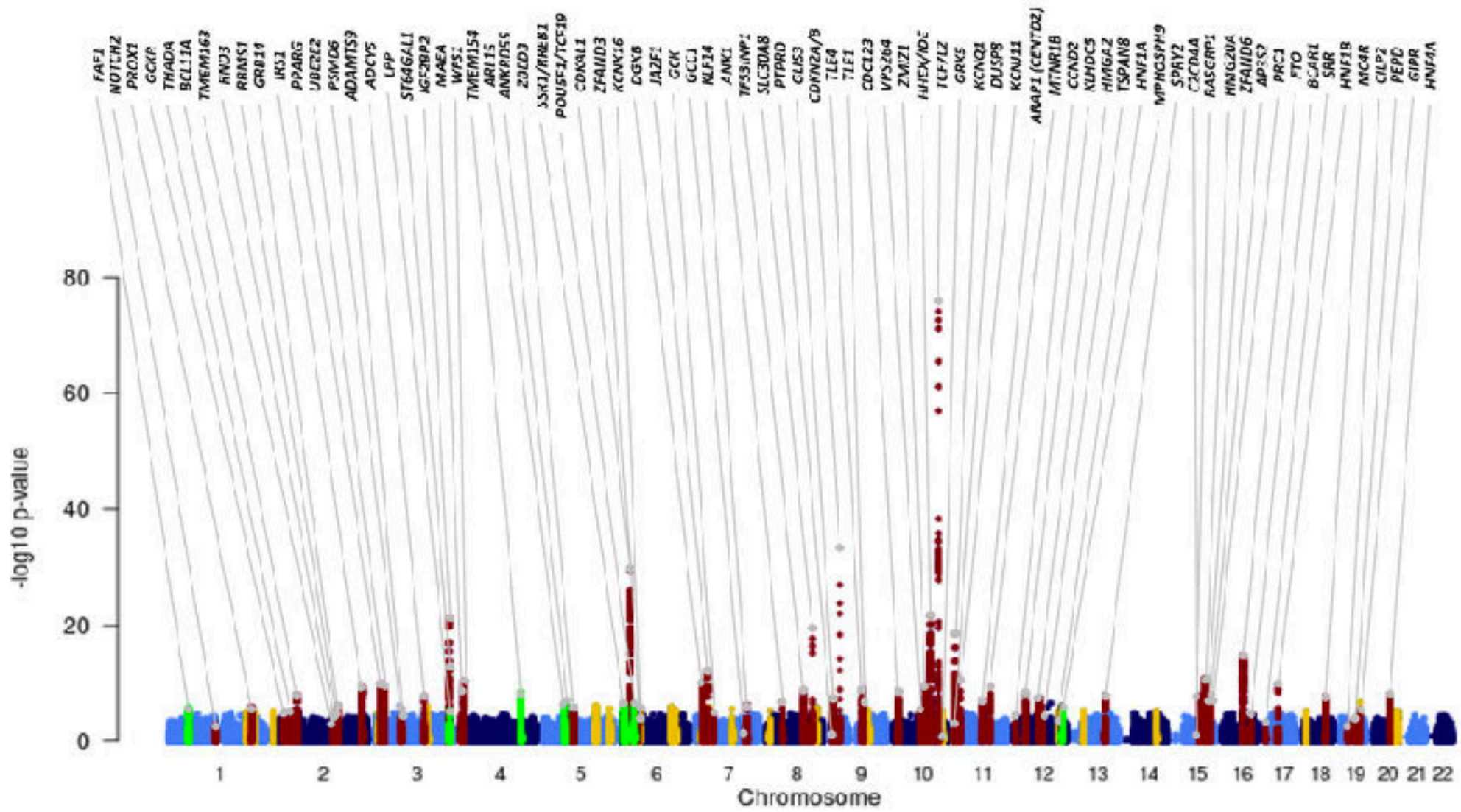
High-dimensional inference from summary statistics

Manuel A. Rivas
Department of Biomedical Data Science
Stanford University
rivaslab.stanford.edu

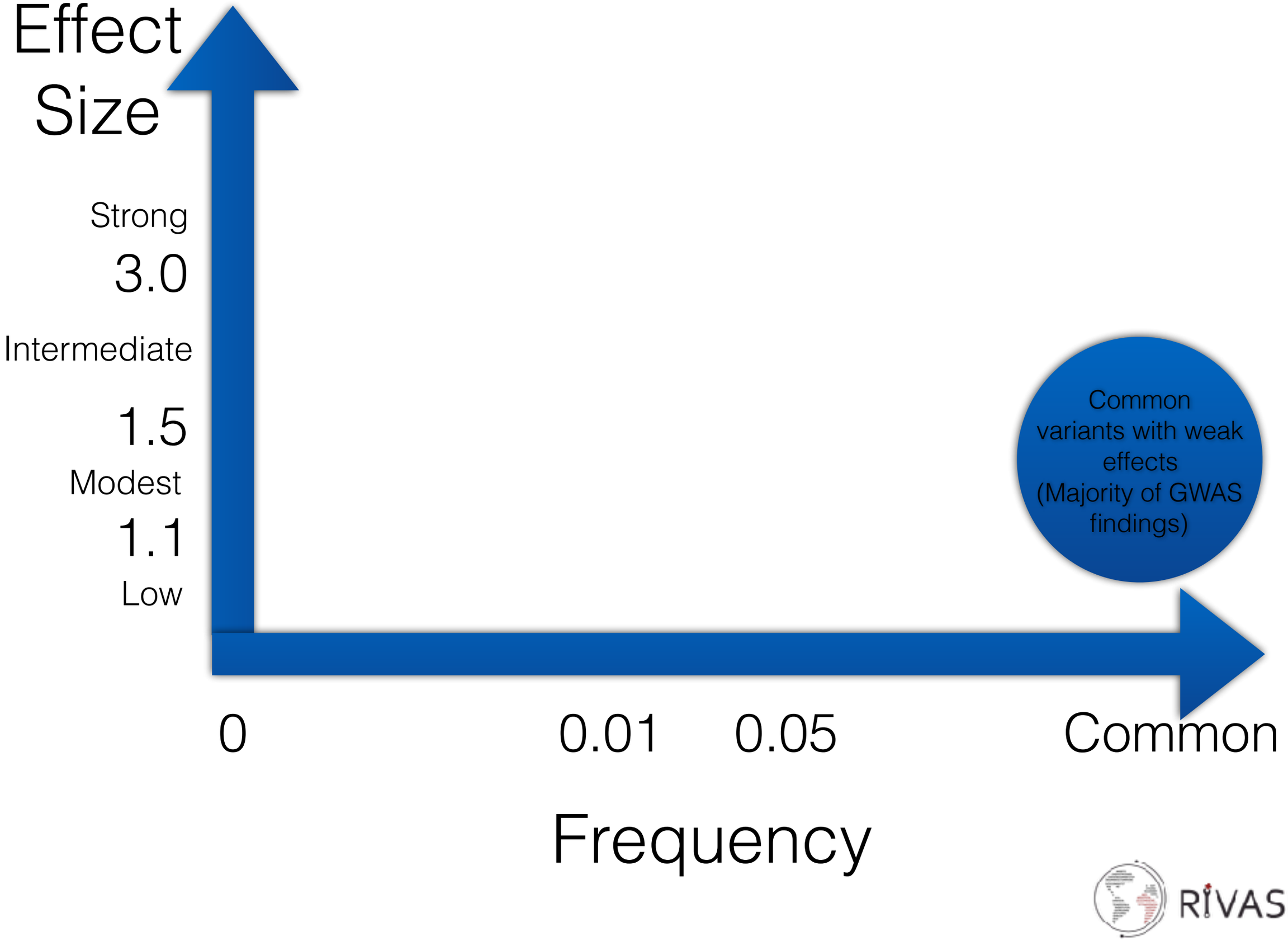
RIVASLAB



Genome-wide association studies have been very successful*

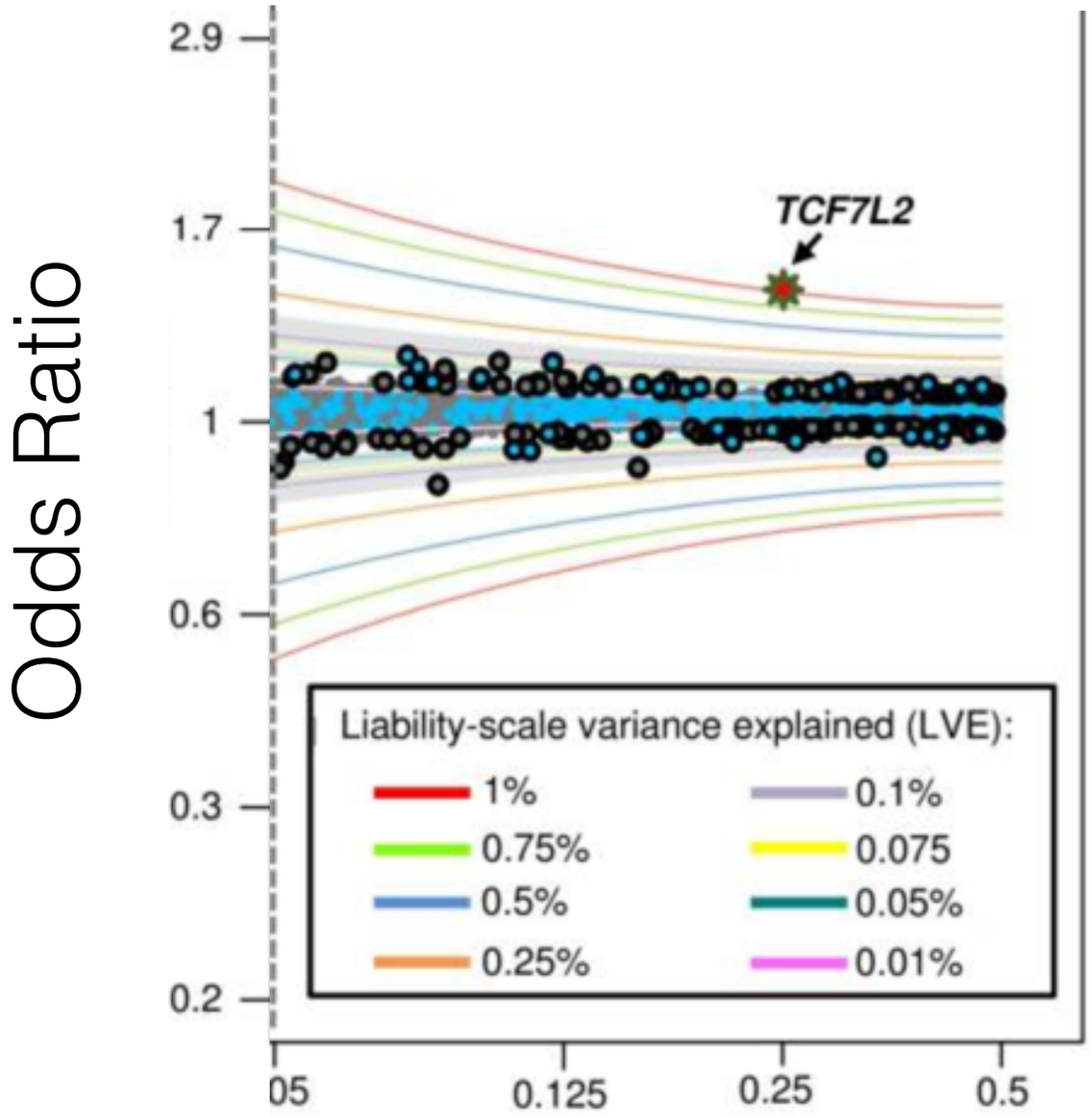
RIVASLAB

Why the “*”?


Effect Size
Strong
3.0
Intermediate
1.5
Modest
1.1
Low
0
0.01
0.05
Common
Frequency
Common variants with weak effects (Majority of GWAS findings)

RIVASLAB

In the context of **type 2 diabetes** all common variant associations were tiny

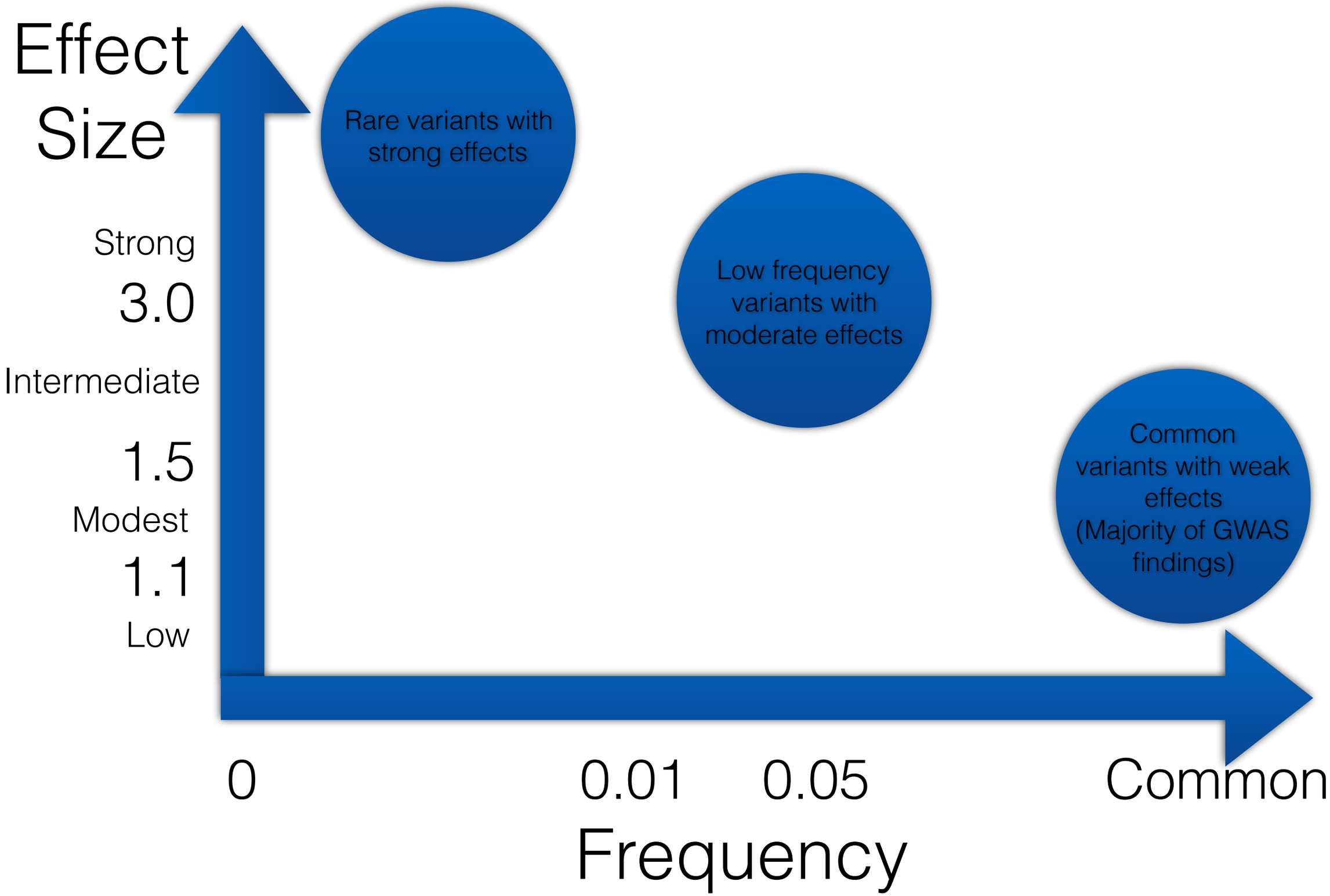


Tiny effect sizes for all associated variants

Fuchsberger et al. Nat 2016

Minor allele frequency

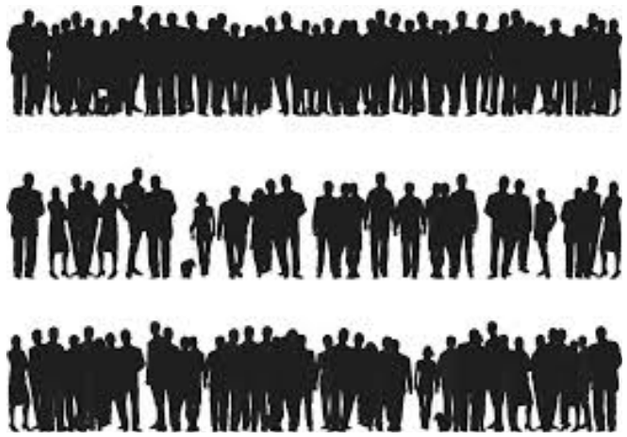
RIVASLAB

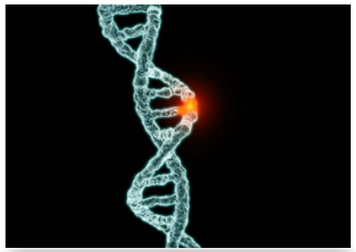


Additional signals started emerging from sequencing studies of rare variants

RIVASLAB

Experiments of Nature

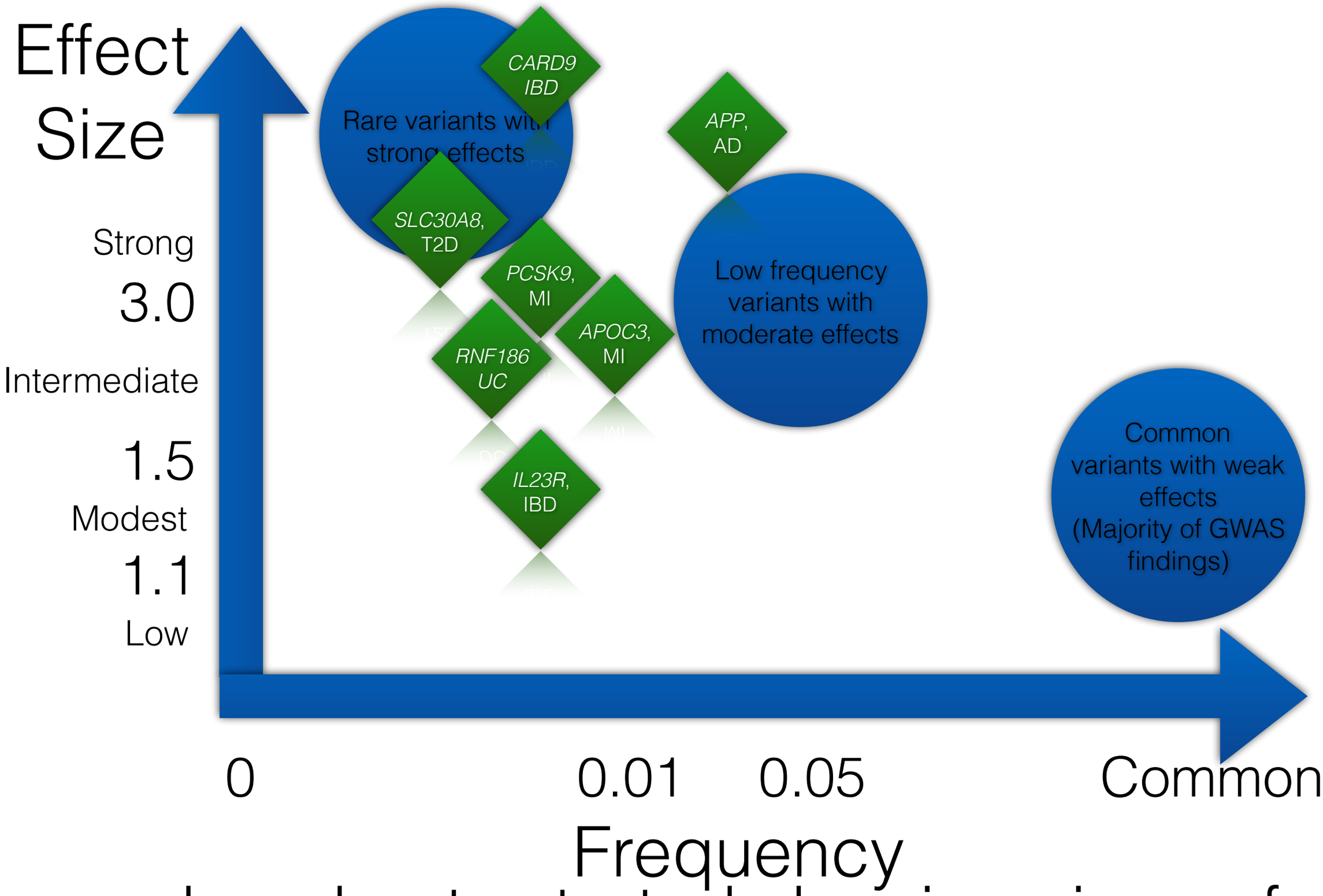




Lower risk
for
disease

Can Guide Selection of Targets

RIVASLAB



Human knockouts started showing signs of protection against disease

RIVASLAB

Population Biobank Studies

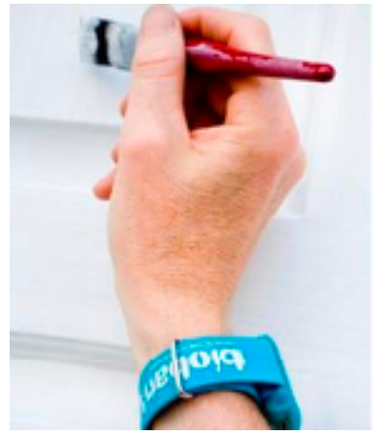
UK Biobank

RIVASLAB

UK Biobank

Health records
n=500,000

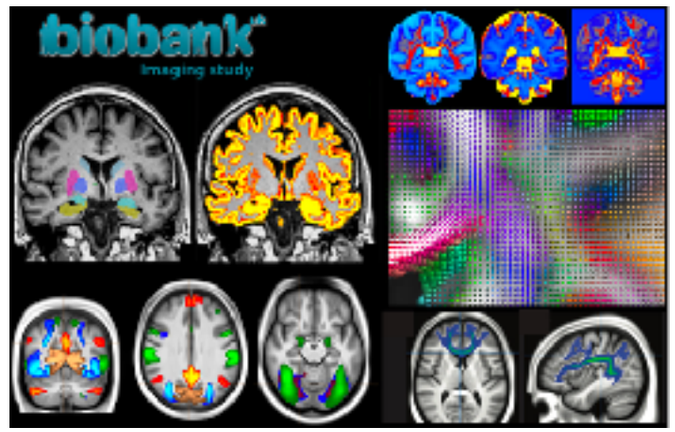
Web-based questionnaire
n=200,000



Physical activity monitor
n=100,000

Genotyping
n=500,000

Baseline Biochemistry
n=500,000
Q1 2018



Imaging
n=100,000
2015-2023

RIVASLAB

Massive Datasets

$q \sim 10^4$

$p \sim 10^7$

$N \sim 10^6$

$N \sim 10^6$

RIVASLAB

Going from ‘BIG DATA’ to ‘SMALL DATA’

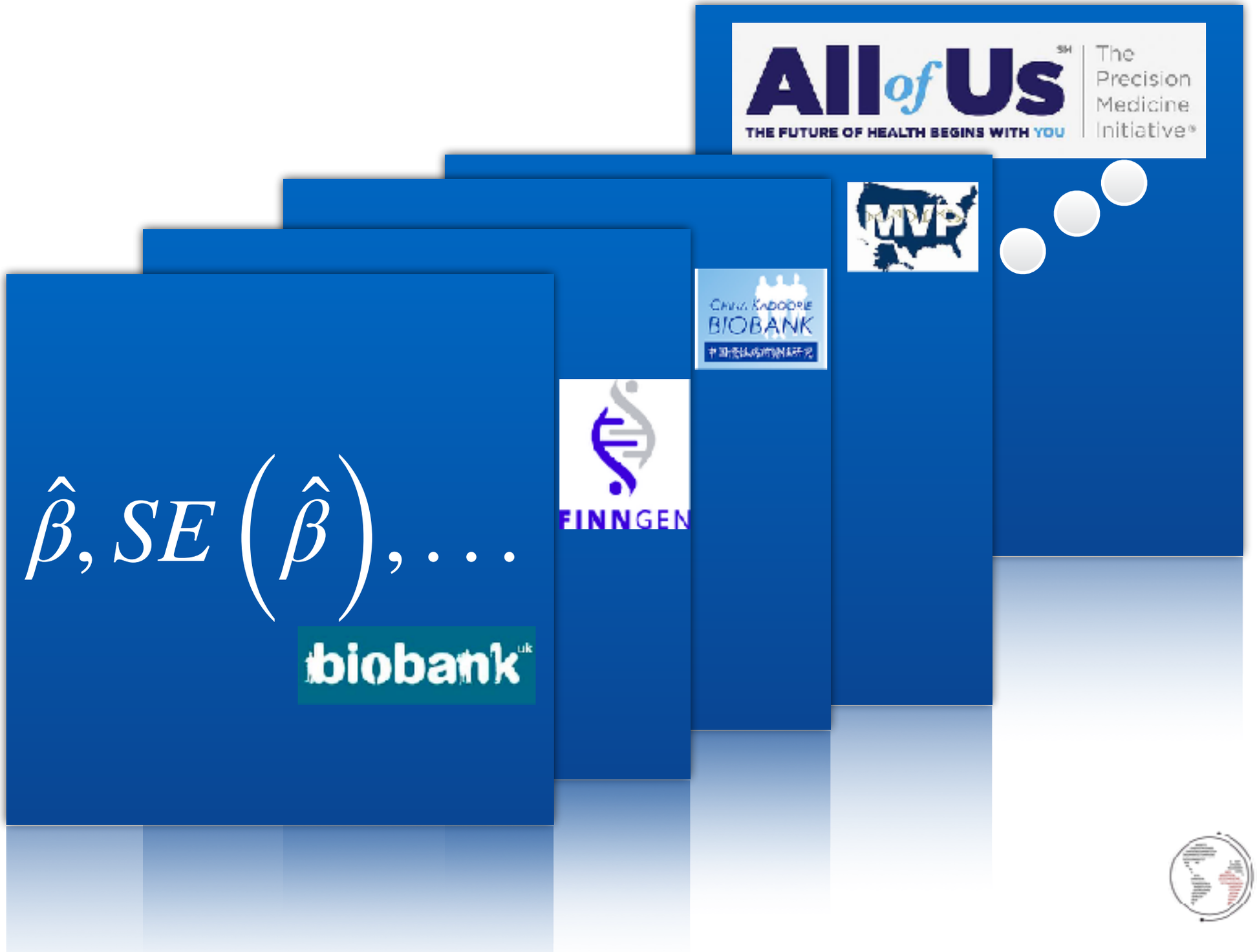
K ~ 10^4

M ~ 10^7

$$\hat{\beta}, SE\left(\hat{\beta}\right), \ldots$$

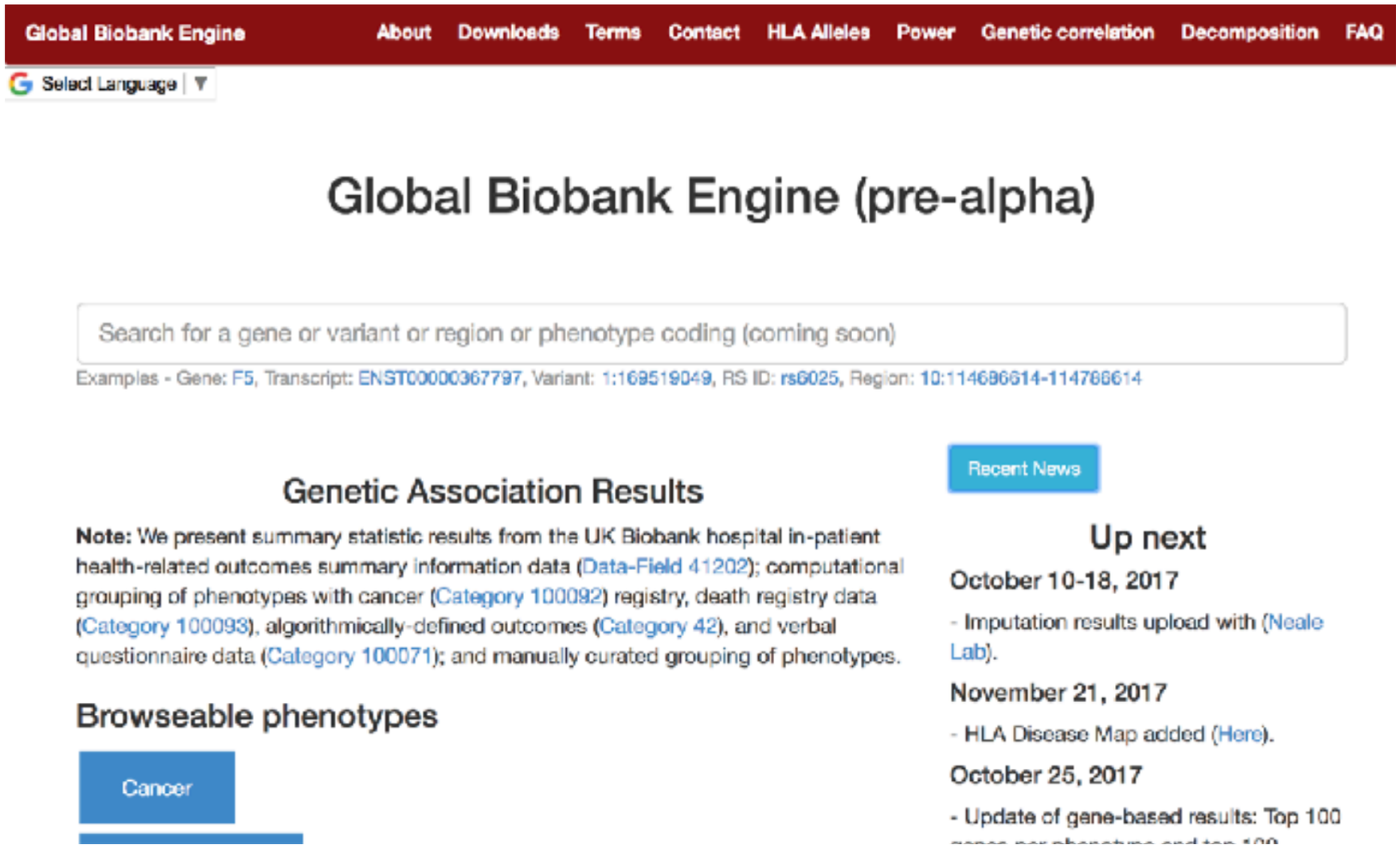
RIVASLAB

Improving inference



RIVASLAB

Global Biobank Engine



https://biobankengine.stanford.edu

RIVASLAB

Bayesian model comparison for rare variant association studies of multiple phenotypes

RIVASLAB

Introduction

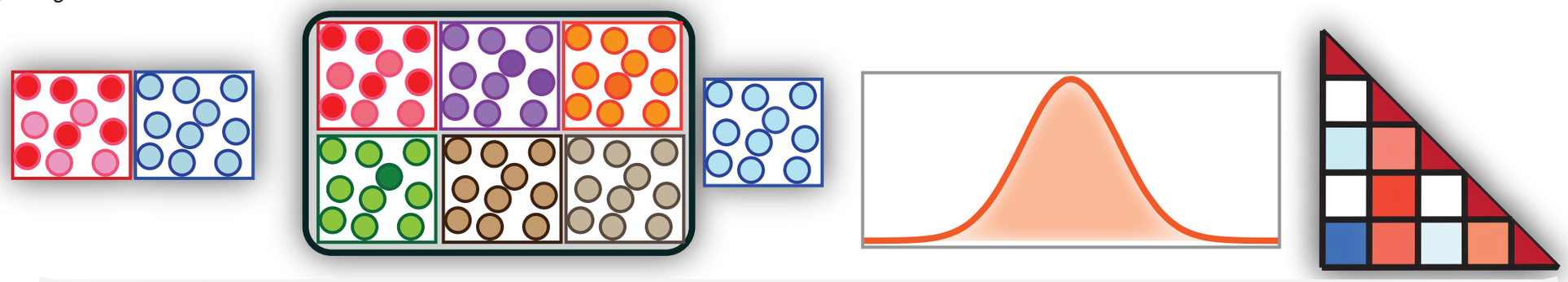
- How can we detect associations for rare variants (MAF < 1%)?
- Common approach: aggregate rare variants by gene or other genomic “unit”
- Can we incorporate multiple genetically-related phenotypes to improve power?
- **M**ultiple **R**are variants and **P**henotypes (**MRP**) Bayesian model comparison:
    - Compare genetic effects for rare variants in a “unit” (e.g. gene) using null model of no effects or alternative model that captures correlation, scale, and location of genetic effects

RIVASLAB

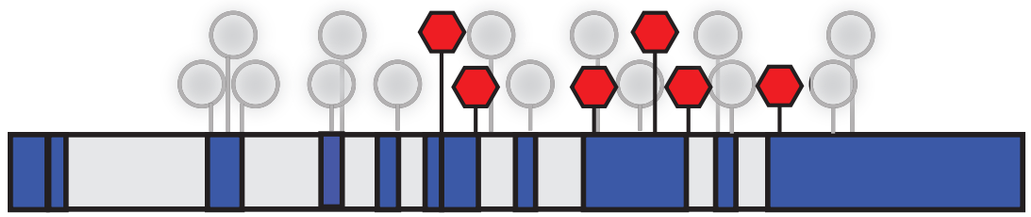
Study designs for multiple phenotypes

- MRP can be applied to different study designs:
  - Case-control
  - Multiple diseases and shared controls
  - Quantitative phenotypes
  - Mixture of case-control and quantitative phenotypes

Study design



Rare variant genetic analysis



RIVASLAB

Bayesian model comparison

Defined as a Bayes factor between the alternative model and the null model.

Bayesian model comparison

Defined as a Bayes factor between the alternative model and the null model.

Where the null model is simply

$$\beta = \mathbf{0}$$

Bayesian model comparison

The Bayes Factor (BF) is obtained as a ratio of the marginal likelihoods of the data for the two models:

$$\mathrm{BF} = \frac{\int_{\boldsymbol{\beta}_1} p\left(\mathrm{Data}|\boldsymbol{\beta}_1\right) p\left(\boldsymbol{\beta}_1\right) d\boldsymbol{\beta}_1}{\int_{\boldsymbol{\beta}_0} p\left(\mathrm{Data}|\boldsymbol{\beta}_0\right) p\left(\boldsymbol{\beta}_0\right) d\boldsymbol{\beta}_0}$$

Bayesian model comparison

The Bayes Factor (BF) is obtained as a ratio of the marginal likelihoods of the data for the two models:

$$\mathrm{BF} = \frac{\int_{\boldsymbol{\beta}_1} p\left(\mathrm{Data}|\boldsymbol{\beta}_1\right) p\left(\boldsymbol{\beta}_1\right) d\boldsymbol{\beta}_1}{\int_{\boldsymbol{\beta}_0} p\left(\mathrm{Data}|\boldsymbol{\beta}_0\right) p\left(\boldsymbol{\beta}_0\right) d\boldsymbol{\beta}_0}$$

Data can correspond to the effect size estimates and the estimated variance-covariance matrix:

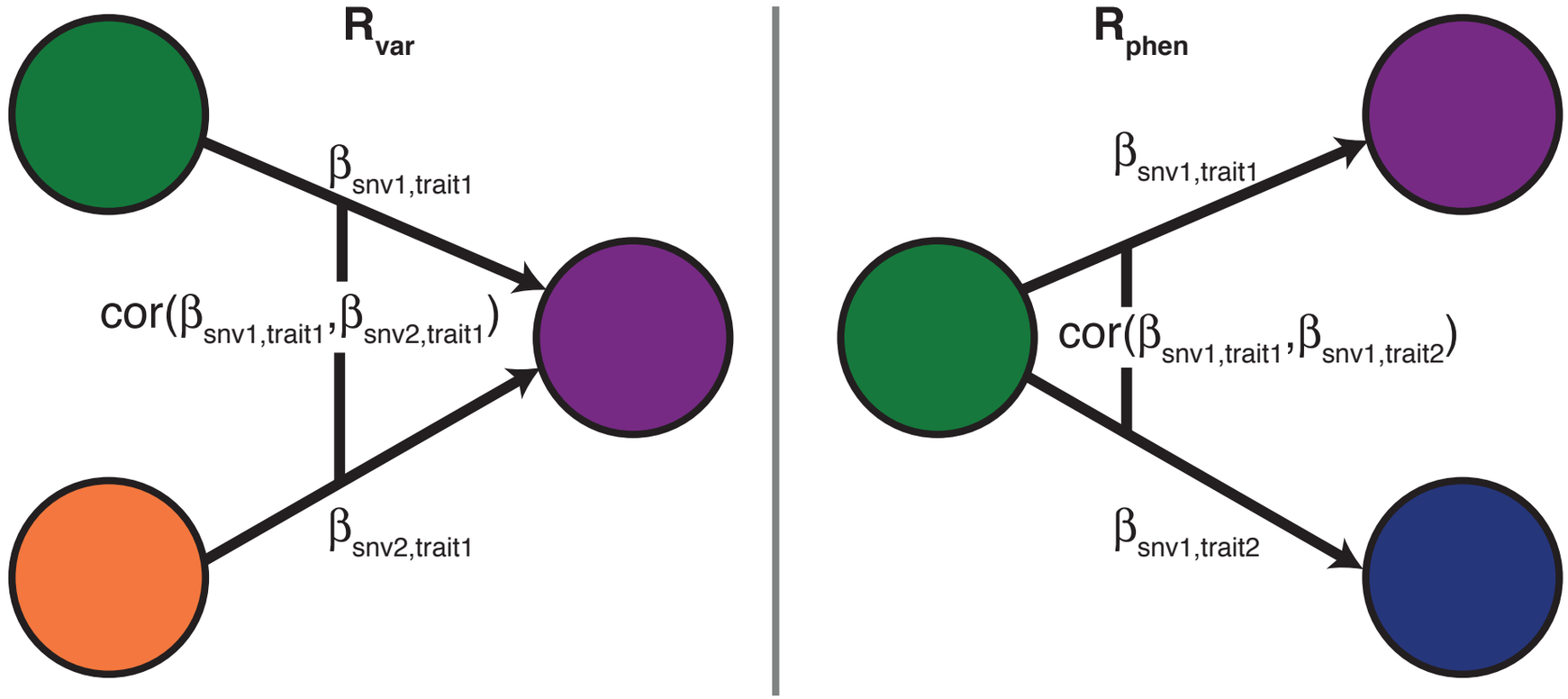
$$\widehat{\beta}, \widehat{\mathbf{V}_B}$$

Bayesian model comparison

In MRP we extend it so that the effect size estimates is across all variants analyzed, across all studies, and all phenotypes. We specify the prior distribution for the alternative model in three steps:

1. Single study
2. Multiple studies
3. Prior mean of genetic effects

Bayesian model comparison: Single study



The prior density for β incorporates the expected correlation of genetic effects among a group of variants (Rvar) and among a group of phenotypes (Rphen).

Bayesian model comparison: Multiple studies

We introduce the matrix

$$\mathbf{R}_{\text{study}}$$

to specify prior on the similarity in effect size across studies

Bayesian model comparison: Multiple studies

We introduce the matrix

$$\mathbf{R}_{\text{study}}$$

to specify prior on the similarity in effect size across studies

$$\mathbf{U} = \mathbf{R}_{\text{study}} \otimes (\mathbf{S}_{\text{var}} \otimes \mathbf{R}_{\text{phen}})$$

and the prior is

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{U})$$

Bayesian model comparison: Multiple studies

Attractive property of Bayesian model comparison is ability to obtain a measure of relative support across a number of models

Additional use case example

- Quality control

Heterogeneity of effects across studies may be representative of QC/technical artifacts!

Bayesian model comparison: prior mean of genetic effects

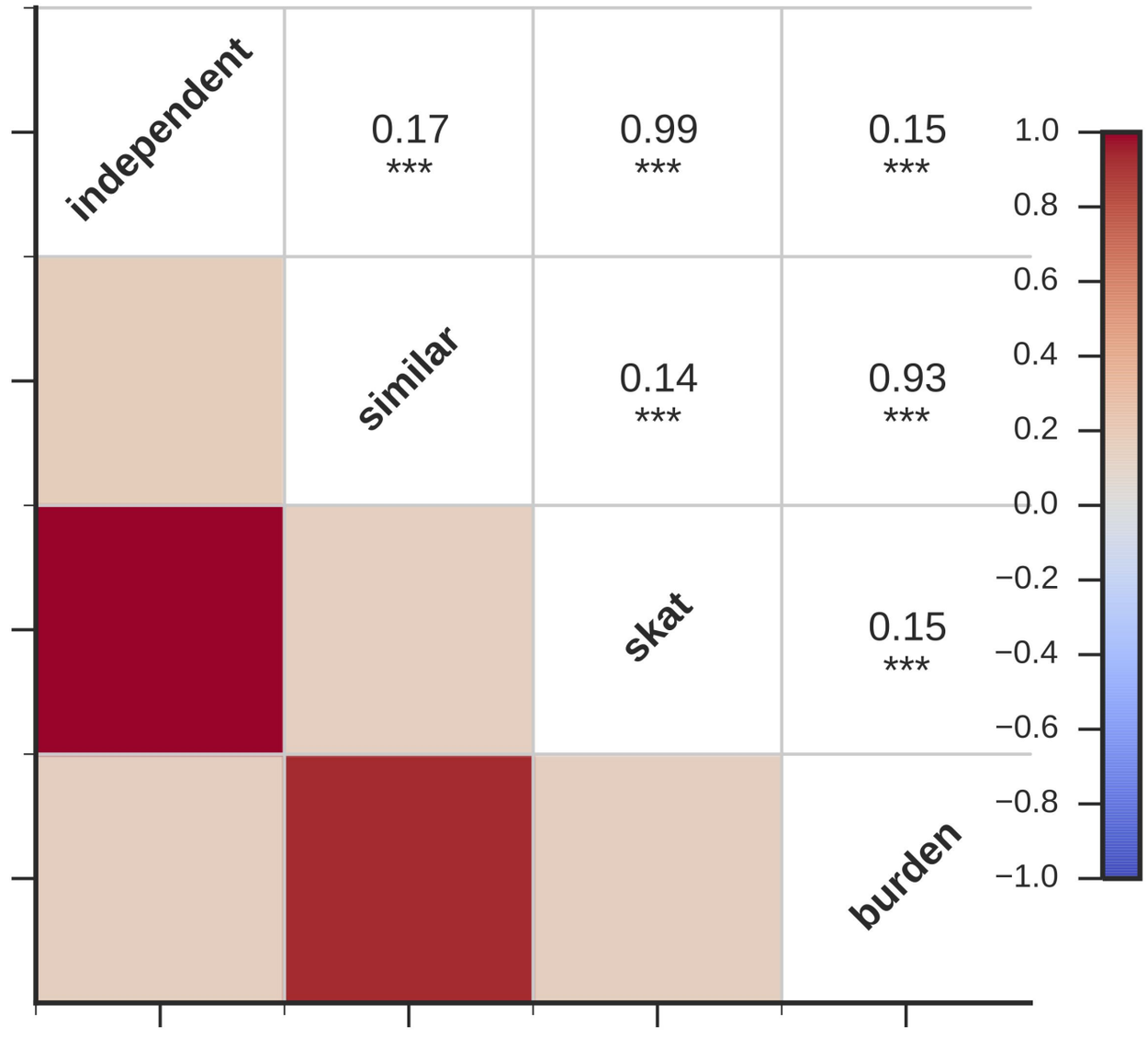
It is straightforward to incorporate prior mean of genetic effects

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{U}\right)$$

Useful for prioritizing protective modifiers of disease risk

Comparison to other methods

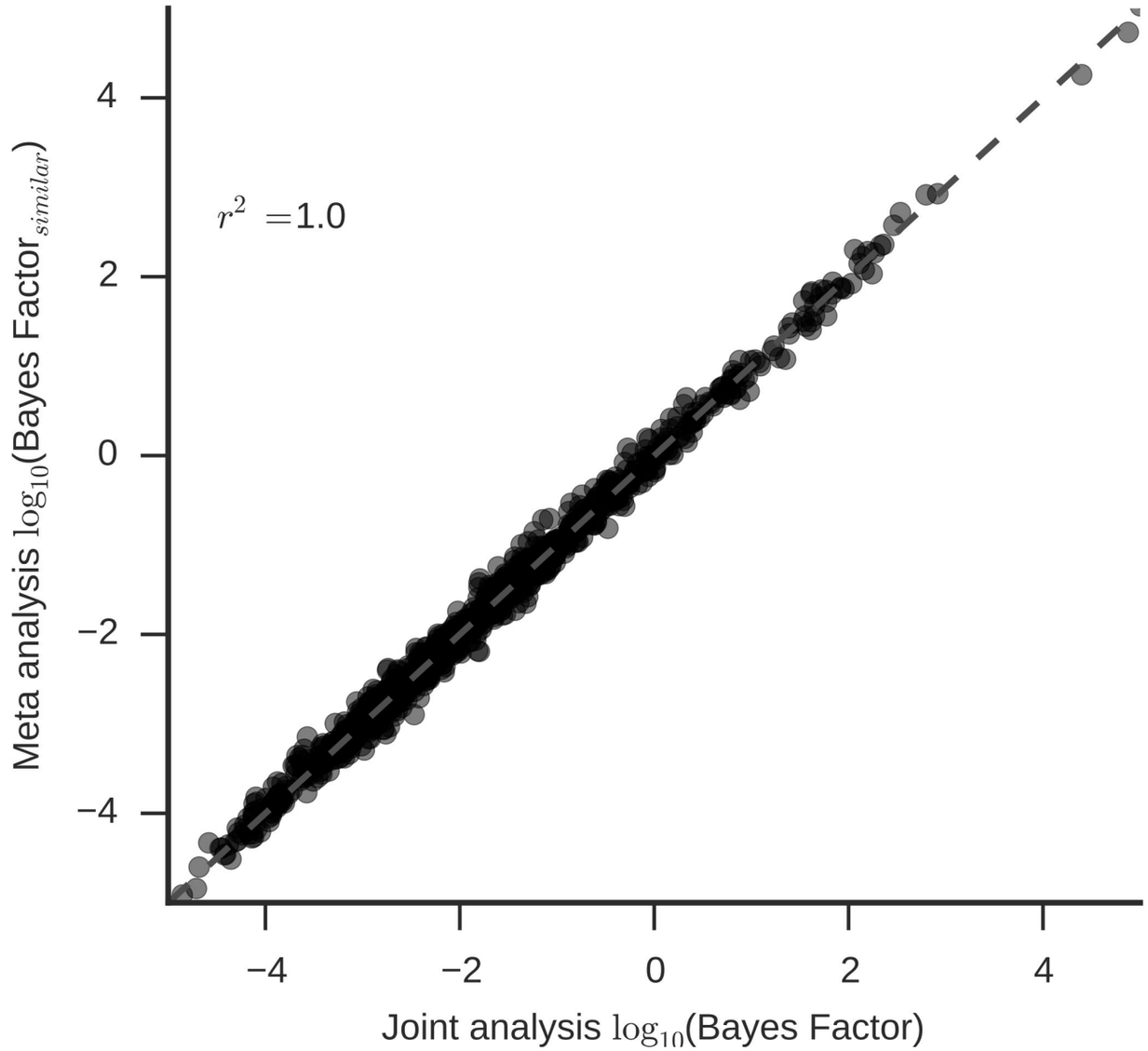
- MRP independent effects: different effects for variants in group
- MRP similar effects: similar effects for variants in group
- MRP independent or similar effects model recover SKAT or burden results



RIVASLAB

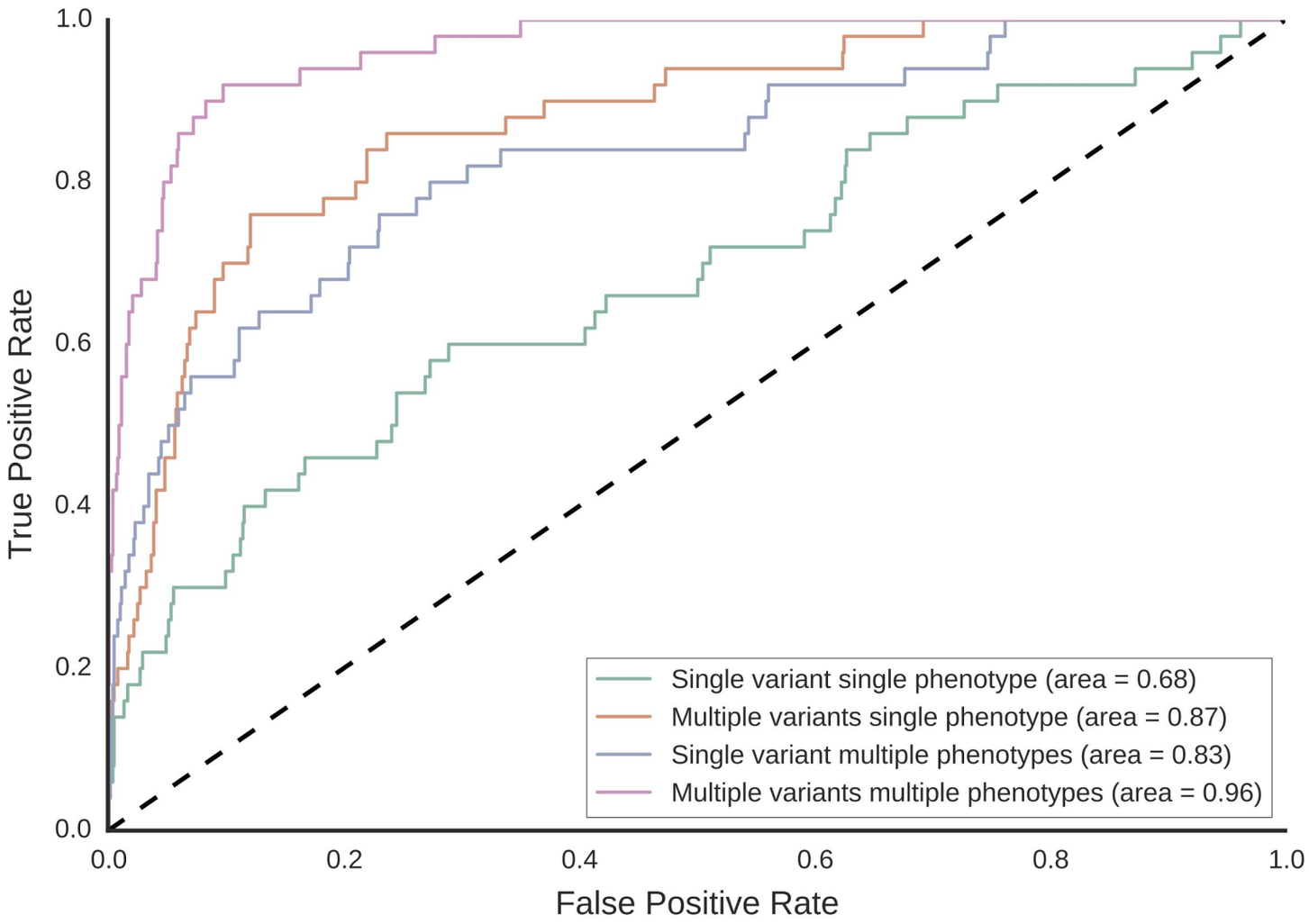
MRP can use GWAS summary statistics

- Results from MRP using raw genotype-phenotype data (x-axis) or GWAS summary statistics (y-axis) are highly correlated



RIVASLAB

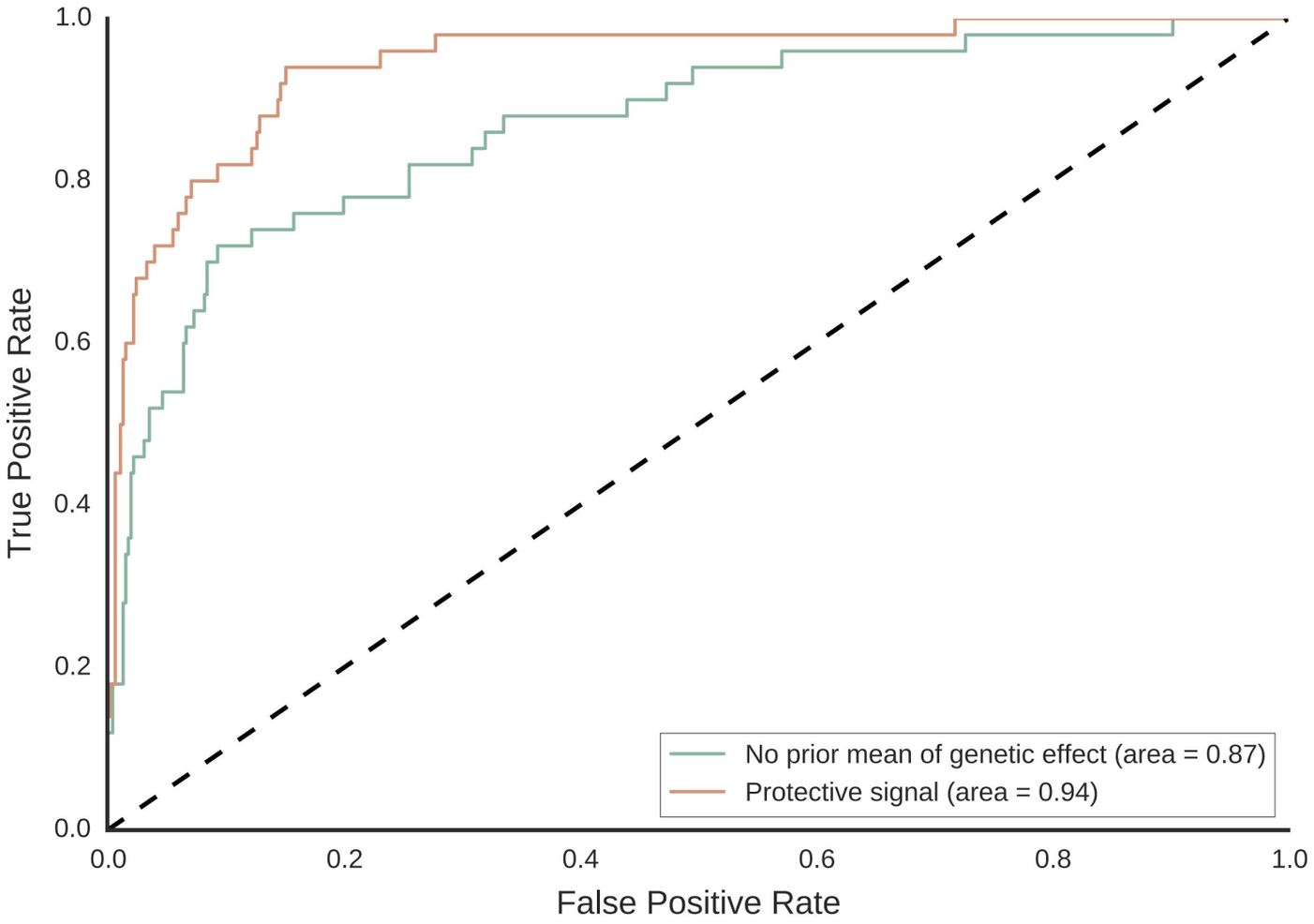
Univariate vs. multivariate analyses



- Using multiple variants and multiple phenotypes (purple line) improves power to detect associations

RIVASLAB

Protective associations

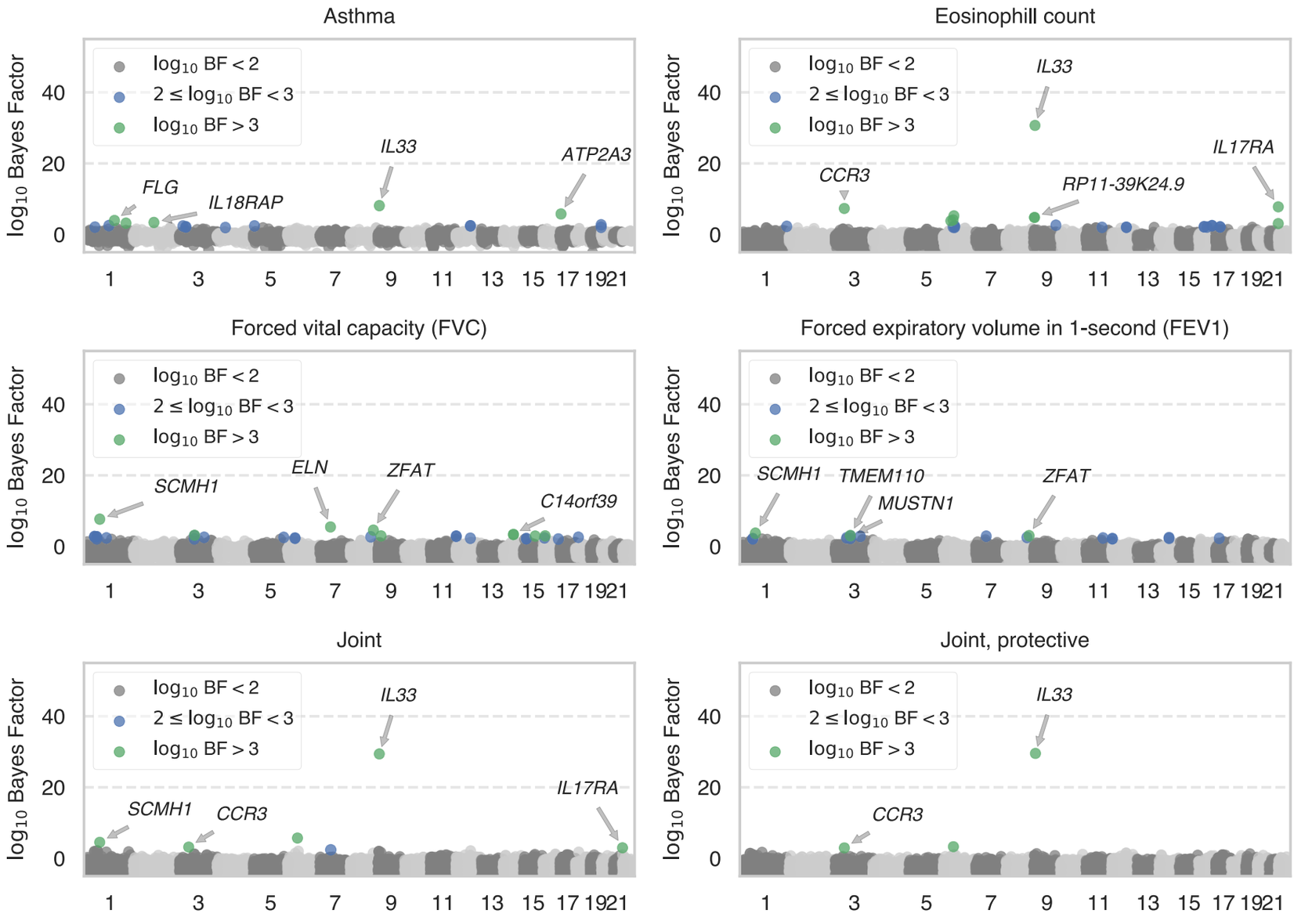


- Adding a prior mean for genetic effects can boost power to identify protective associations

RIVASLAB

Asthma application

- Applied MRP to asthma and related phenotypes from UK Biobank
- Identified association for rare variants in *IL33* that protect against asthma
- Also observed moderate evidence for protective associations for drug target *CCR3*

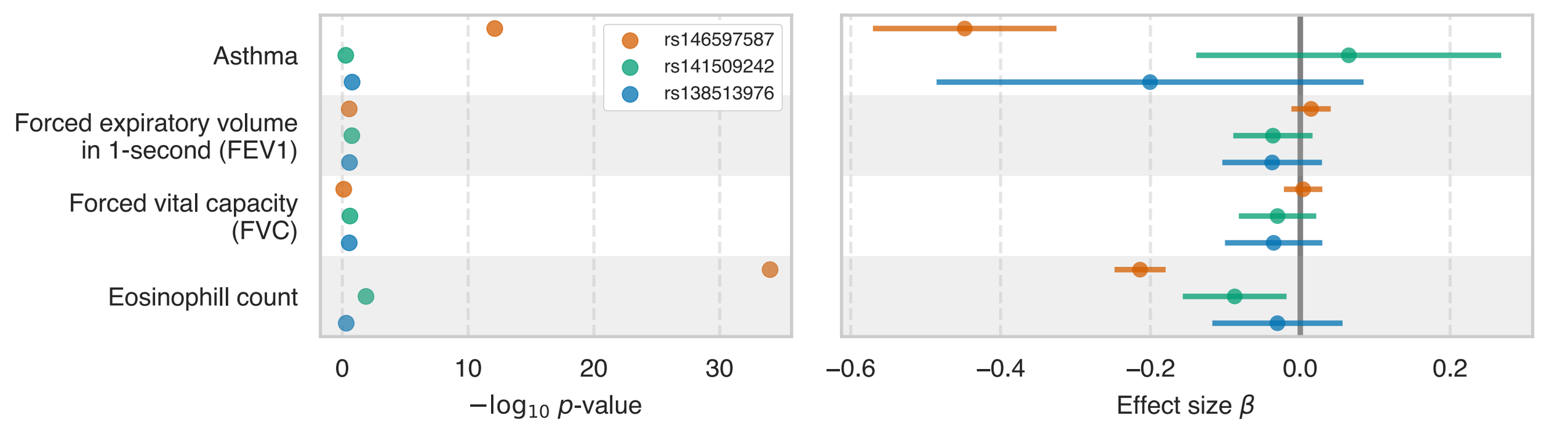


RIVASLAB

Asthma application

- Protective association in *IL33* driven by PTV rs146597587 (Smith *et al.* 2017, DeBoever *et al.* 2017)

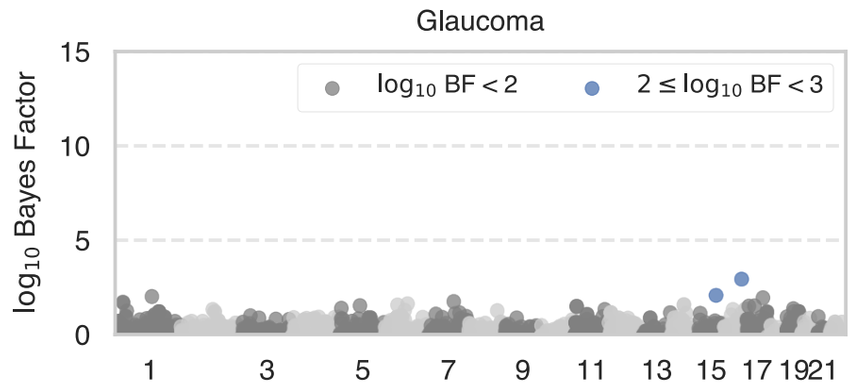
- PTV also has strong effect on eosinophil counts

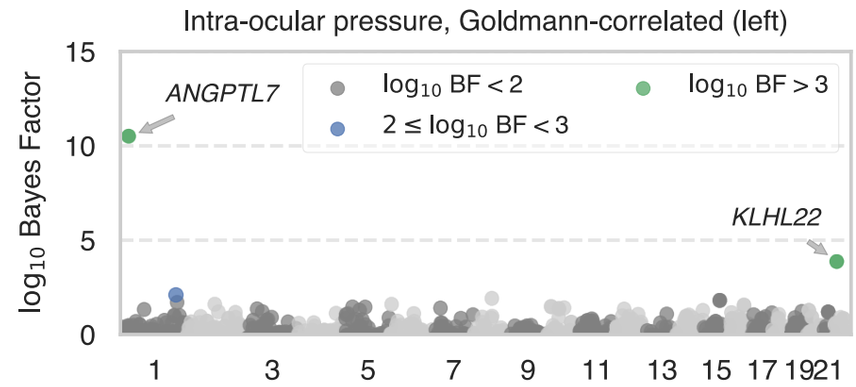


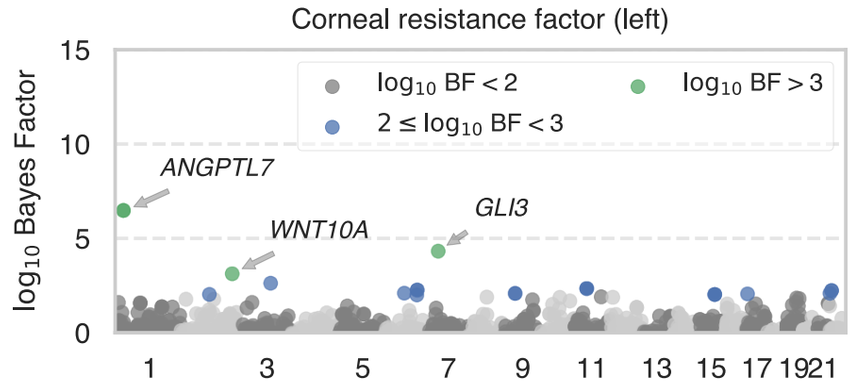
RIVASLAB

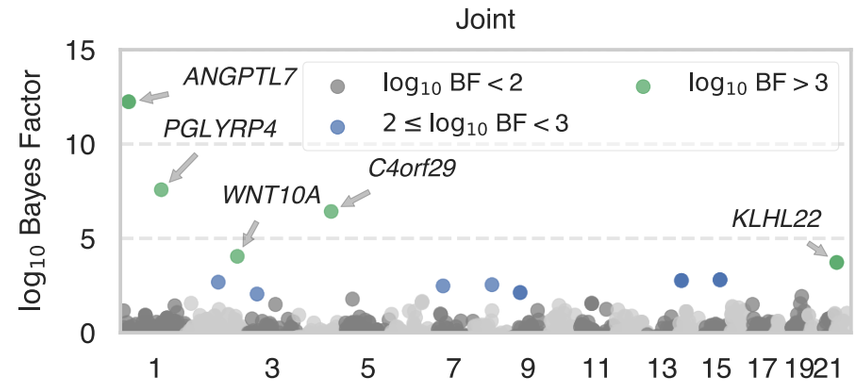
Glaucoma application

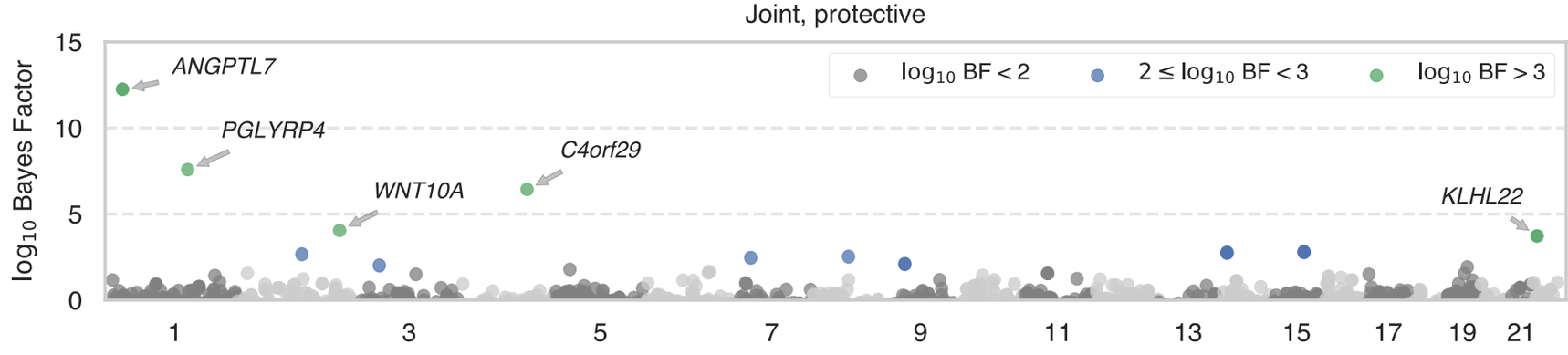
- Applied MRP to glaucoma and two related traits from the UK Biobank
- Identified protective association for *ANGPTL7*
- *ANGPTL7* is unregulated in glaucoma and the protein expression is unique to the vitreous humor of the eye











RIVASLAB

Conclusions

- MRP
  - Can use raw genotype/phenotype data or GWAS summary statistics
  - Uses information from multiple variants and phenotypes to boost power to detect rare variant associations
  - Can prioritize associations that protect against disease

RIVASLAB

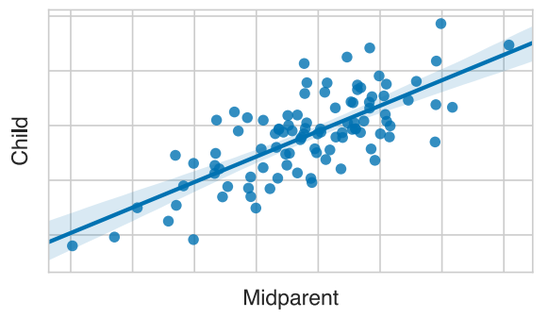
Multivariate Polygenic Mixture Models

RIVASLAB

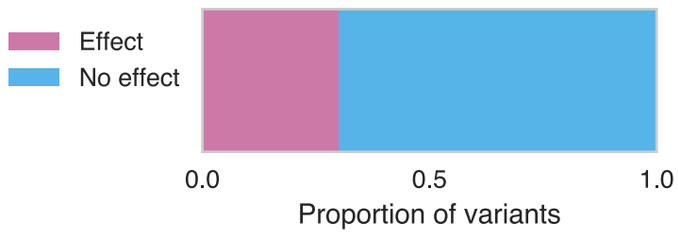
Motivation

- Most common diseases are polygenic
- Large number of phenotypes and loci tested by GWAS
- Opportunity to estimate genetic parameters

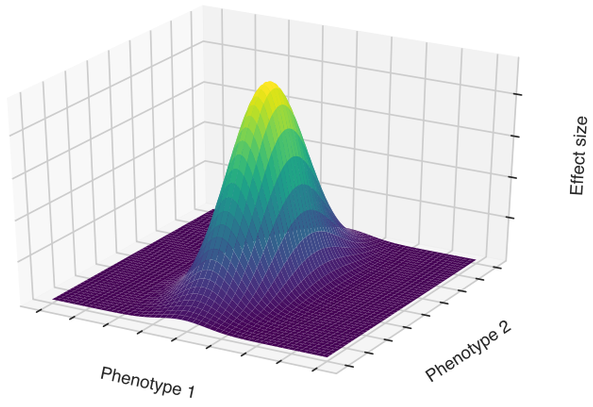
- h^2 : heritability



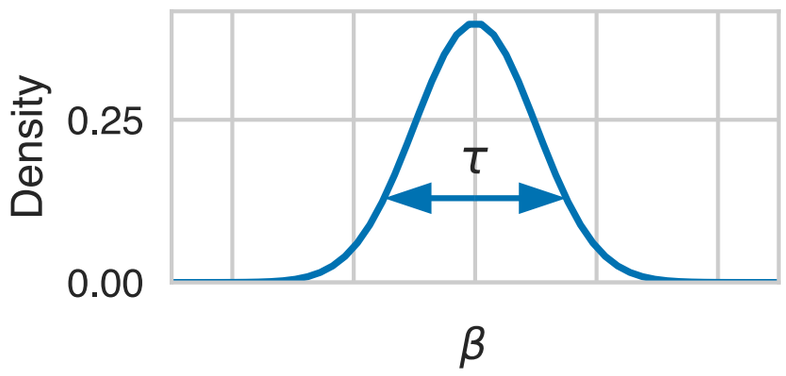
- π : membership



- Ω : genetic correlation



- τ : spread/scale of effects



RIVASLAB

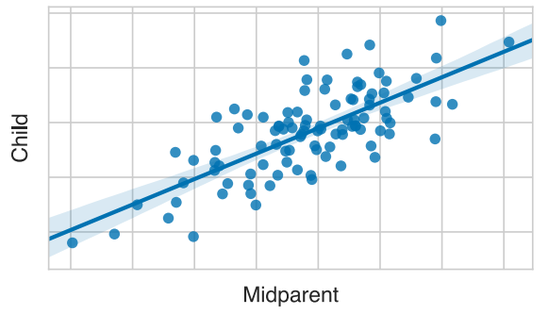
Motivation

- Most common diseases are polygenic
- Large number of phenotypes and loci tested by GWAS
- Opportunity to estimate genetic parameters

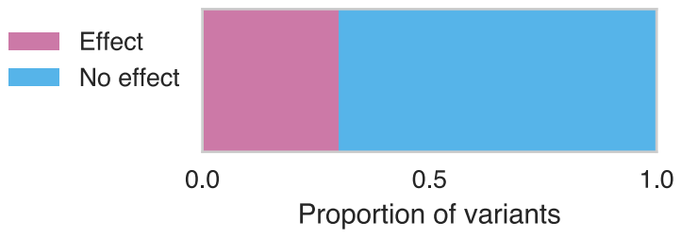
Phenotype 1		Phenotype 2	
BETA	**SE**	**BETA**	**SE**
1.2	0.03	1.15	0.04
1.5	0.9	1.3	0.1
-1.1	0.006	-1.2	0.05
…	…	…	…

MVPMM

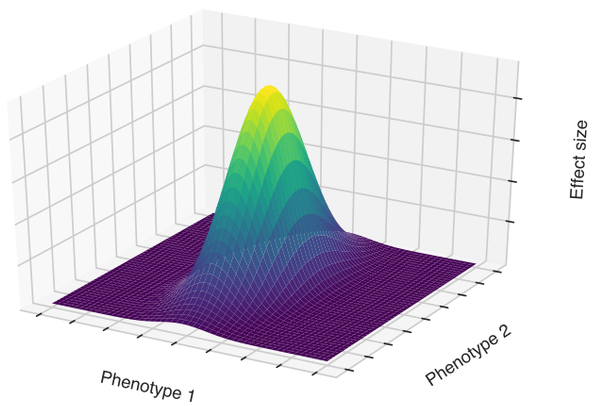
- h^2 : heritability



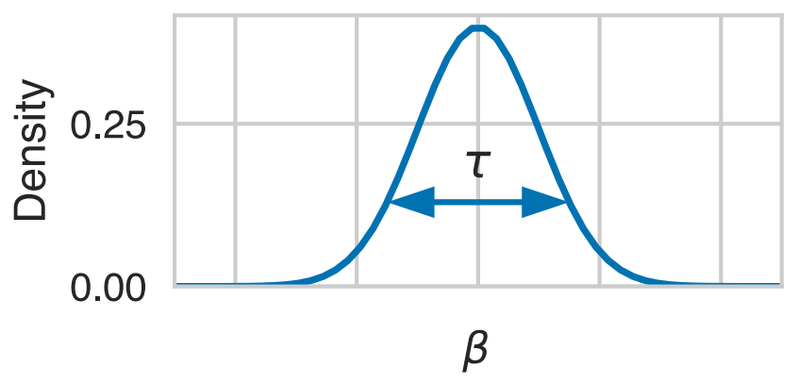
- π : membership



- Ω : genetic correlation



- τ : spread/scale of effects



RIVASLAB

Estimating genetic parameters

- Model GWAS summary statistics as generated from one of two components
  - Null component with correlated errors
  - Non-null component with correlated genetics effects and errors

$\hat{\beta}_i$: regression effect size for locus i

$\hat{\sigma}_i$: regression SE for locus i

π : mixture proportion

Null component

$$\hat{\beta}_i \sim \mathrm{MVN}\left(0, \Sigma_{\Theta i}\right)$$

Correlated errors

$$\Sigma_{\Theta i} = \mathrm{diag}\left(\hat{\sigma}_i\right) \cdot \Theta \cdot \mathrm{diag}\left(\hat{\sigma}_i\right)$$

Non-null component

$$\hat{\beta}_i \sim \mathrm{MVN}\left(0, \Sigma_{\Theta i} + \Sigma_{\Omega}\right)$$

Correlated genetic effects

$$\Sigma_{\Omega} = \mathrm{diag}\left(\tau\right) \cdot \Omega \cdot \mathrm{diag}\left(\tau\right)$$

RIVASLAB

Can we use genetic parameters to evaluate different phenotyping methods for population biobanks?

What are the estimates of genetic parameters across diseases?

UK Biobank

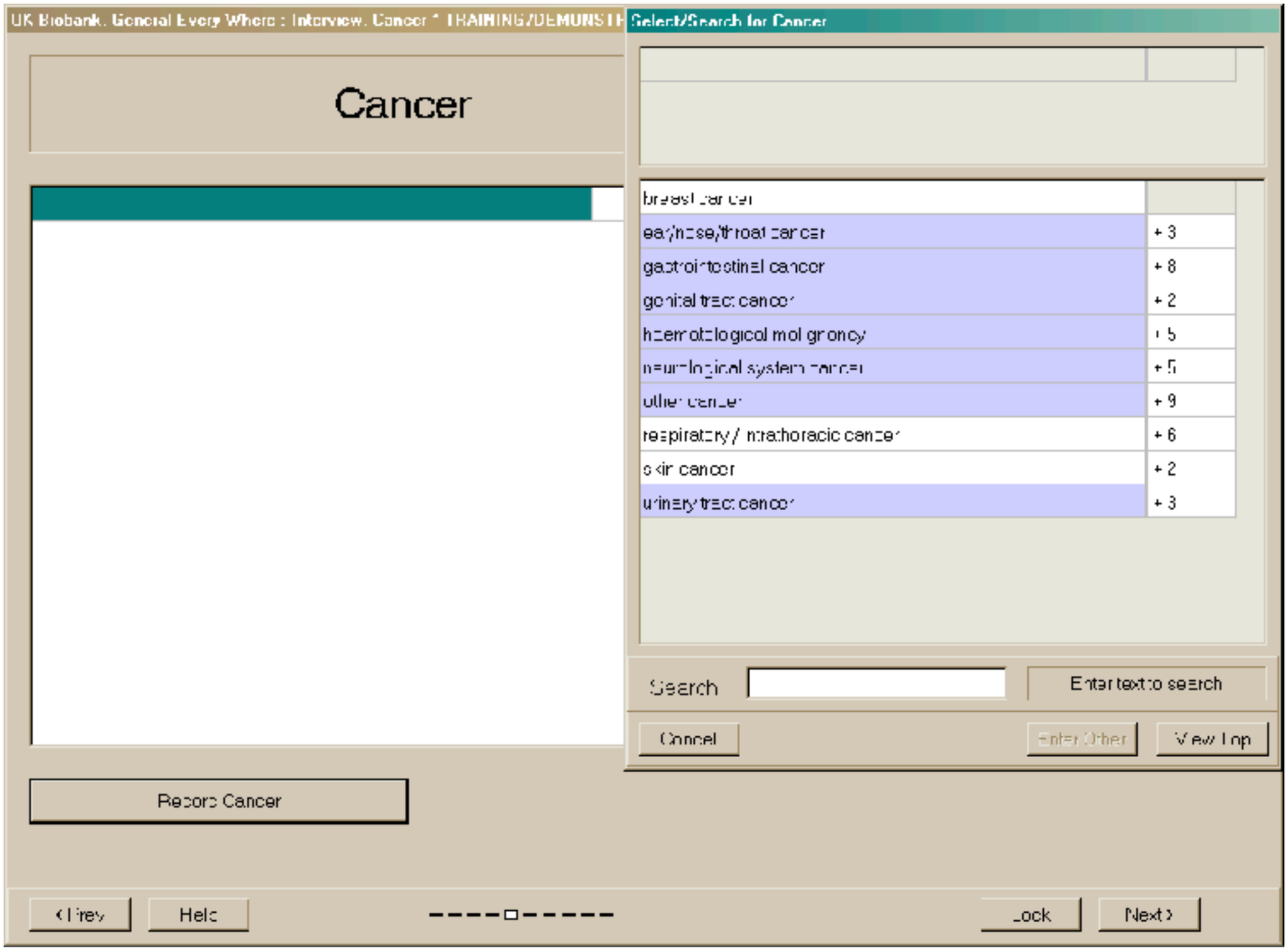
- Diverse data collected on ~500,000 participants
- Array genotyping for ~800,000 variants
- Hospital in-patient records, cancer registry, death registry

J45 Asthma	-
J45.0 Predominantly allergic asthma	79
J45.1 Nonallergic asthma	1
J45.8 Mixed asthma	2
J45.9 Asthma, unspecified	2659
J46 Status asthmaticus	300
J47 Bronchiectasis	631

http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=41202

RIVASLAB

- Verbal questionnaire data



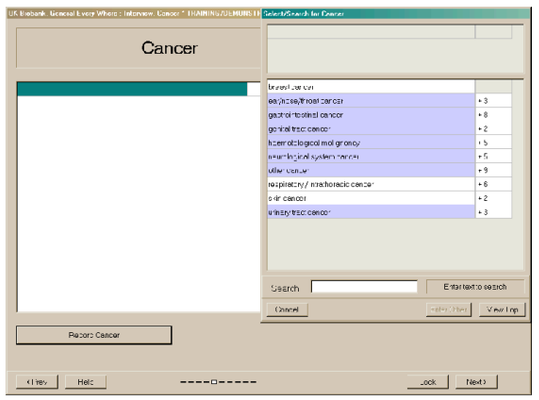
http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100074

RIVASLAB

How can we use genetic parameters to evaluate different phenotyping methods for population biobanks?

J45 Asthma	-
J45.0 Predominantly allergic asthma	79
J45.1 Nonallergic asthma	1
J45.8 Mixed asthma	2
J45.9 Asthma, unspecified	2659
J46 Status asthmaticus	300
J47 Bronchiectasis	631

VS.

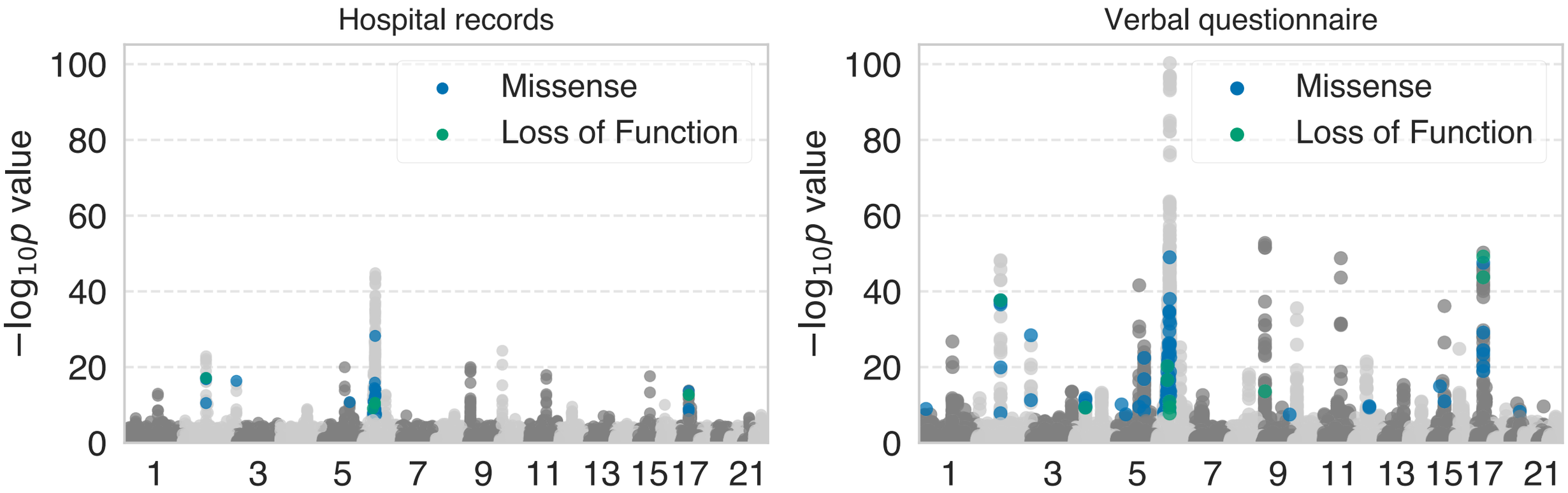


What are the estimates of genetic parameters across diseases?

RIVASLAB

Verbal questionnaire vs. hospital records

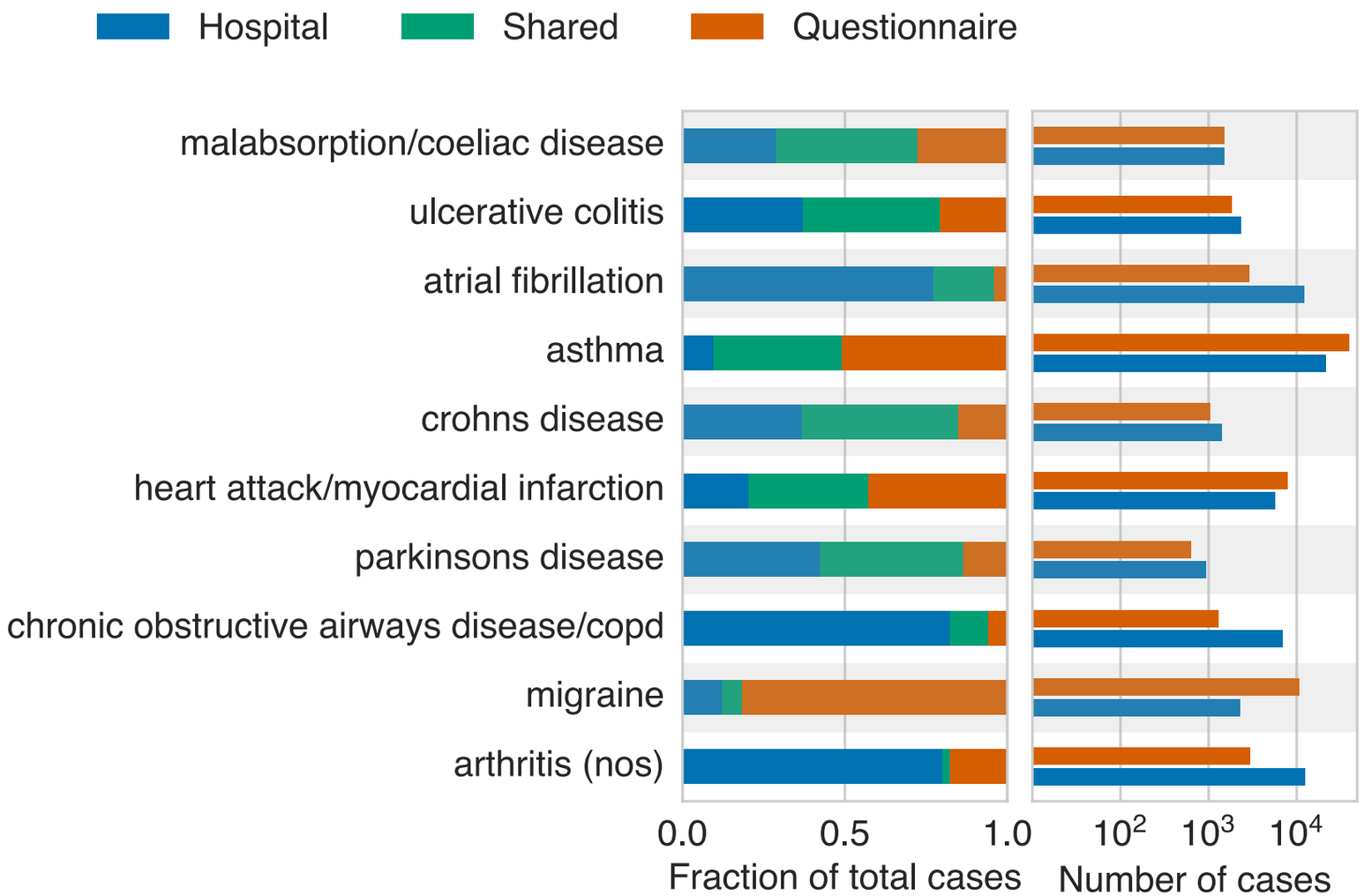
- Asthma cases defined according to either hospital records ***or*** questionnaire data



RIVASLAB

Verbal questionnaire vs. hospital records

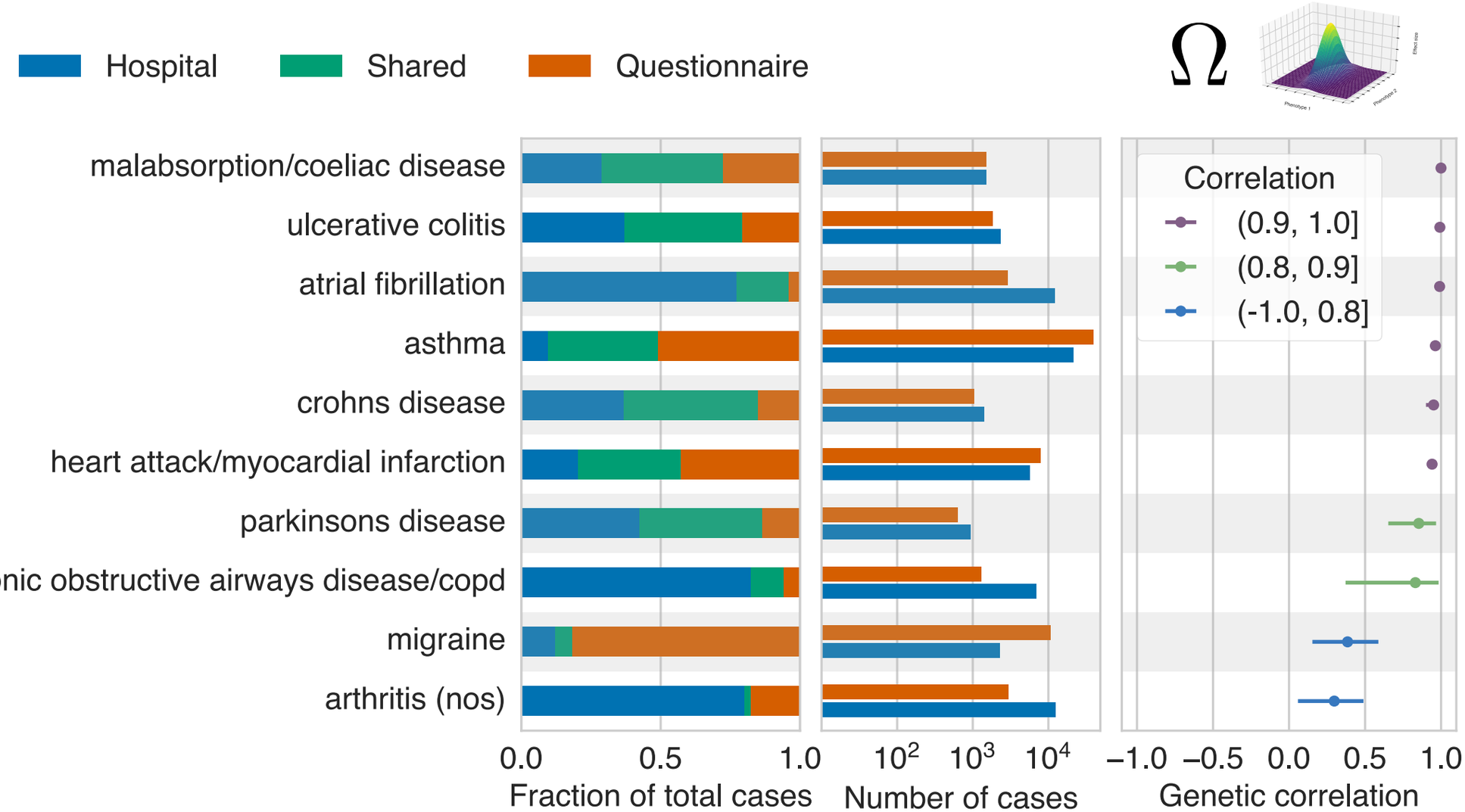
- Genetic correlation for phenotypes defined using either hospital and registry records ***or*** questionnaire data



RIVASLAB

Verbal questionnaire vs. hospital records

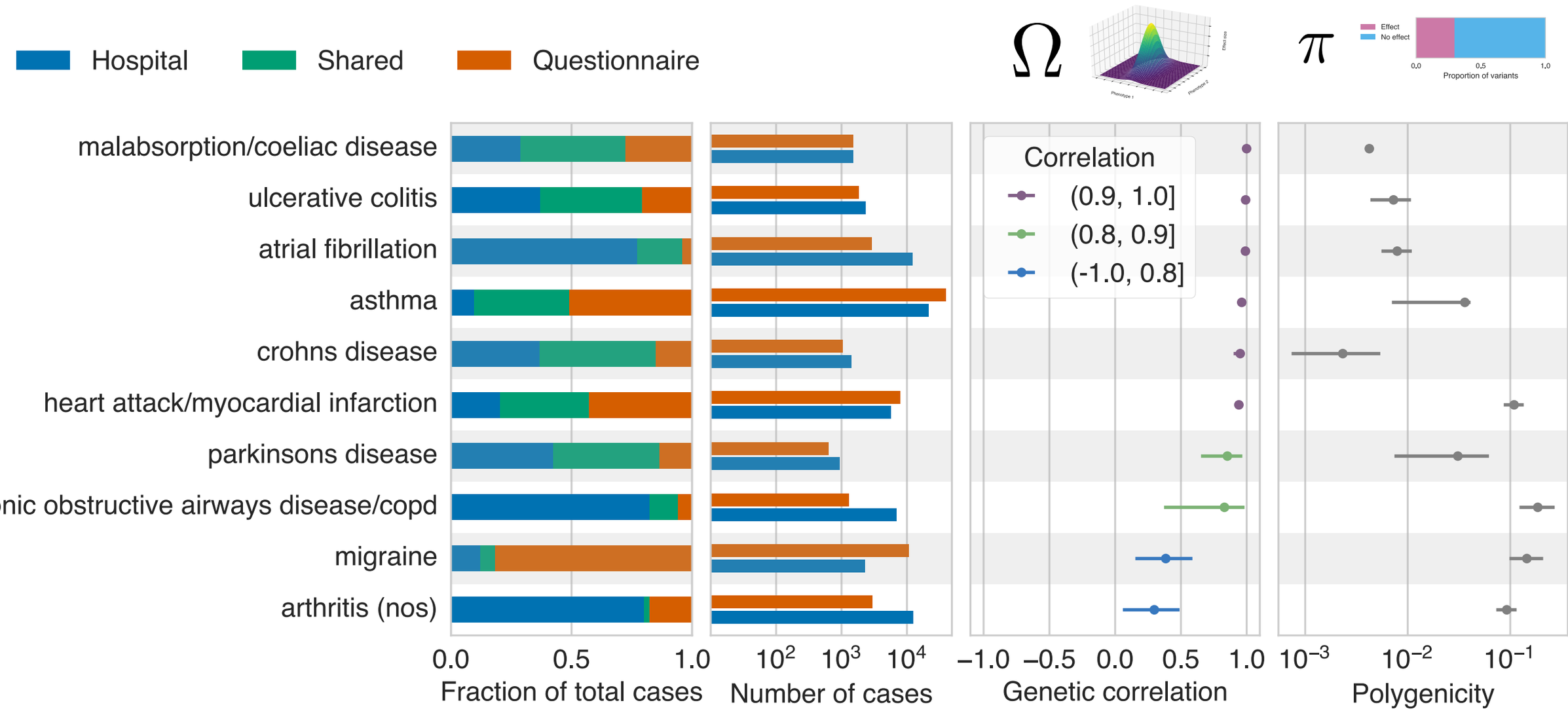
- Genetic correlation for phenotypes defined using either hospital and registry records ***or*** questionnaire data



RIVASLAB

Verbal questionnaire vs. hospital records

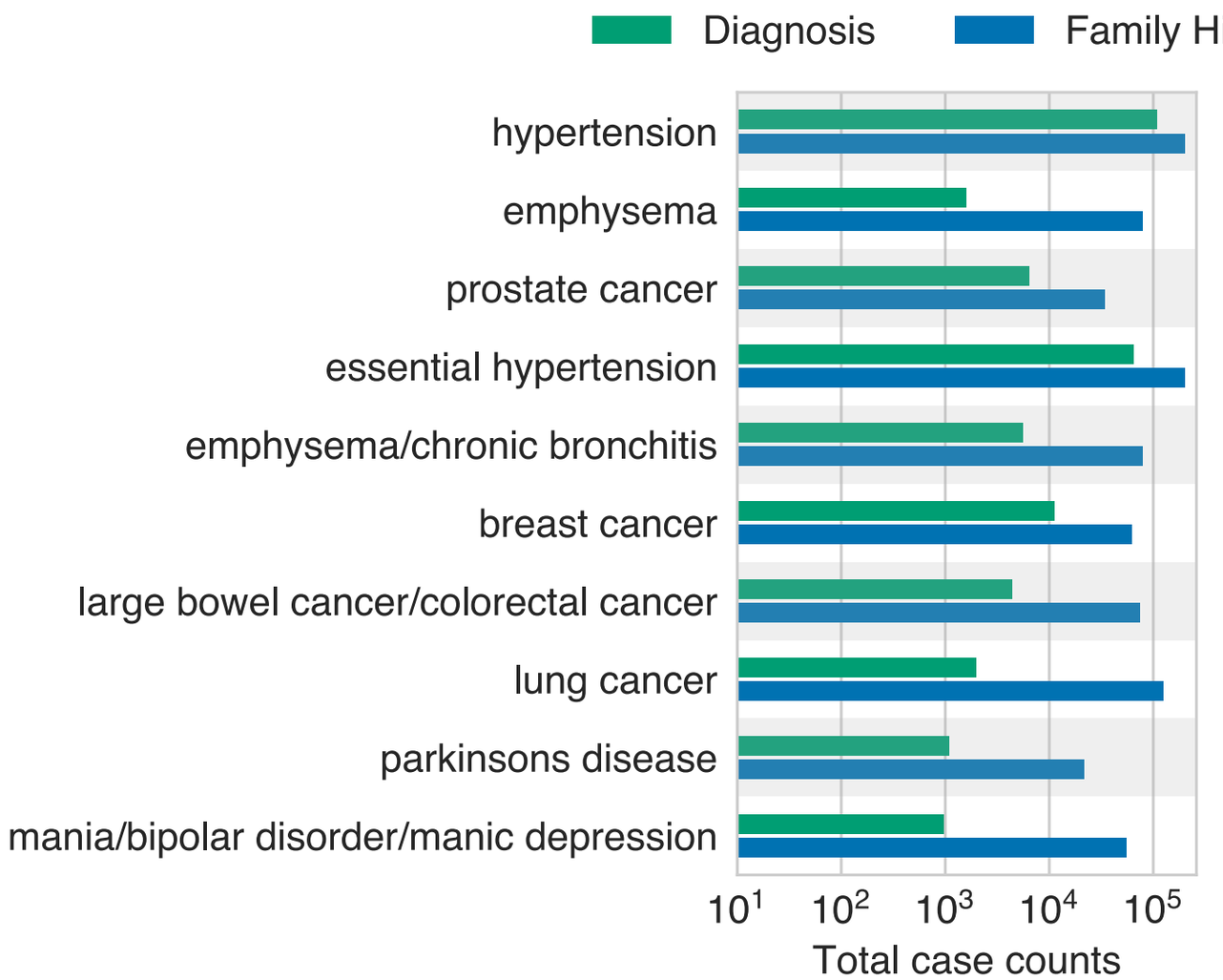
- Genetic correlation for phenotypes defined using either hospital and registry records ***or*** questionnaire data



RIVASLAB

Family history captures disease genetics

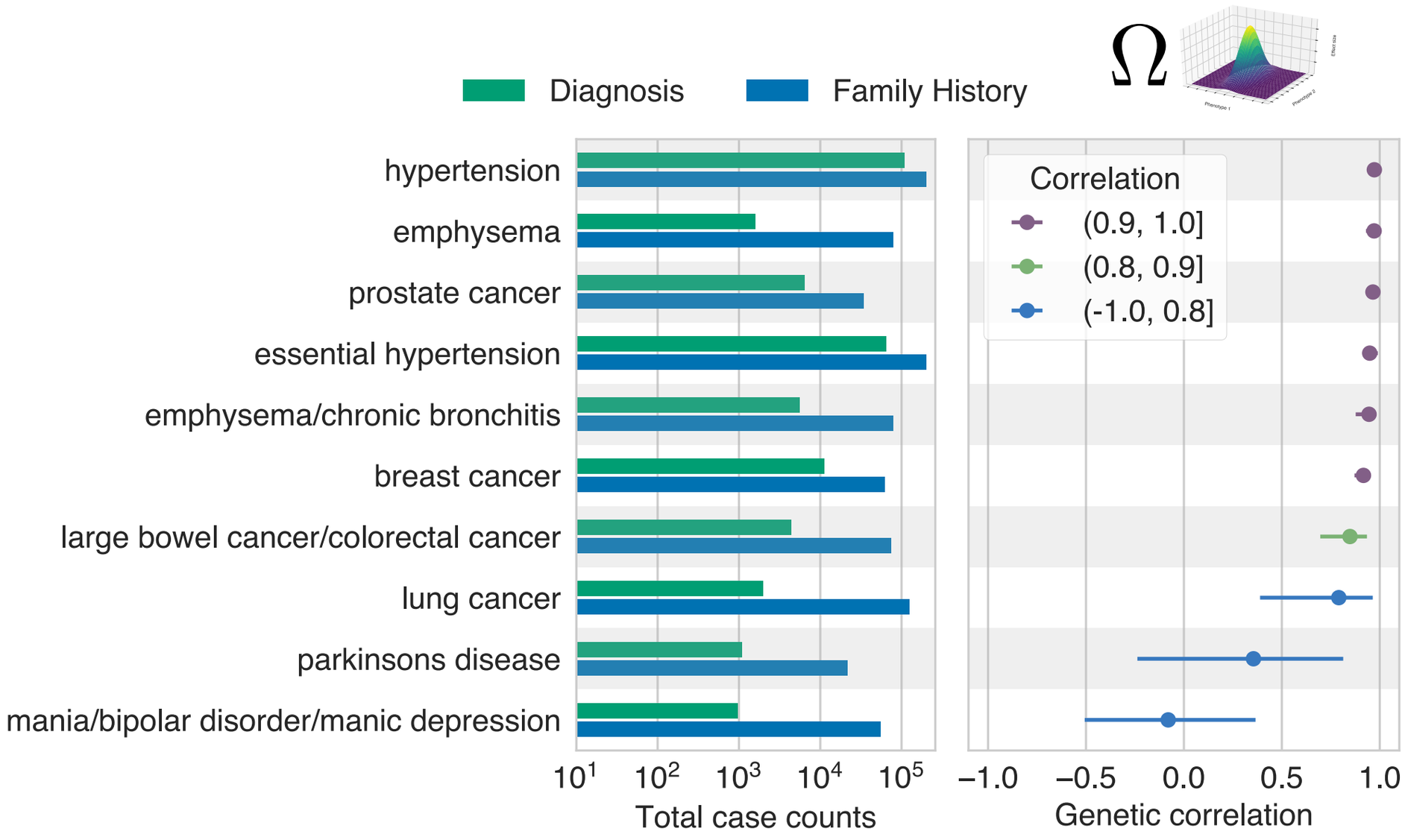
- Family history of disease from questionnaire vs. diagnosis



RIVASLAB

Family history captures disease genetics

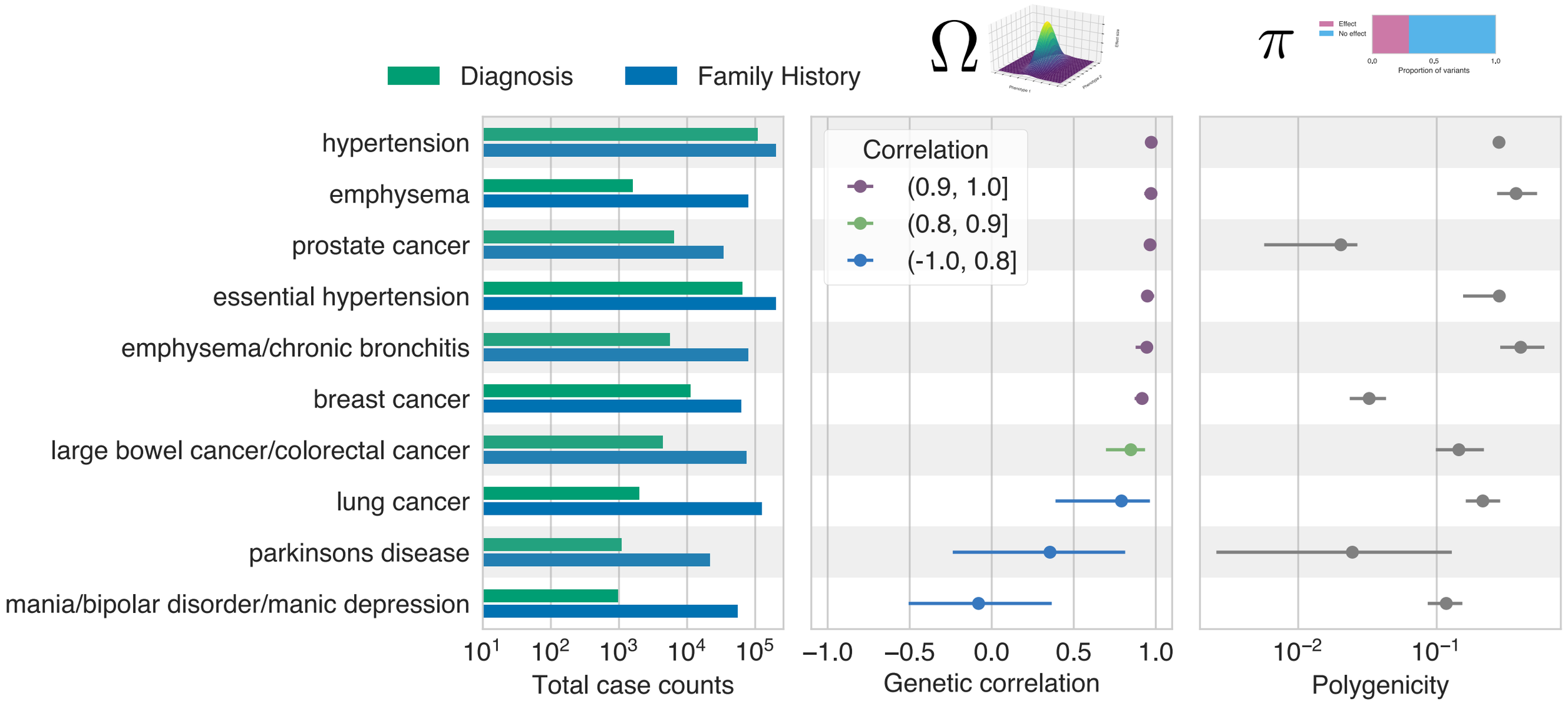
- Family history of disease from questionnaire vs. diagnosis



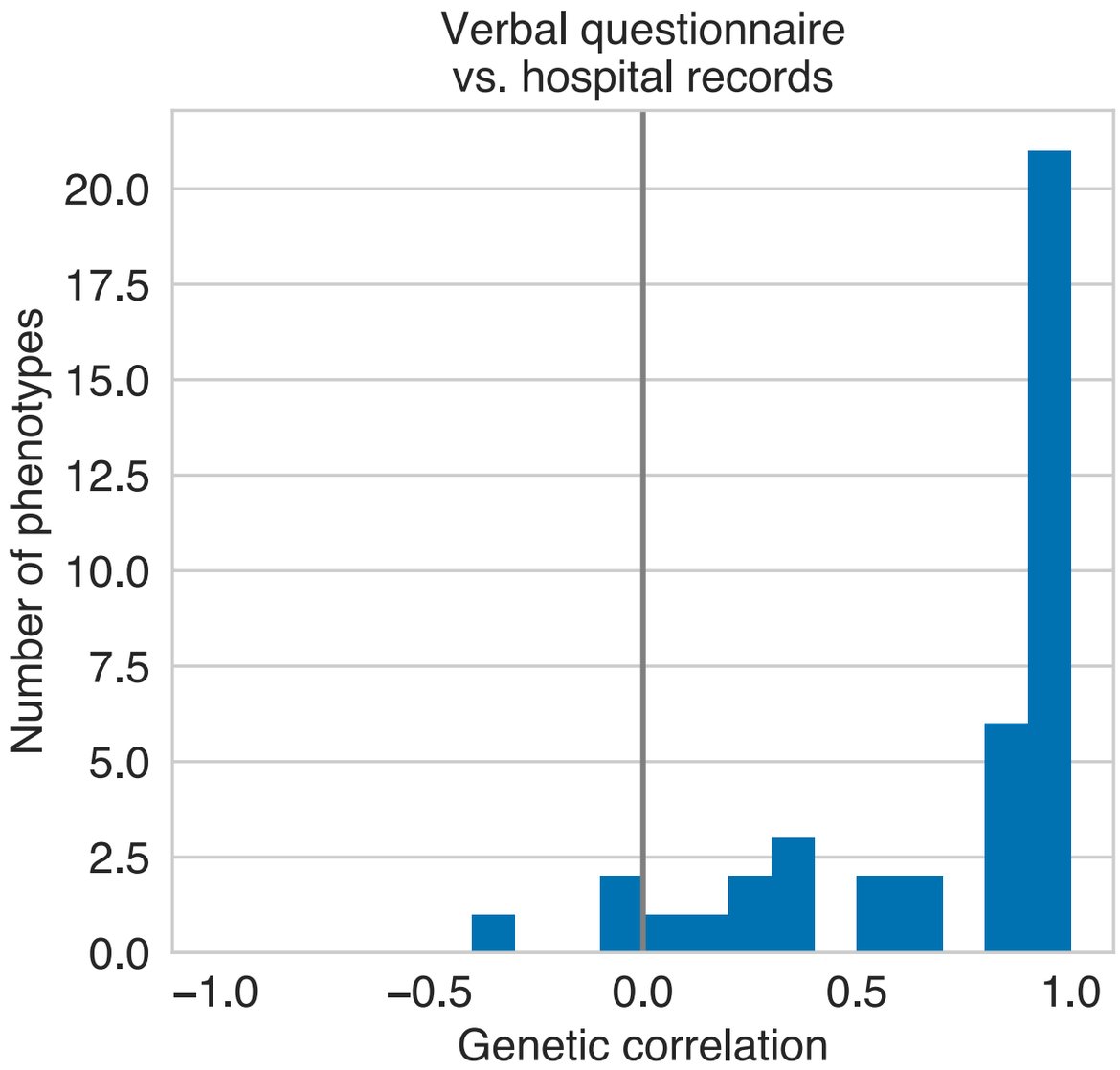
RIVASLAB

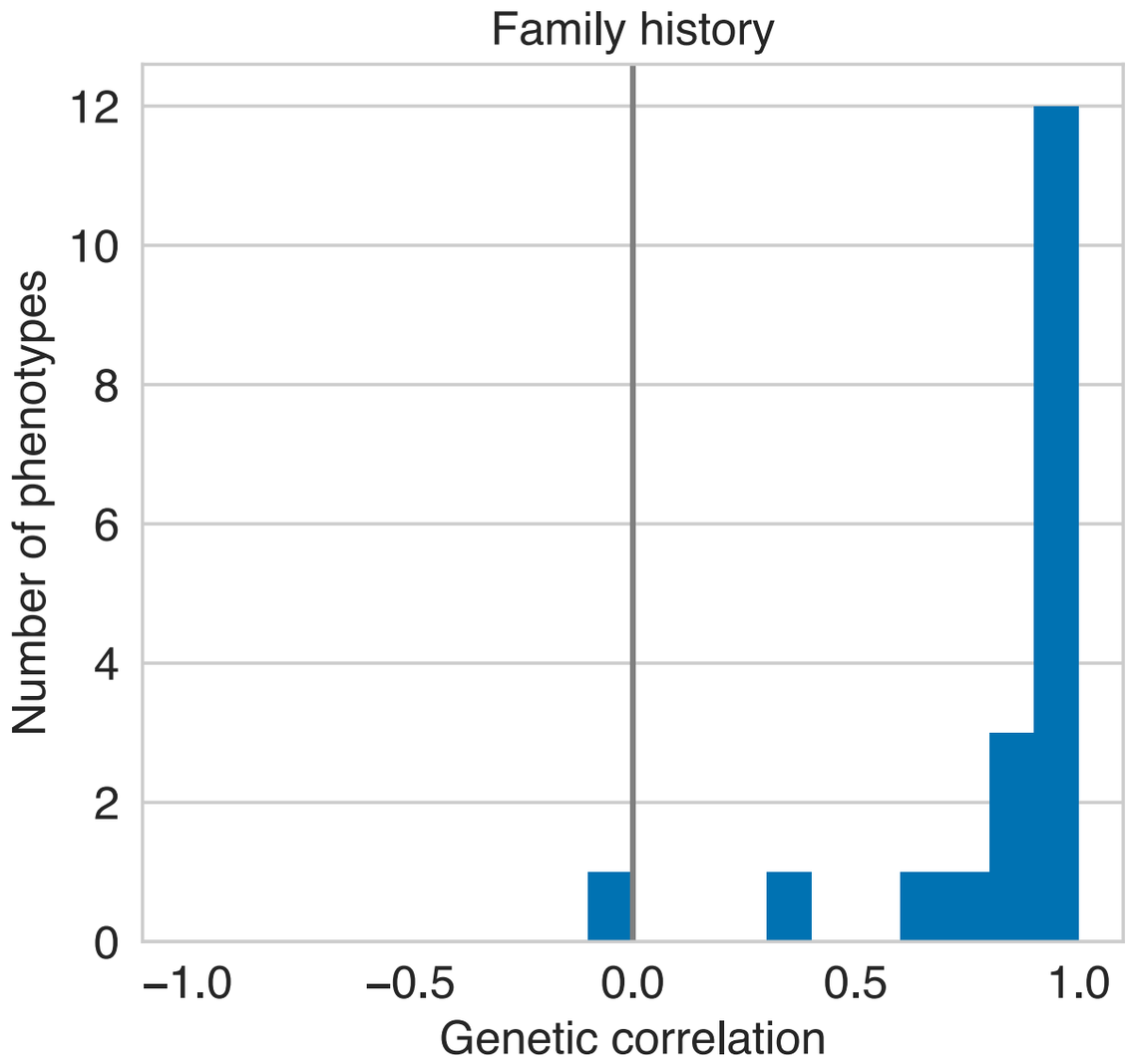
Family history captures disease genetics

- Family history of disease from questionnaire vs. diagnosis



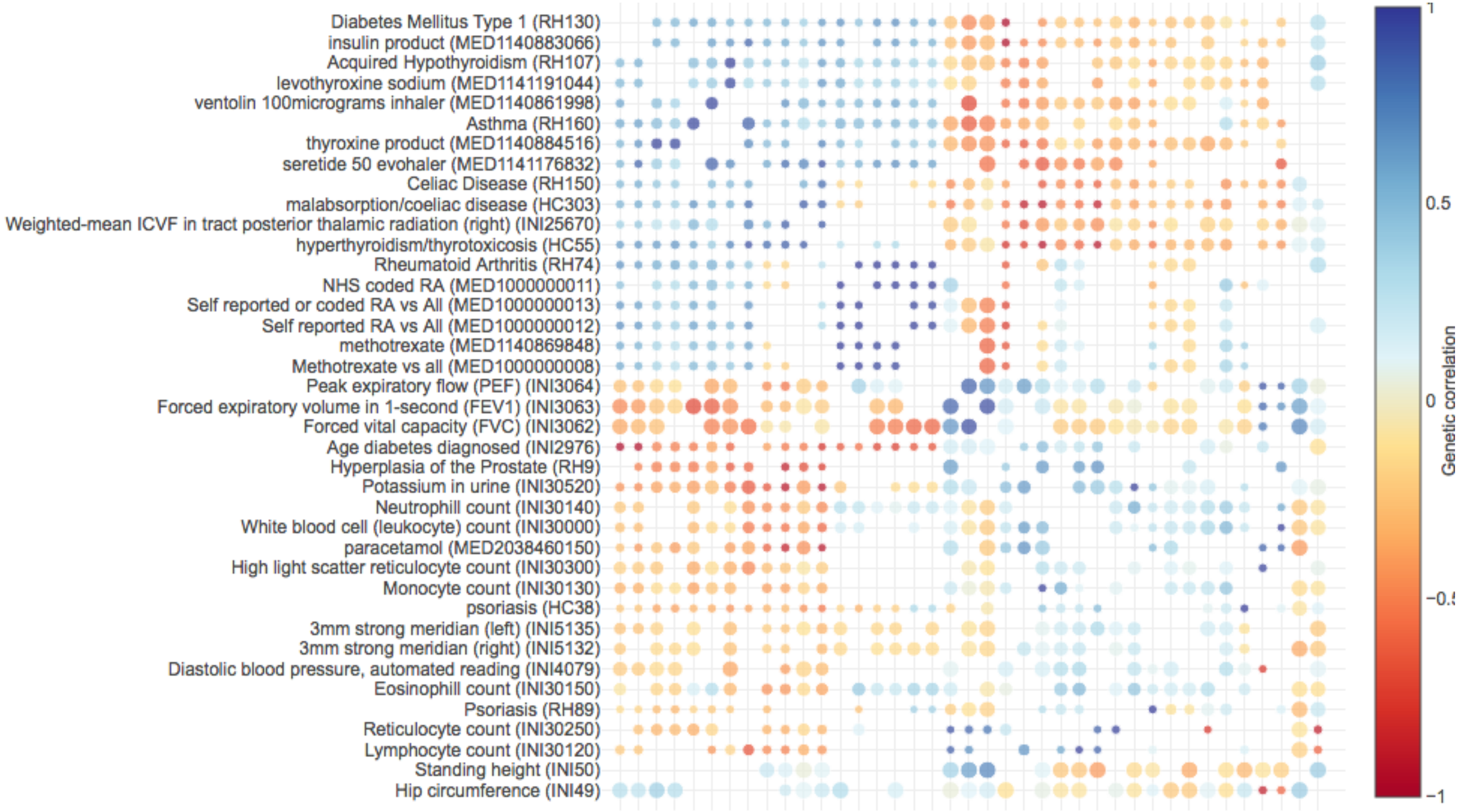
RIVASLAB





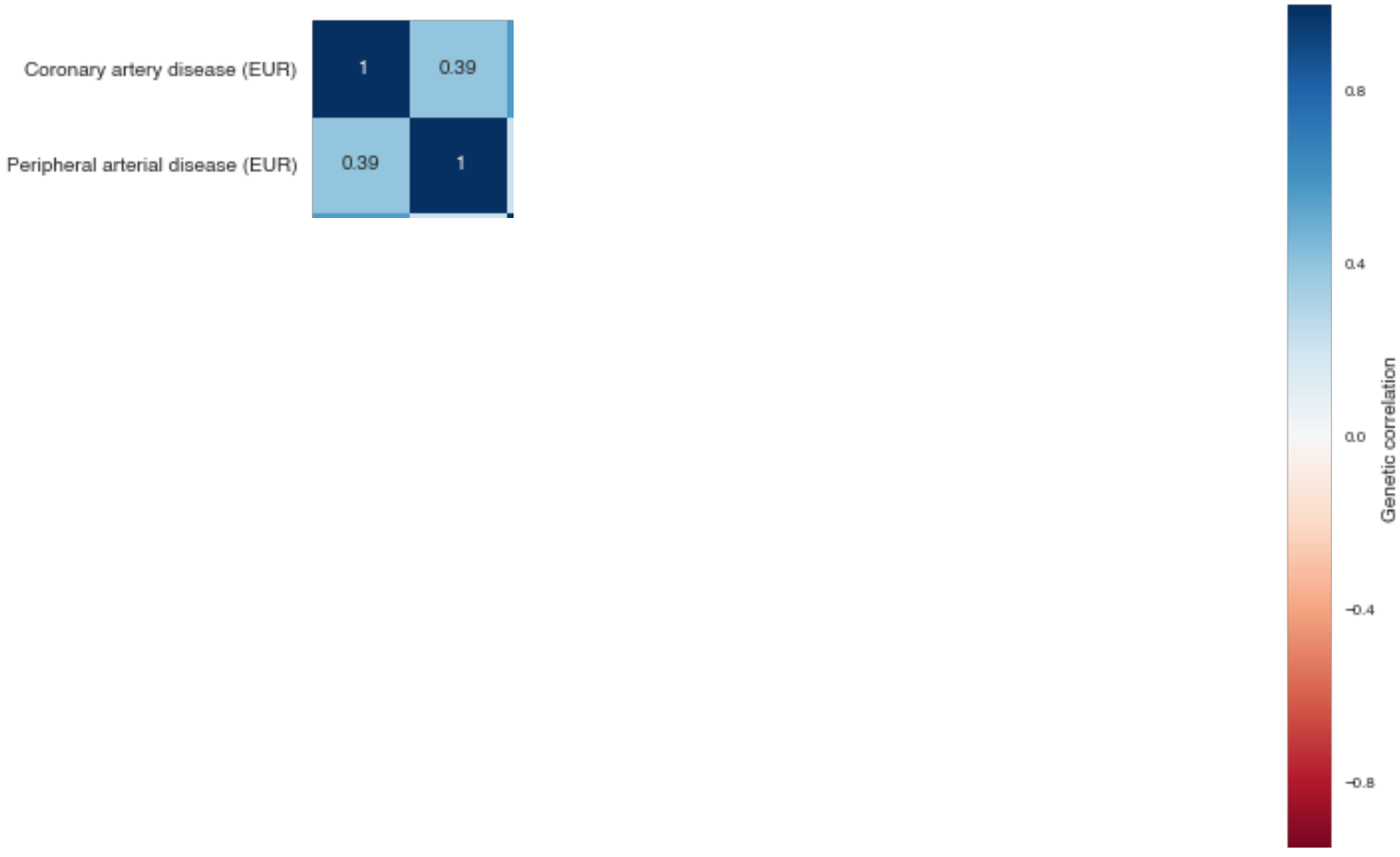
- Compared 41 phenotypes for hospital records vs. verbal questionnaire
- Compared 19 family history phenotypes
- Good agreement for most phenotypes

RIVASLAB



https://biobankengine.stanford.edu/gcorr

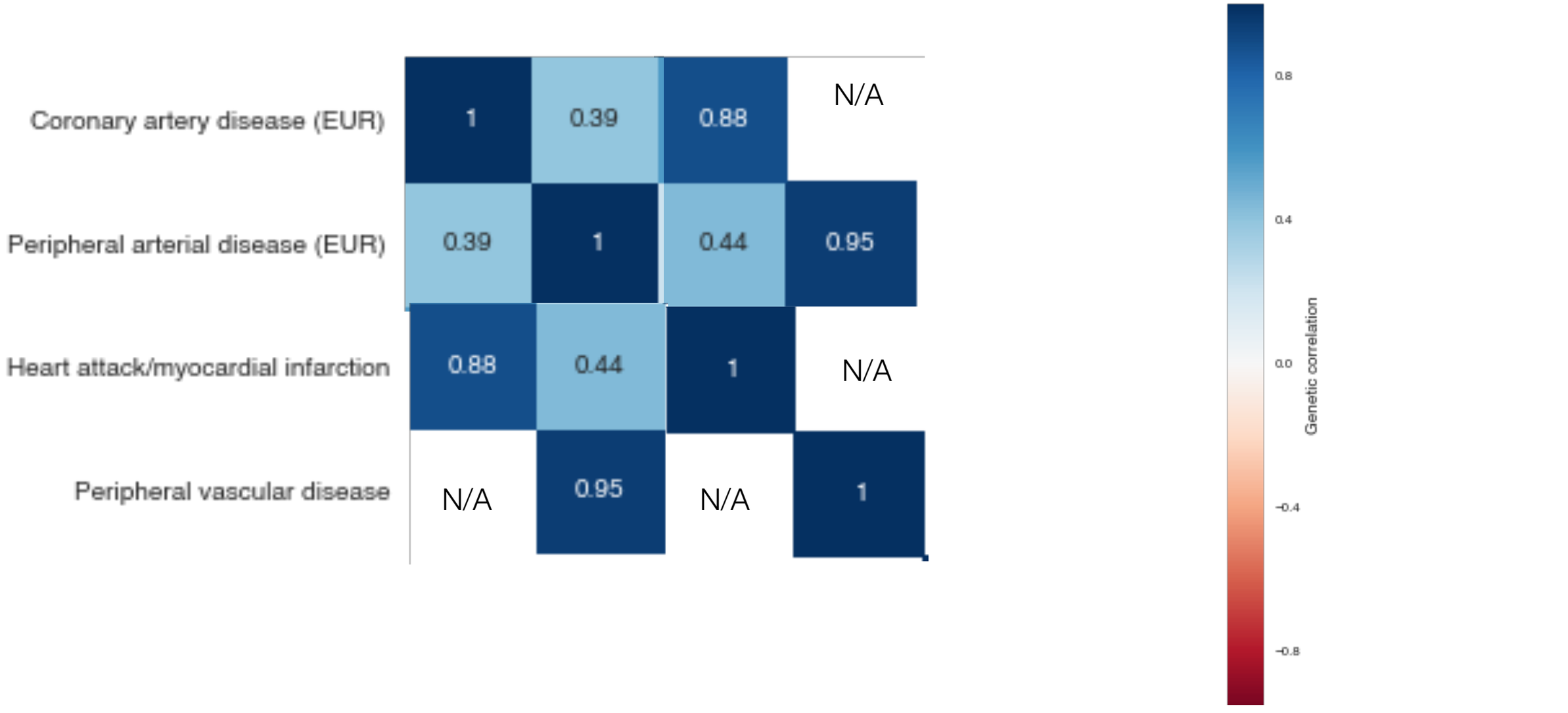
Bringing in multiple studies



From one biobank to the next…

RIVASLAB

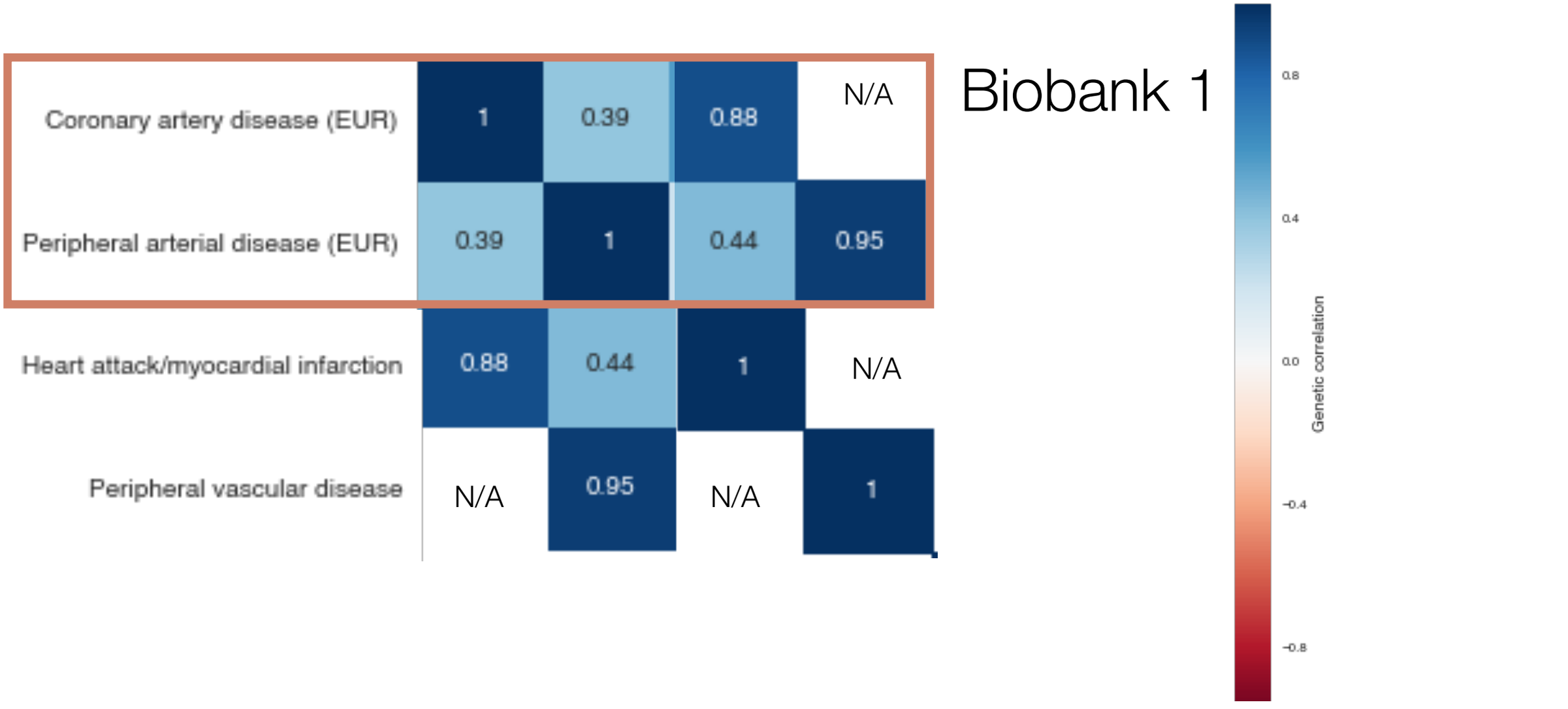
Bringing in multiple studies



From one biobank to the next…

RIVASLAB

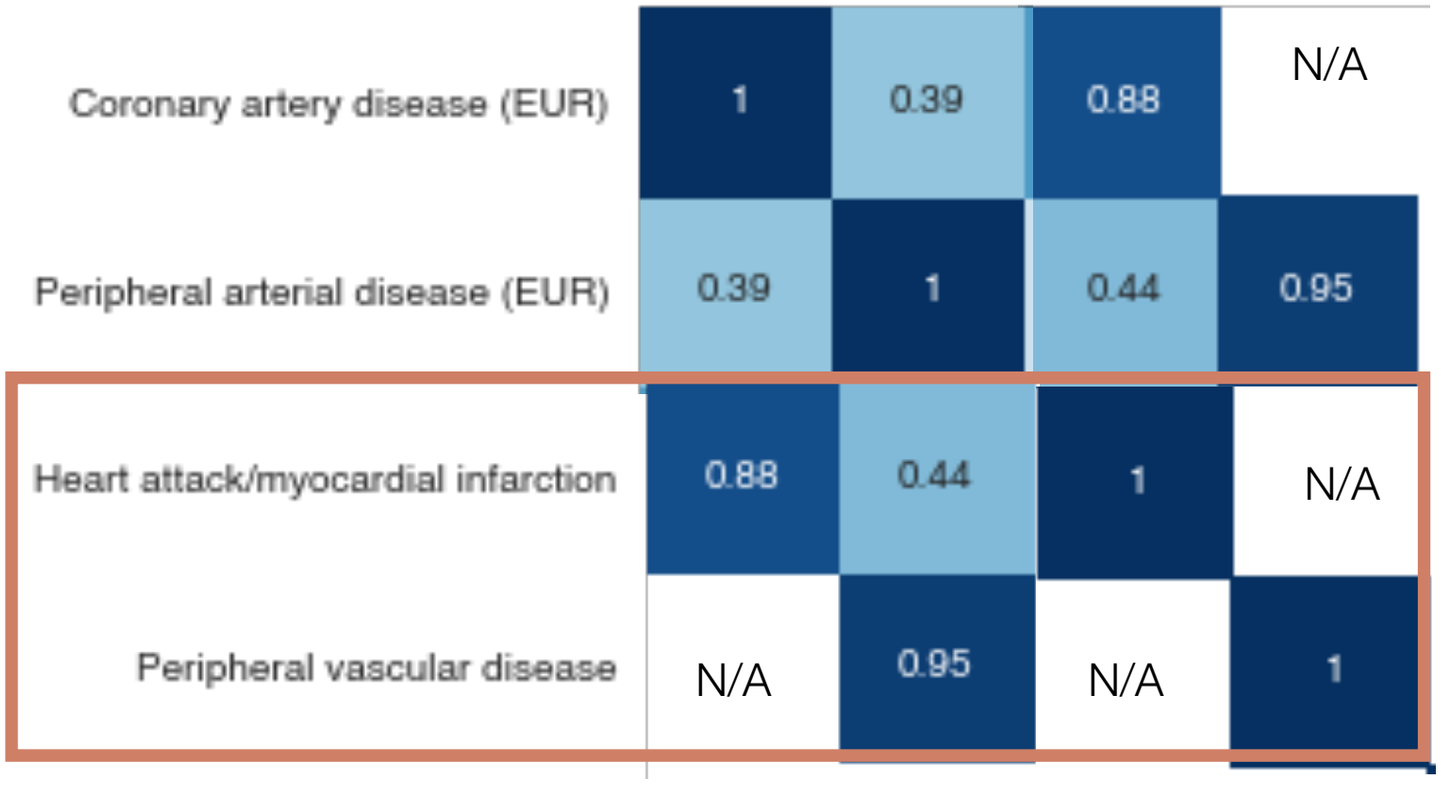
Bringing in multiple studies



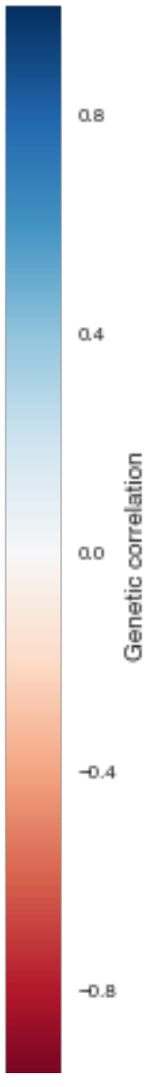
From one biobank to the next…

RIVASLAB

Bringing in multiple studies



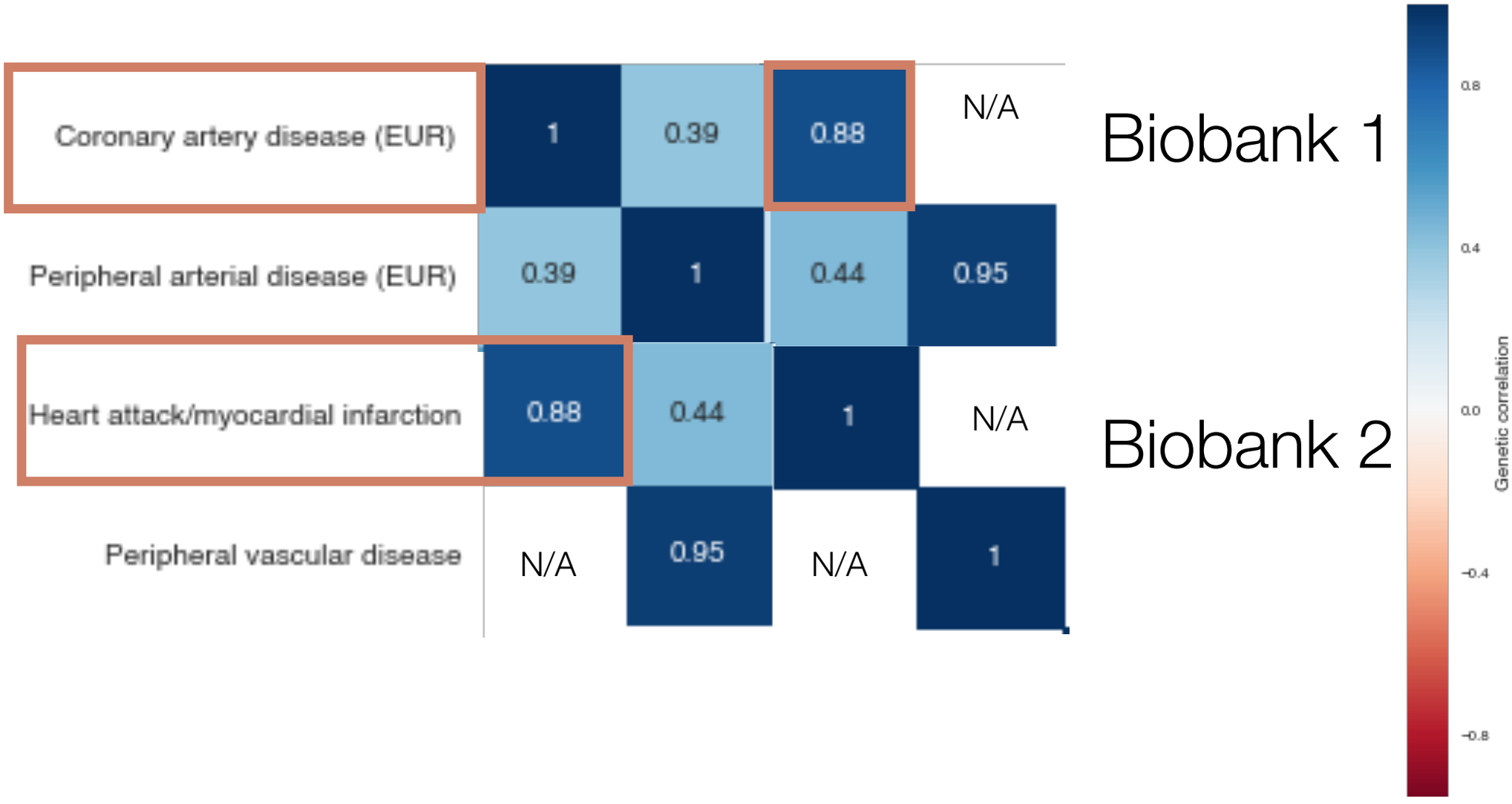
Biobank 2



From one biobank to the next…

RIVASLAB

Coronary artery disease & heart attack/MI



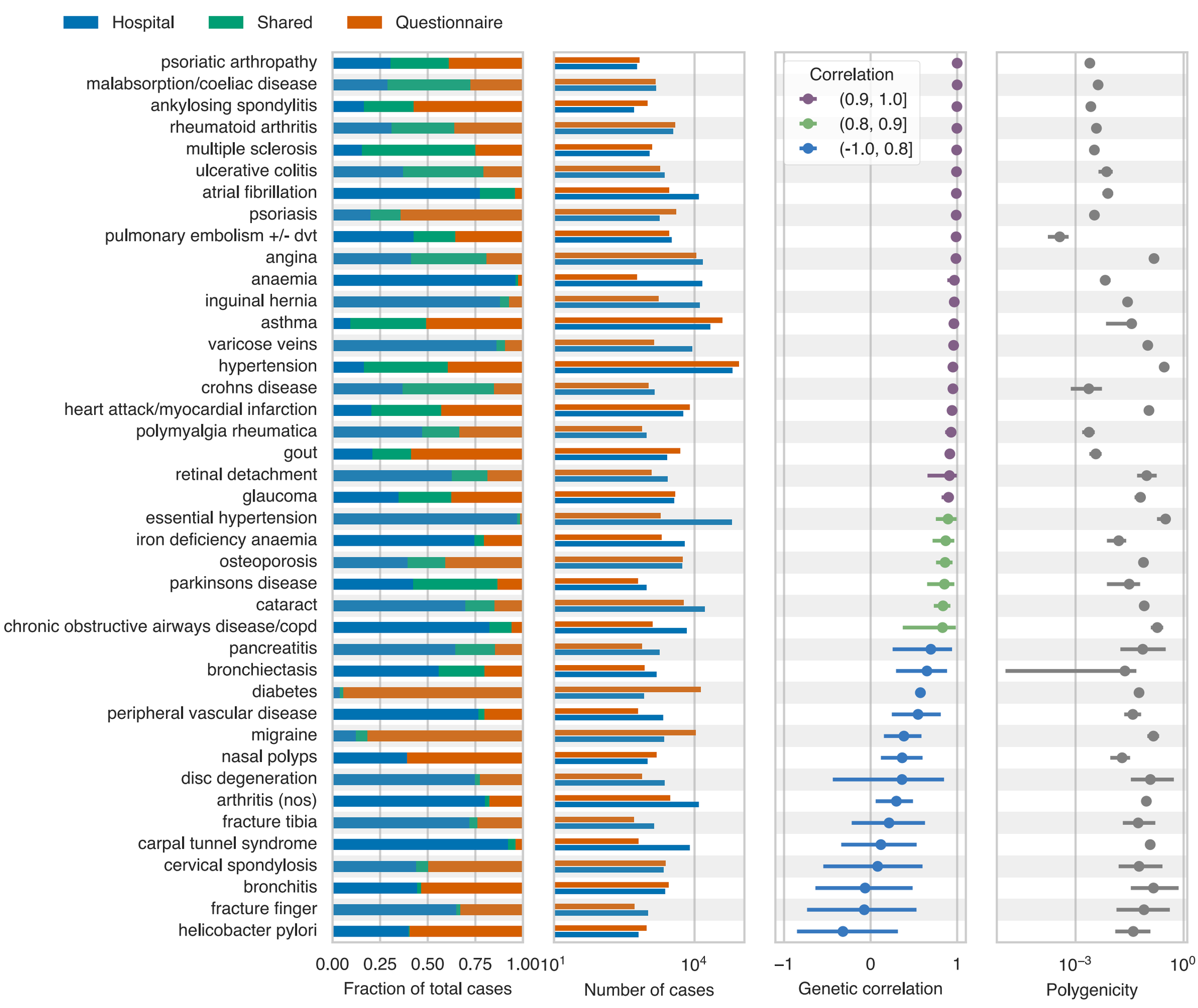
From one biobank to the next…

RIVASLAB

Peripheral arterial disease & Peripheral vascular disease

From one biobank to the next…

RIVASLAB


Hospital
Shared
Questionnaire
psoriatic arthropathy
malabsorption/coeliac disease
ankylosing spondylitis
rheumatoid arthritis
multiple sclerosis
ulcerative colitis
atrial fibrillation
psoriasis
pulmonary embolism +/- dvt
angina
anaemia
inguinal hernia
asthma
varicose veins
hypertension
crohns disease
heart attack/myocardial infarction
polymyalgia rheumatica
gout
retinal detachment
glaucoma
essential hypertension
iron deficiency anaemia
osteoporosis
parkinsons disease
cataract
chronic obstructive airways disease/copd
pancreatitis
bronchiectasis
diabetes
peripheral vascular disease
migraine
nasal polyps
disc degeneration
arthritis (nos)
fracture tibia
carpal tunnel syndrome
cervical spondylosis
bronchitis
fracture finger
helicobacter pylori
Correlation
(0.9, 1.0]
(0.8, 0.9]
(-1.0, 0.8]
0.00
0.25
0.50
0.75
1.00
Fraction of total cases
10^1
10^4
Number of cases
−1
0
1
Genetic correlation
10^{-3}
10^0
Polygenicity

Diagnosis
Family History
type 2 diabetes (HC165)
Diabetes (FH1220)
Emphysema (RH108)
Chronic bronchitis/emphysema (FH1113)
hypertension (HC215)
High blood pressure (FH1065)
emphysema (HC431)
Chronic bronchitis/emphysema (FH1113)
Diabetes Mellitus Type 2 (RH134)
Diabetes (FH1220)
prostate cancer (cancer1044)
Prostate cancer (FH1044)
Prostate Cancer (RH45)
Prostate cancer (FH1044)
diabetes (HC221)
Diabetes (FH1220)
Diabetes diagnosed by doctor (BIN2443)
Diabetes (FH1220)
essential hypertension (HC273)
High blood pressure (FH1065)
emphysema/chronic bronchitis (HC106)
Chronic bronchitis/emphysema (FH1113)
breast cancer (cancer1002)
Breast cancer (FH1002)
Colorectal Cancer (RH95)
Bowel cancer (FH1019)
large bowel cancer/colorectal cancer (cancer1020)
Bowel cancer (FH1019)
Lung Cancer (RH145)
Lung cancer (FH1001)
lung cancer (cancer1001)
Lung cancer (FH1001)
Diabetes Mellitus Type 1 (RH130)
Diabetes (FH1220)
parkinsons disease (HC206)
Parkinson's disease (FH1262)
mania/bipolar disorder/manic depression (HC233)
Severe depression (FH1286)

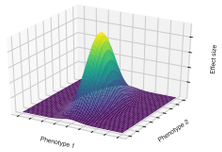
Correlation
(0.9, 1.0]
(0.8, 0.9]
(-1.0, 0.8]

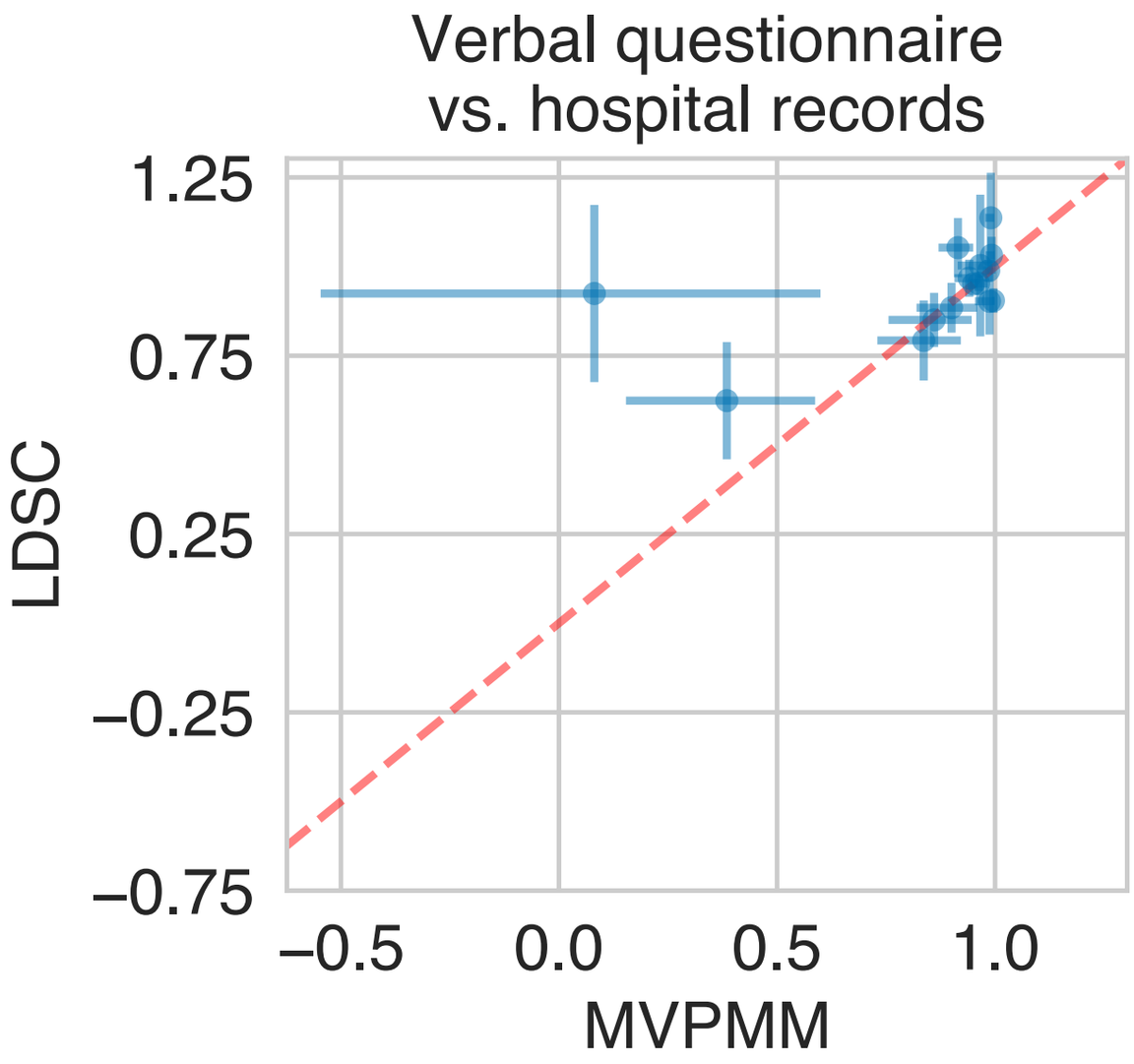
10^1 10^3 10^5
Total case counts

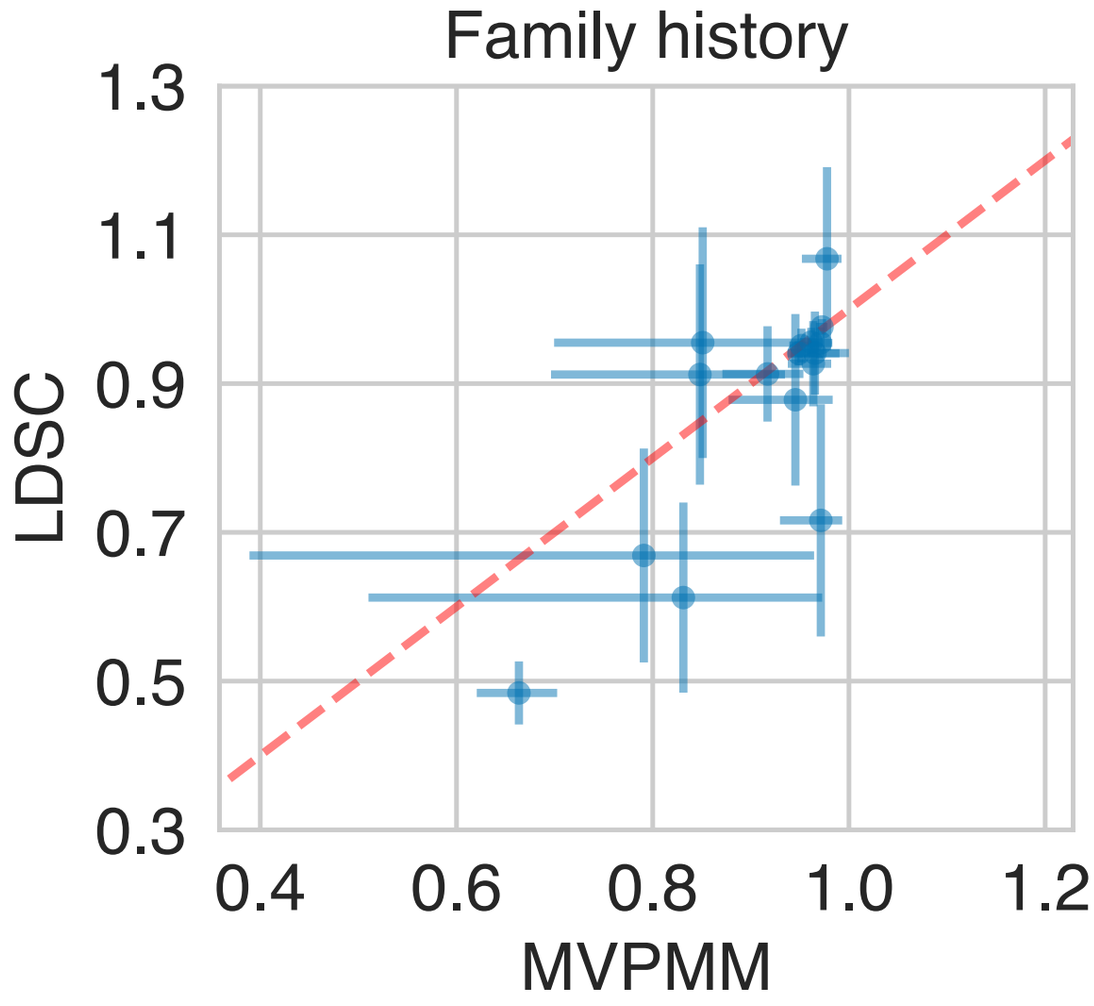
−1 0 1
Genetic correlation

10^{-2} 10^{-1}
Polygenicity

Comparison to LD Score Regression

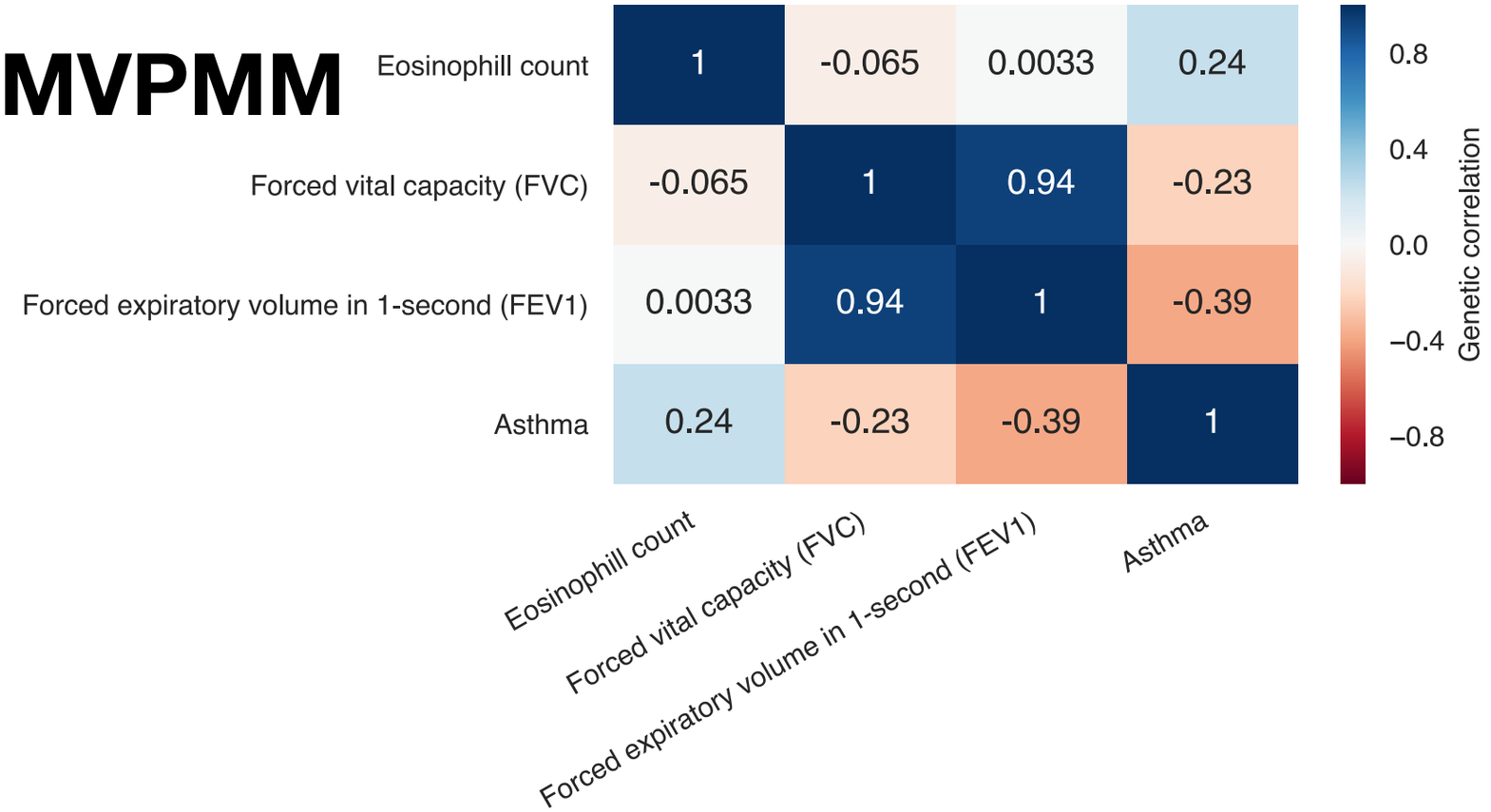
Ω 



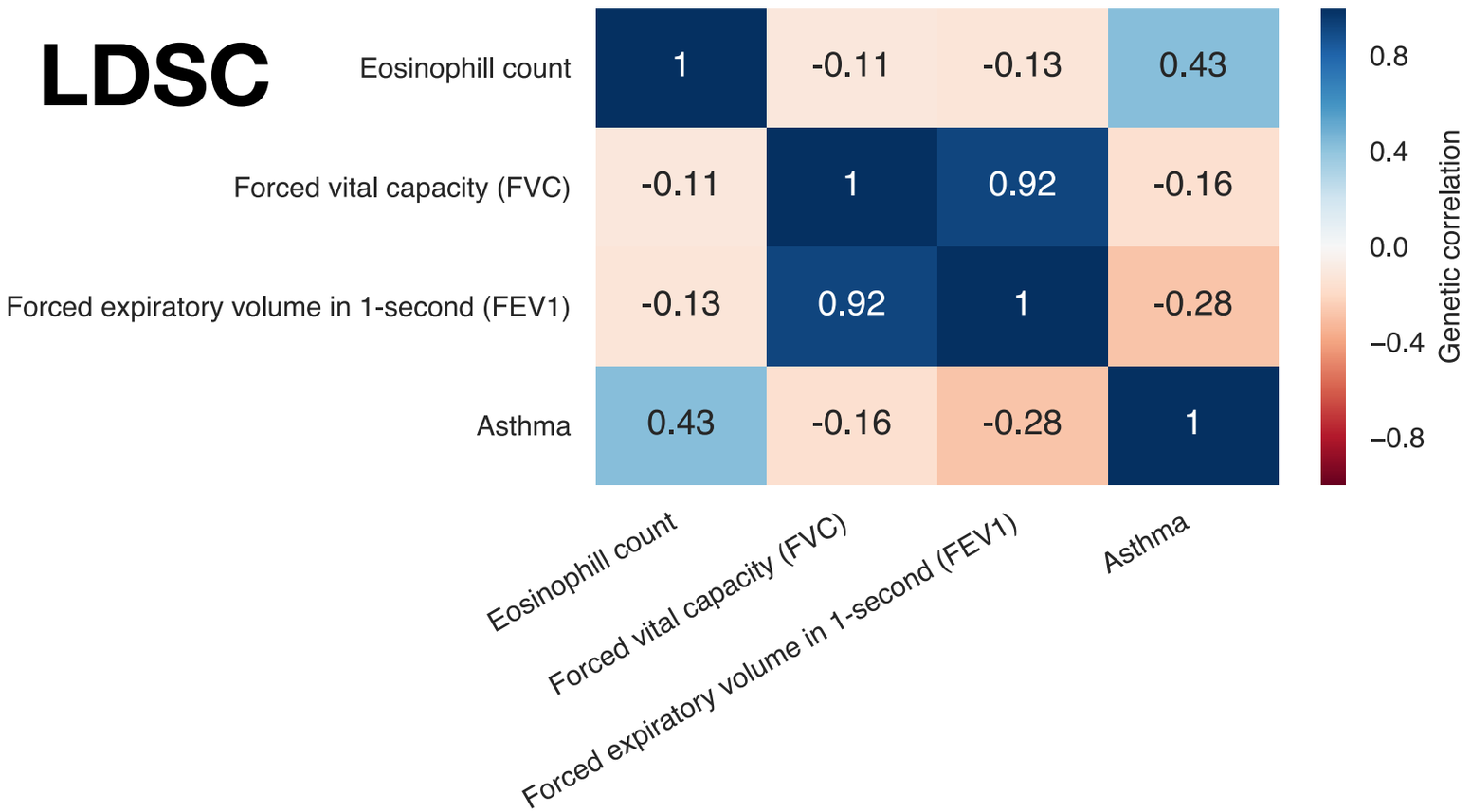


RIVASLAB

MVPMM



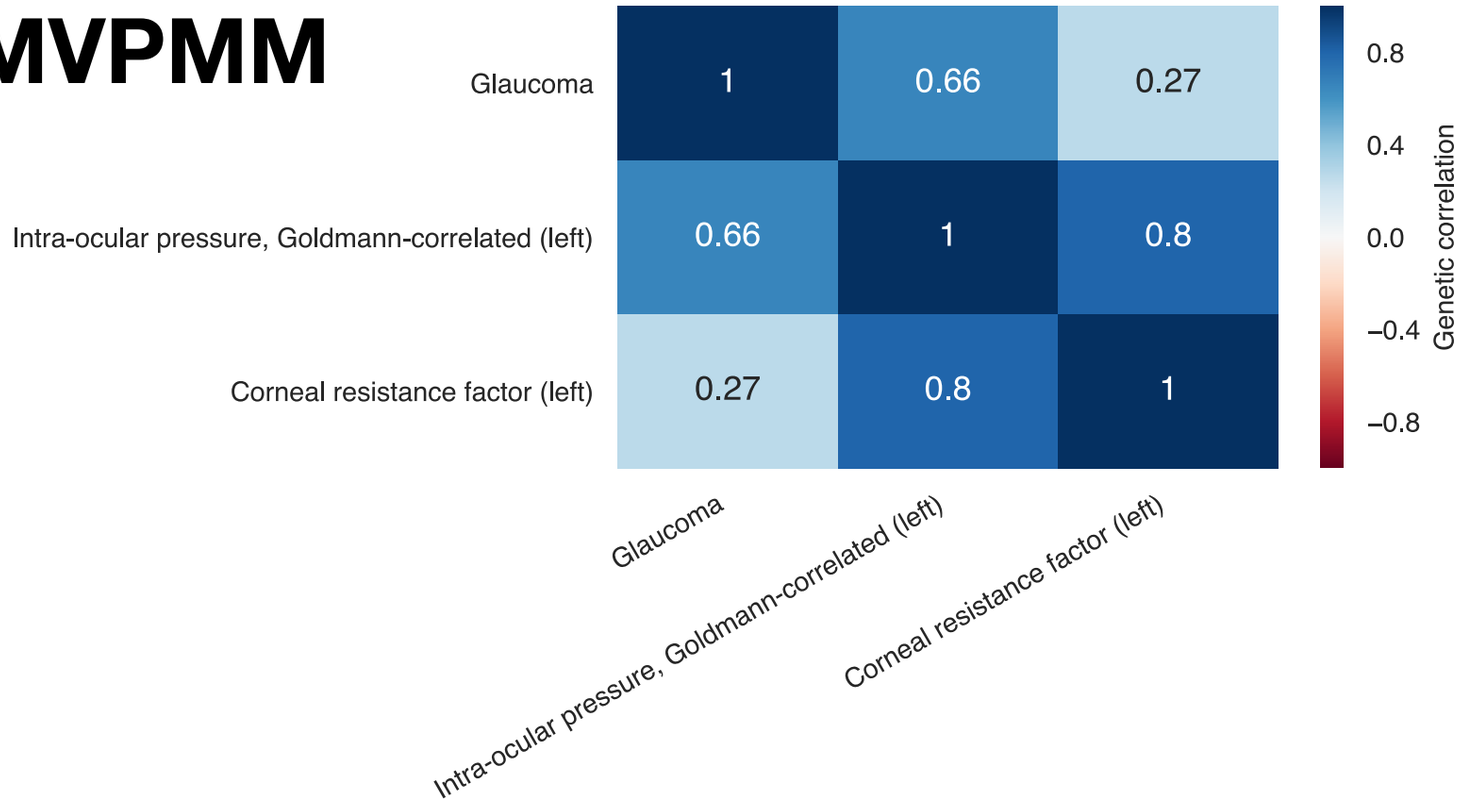
LDSC



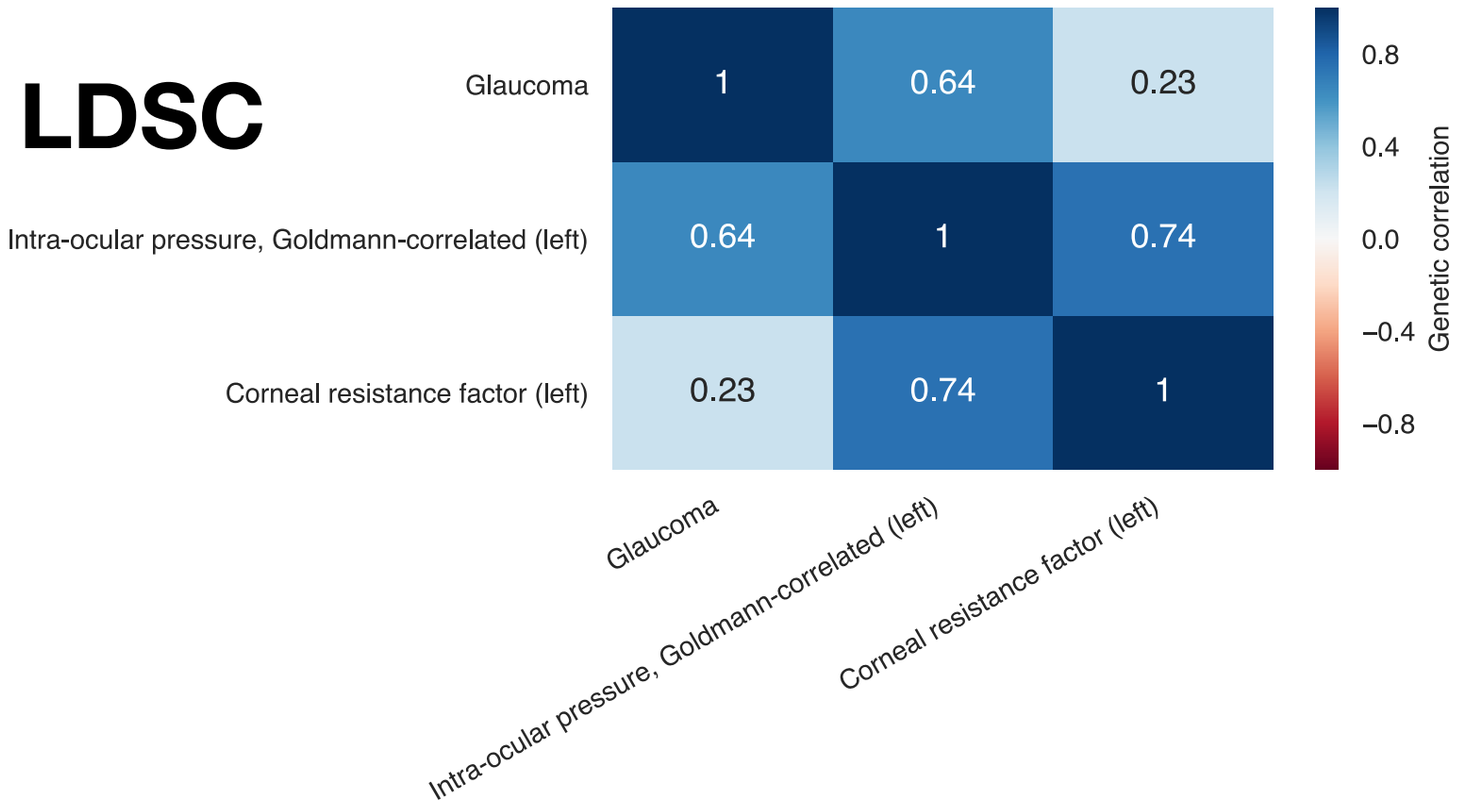
Genetic correlation estimates: asthma

- Genetic correlation estimates from MVPMM and LD score-regression

MVPMM



LDSC



Genetic correlation estimates: glaucoma

- Genetic correlation estimates from MVPMM and LD score-regression