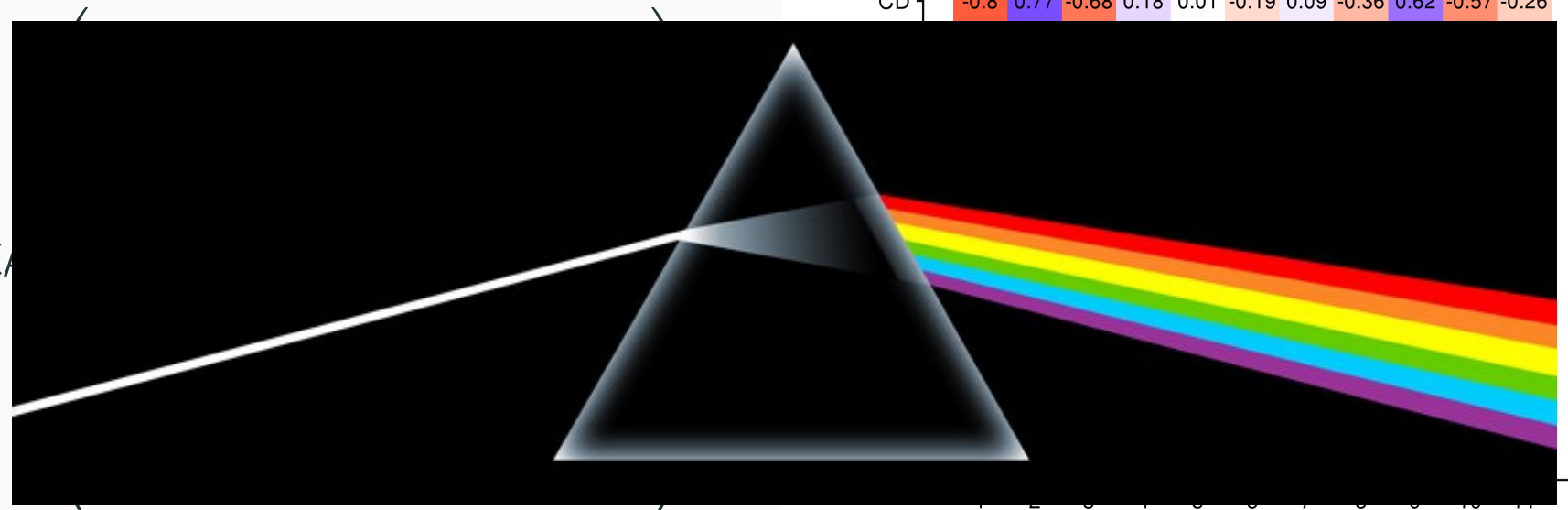
PCA



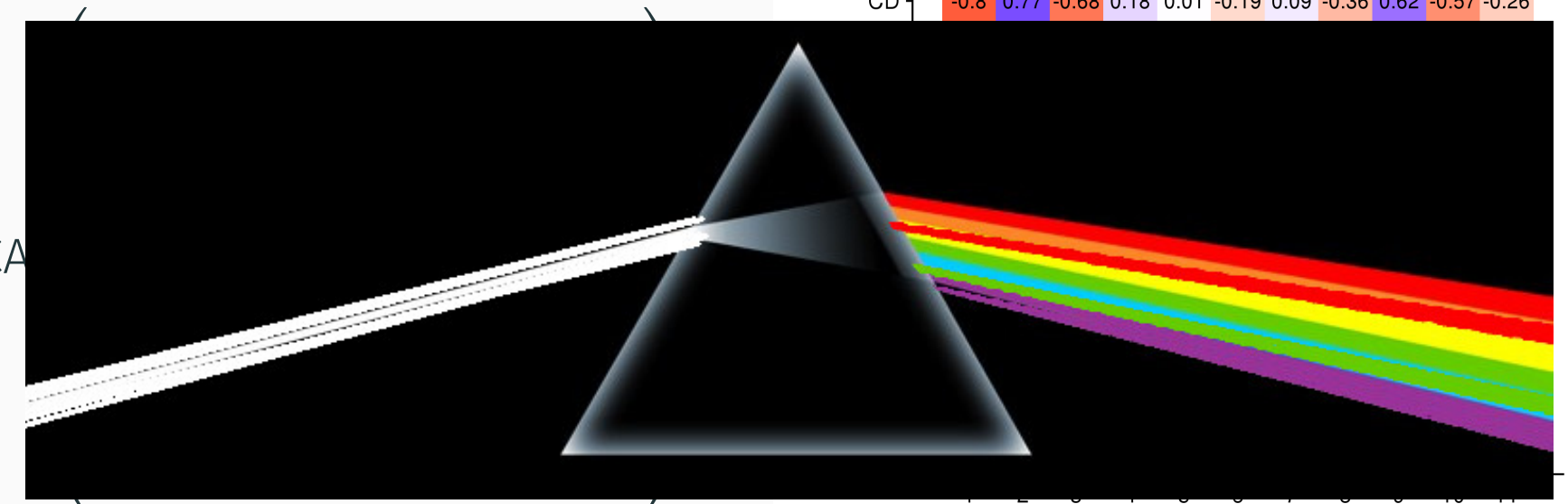
=



Solution: focus data before principal components analysis

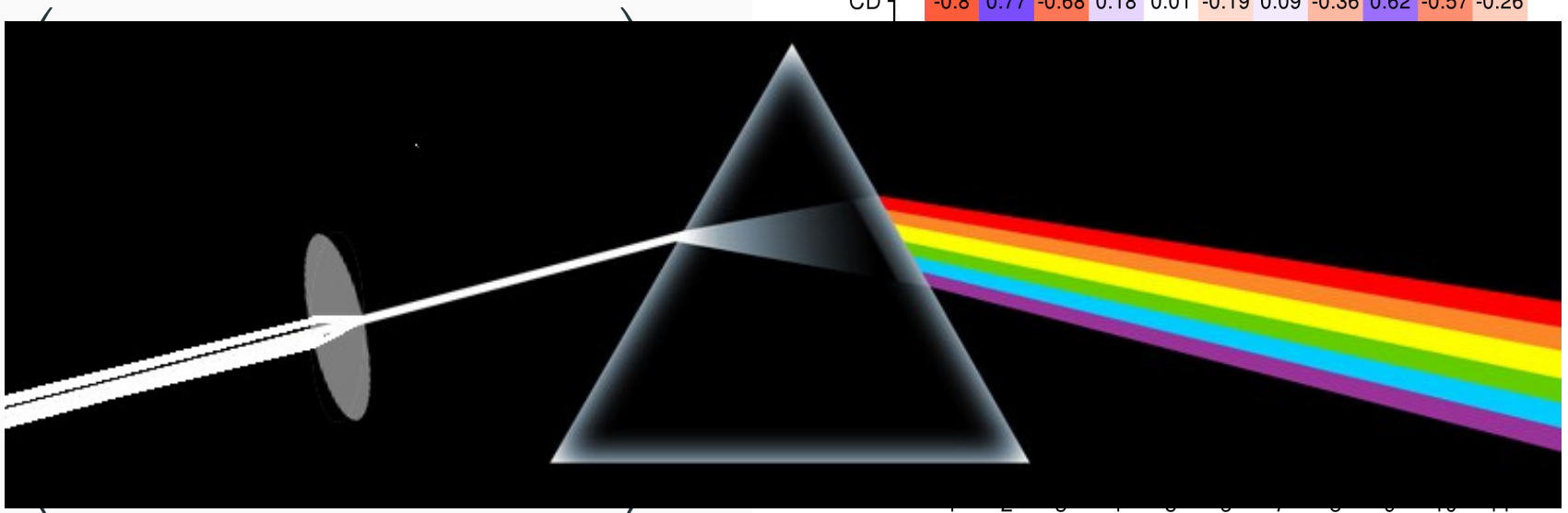


Solution: focus data before principal components analysis



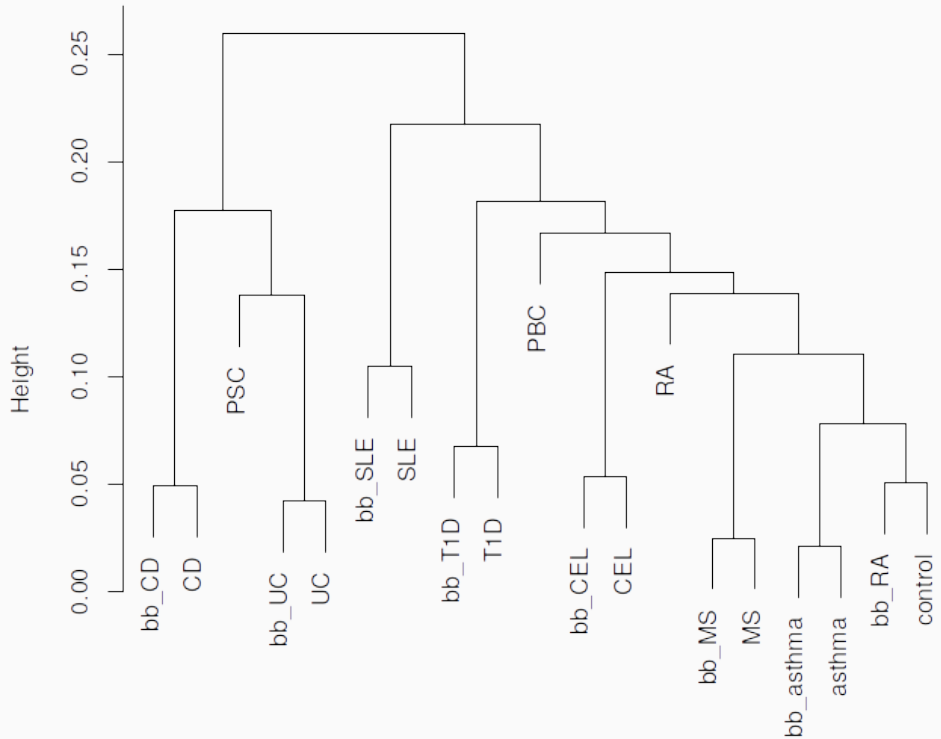
Solution: focus data before principal components analysis

PCA

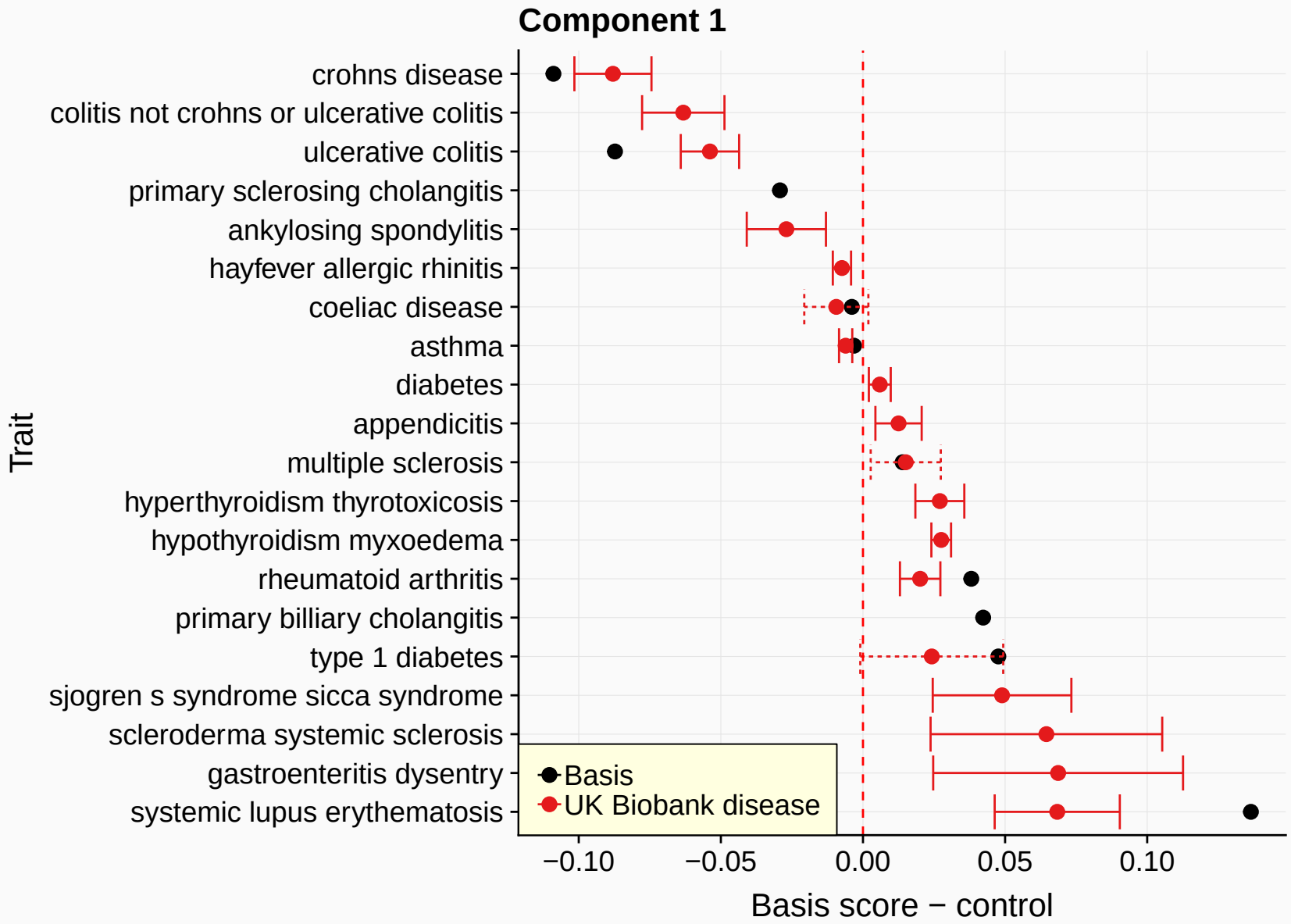


Focused basis captures disease information

UK Biobank	Cases
Type 1 diabetes	286
Systemic lupus erythematosus	366
Multiple sclerosis	1,228
Coeliac disease	1,452
Ulcerative colitis	1,795
Crohn's disease	1,032
Rheumatoid arthritis	3,730
Asthma	39,049

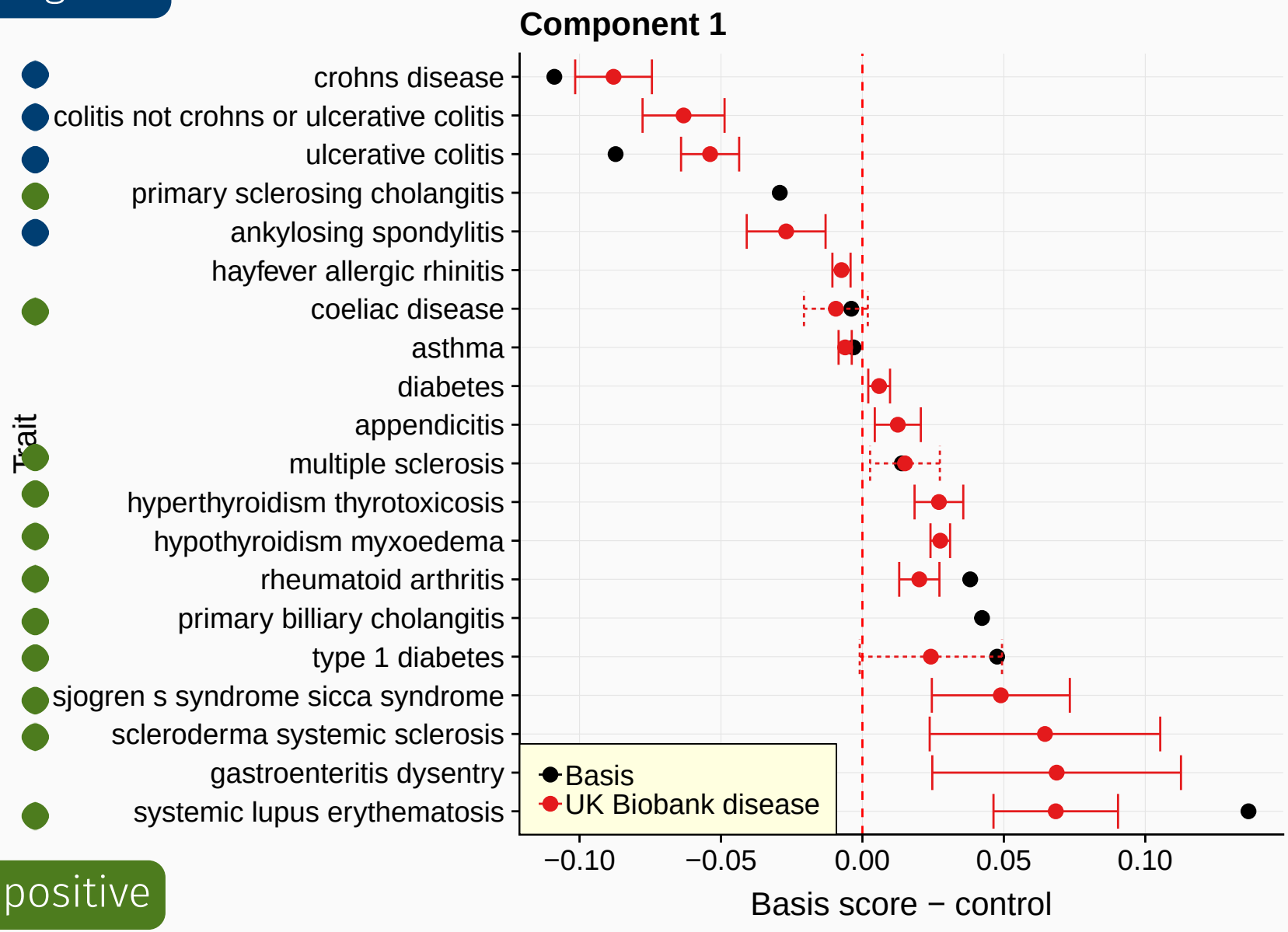


Component 1: sero+/- disease, innate/adaptive



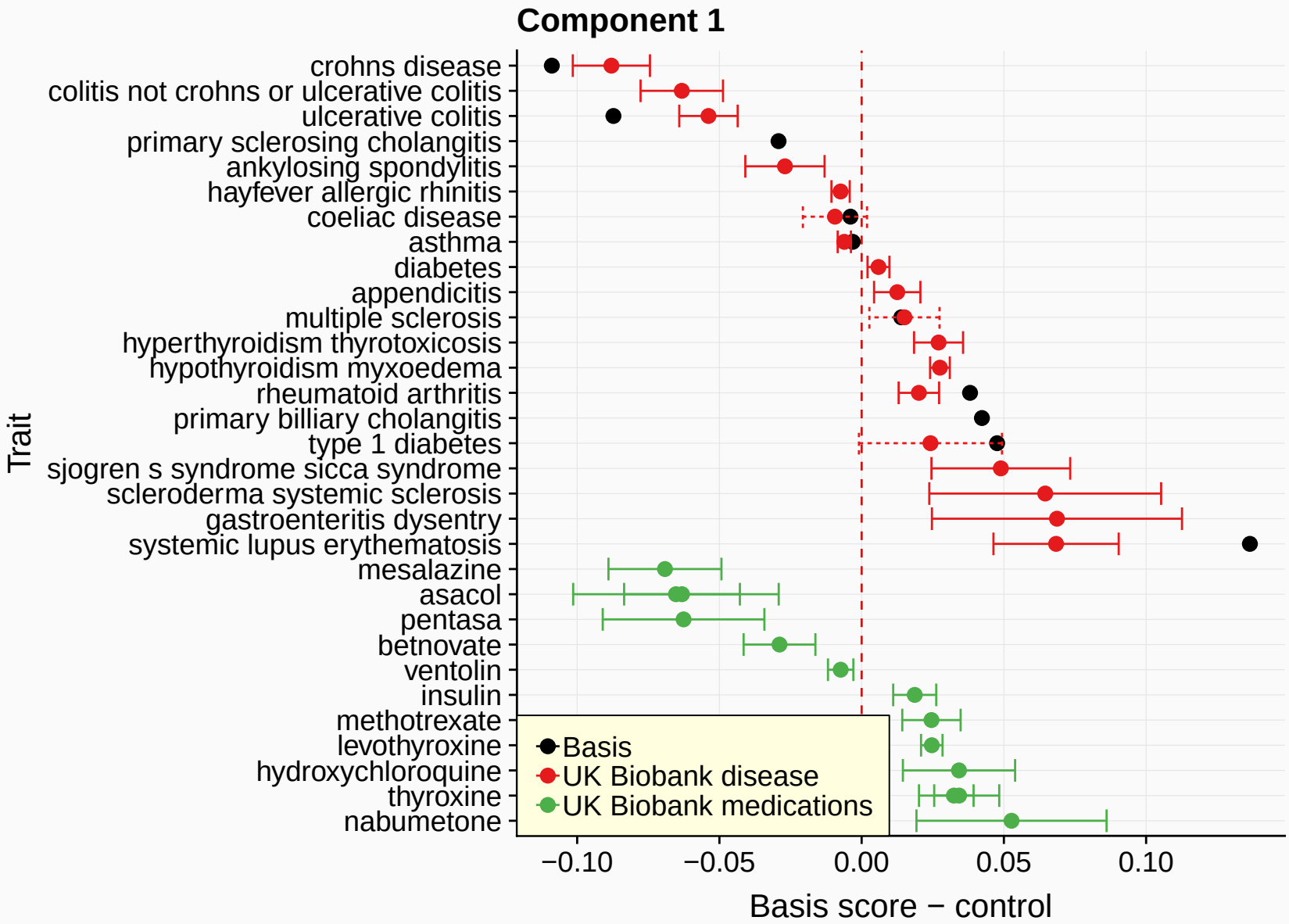
Component 1: sero+/- disease, innate/adaptive

seronegative

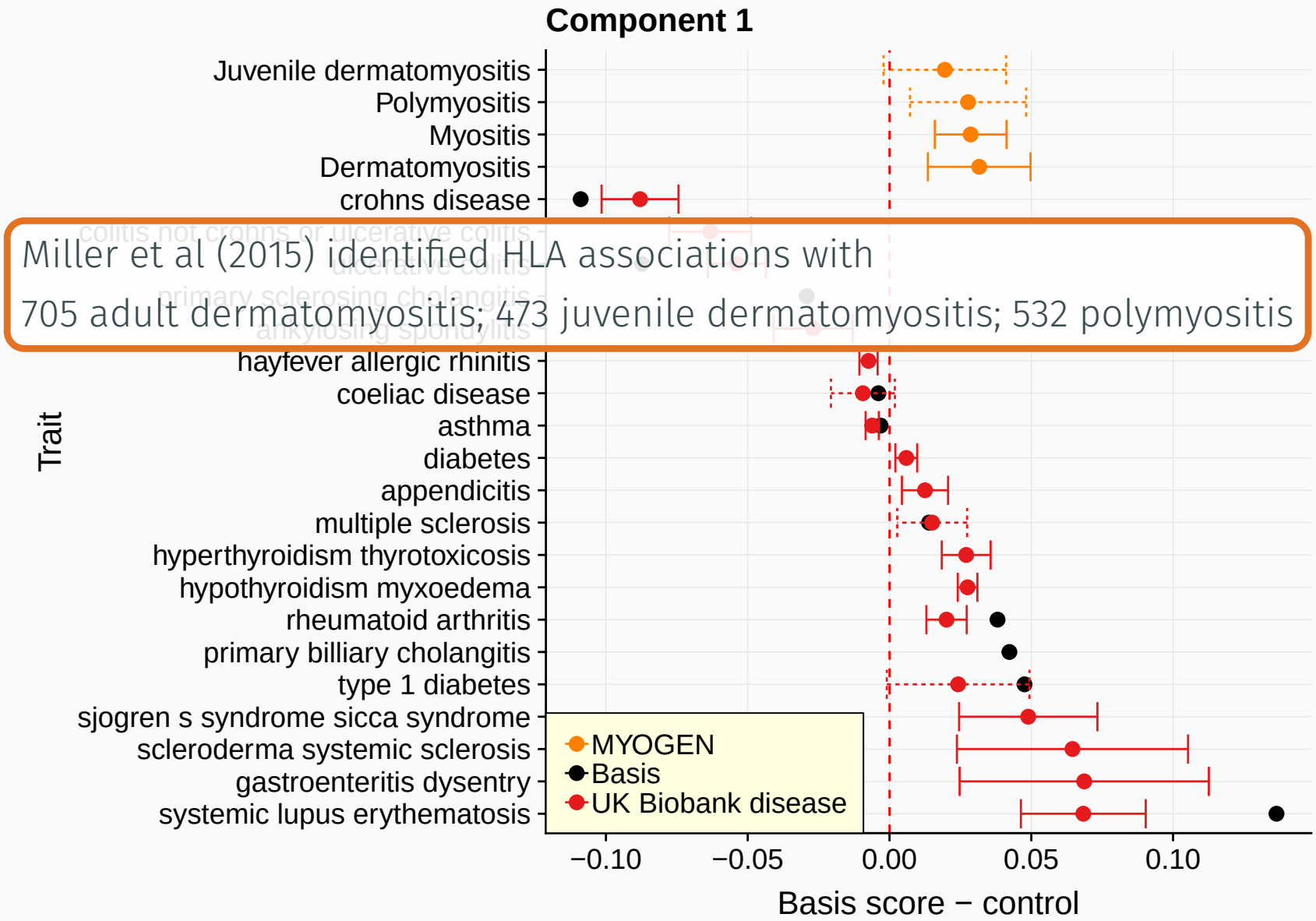


seropositive

Component 1: sero+/- disease, innate/adaptive



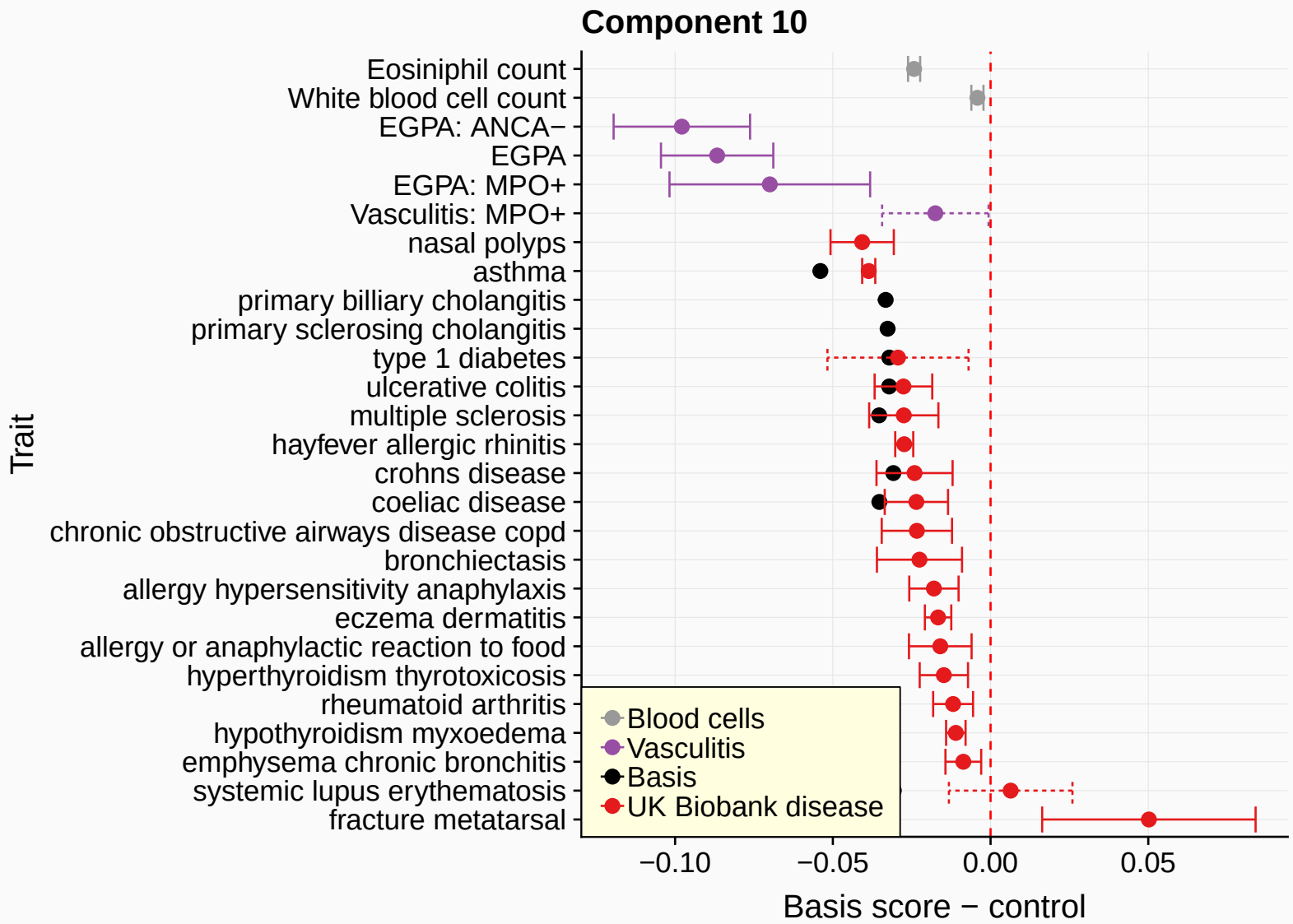
Component 1: sero+/- disease, innate/adaptive



Miller et al (2015) identified HLA associations with

705 adult dermatomyositis; 473 juvenile dermatomyositis; 532 polymyositis

Vasculitis



Astle et al, 2016; Lyons et al 2019

Summary

- Genetic sharing across immune mediated diseases is pervasive, but complex

Summary

- Genetic sharing across immune mediated diseases is pervasive, but complex
- Sharing **should** be exploited to learn about less common diseases

Summary

- Genetic sharing across immune mediated diseases is pervasive, but complex
- Sharing **should** be exploited to learn about less common diseases

Summary

- Genetic sharing across immune mediated diseases is pervasive, but complex
- Sharing **should** be exploited to learn about less common diseases
- Dimension reduction obscures individual effects, offers manageable holistic view of multiple diseases

Summary

- Genetic sharing across immune mediated diseases is pervasive, but complex
- Sharing **should** be exploited to learn about less common diseases
- Dimension reduction obscures individual effects, offers manageable holistic view of multiple diseases
  - Focusing on OR rather than significance puts everything on same scale
  - Improves power: **many** fewer tests, stronger prior for association
  - Components are directions, from x to y, but interpretation difficult
  - Given difference on component, can identify driving SNPs - “zoom out”

Thanks to...

Olly Burren

Patients, families, study PIs who shared data

Sample Cohorts

MYOGEN consortium

European Vasculitis Genetics Consortium

Cambridge

Ken Smith Paul Lyons

MYOGEN consortum

Fred Miller Chris Amos

University College London

Lucy Wedderburn Claire Deakin

wellcome

UNIVERSITY OF CAMBRIDGE

MRC Biostatistics Unit

Genetic correlation is a conservative estimate of $q_{1|2}$

Assume two traits Y_1, Y_2 can be modelled as

$$Y_1 = \sum_{i=1}^{n_{12}} \alpha_i G_i + \sum_{i=1}^{n_1} \beta_i H_i + E_1 \qquad Y_2 = \sum_{i=1}^{n_{12}} \alpha_i' G_i + \sum_{i=1}^{n_2} \gamma_i J_i + E_2$$

Genotypes:

$$G_i, H_i, J_i \overset{\text{iid}}{\sim} f$$

Effects:

$$\beta_i, \gamma_i \overset{\text{iid}}{\sim} N(0, w^2).$$

$$\begin{bmatrix} \alpha_i \\ \alpha_i' \end{bmatrix} \overset{\text{iid}}{\sim} MNV\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} w^2 & \rho w^2 \\ \rho w^2 & w^2 \end{bmatrix} \right)$$

Genetic correlation is a conservative estimate of $q_{1|2}$

Extreme case $\rho = 1$, ie $\alpha_i' = \alpha_i$. Genetic correlation between Y_1 and Y_2

$$r_g = \frac{\sum_i^{n_{12}} \text{var}(\alpha_i G_i)}{\sqrt{\left[\sum_i^{n_{12}} \text{var}(\alpha_i G_i) + \sum_i^{n_1} \text{var}(\beta_i H_i)\right]\left[\sum_i^{n_{12}} \text{var}(\alpha_i G_i) + \sum_i^{n_2} \text{var}(\gamma_i J_i)\right]}} \tag{1}$$

All variants, effects are iid

$$\text{var}(\alpha_i G_i) = \text{var}(\beta_i H_i) = \text{var}(\gamma_i J_i) = \nu$$

So

$$r_g = \frac{n_{12}\nu}{\sqrt{(n_{12} + n_1)\nu(n_{12} + n_2)\nu}} = \frac{n_{12}}{\sqrt{(n_{12} + n_1)(n_{12} + n_2)}} \tag{2}$$

Genetic correlation is a conservative estimate of $q_{1|2}$

Can show $|r_g|$ maximal when $|\rho| = 1$

$$|r_g| \geq \frac{n_{12}}{\sqrt{(n_{12} + n_1)(n_{12} + n_2)}} = \frac{p_{12}}{\sqrt{q_1 q_2}}$$

Therefore conservative point estimates are

$$p_{12} = \sqrt{q_1 q_2}|r_g| \qquad p_1 = q_1 - p_{12} \qquad p_1 = q_2 - p_{12}$$