# Integrative Genomics via Colocalisation: informing choice of prior parameter values

Chris Wallace

chr1swallace chr1swallace.github.io

7th May 2019

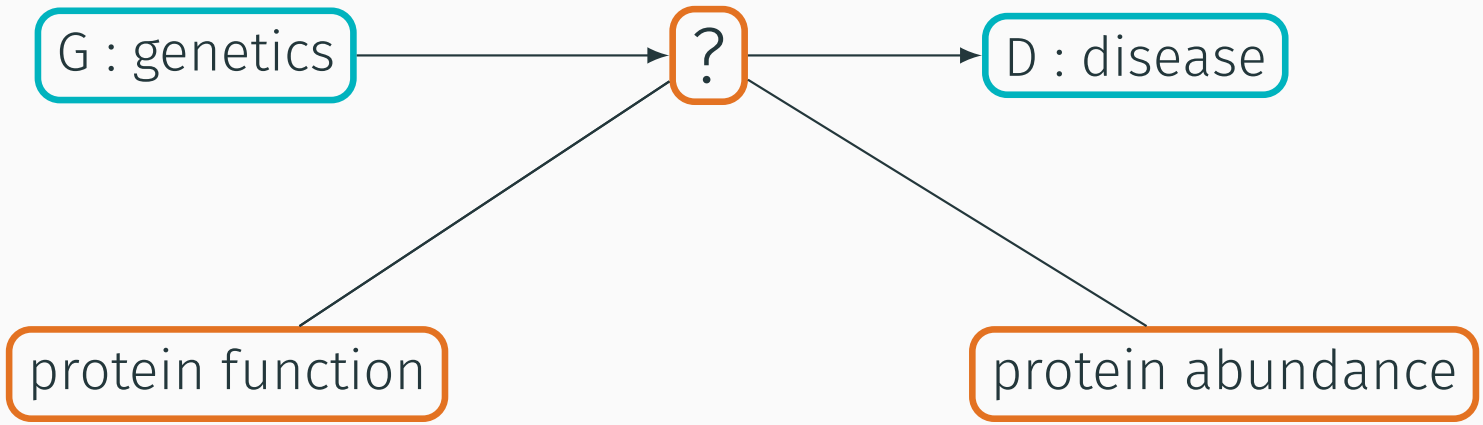
University of Cambridge

Stasia Grinberg

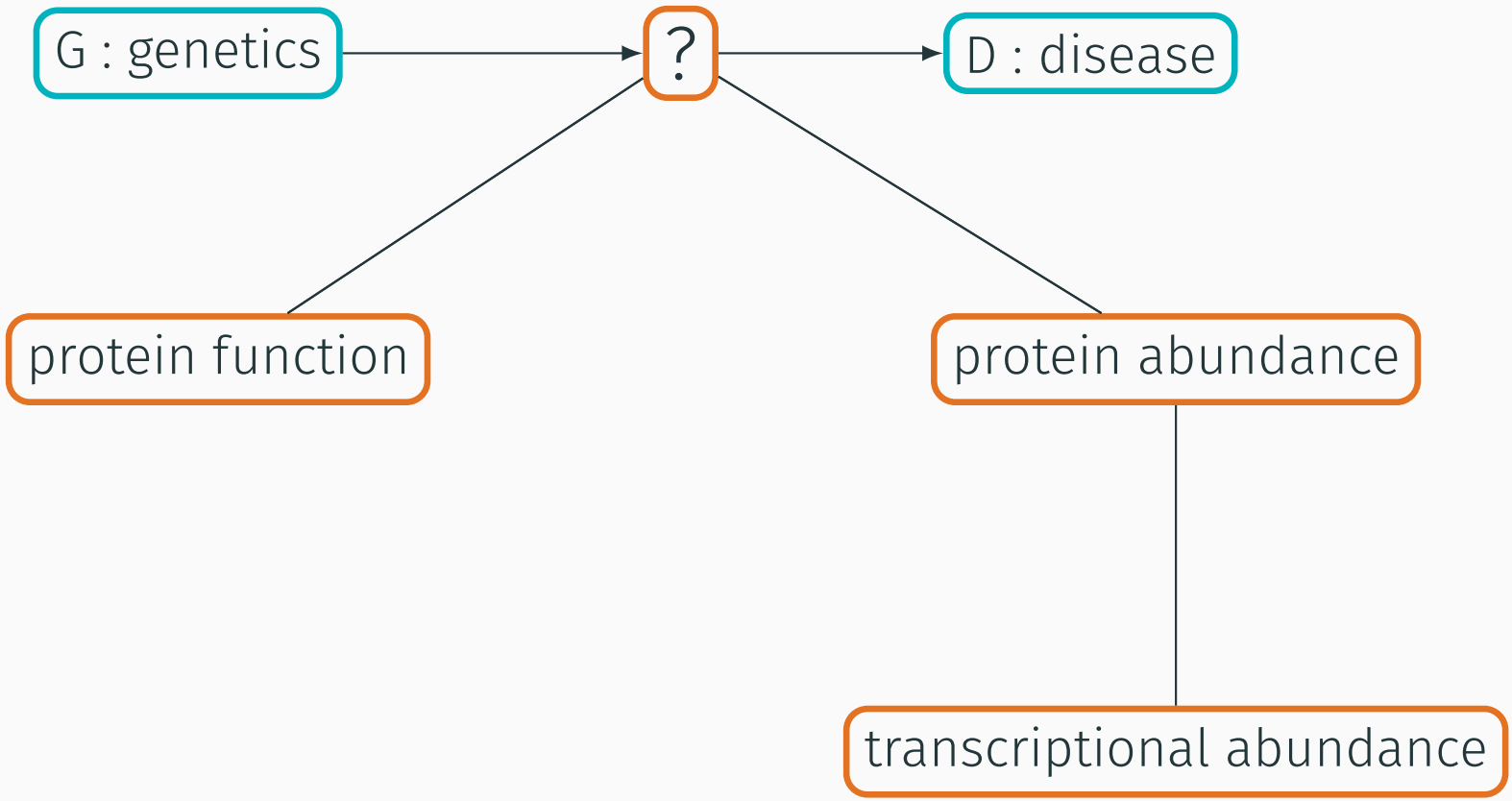
# Identifying molecular traits that mediate a disease association

# Identifying molecular traits that mediate a disease association

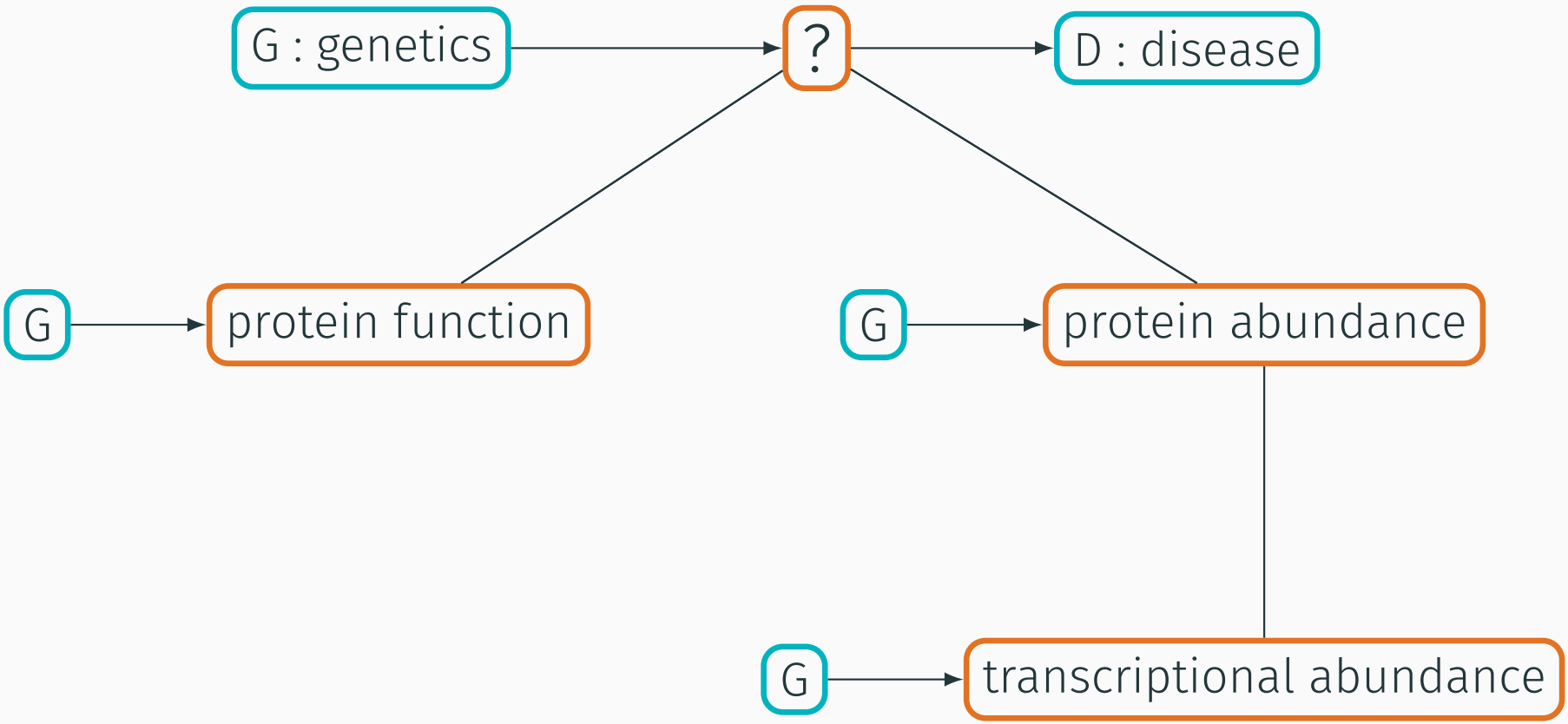
# Identifying molecular traits that mediate a disease association



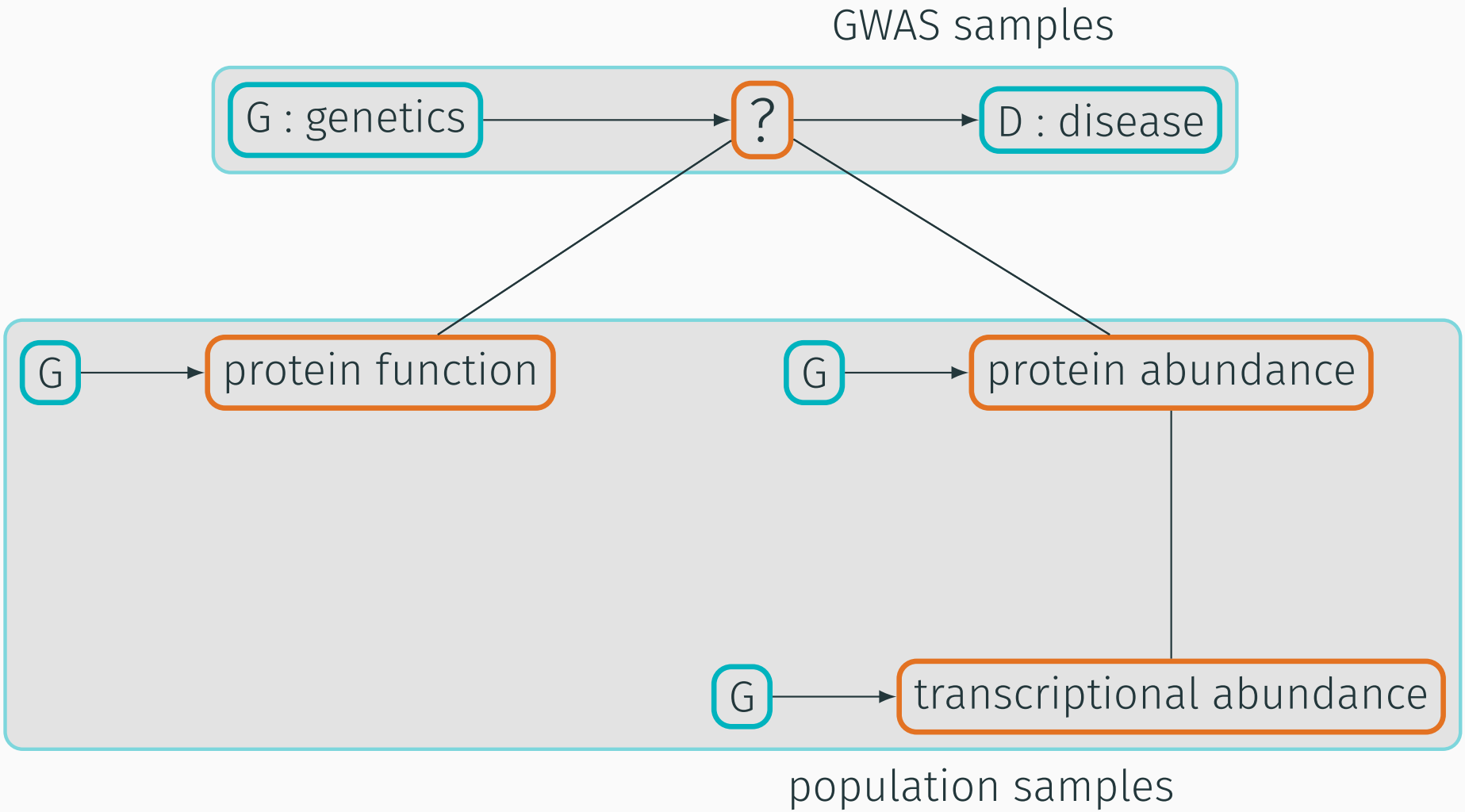
# Identifying molecular traits that mediate a disease association



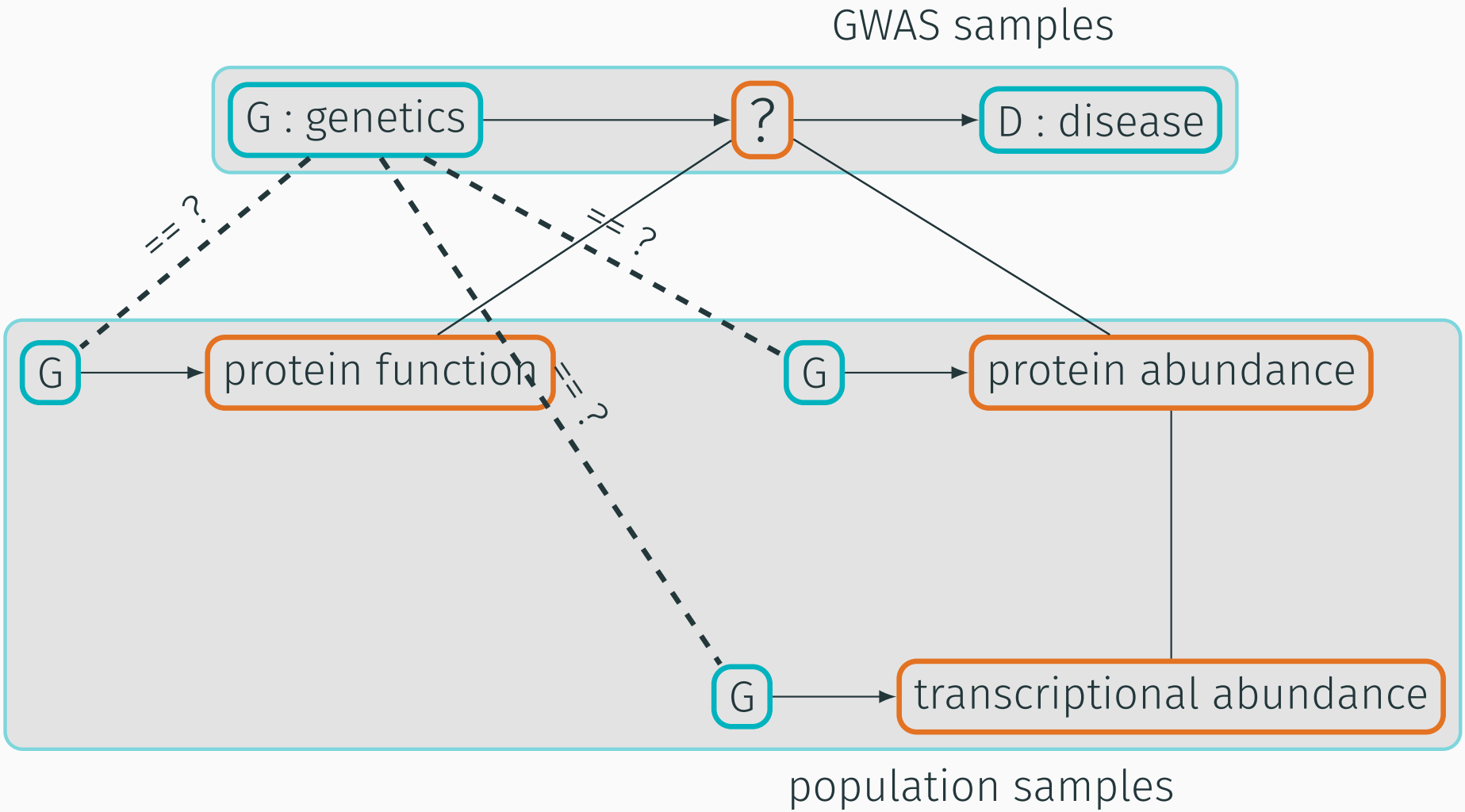
# Identifying molecular traits that mediate a disease association



# Identifying molecular traits that mediate a disease association



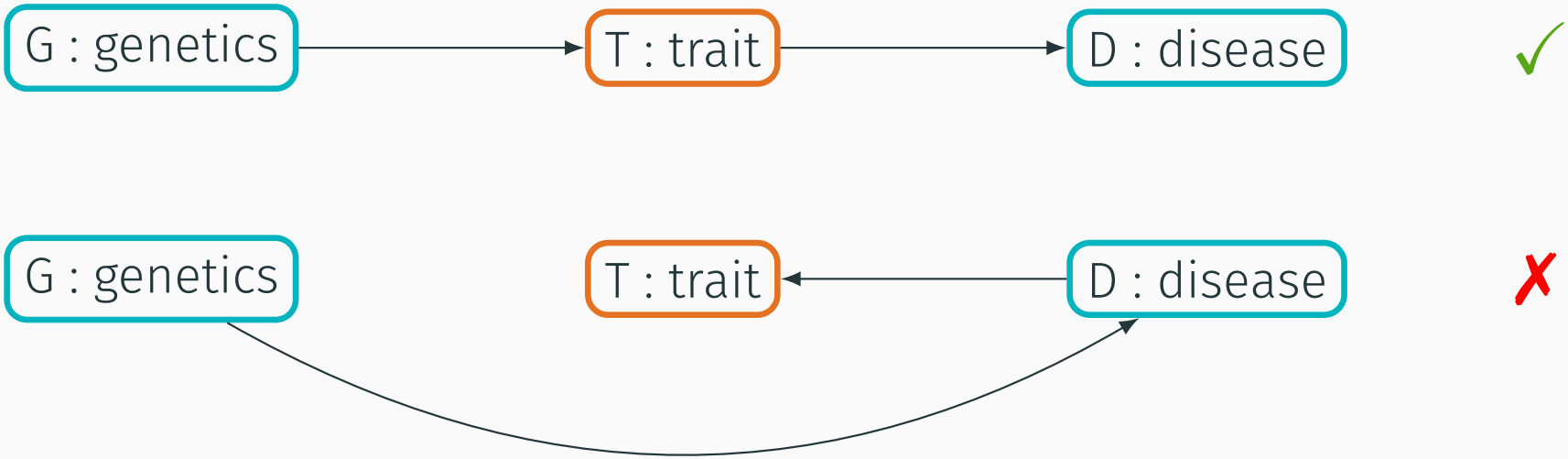
# Identifying molecular traits that mediate a disease association



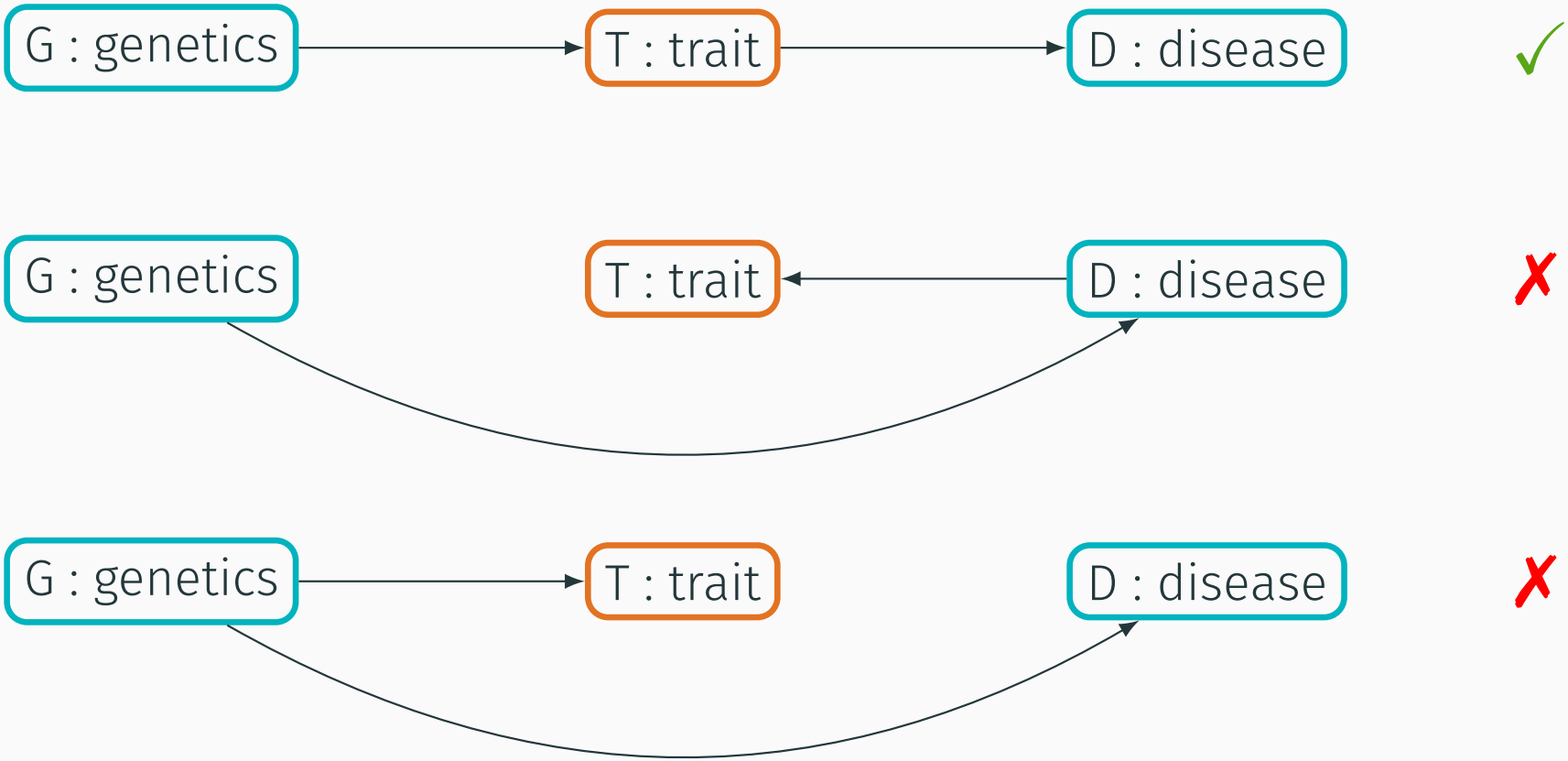
# Identifying molecular traits that mediate a disease association


✓

# Identifying molecular traits that mediate a disease association



# Identifying molecular traits that mediate a disease association

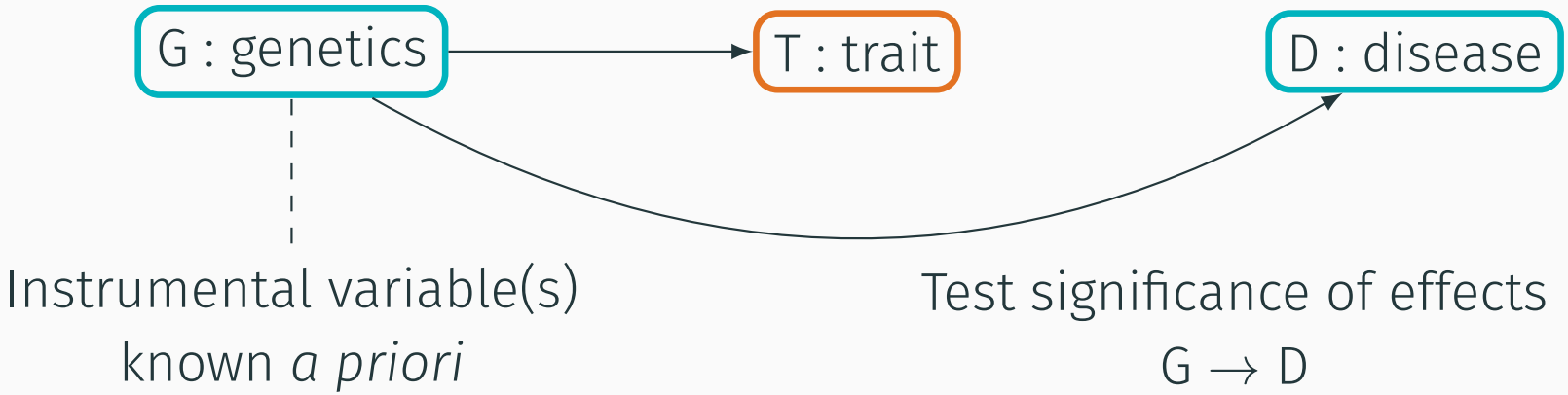


# Identifying molecular traits that mediate a disease association

Mendelian Randomisation

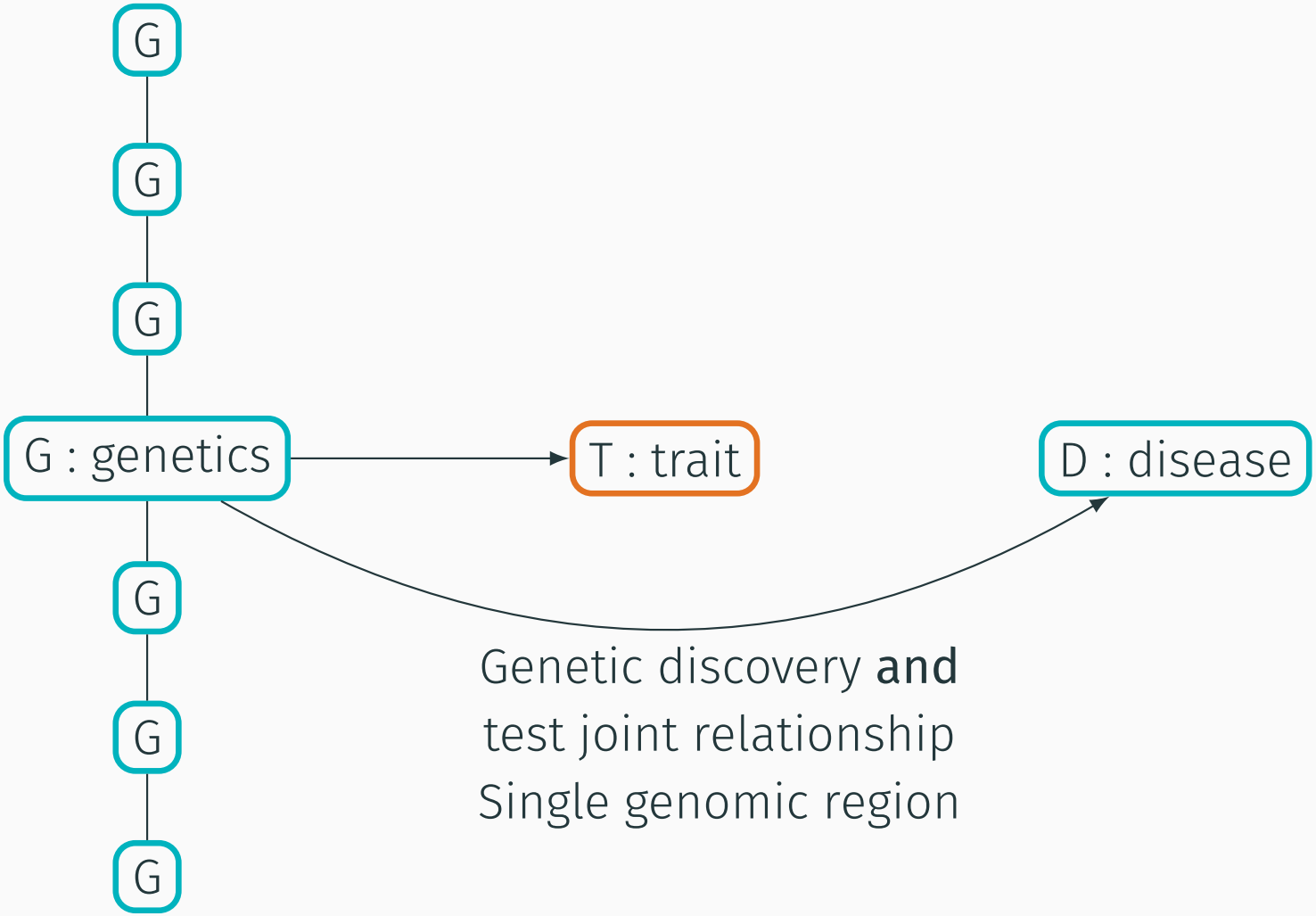
# Identifying molecular traits that mediate a disease association

## Mendelian Randomisation



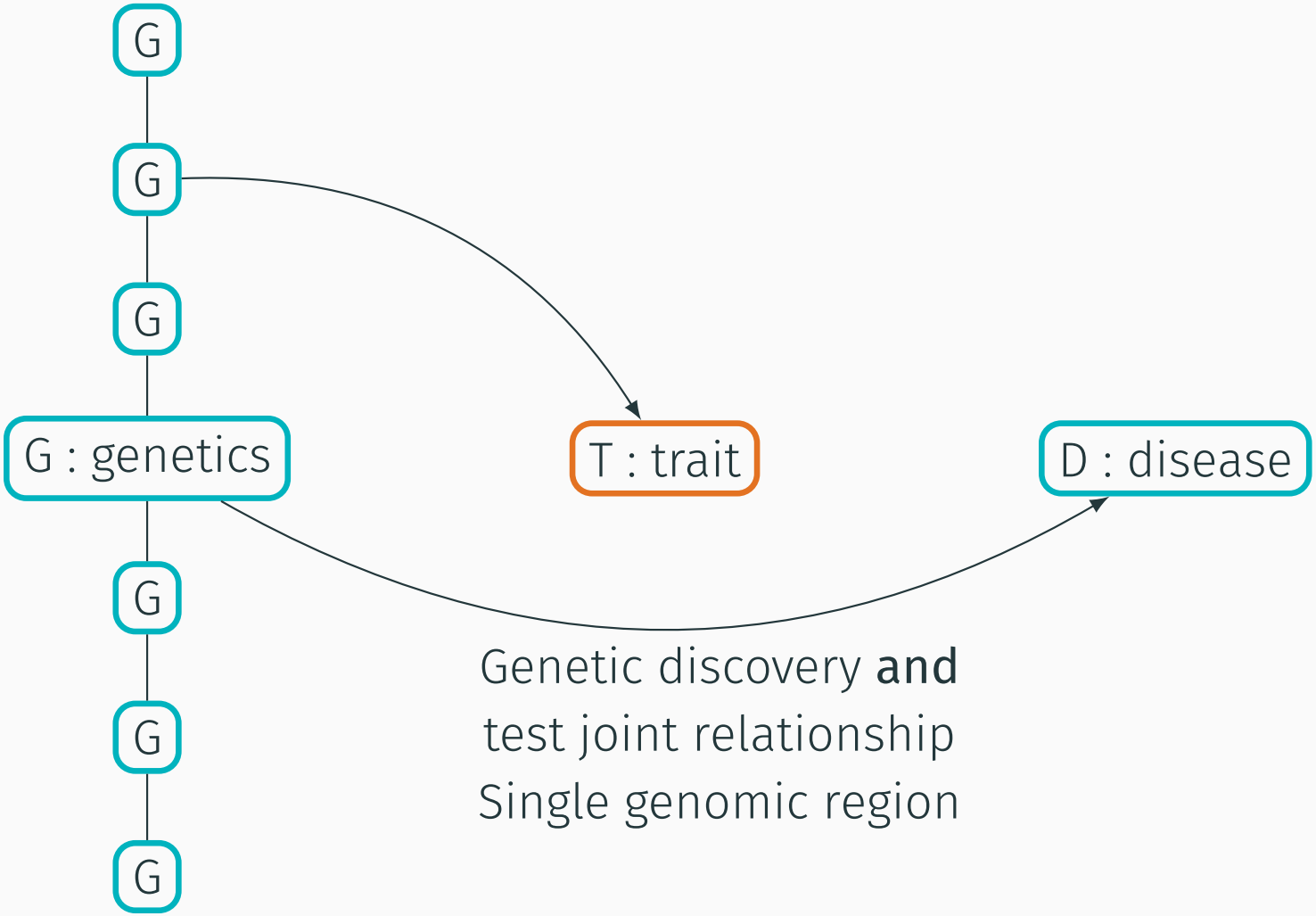
# Identifying molecular traits that mediate a disease association

## Colocalisation



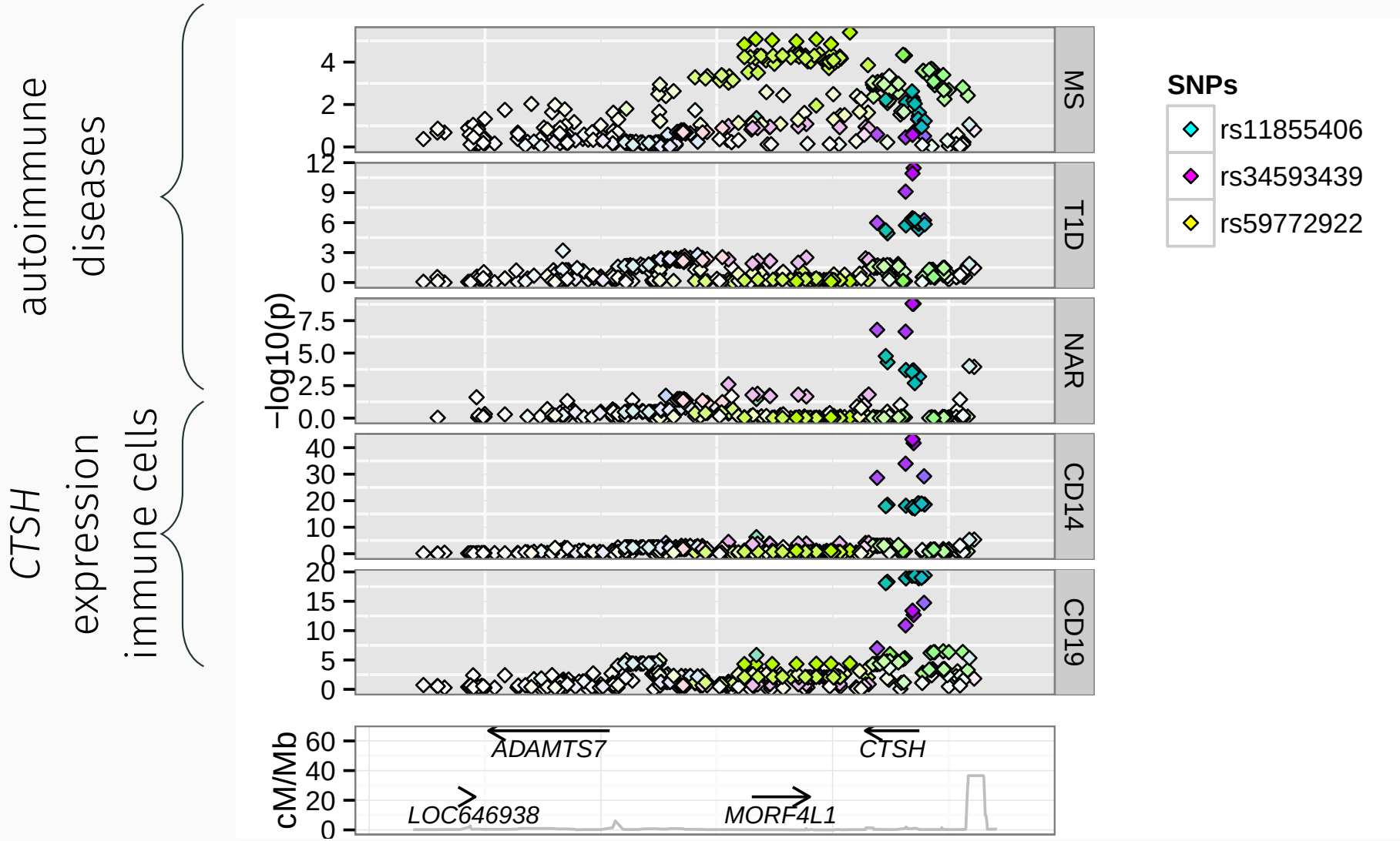
# Identifying molecular traits that mediate a disease association

## Colocalisation

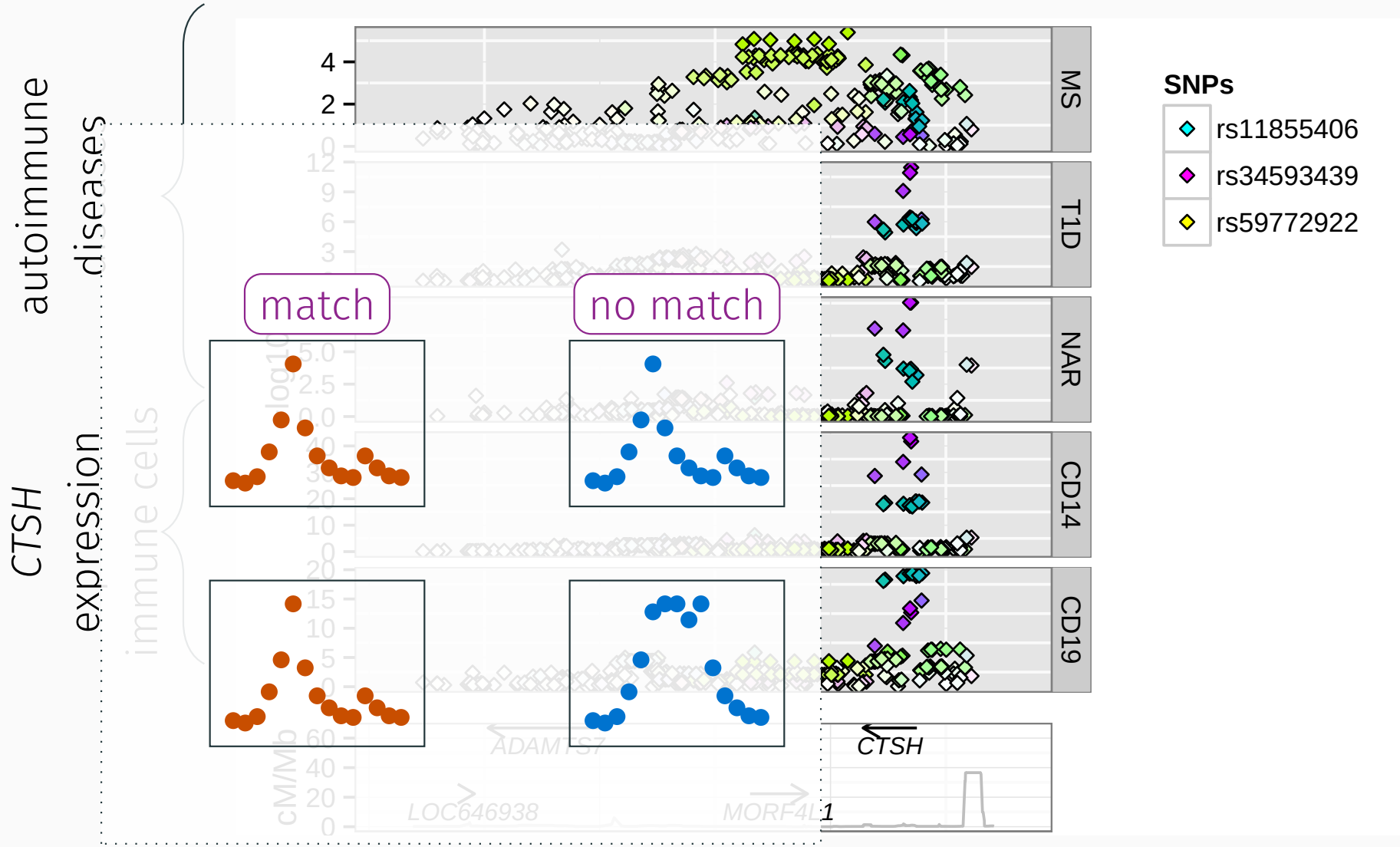


# Colocalisation:
# the coloc approach

# 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

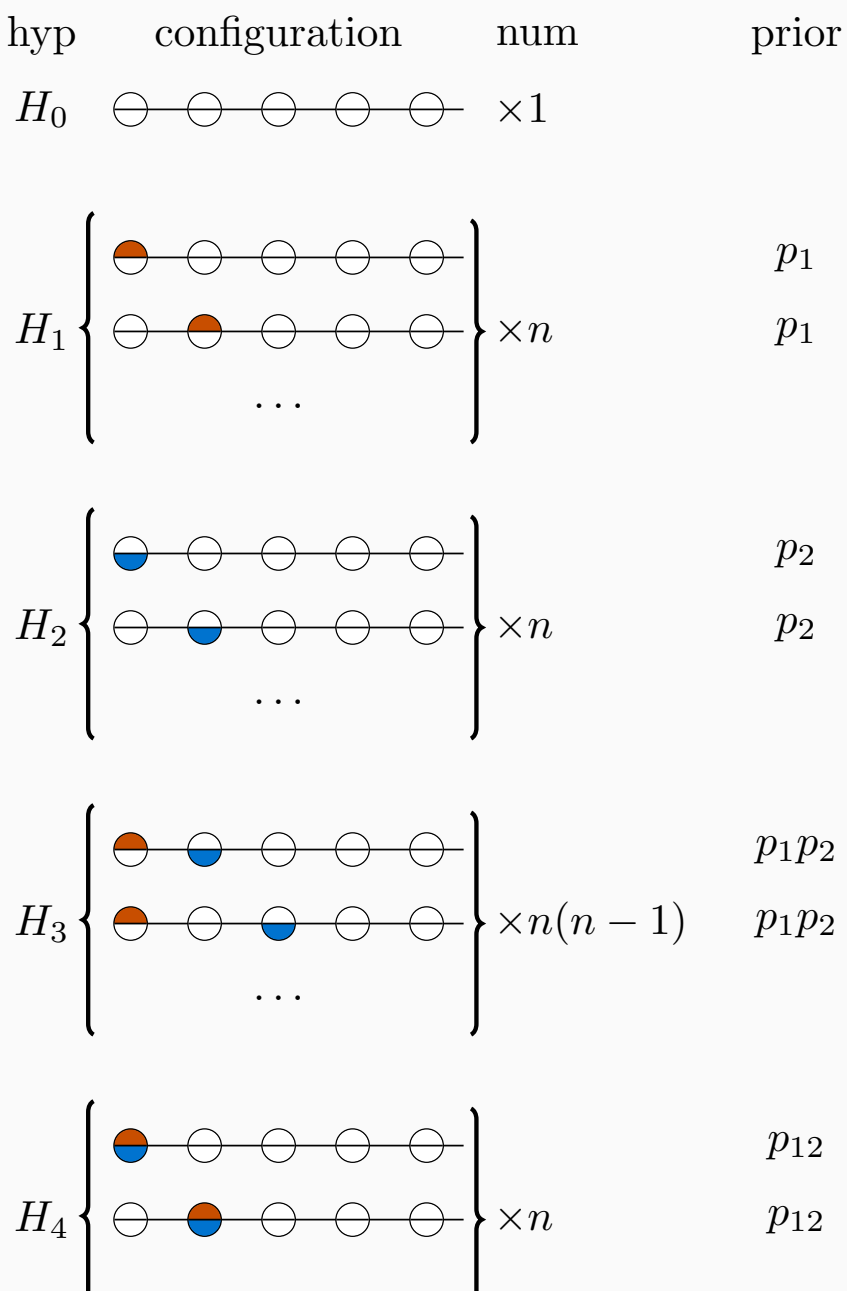
$H_1$ : association to trait 1 only

$H_2$ : 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




Prob. causal

$q_1 = p_1 + p_{12}$

$q_2 = p_2 + p_{12}$

Cond. Prob. jointly causal

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

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

# 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 ...

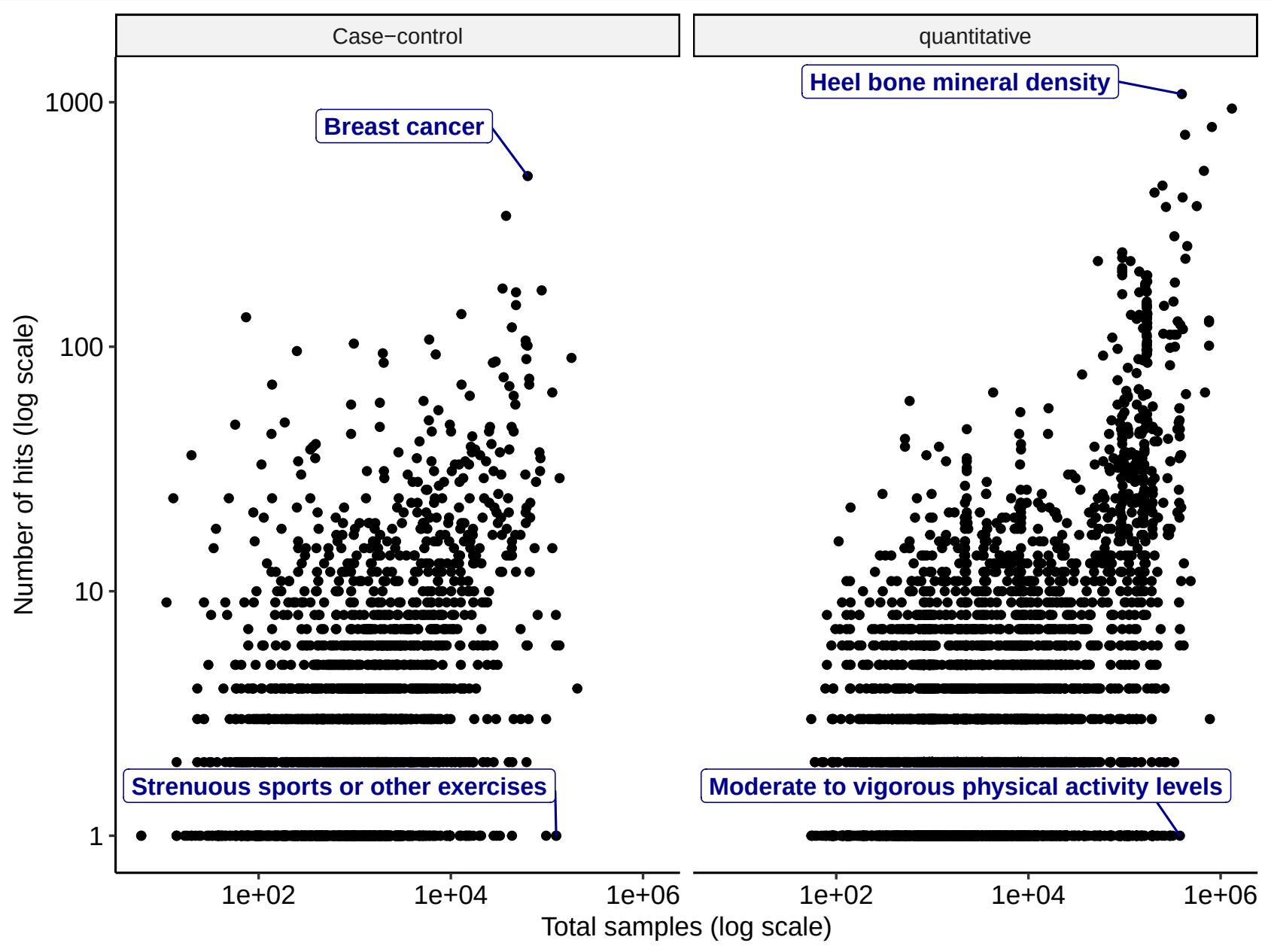
... 22 used software default priors

Non-default priors:

| | |
|---|---|
| Gusev et al | Considered range of $p_{12}$ to test robustness |
| Dobbyn et al | SNP-wise analysis to estimate $p_{12}$ |
| Yao et al | Subjective informed prior: $p_{12}/(p_{12} + p_1) = 0.75$ |

# 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:

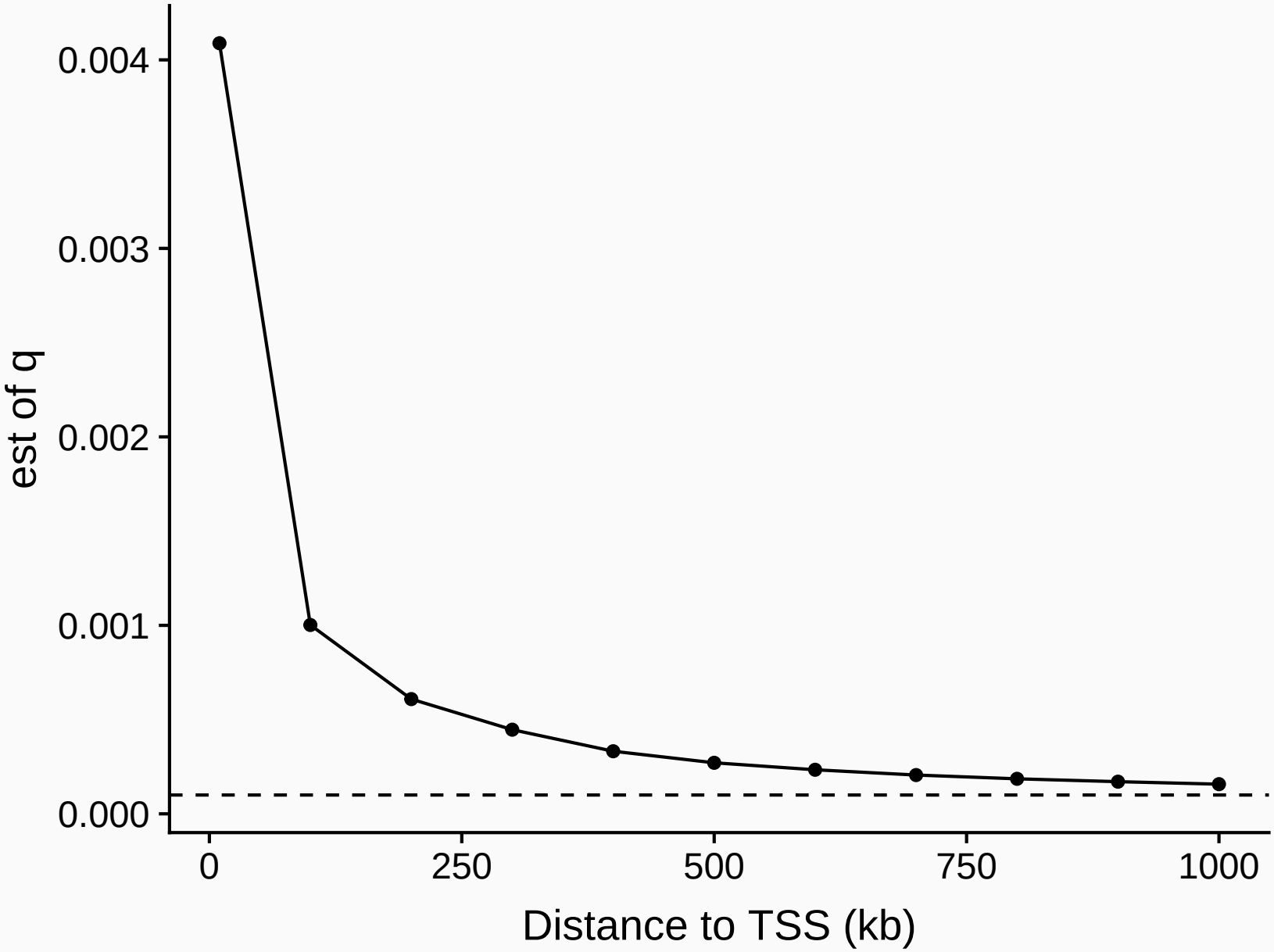
eg IBD, 60,000 subjects

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

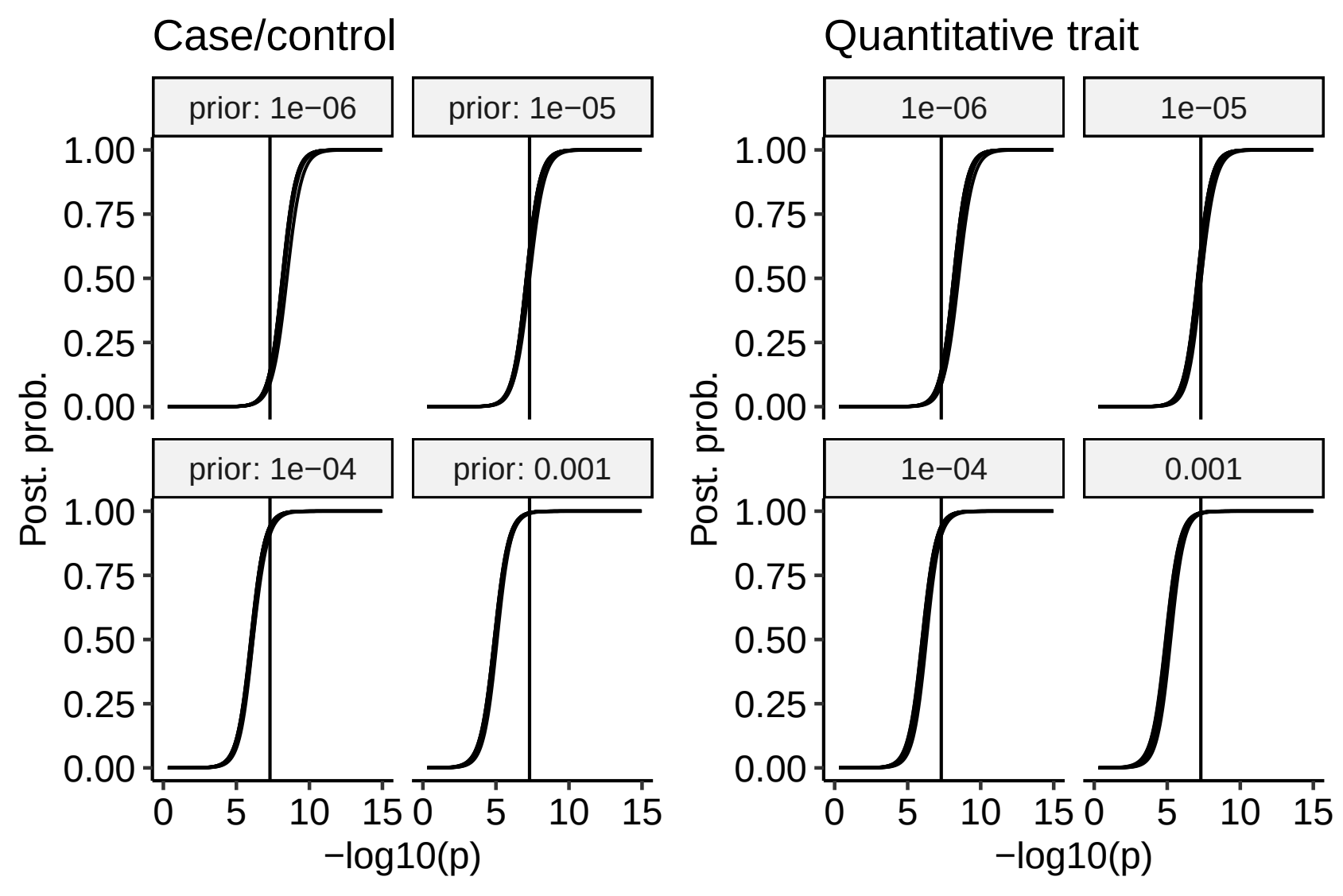
$$q_i = \frac{200}{2,000,000} = \frac{1}{10,000}$$

*de Lange et al, Nat Genet 2017*

# Empirical prior for single trait, GTeX



# Pragmatic prior for single trait



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

# Empirical priors

?

# 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 $q_{1|2}$

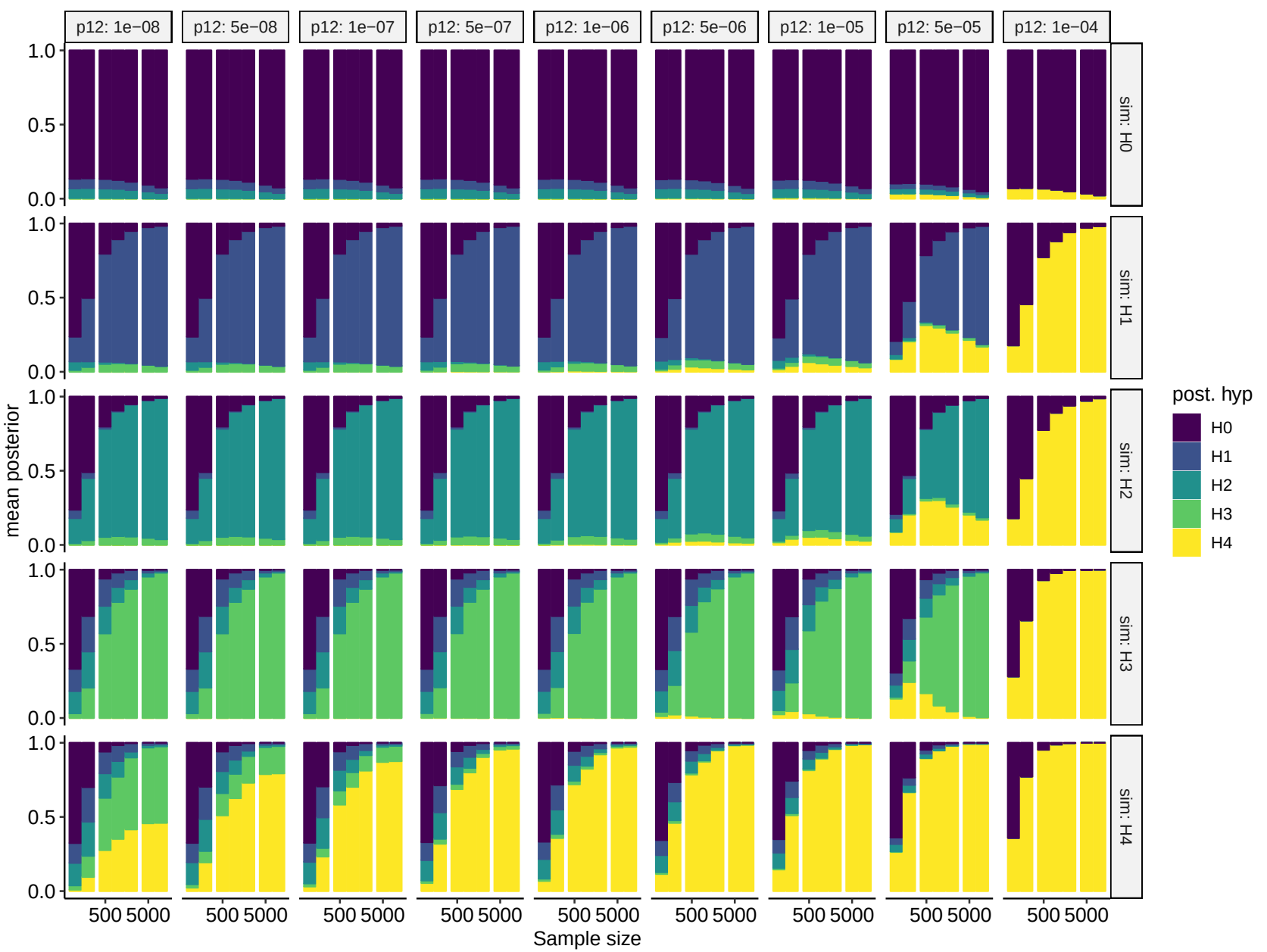
Can show $|r_g|$ maximal when $|\rho| = 1$

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

Therefore conservative point estimates are

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

# Pragmatic $p_{12}$



# Simplifying assumption

# Simplifying assumption: Single Causal Variant

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*

# Simplifying assumption: Single Causal Variant

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*

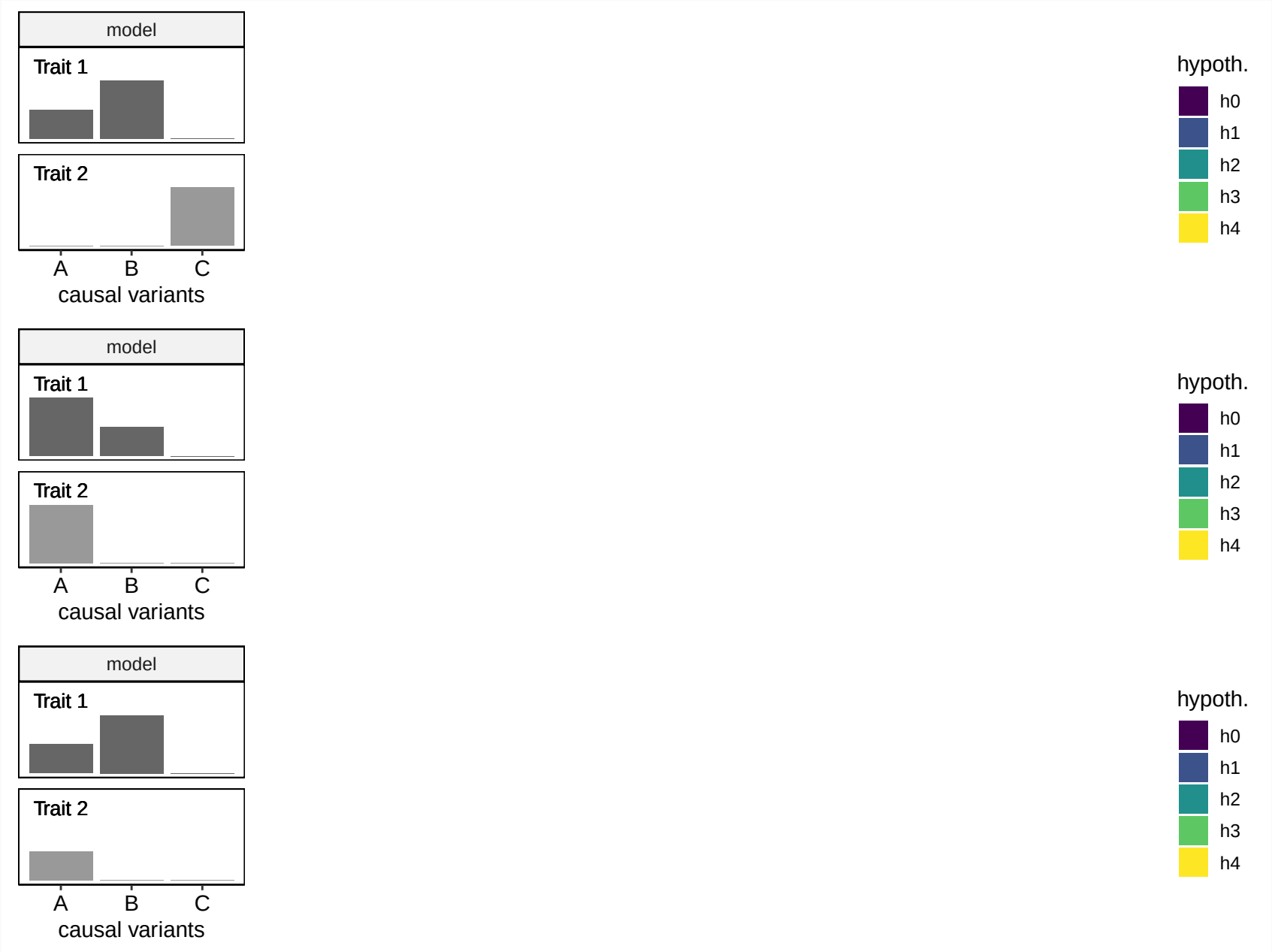
# Review of current practice

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

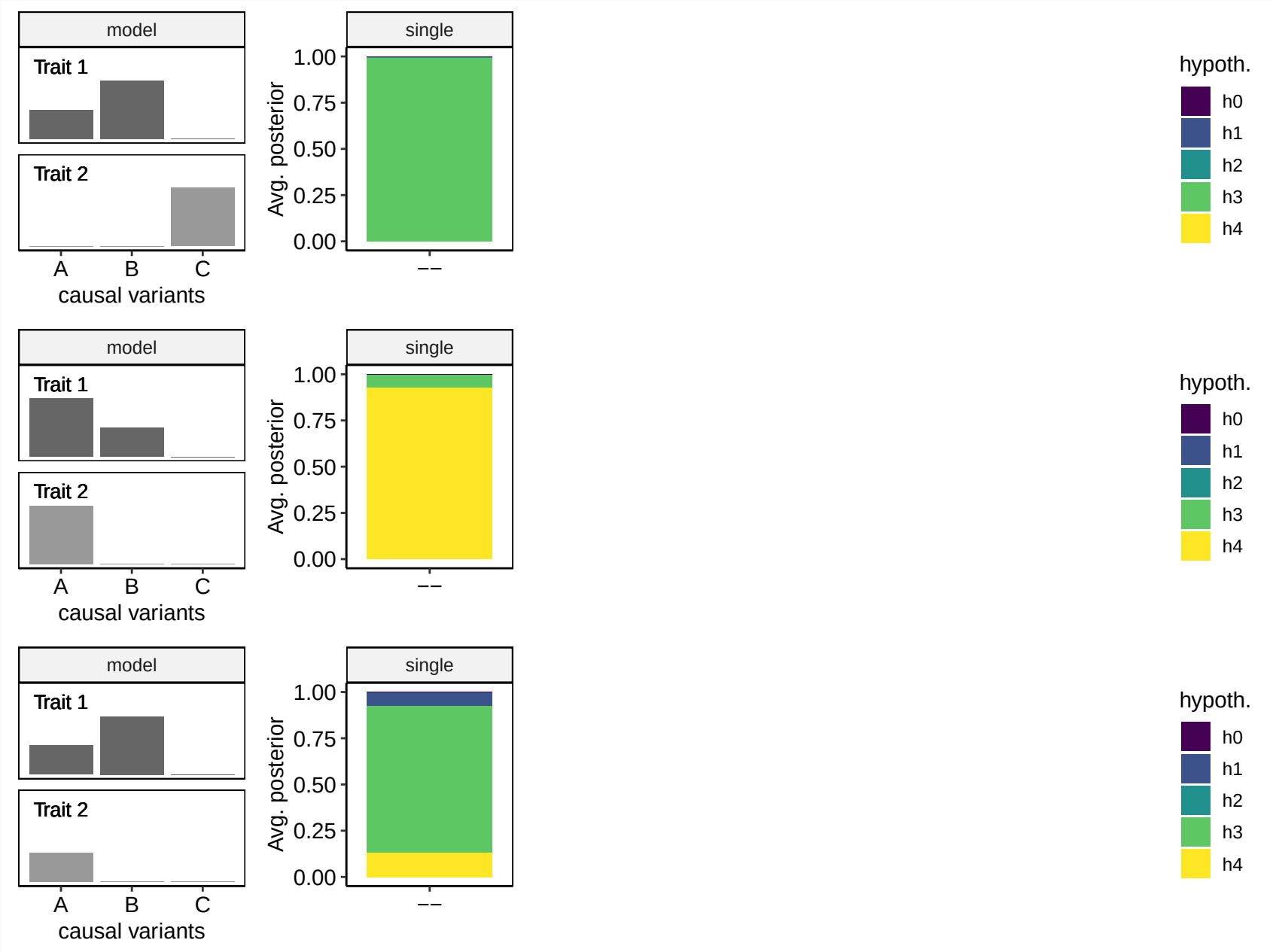
... 1 used conditioning*

*on one trait for which they had full data

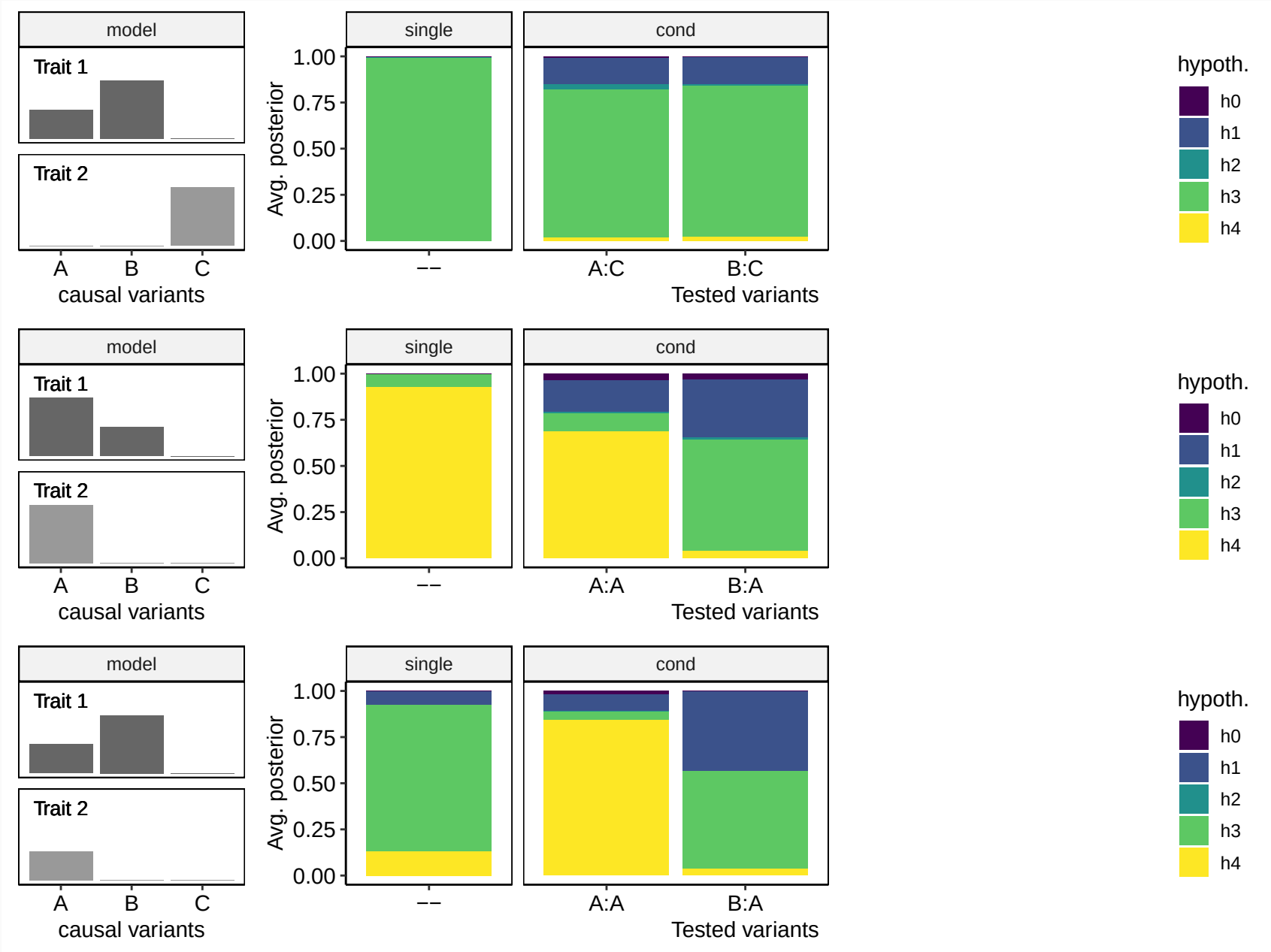
# Violating the single causal variant assumption



# Violating the single causal variant assumption



# Violating the single causal variant assumption



# Conditioning and masking

**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

# Conditioning and masking

**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

# Conditioning and masking

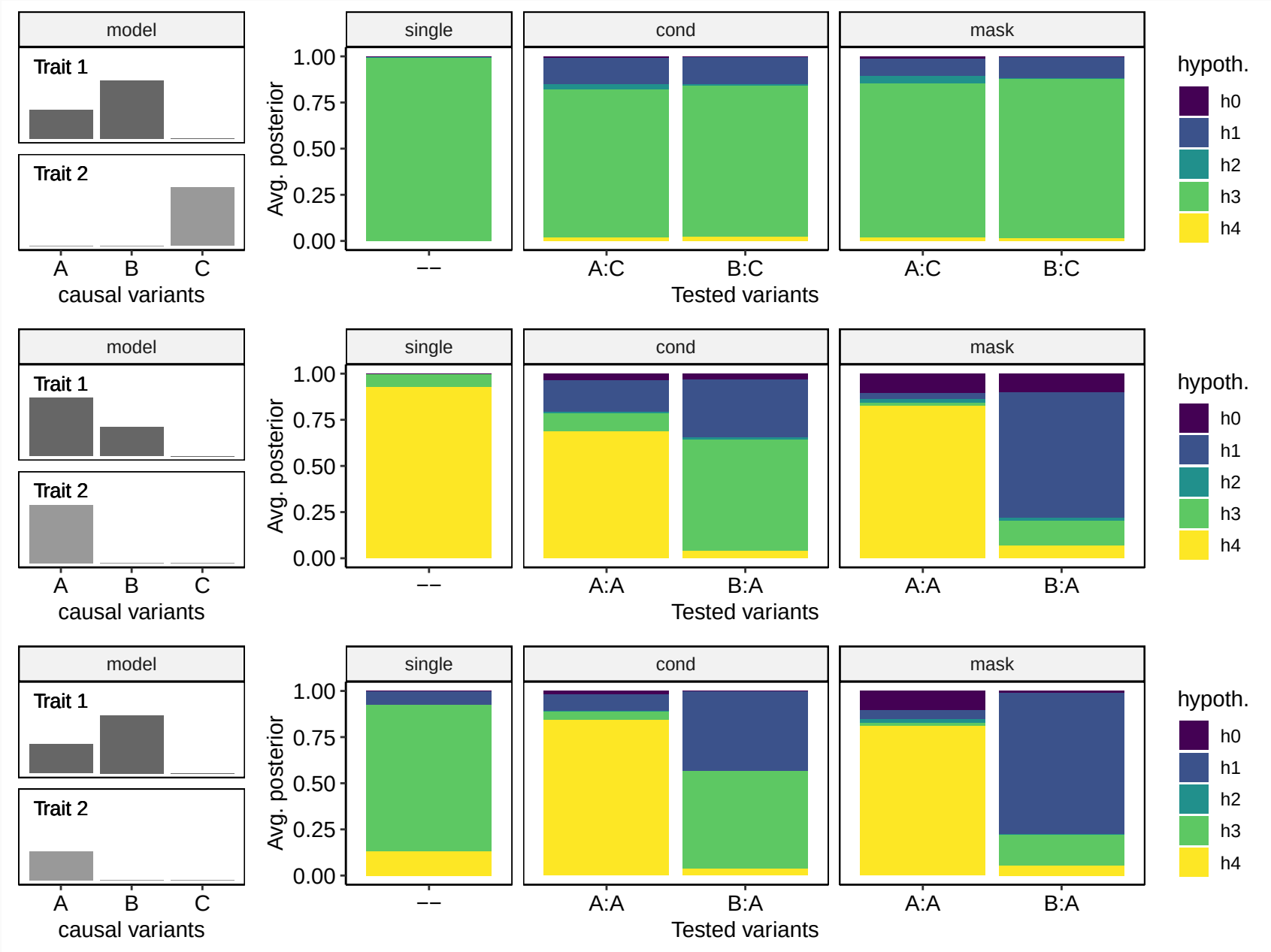
**Algorithm 1:** Conditioning

**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

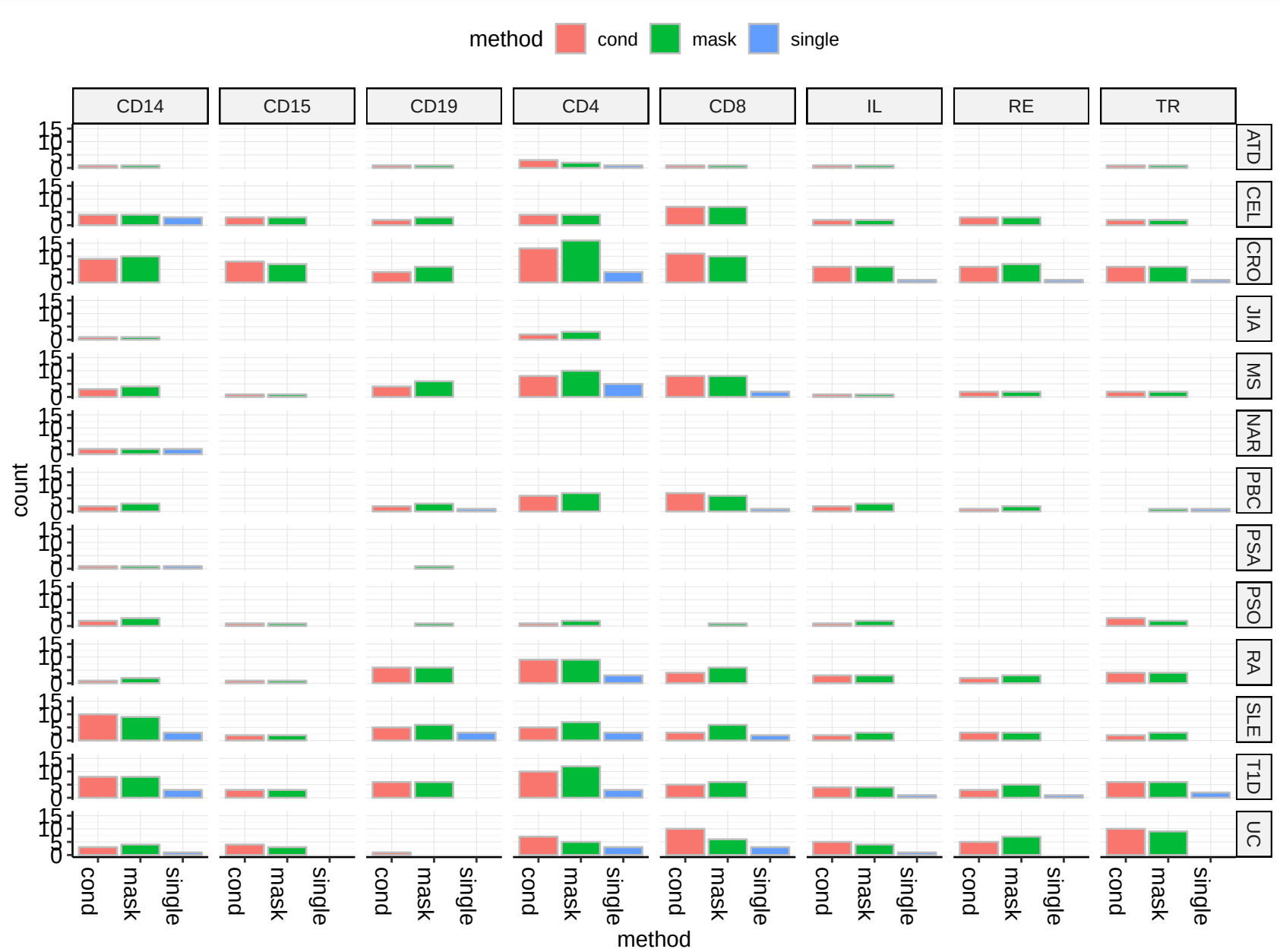
**Result:** $S$ = a set of selected SNPs
Initialise S as an empty set;
**while** *TRUE* **do**
    Consider $T$ = all SNPs not in LD with any SNP in S;
    Fit a model to each SNP in T;
    Choose the *best SNP* among these;
    **if** *the best SNP is significant* **then**
        add it to S;
    **else**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

# Conditioning and masking allow multiple comparisons

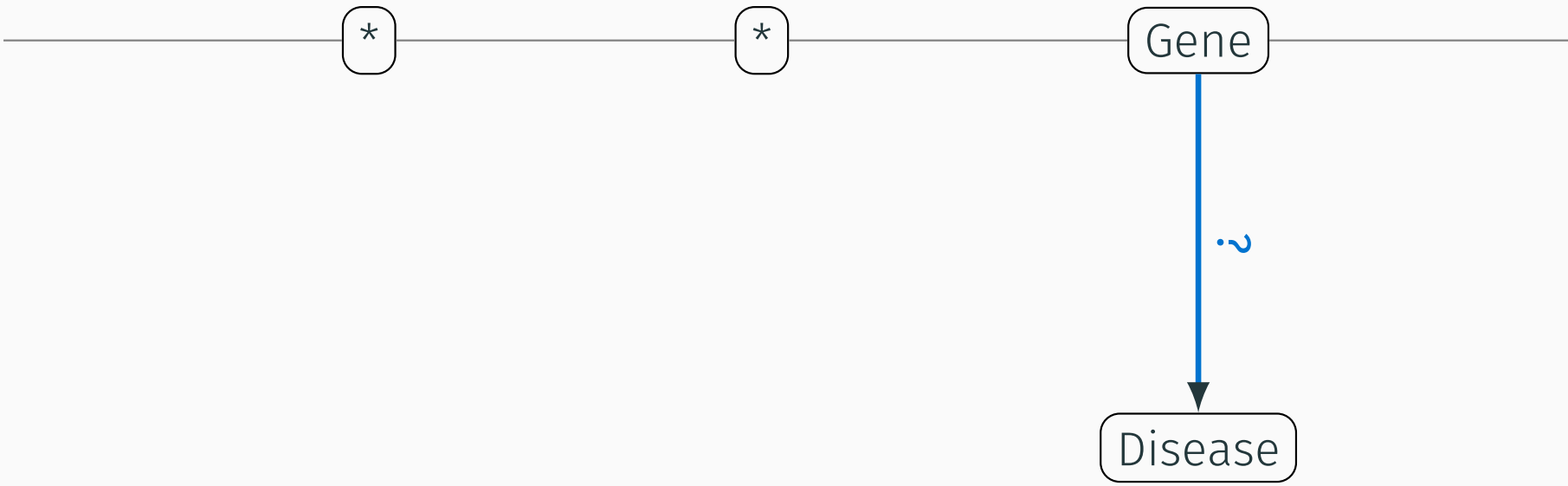


# Impact on results

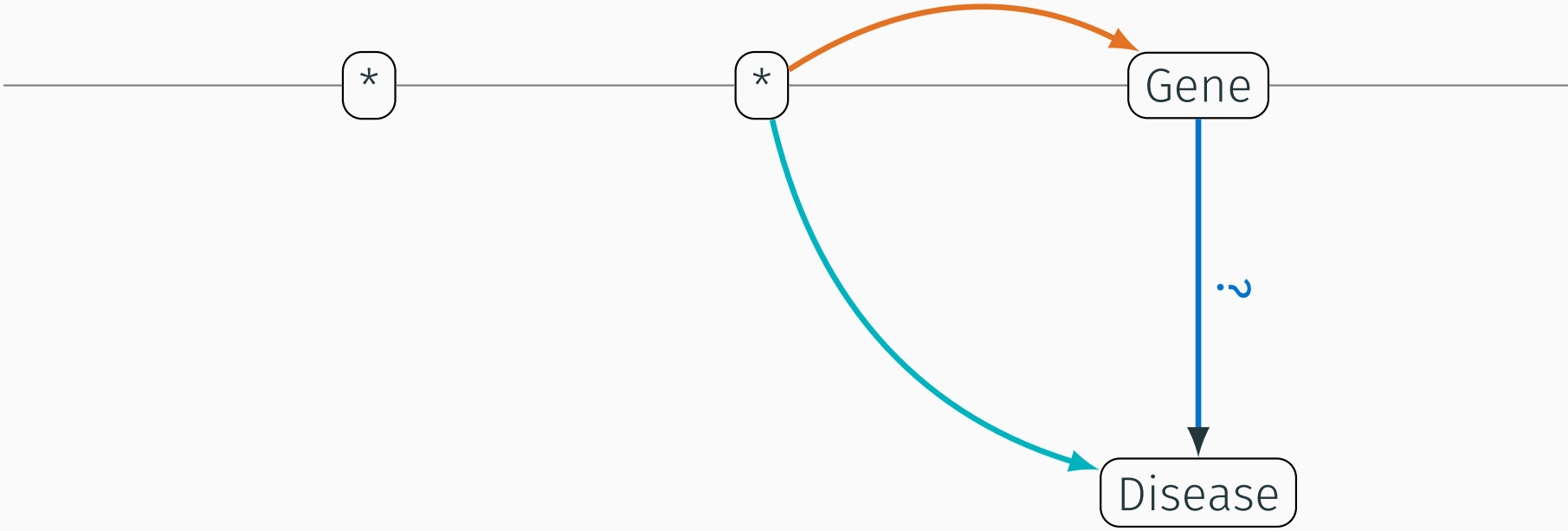
# Colocalisation of eQTL (8 cell types) vs 13 diseases



# Impact on interpretation

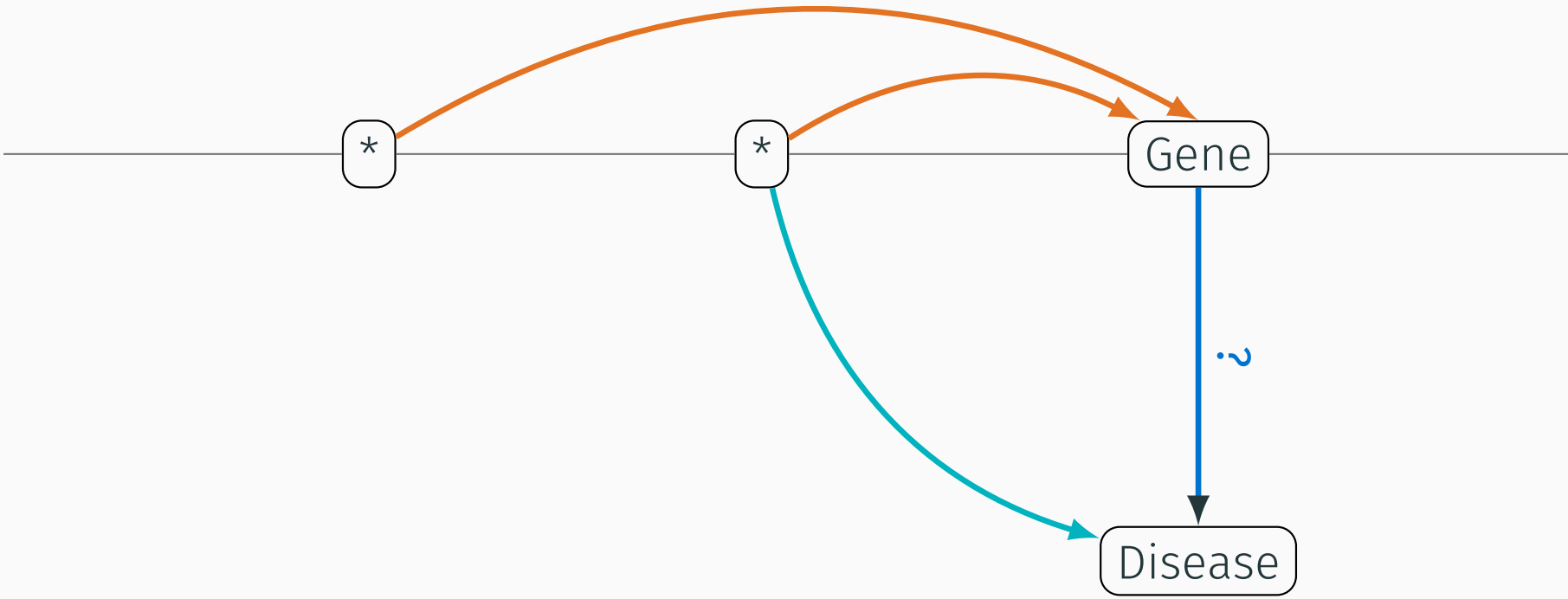


# Impact on interpretation



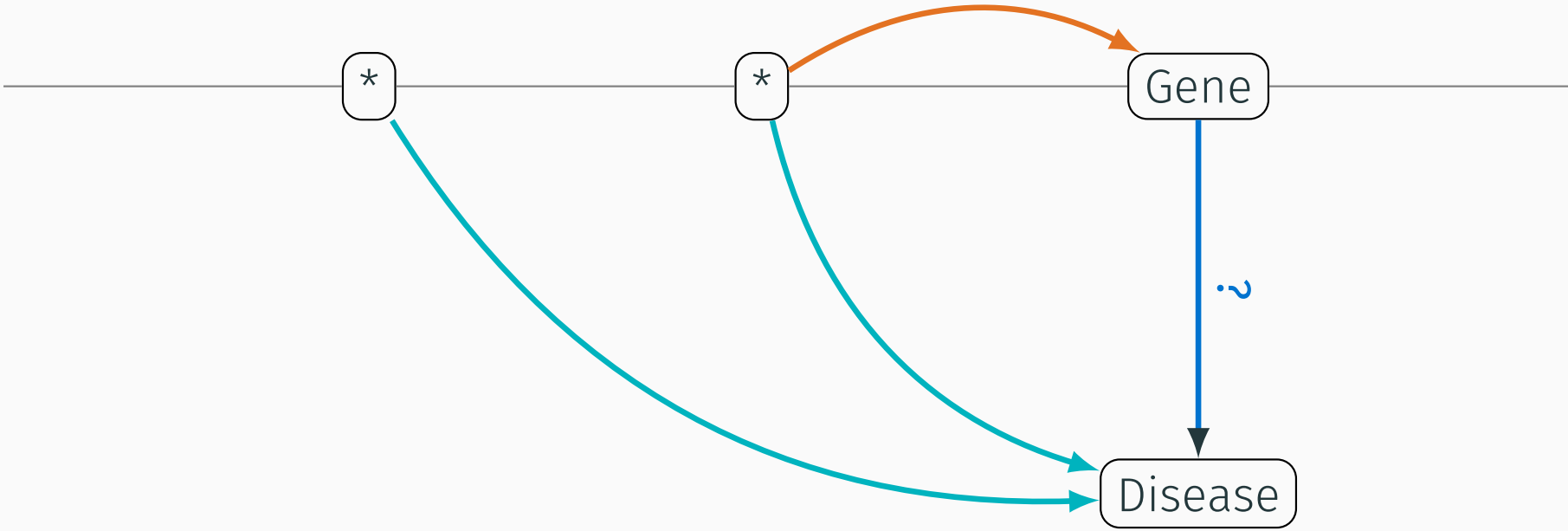

Strong evidence

# 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

- Priors matter!
- Conditioning helps, but need to be more nuanced in interpretation of partial colocalisation
- Masking can approximate conditioning when aligned alleles not known

**Question:**

Should I remove or change default $p_{12}$ in software?

# Thanks to...

Stasia Grinberg

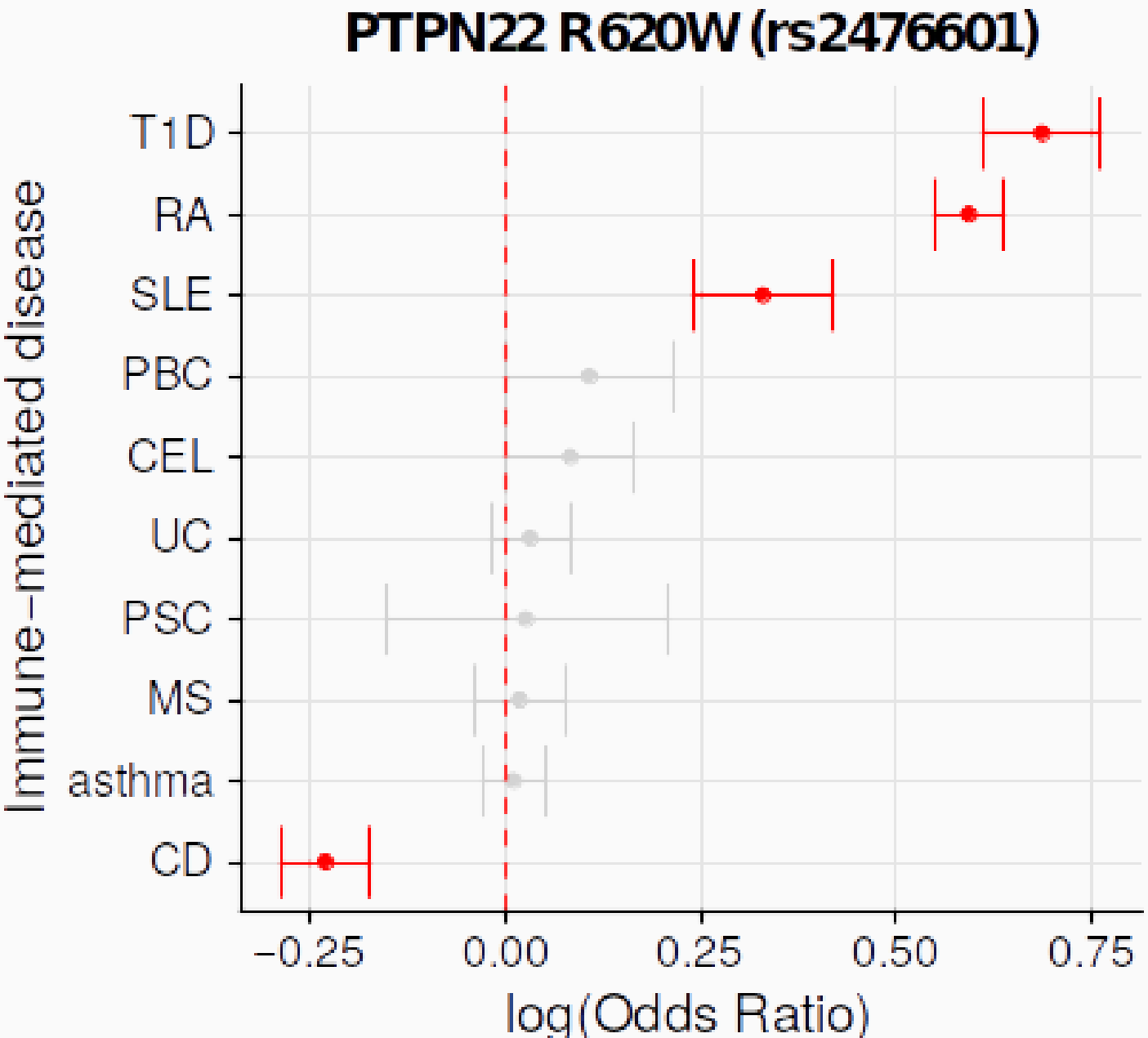
wellcome

UNIVERSITY OF CAMBRIDGE

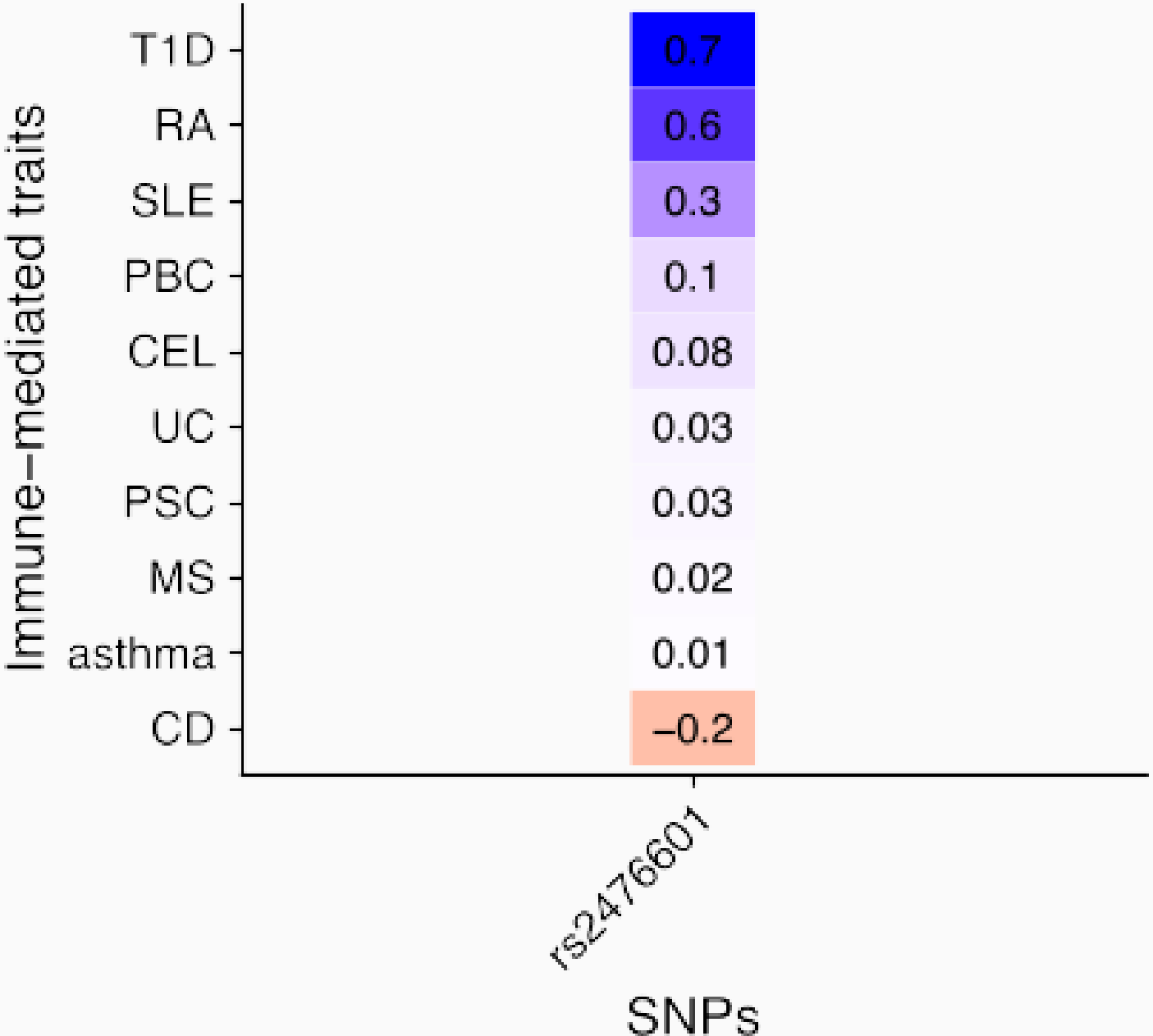
MRC | Biostatistics Unit

# Informed dimension reduction

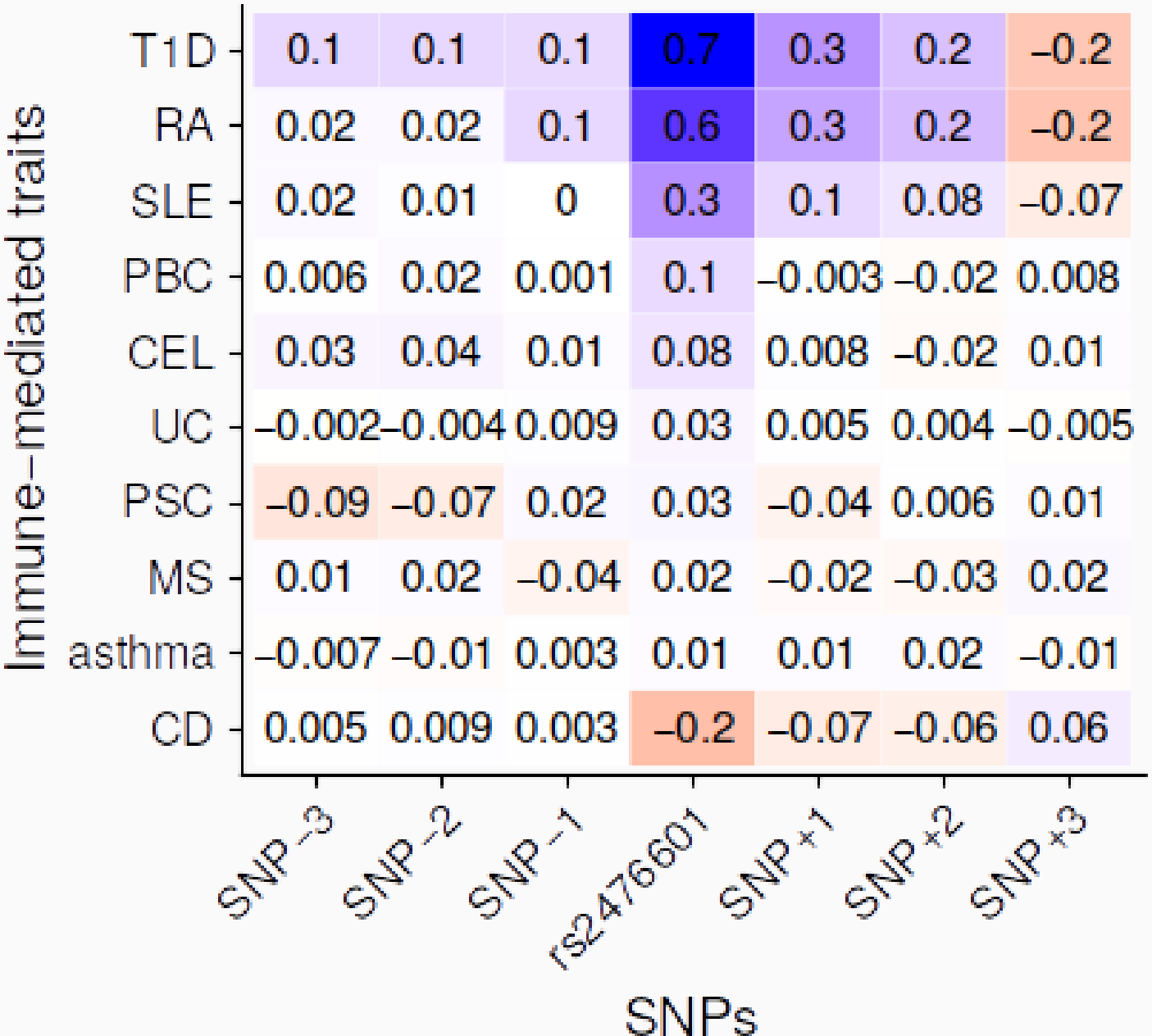
# Heterogeneity in effects across diseases



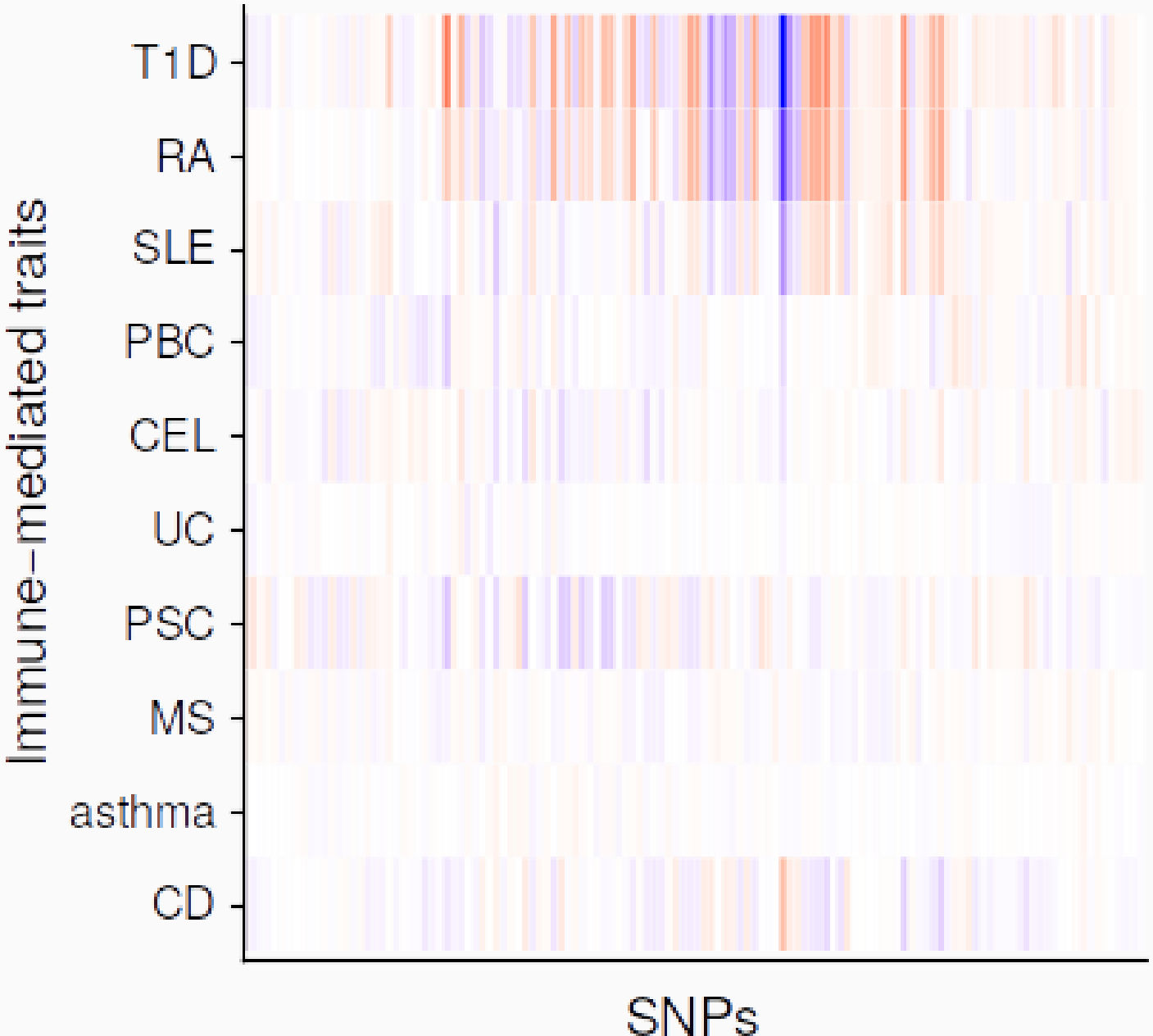
# Heterogeneity in effects across diseases



# Heterogeneity in effects across diseases

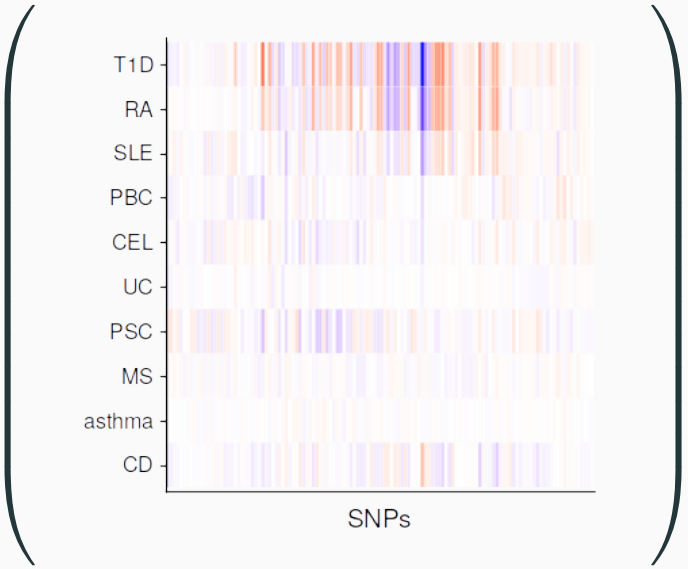


# Heterogeneity in effects across diseases

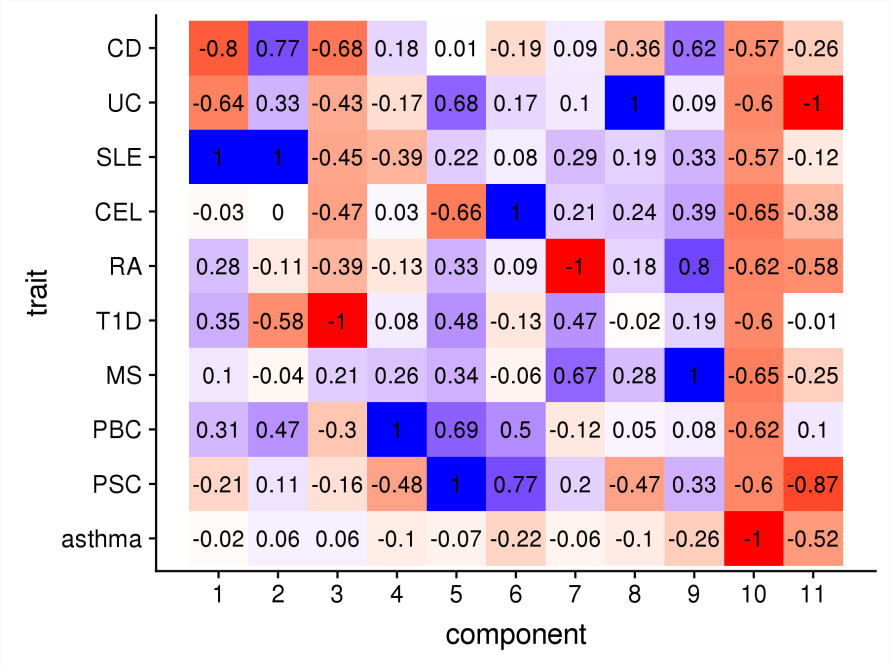


# Principal components analysis to generate a new “basis”

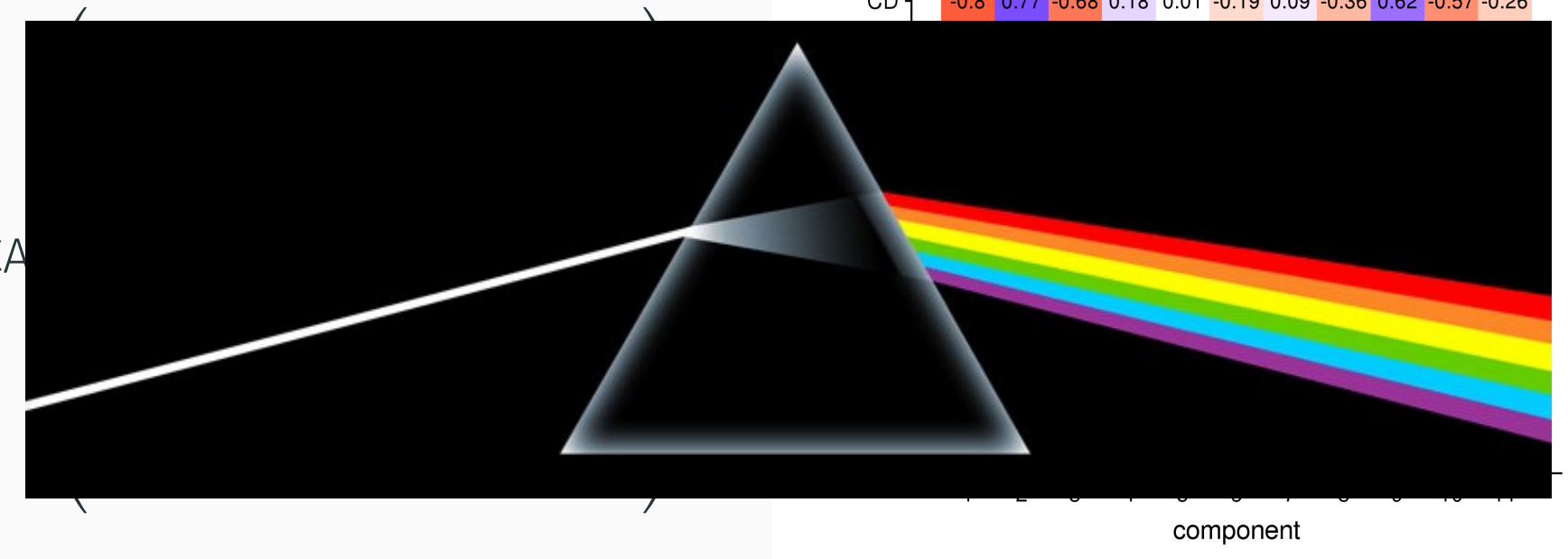
PCA




=



# Principal components analysis to generate a new “basis”

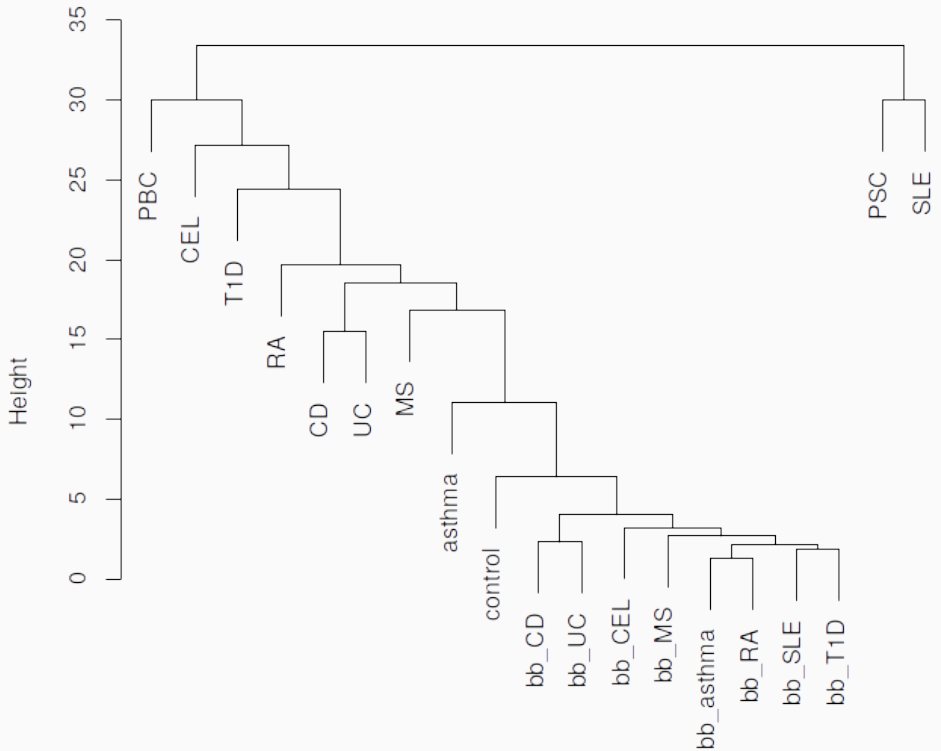


# Basis diseases

| 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 |

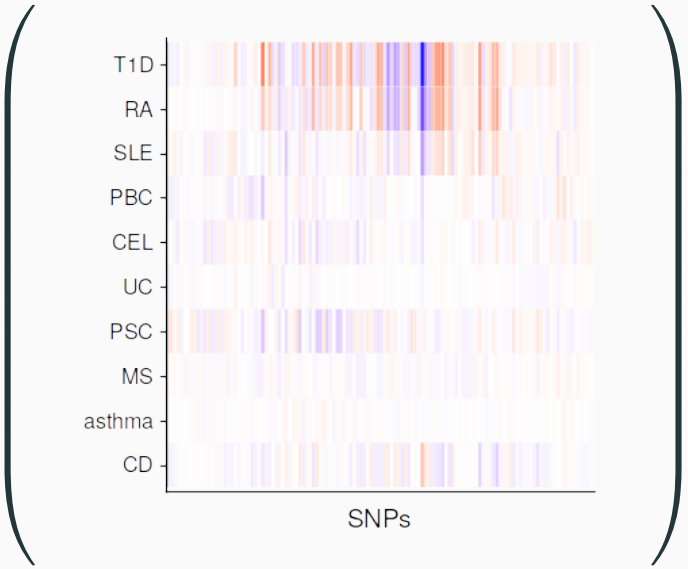
# Naive basis captures dataset, not disease

| 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 |

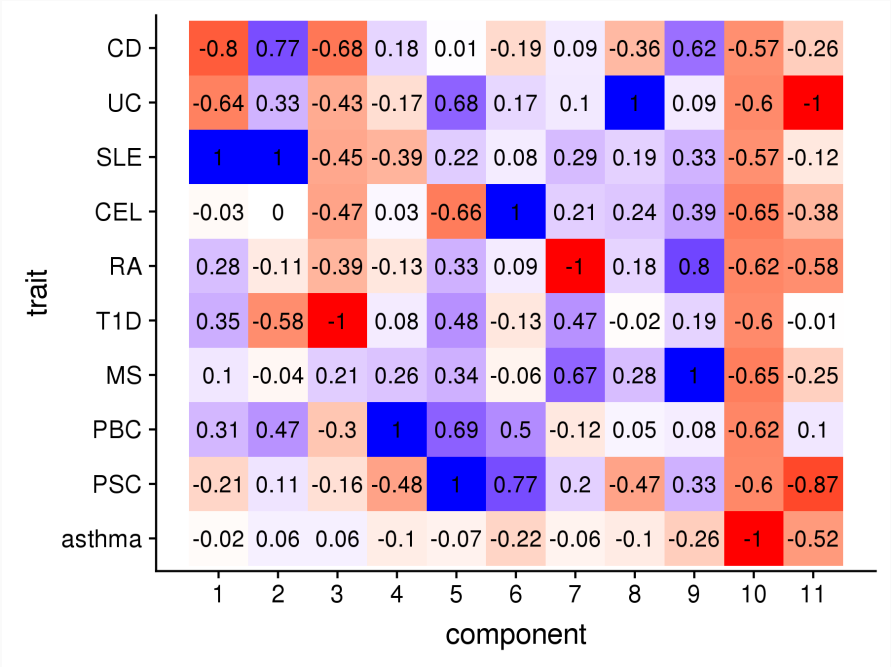


# Solution: focus data before principal components analysis

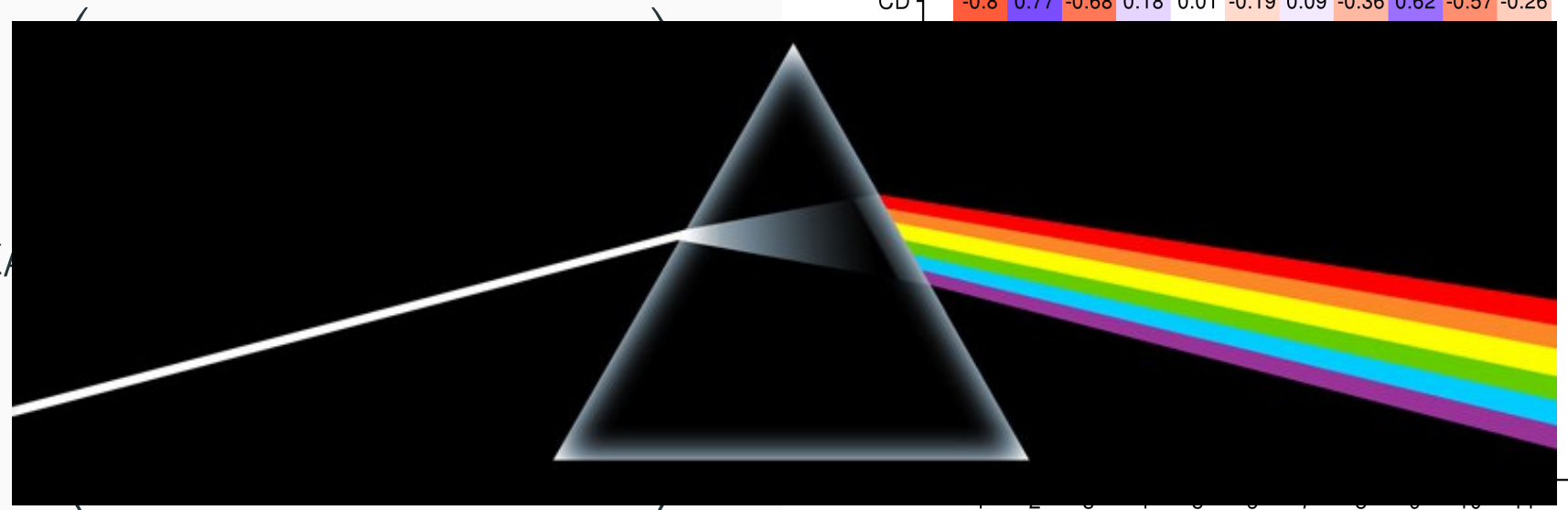
PCA



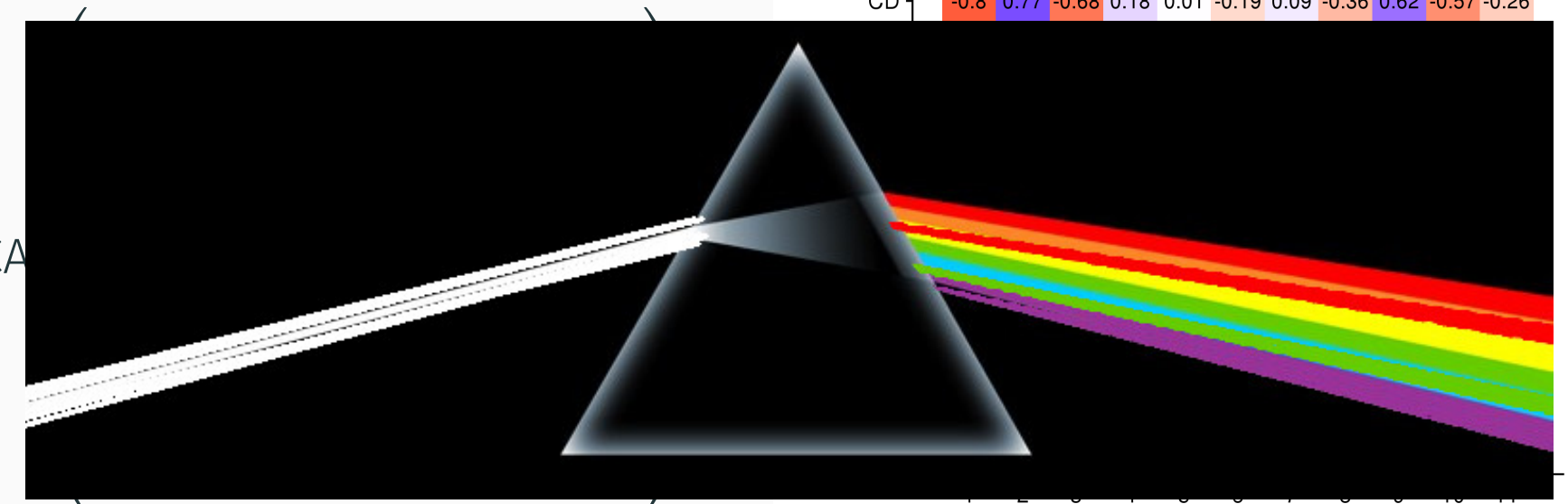

=



# Solution: focus data before principal components analysis

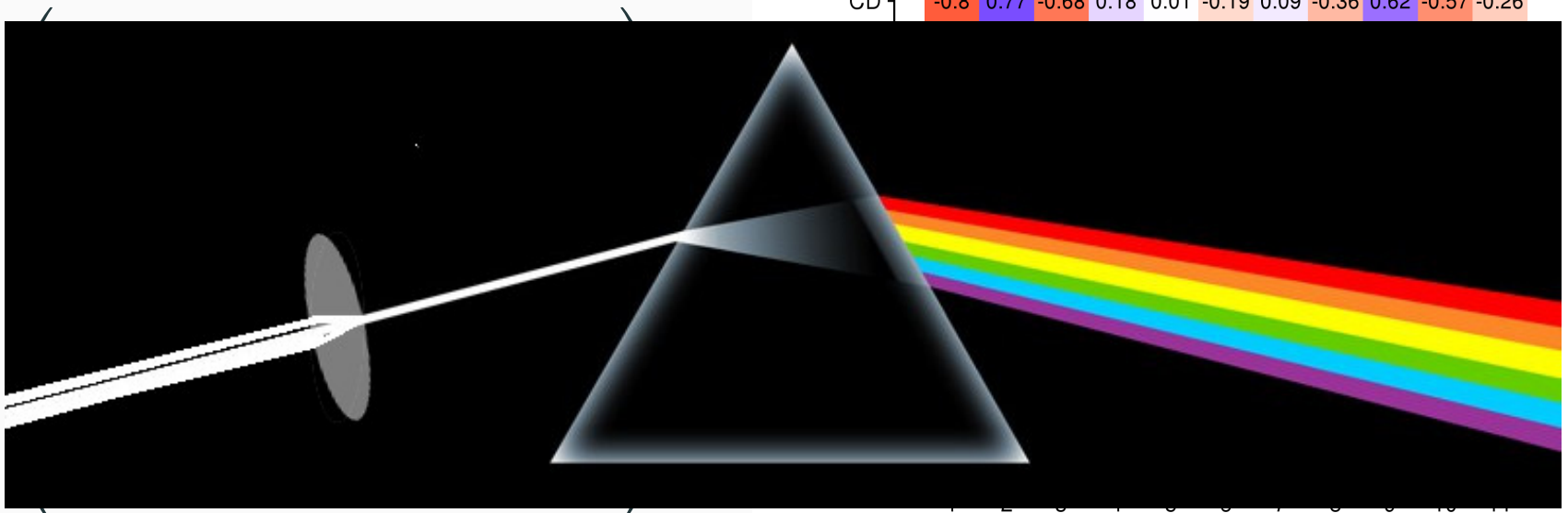


# Solution: focus data before principal components analysis



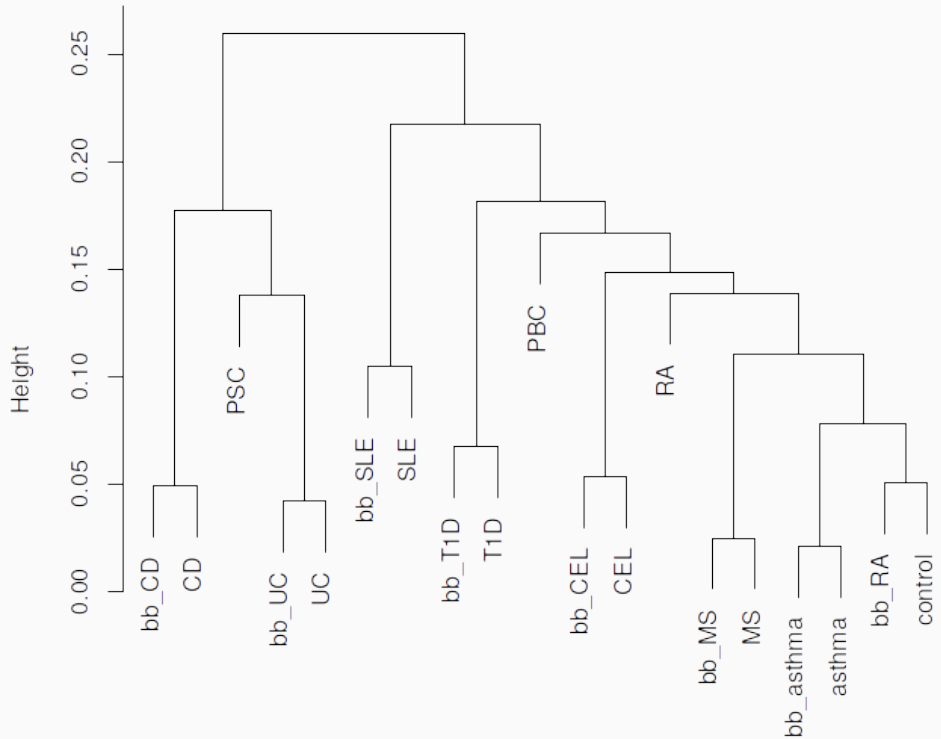
# Solution: focus data before principal components analysis

PCA

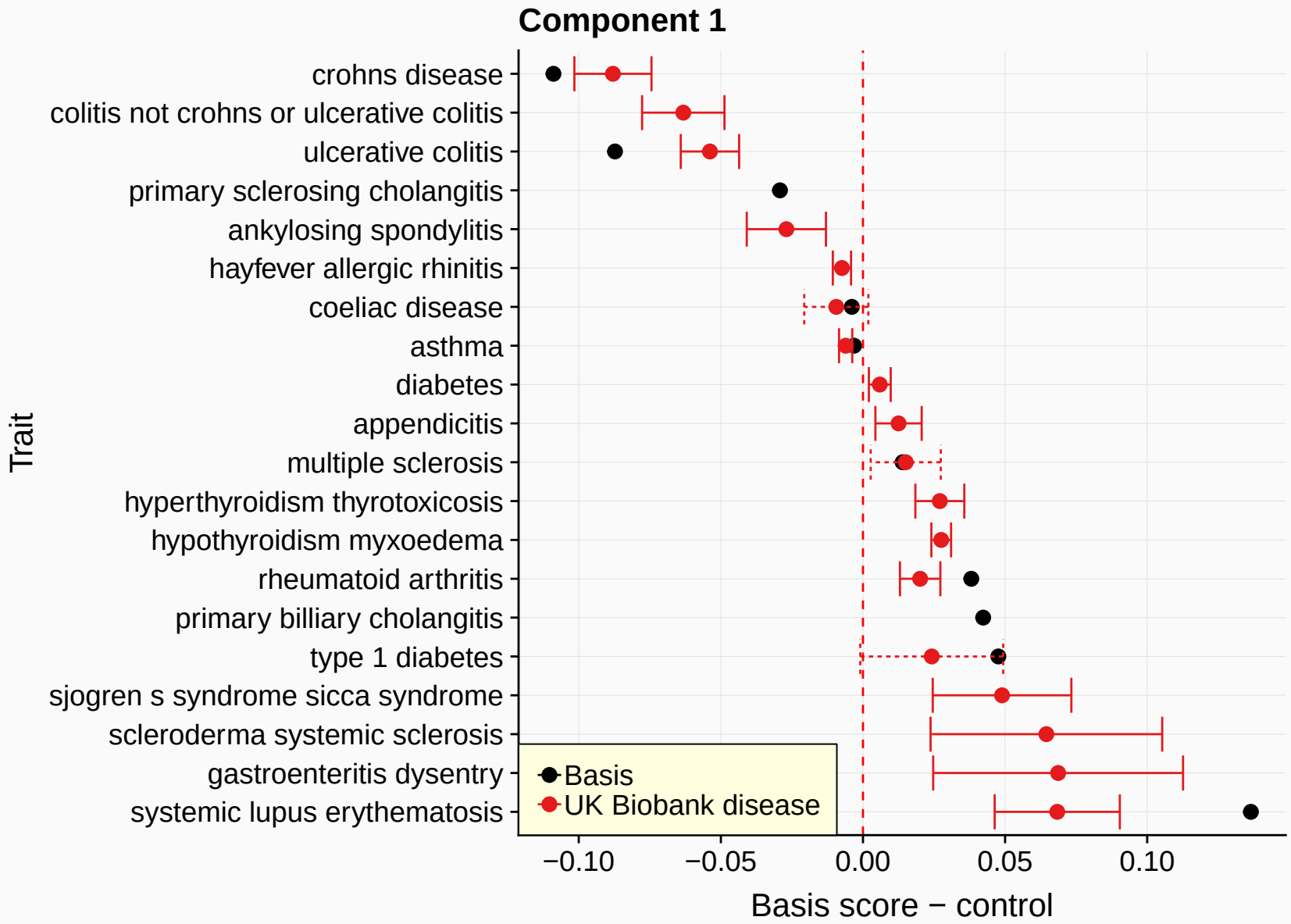


# Focused basis captures disease information

| 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 |

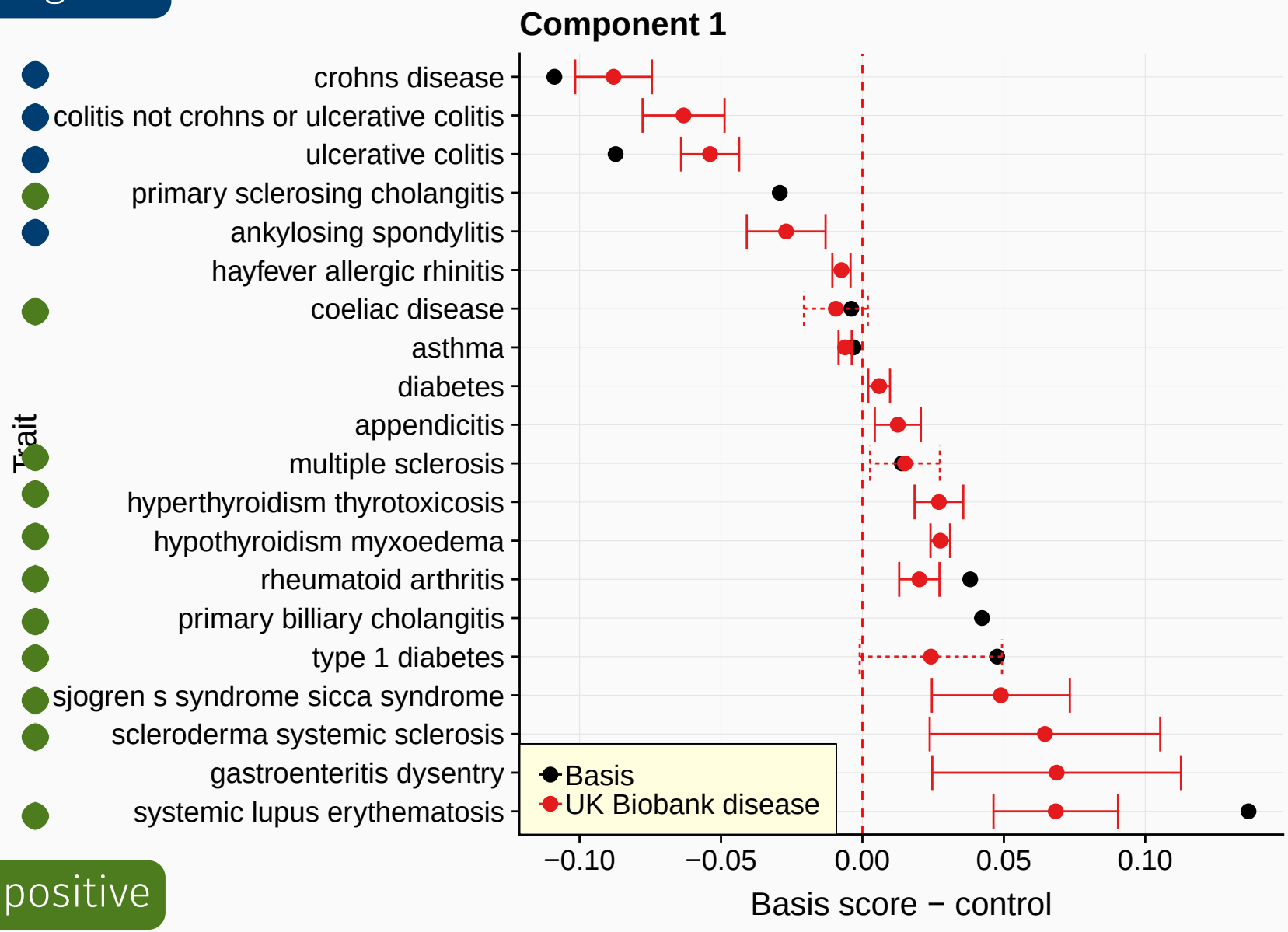


# Component 1: sero+/- disease, innate/adaptive



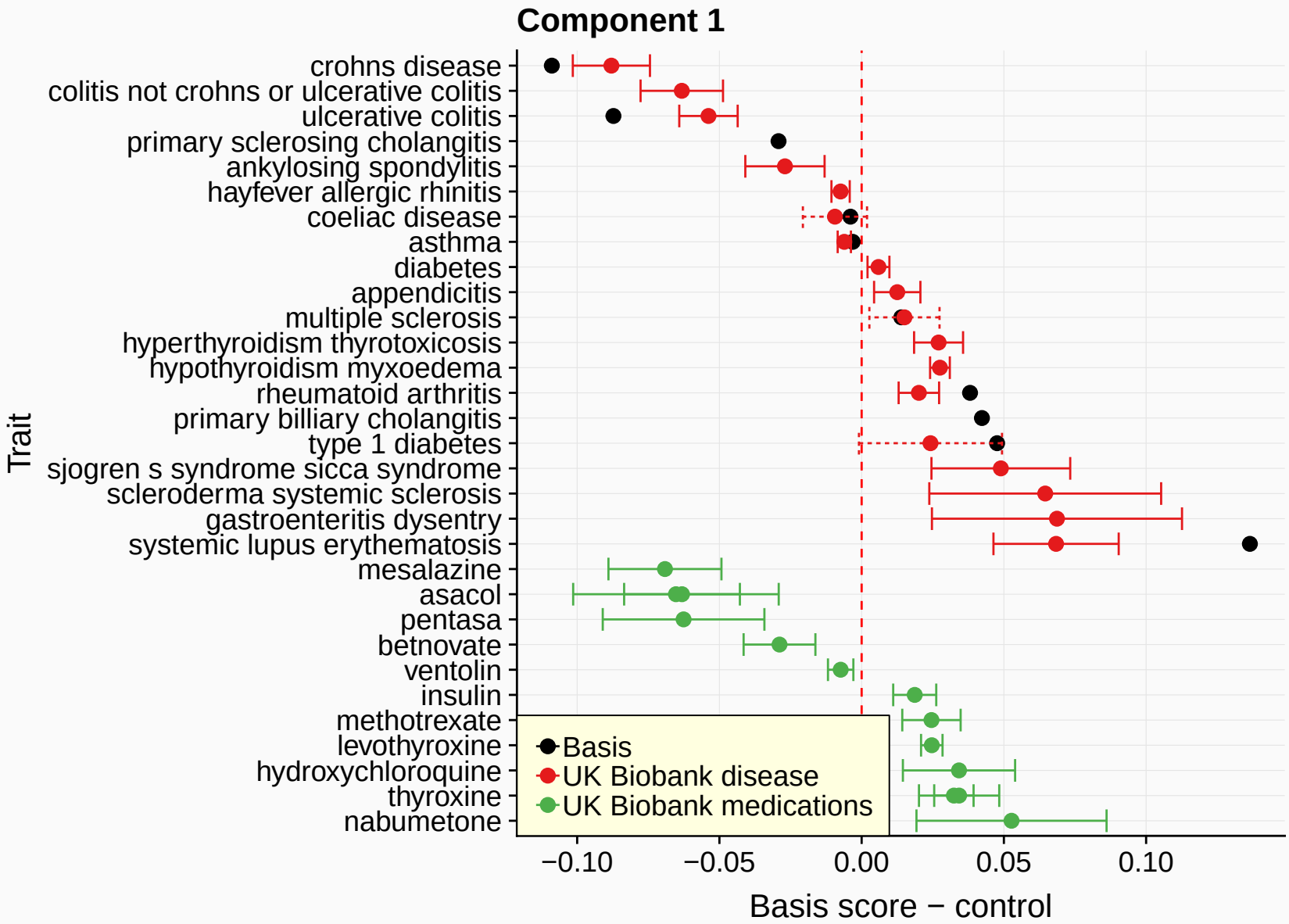
# Component 1: sero+/- disease, innate/adaptive

seronegative

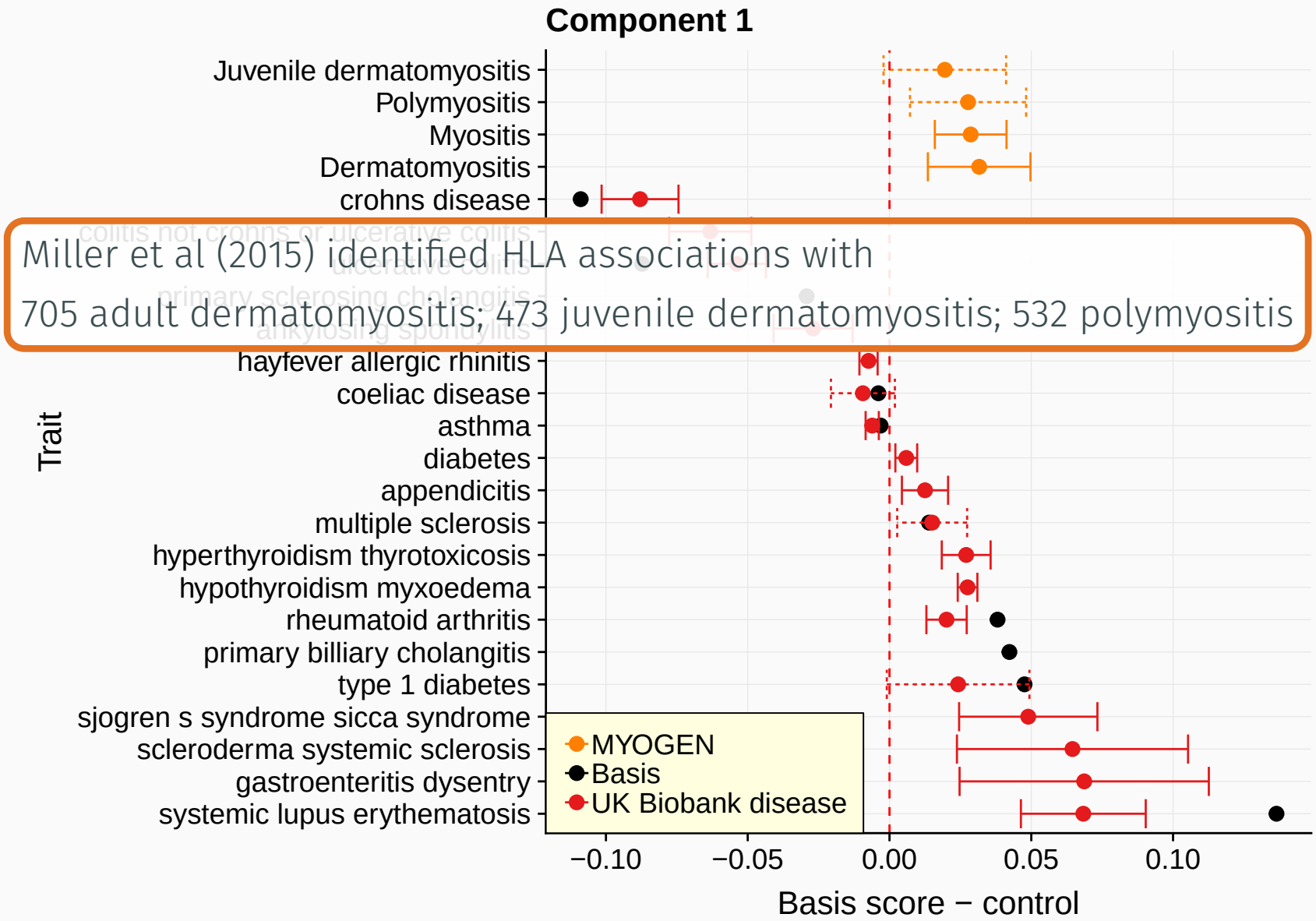



seropositive

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



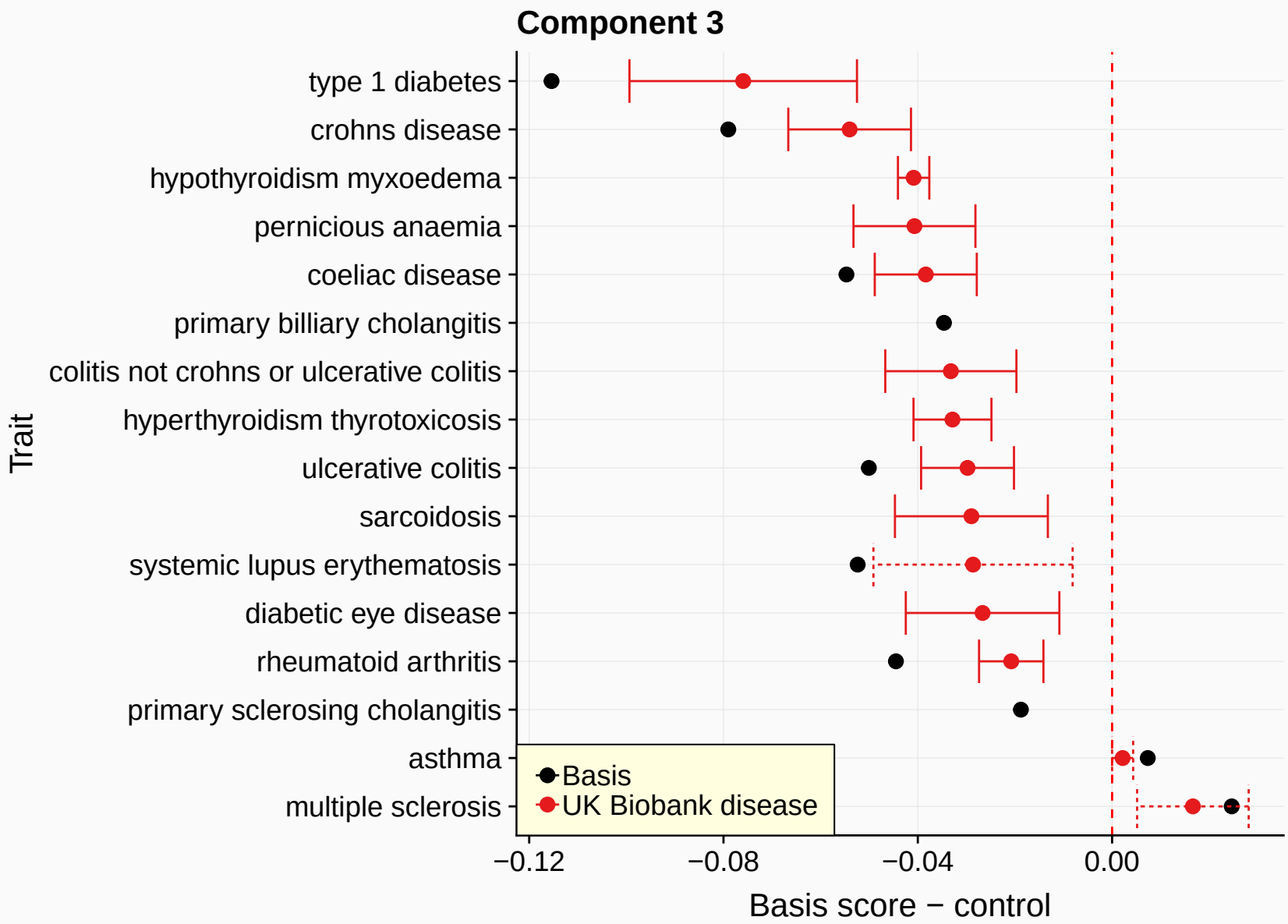
# Component 1: sero+/- disease, innate/adaptive



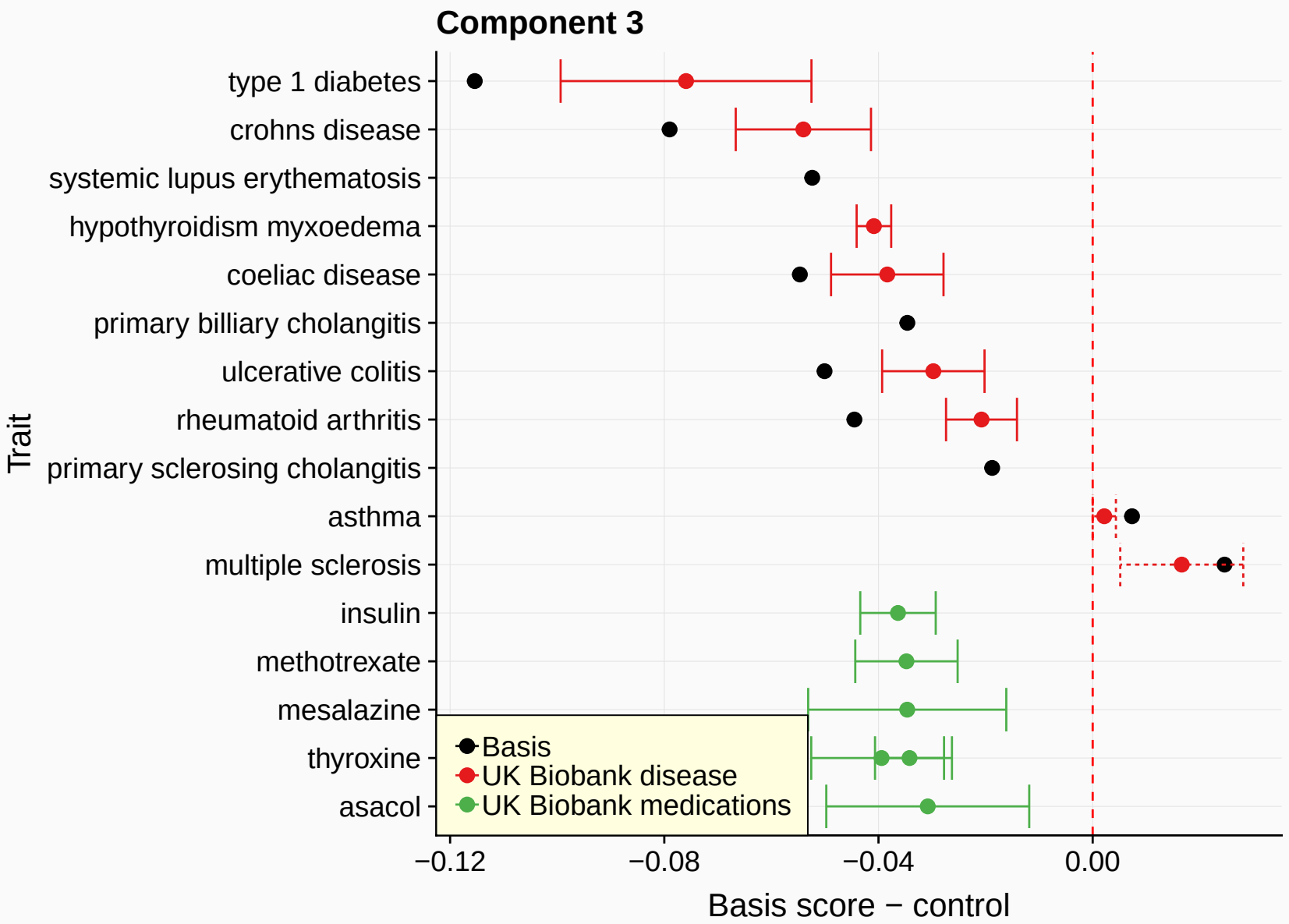

Miller et al (2015) identified HLA associations with

705 adult dermatomyositis; 473 juvenile dermatomyositis; 532 polymyositis

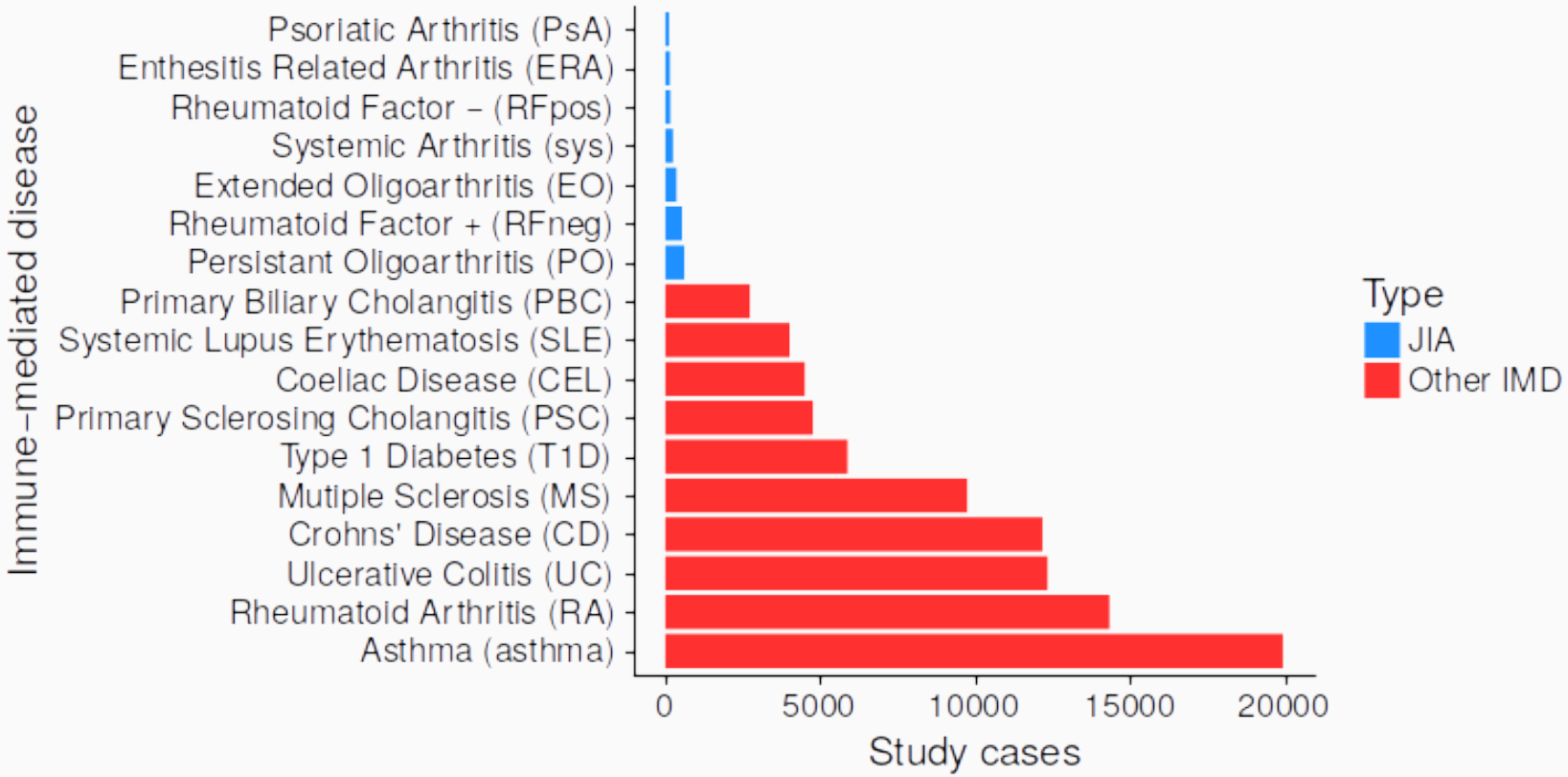
# Component 3


Component 3
type 1 diabetes
crohns disease
hypothyroidism myxoedema
pernicious anaemia
coeliac disease
primary billiary cholangitis
colitis not crohns or ulcerative colitis
hyperthyroidism thyrotoxicosis
ulcerative colitis
sarcoidosis
systemic lupus erythematosis
diabetic eye disease
rheumatoid arthritis
primary sclerosing cholangitis
asthma
multiple sclerosis
Trait
Basis
UK Biobank disease
−0.12
−0.08
−0.04
0.00
Basis score − control

# Component 3

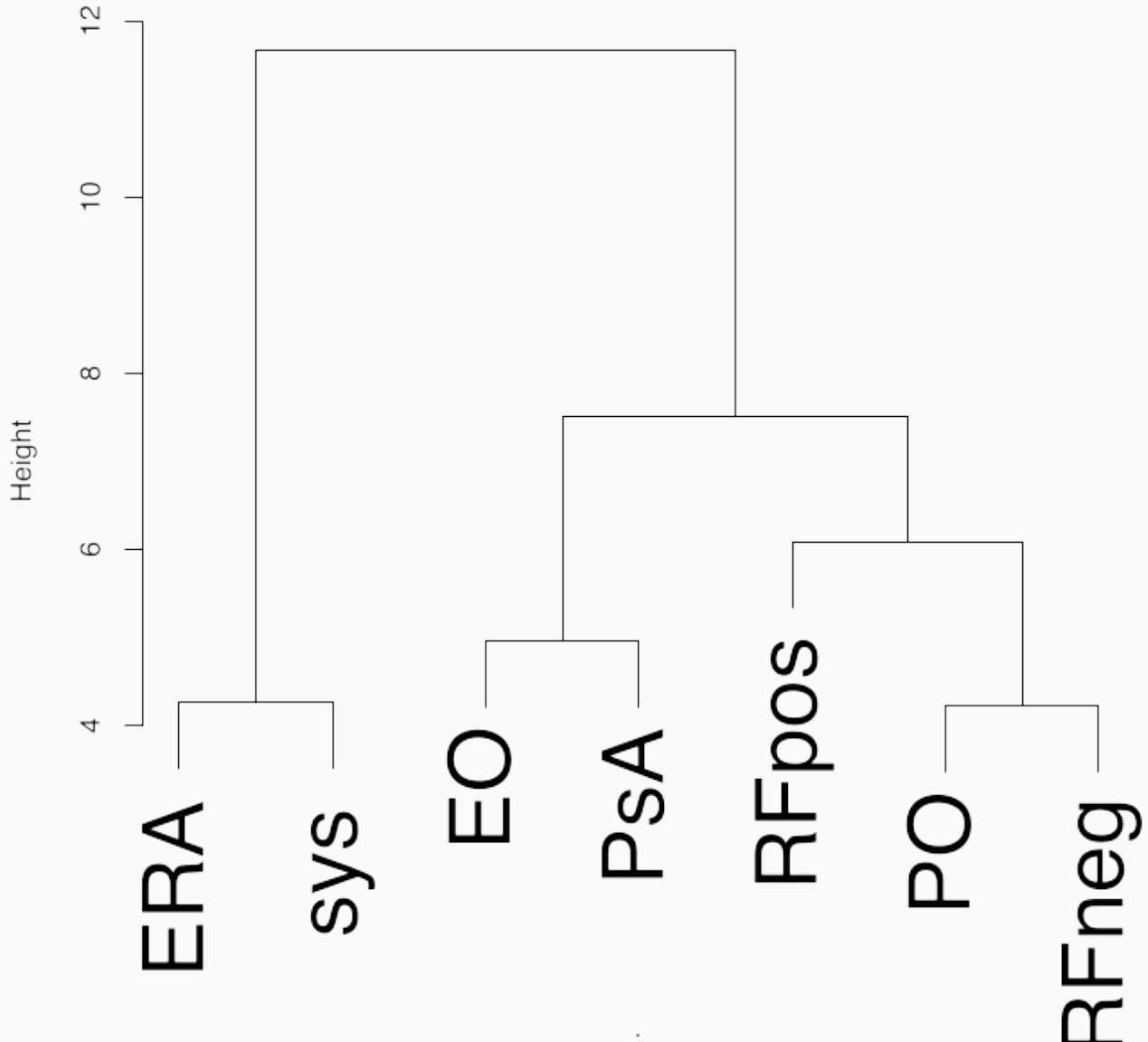

Component 3
type 1 diabetes
crohns disease
systemic lupus erythematosis
hypothyroidism myxoedema
coeliac disease
primary billiary cholangitis
ulcerative colitis
rheumatoid arthritis
primary sclerosing cholangitis
asthma
multiple sclerosis
insulin
methotrexate
mesalazine
thyroxine
asacol
Trait
Basis
UK Biobank disease
UK Biobank medications
−0.12
−0.08
−0.04
0.00
Basis score − control

# Component 3

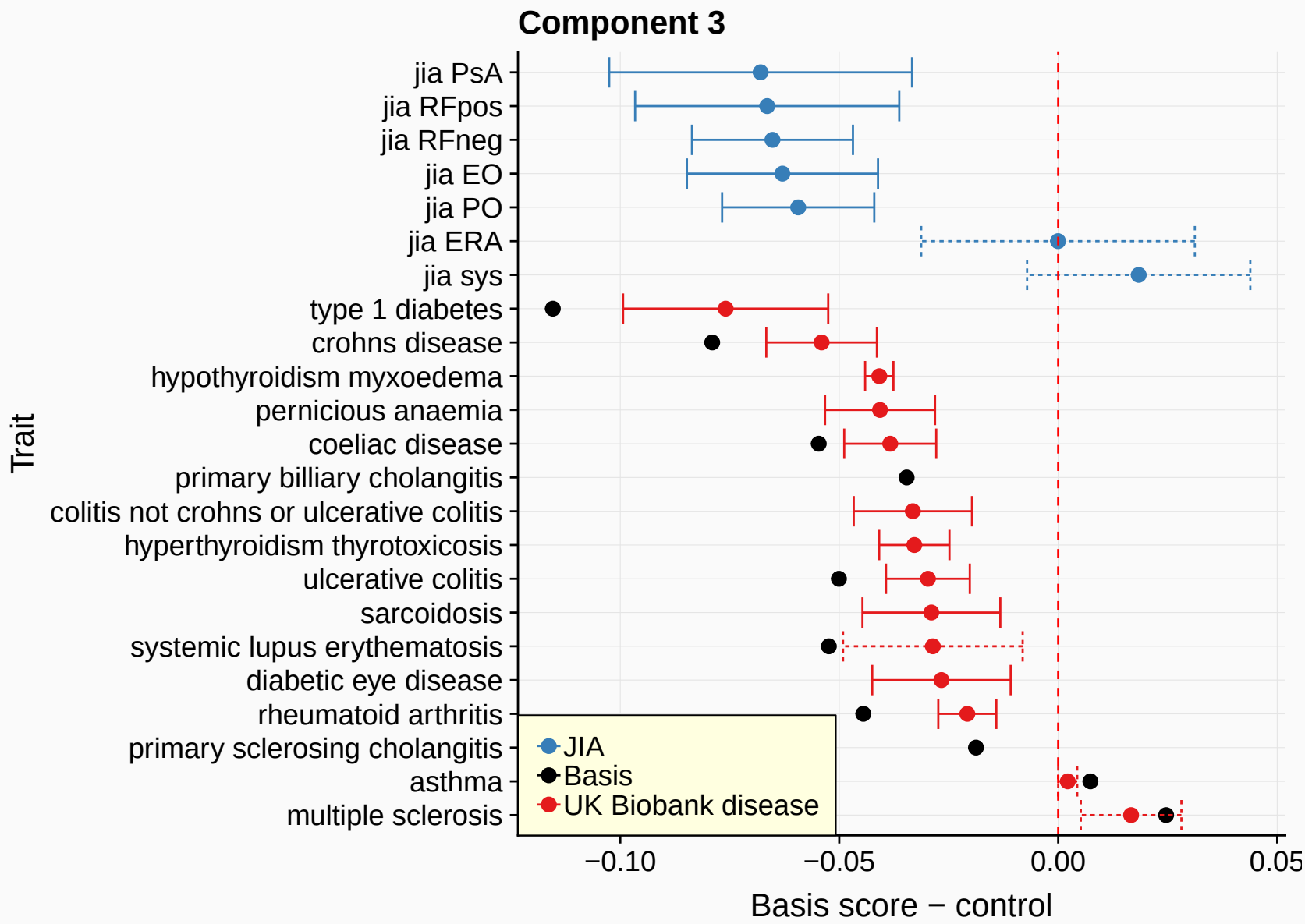



UK JIA genetics consortium

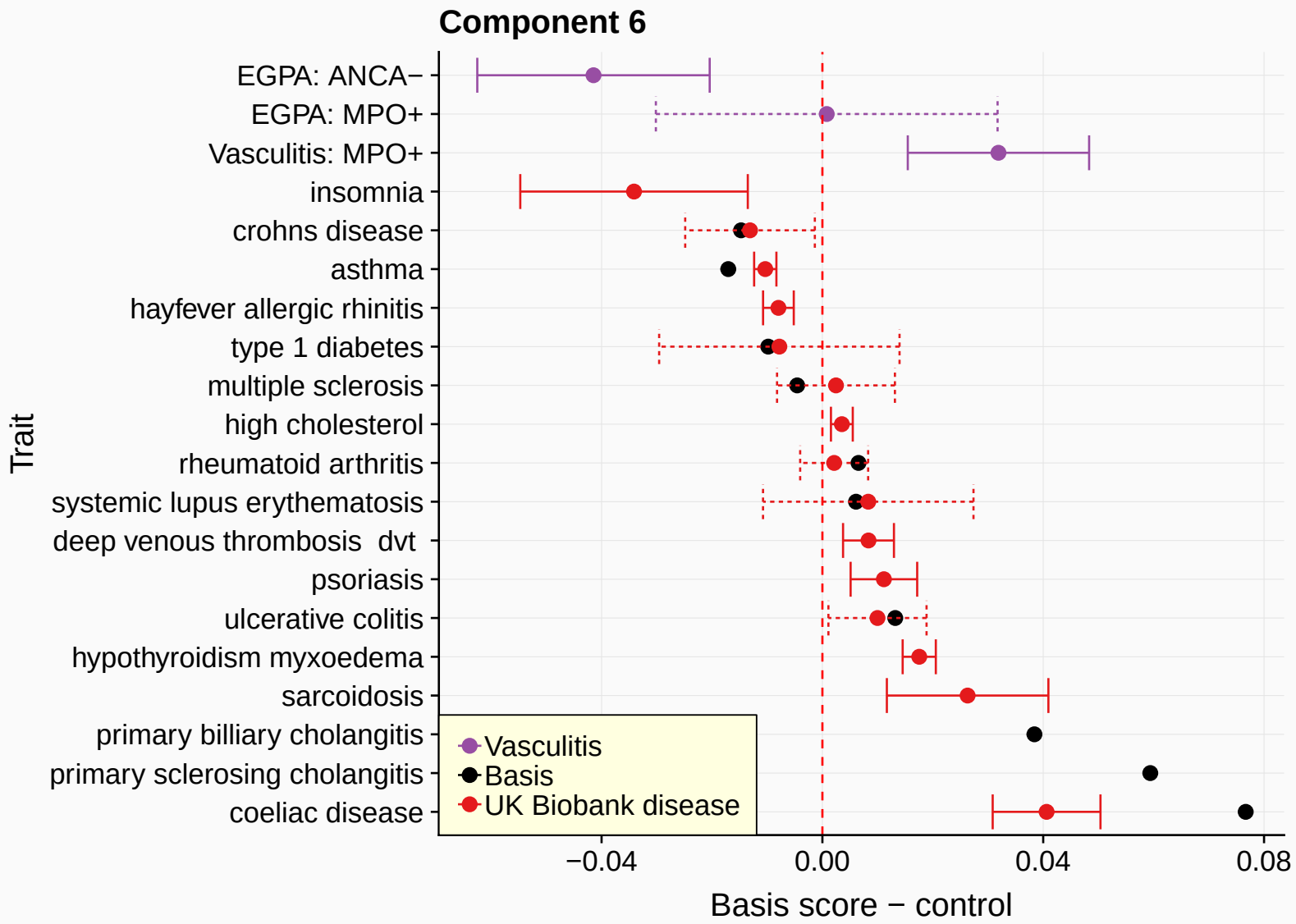
# Component 3



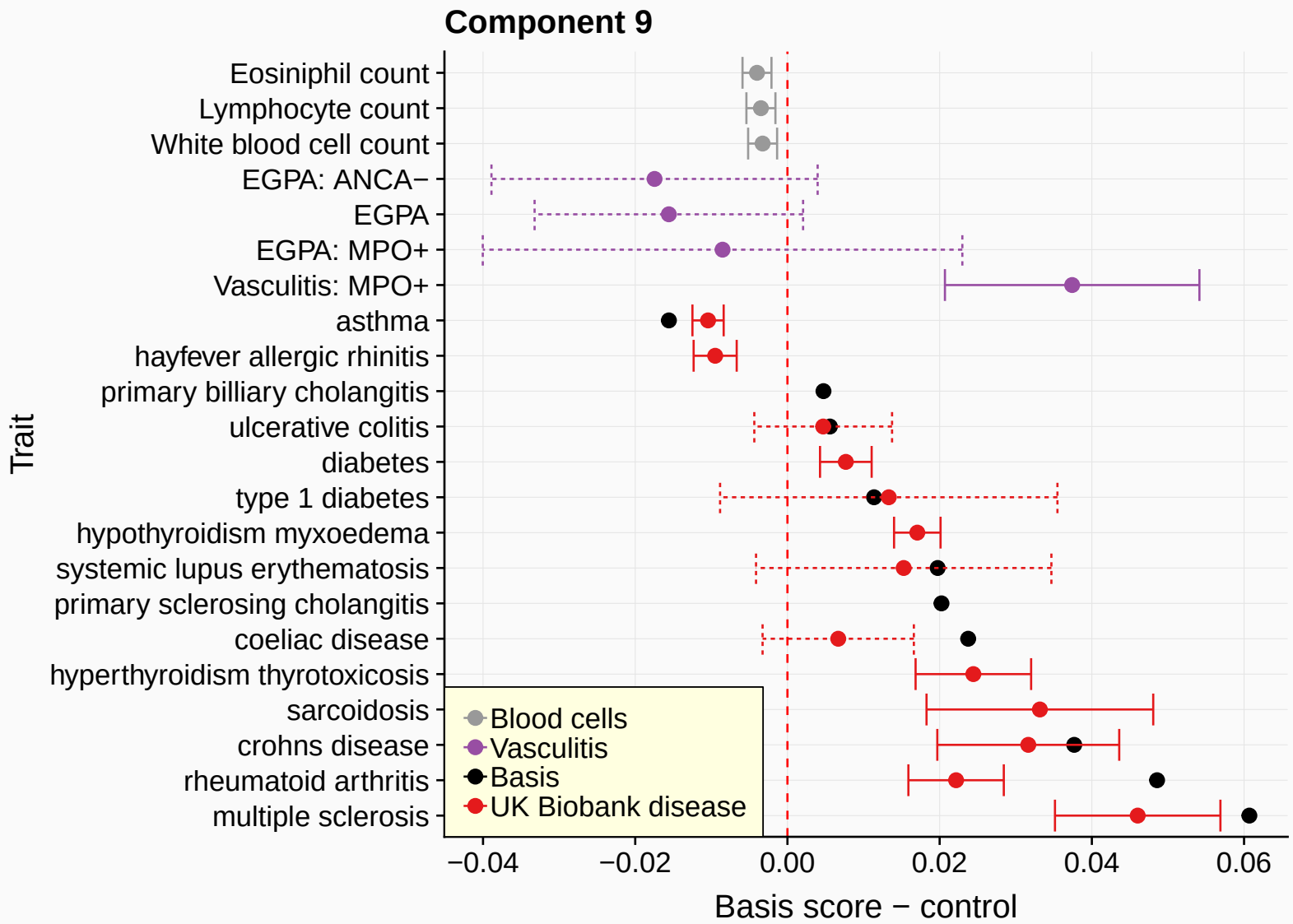
# Component 3


Component 3
jia PsA
jia RFpos
jia RFneg
jia EO
jia PO
jia ERA
jia sys
type 1 diabetes
crohns disease
hypothyroidism myxoedema
pernicious anaemia
coeliac disease
primary billiary cholangitis
colitis not crohns or ulcerative colitis
hyperthyroidism thyrotoxicosis
ulcerative colitis
sarcoidosis
systemic lupus erythematosis
diabetic eye disease
rheumatoid arthritis
primary sclerosing cholangitis
asthma
multiple sclerosis
Trait
JIA
Basis
UK Biobank disease
−0.10
−0.05
0.00
0.05
Basis score − control

# Vasculitis

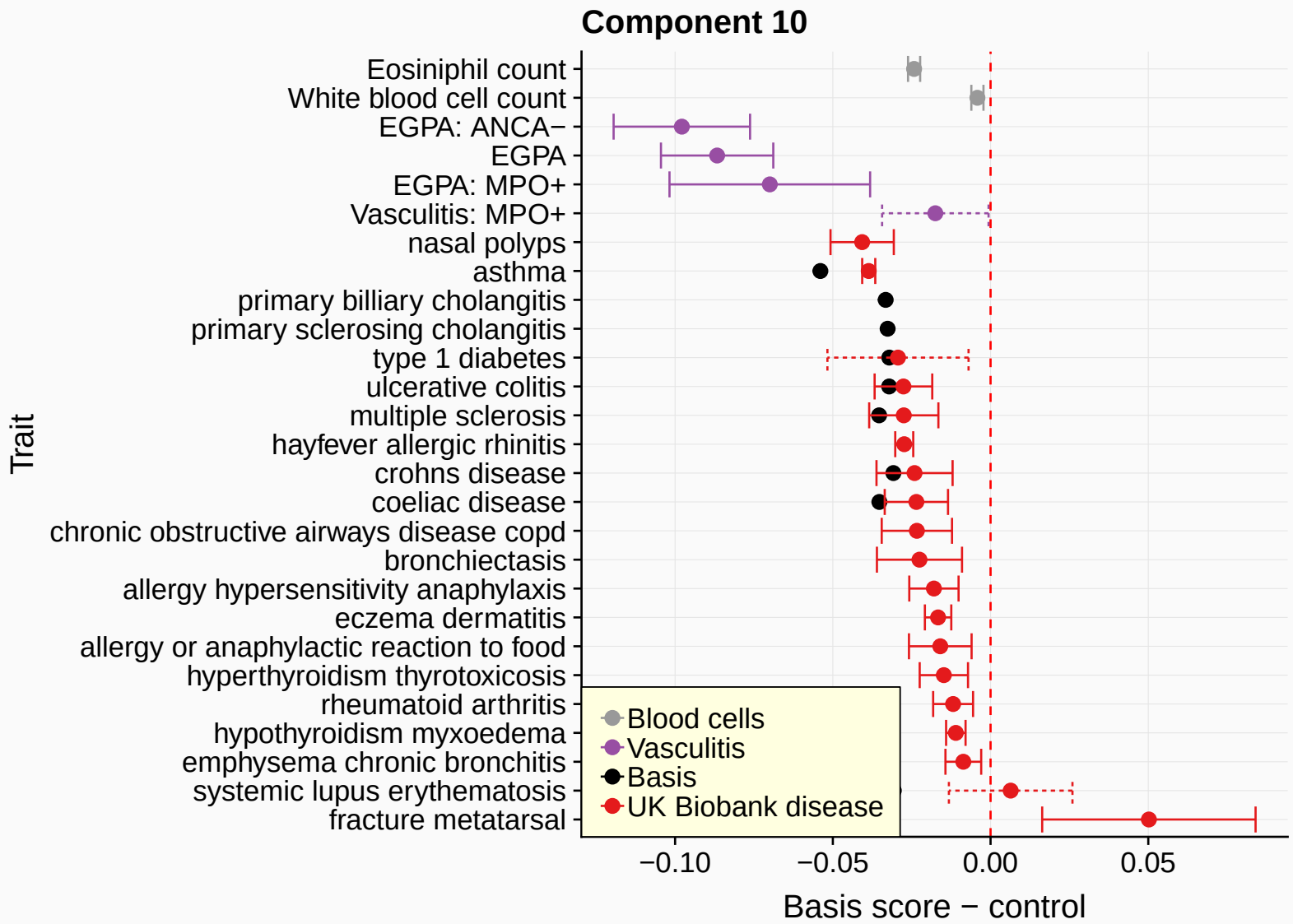

Component 6
EGPA: ANCA−
EGPA: MPO+
Vasculitis: MPO+
insomnia
crohns disease
asthma
hayfever allergic rhinitis
type 1 diabetes
multiple sclerosis
high cholesterol
rheumatoid arthritis
systemic lupus erythematosis
deep venous thrombosis dvt
psoriasis
ulcerative colitis
hypothyroidism myxoedema
sarcoidosis
primary billiary cholangitis
primary sclerosing cholangitis
coeliac disease
Trait
Vasculitis
Basis
UK Biobank disease
−0.04
0.00
0.04
0.08
Basis score − control

# Vasculitis




*Astle et al, 2016*

# Vasculitis




*Astle et al, 2016*

# 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
- Region-wise overlap analysis useful to “zoom in”

# Summary

- Genetic sharing across immune mediated diseases is pervasive, but complex
- Sharing **should** be exploited to learn about less common diseases
- Region-wise overlap analysis useful to “zoom in”
- 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
- Region-wise overlap analysis useful to “zoom in”
- 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...

James Liley

Stasia Grinberg

Olly Burren

## Patients, families, study PIs who shared data

**Sample Cohorts**

- UK JIA Genetics Consortium
- Childhood Arthritis Prospective Study (CAPS)
- Childhood Arthritis Response to Medication Study (CHARMS)
- MYOGEN consortium
- European Vasculitis Genetics Consortium

**Cambridge**

Ken Smith  Paul Lyons

**MYOGEN consortum**

Fred Miller  Chris Amos

**Manchester**

Wendy Thomson  Anne Hinks
Joanna Cobb  John Bowes
Annie Yarwood  Sam Smith
Kimme Hyrich

**University College London**

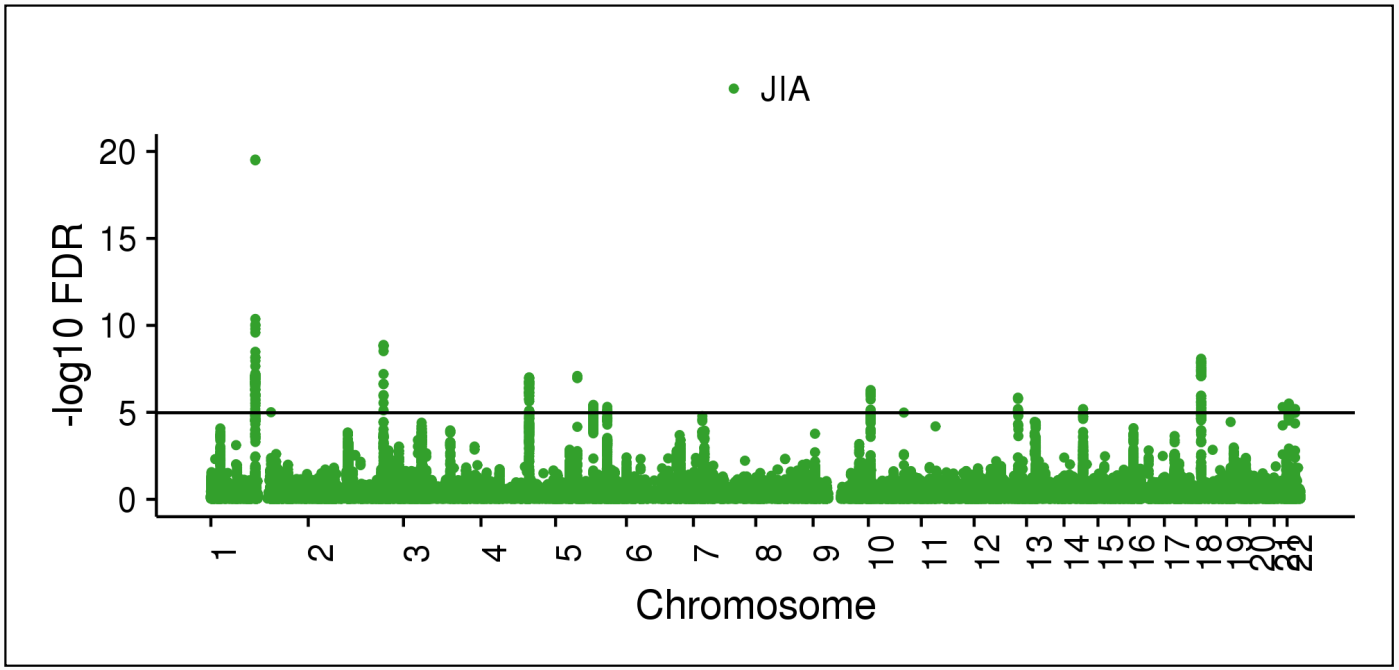
Lucy Wedderburn  Claire Deakin

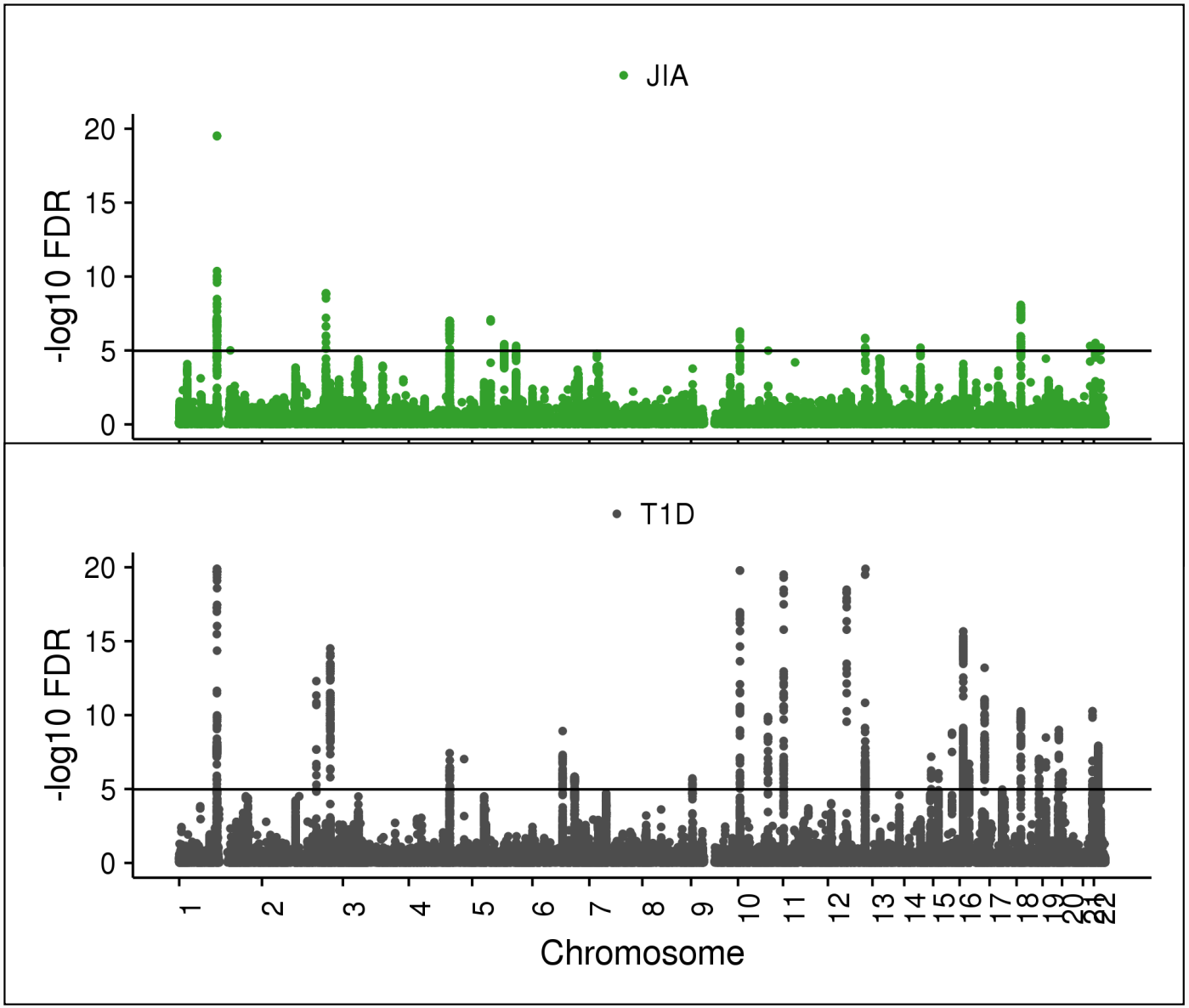
wellcome

UNIVERSITY OF CAMBRIDGE

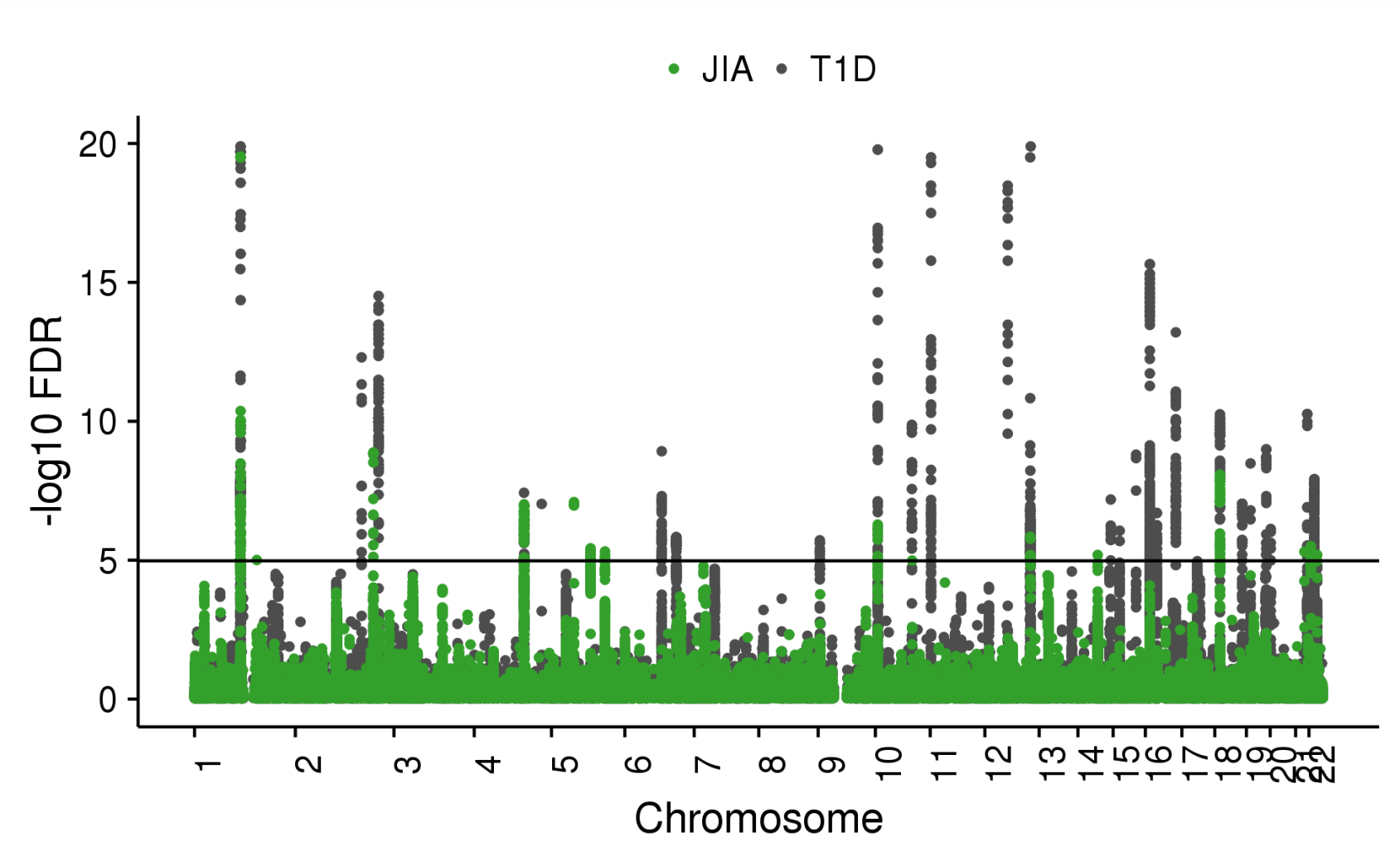
MRC Biostatistics Unit

# Sharing information between diseases to increase power


JIA
-log10 FDR
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

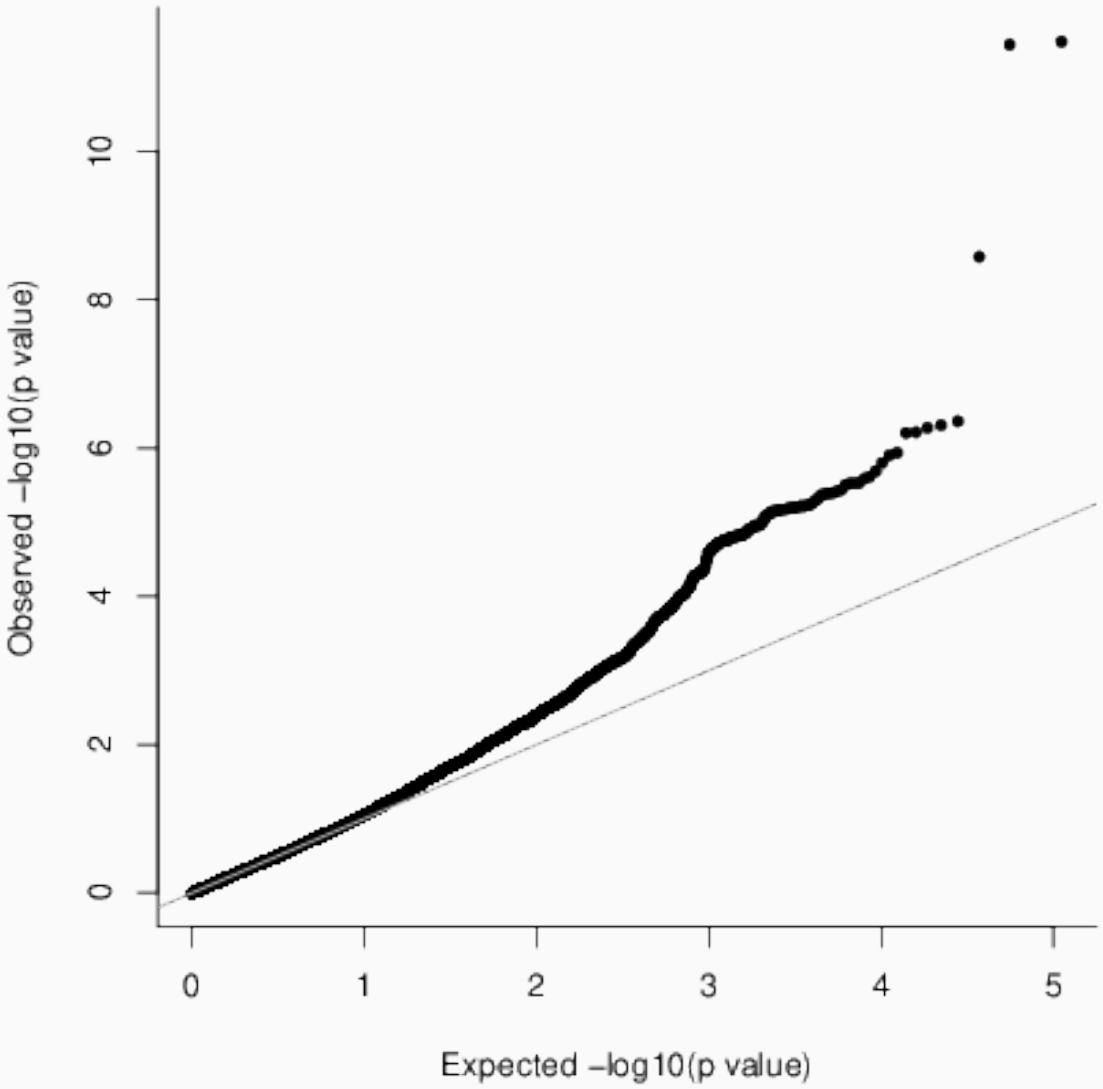



*Hinks et al, 2013; Onengut et al 2015*

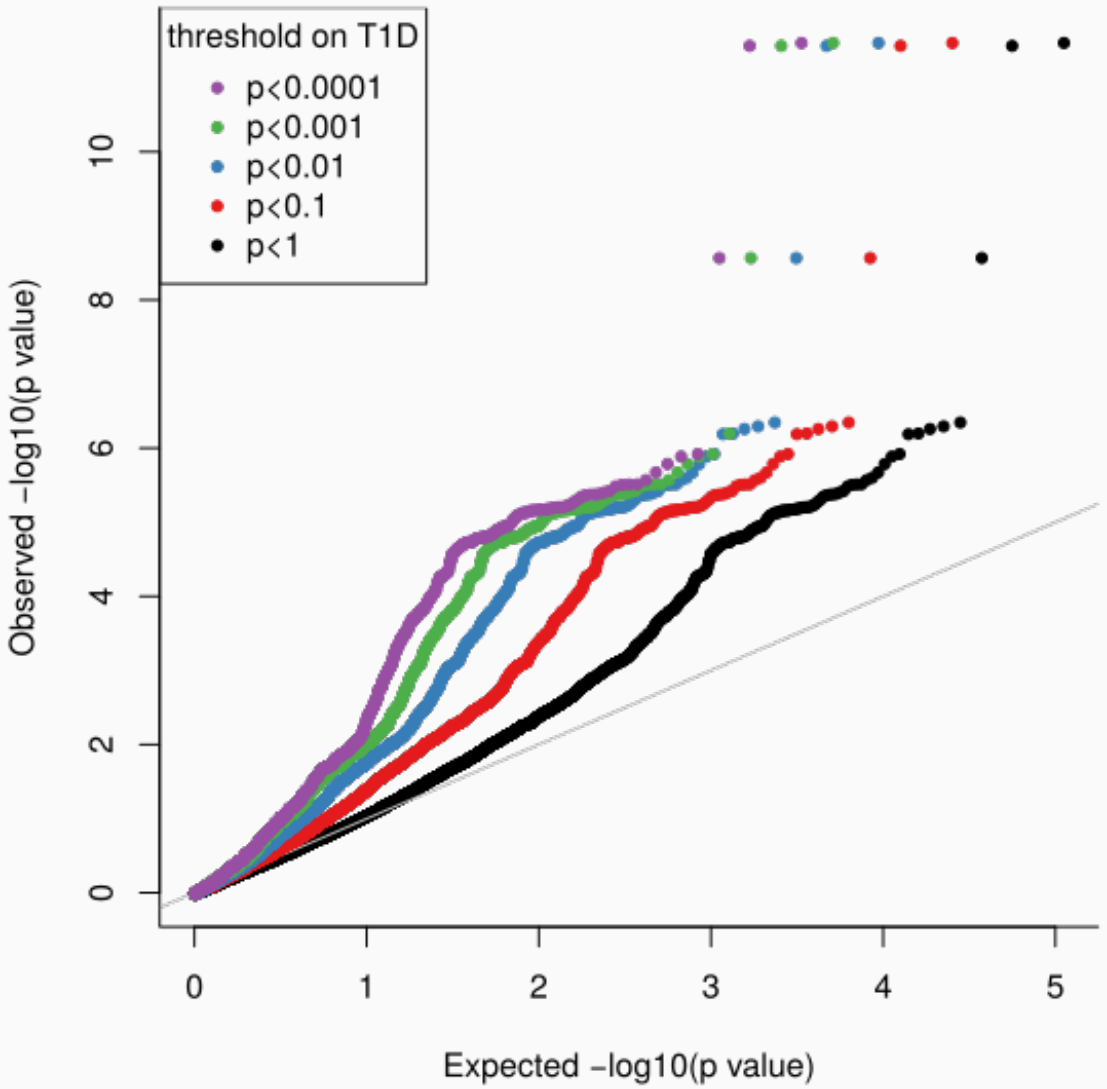



*Hinks et al, 2013; Onengut et al 2015*

# T1D association is informative for JIA



# T1D association is informative for JIA



# False discovery rates $\rightarrow$ conditional false discovery rates

False discovery rate, FDR

$$Pr(H_0|P < \alpha) \propto \alpha \times Pr(H_0)$$

# False discovery rates → conditional false discovery rates

False discovery rate, FDR

$$Pr(H_0|P < \alpha) \propto \alpha \times Pr(H_0)$$

Conditional false discovery rate, cFDR

$$Pr(H_0|P < \alpha, Q < \gamma) \propto \alpha \times Pr(H_0|Q < \gamma)$$

# False discovery rates → conditional false discovery rates

False discovery rate, FDR

$$Pr(H_0|P < \alpha) \propto \alpha \times Pr(H_0)$$

Conditional false discovery rate, cFDR

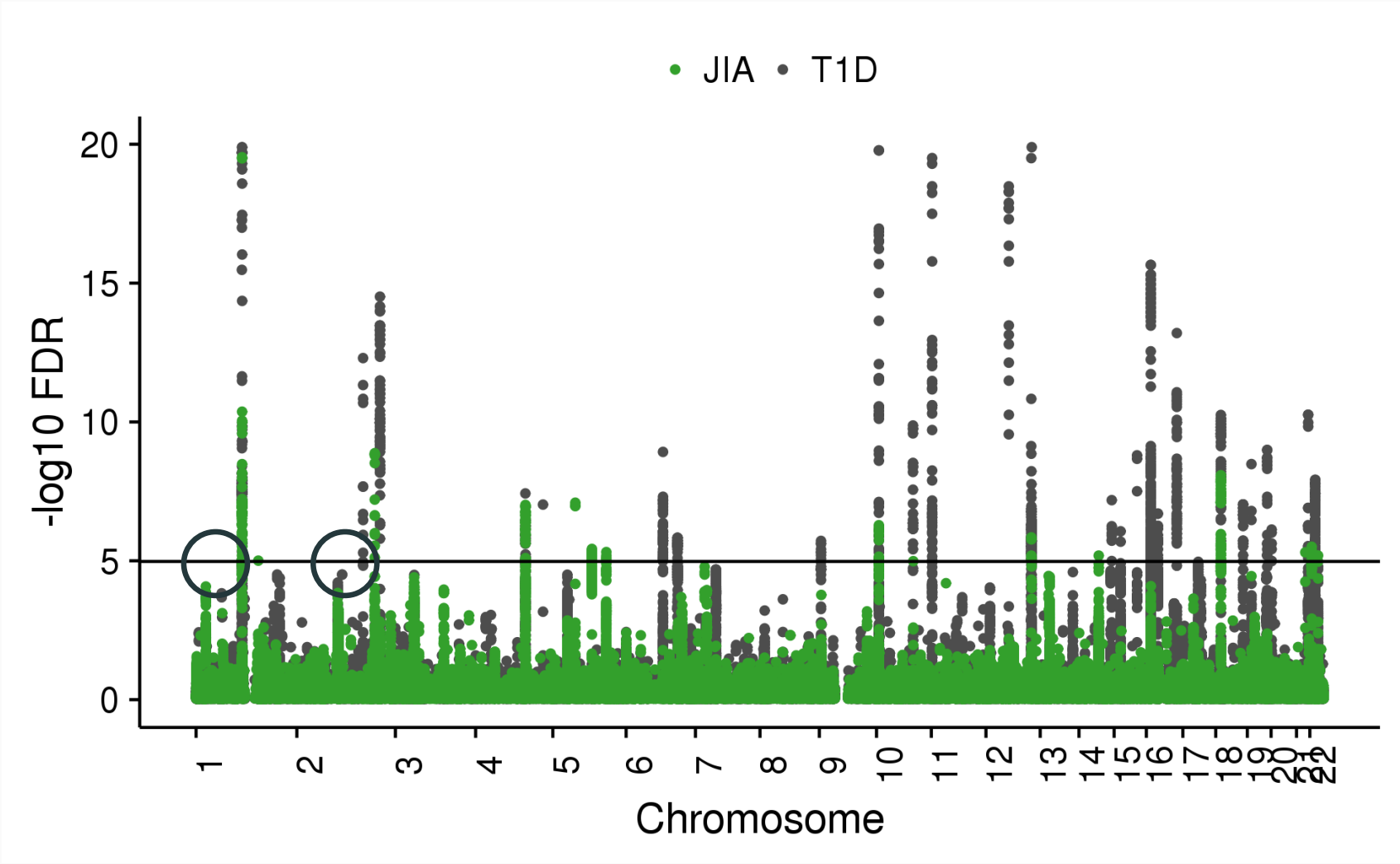
$$Pr(H_0|P < \alpha, Q < \gamma) \propto \alpha \times Pr(H_0|Q < \gamma)$$

If Q is independent of P then

$$cFDR = FDR$$

# False discovery rates → conditional false discovery rates

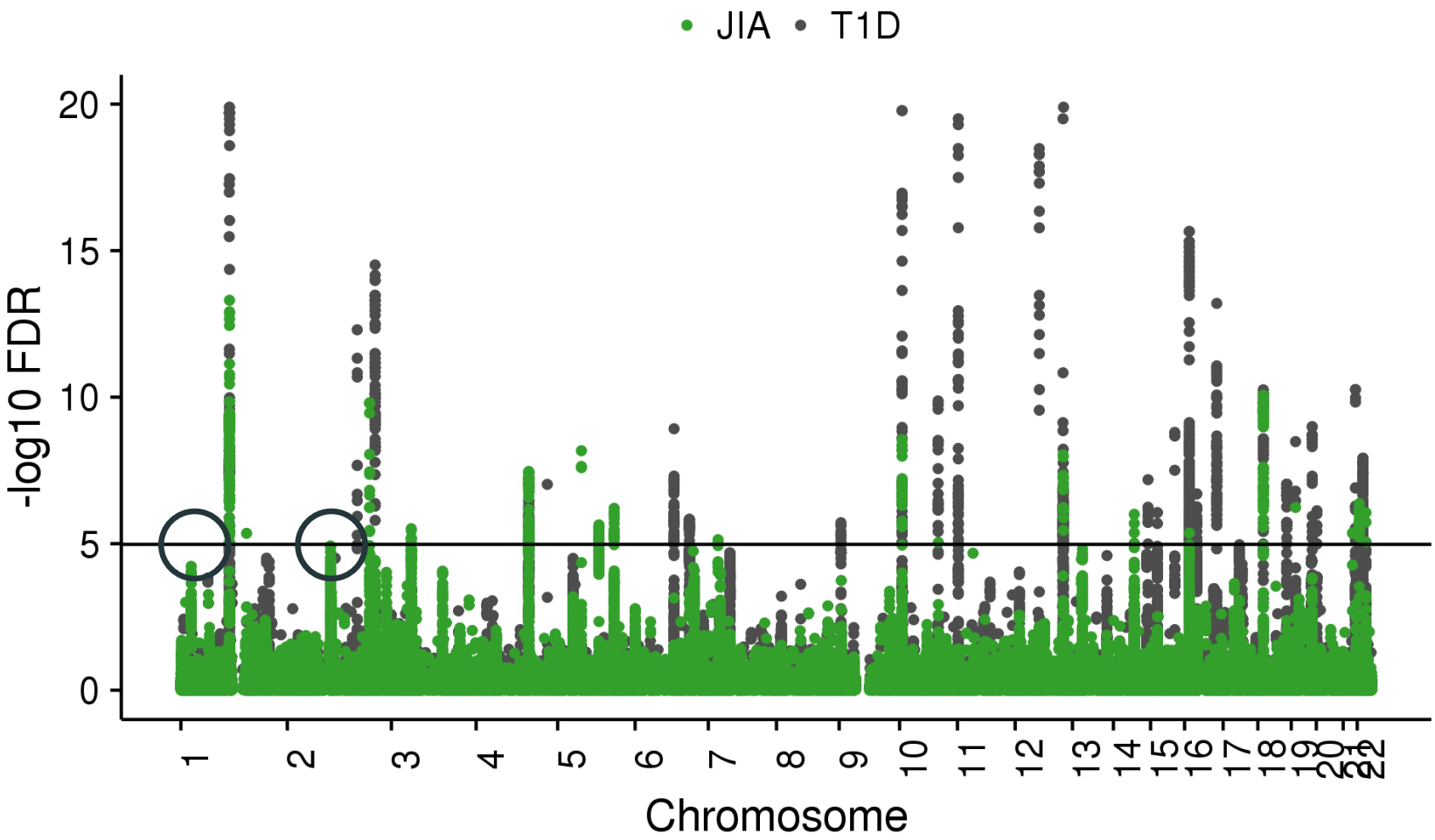
FDR




*Hinks et al, 2013; Onengut et al 2015*

# False discovery rates → conditional false discovery rates

conditional FDR




*Hinks et al, 2013; Onengut et al 2015*

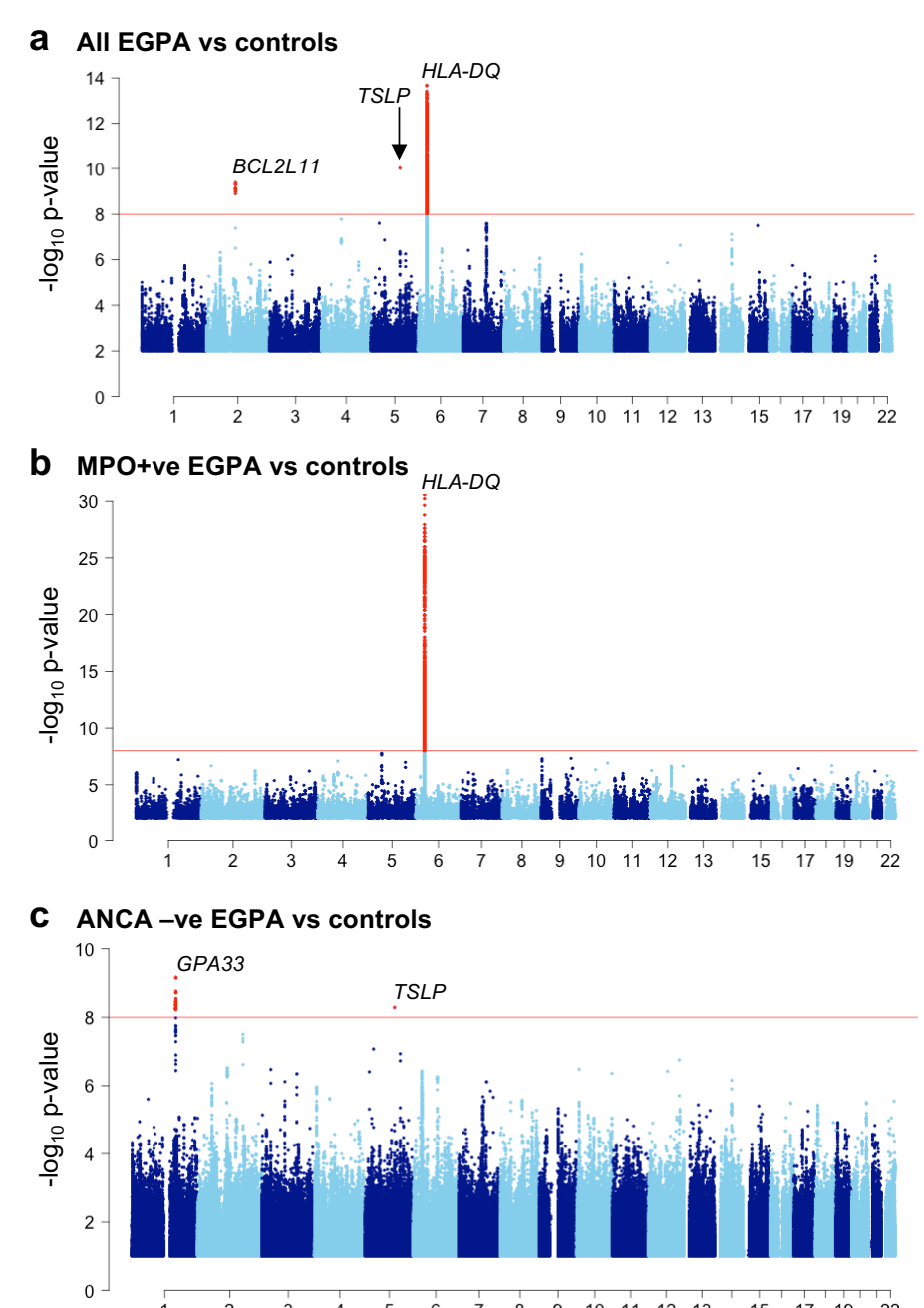
# Example: Eosinophilic Granulomatosis with Polyangiitis (EGPA)

- ANCA-associated vasculitis + eosiniphilia, asthma
- incidence of 0.5-3.7 per million
- considered a single disease
- not all patients are ANCA positive

| | |
|---|---|
| ANCA positive (MPO) | 161 |
| ANCA negative | 358 |
| Controls | 6717 |

*Lyons et al, biorxiv 491837*

# Eosinophilic Granulomatosis with Polyangiitis (EGPA)

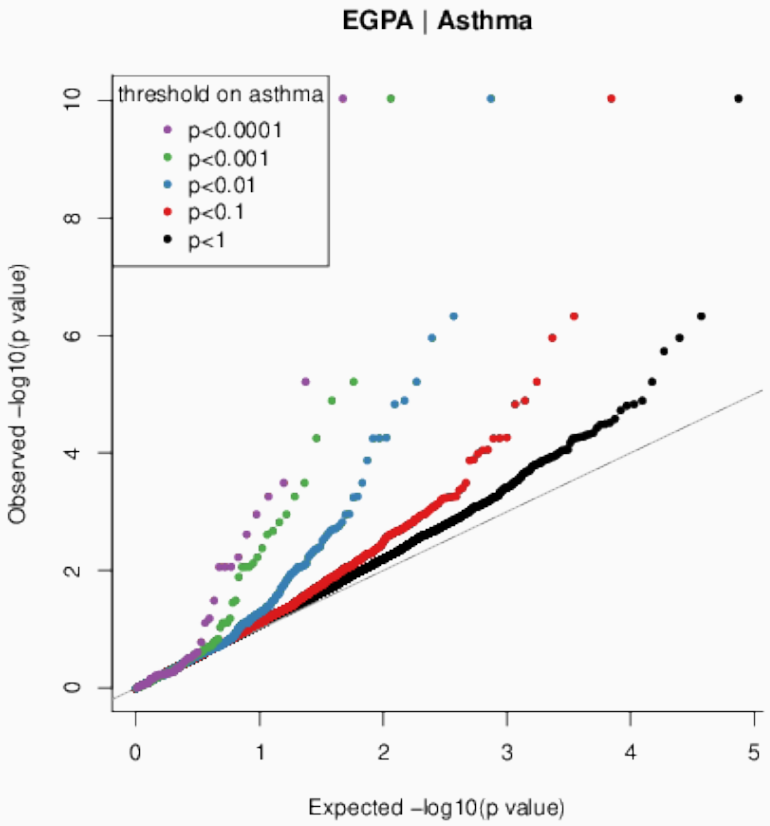


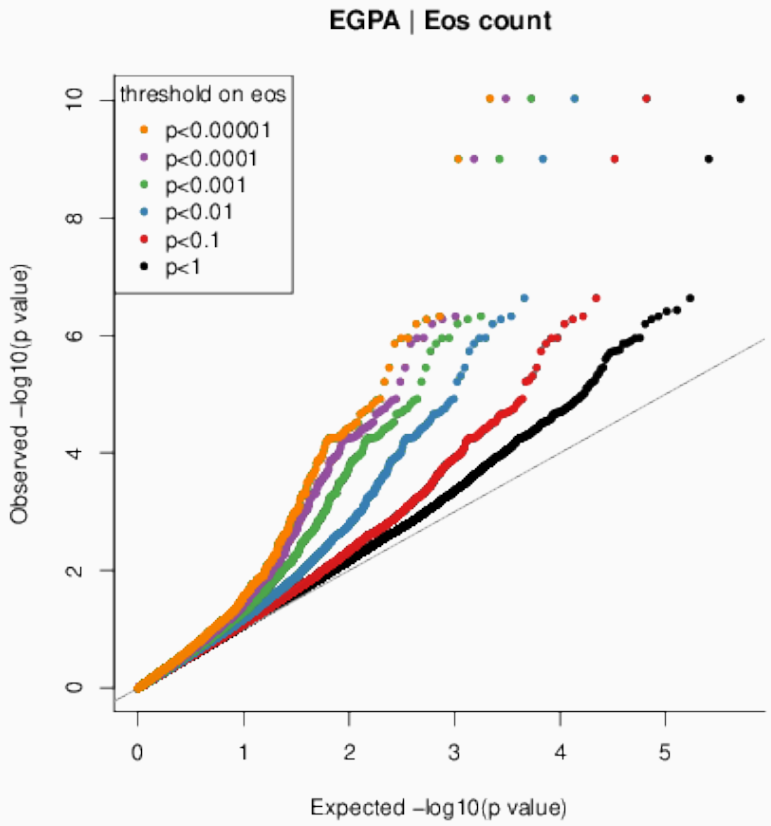

all EGPA

MPO+ EGPA

ANCA- EGPA

# Both asthma and eosinophilia are informative for EGPA



# EGPA: Genetically distinct clinical subsets, shared associations with asthma and eosinophil count

| | MPO+ EGPA | ANCA- EGPA |
|---|---|---|
| *Distinct associations* | HLA-DQ | GPA33, (IL5) |
| *Shared associations* | BIM, TSLP | |
| *Conditional associations* | | |
| Asthma | IL5, BACH2, 10p14 (GATA3?) | |
| Eosiniphilia | CDK6, SOCS1, LPP, TBX3 | |
| *Differential symptoms (%)* | | |
| Glomerulonephritis | 29 | 9 |
| Neuropathy | 79 | 57 |
| *Treatment response (%)* | | |
| **Rituximab response** | 80 | 38 |