Stan applications in Human Genetics

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

RIVASLAB

Rare diseases run in families

If you have **cystic fibrosis**, what is the risk to:

Your neighbor?	**0.04%**
Your sibling?	**25%**
Your identical twin?	**~99%**

Variation in DNA influences disease risk

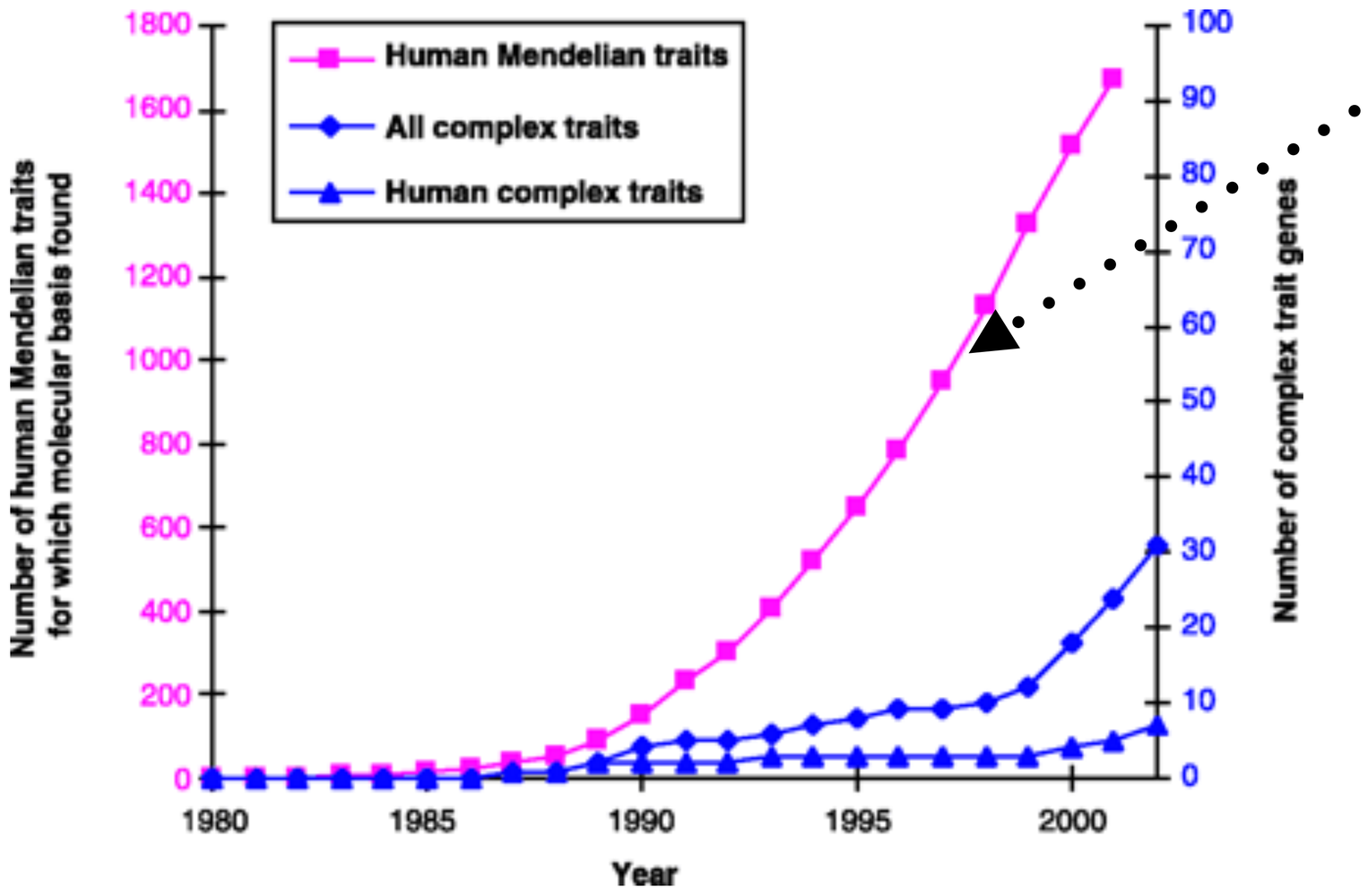
Common diseases run in families

If you have **type 2 diabetes**, what is the risk to:

Your neighbor?	**~10%**
Your sibling?	**~30%**
Your identical twin?	**>50%**

Variation in DNA influences disease risk

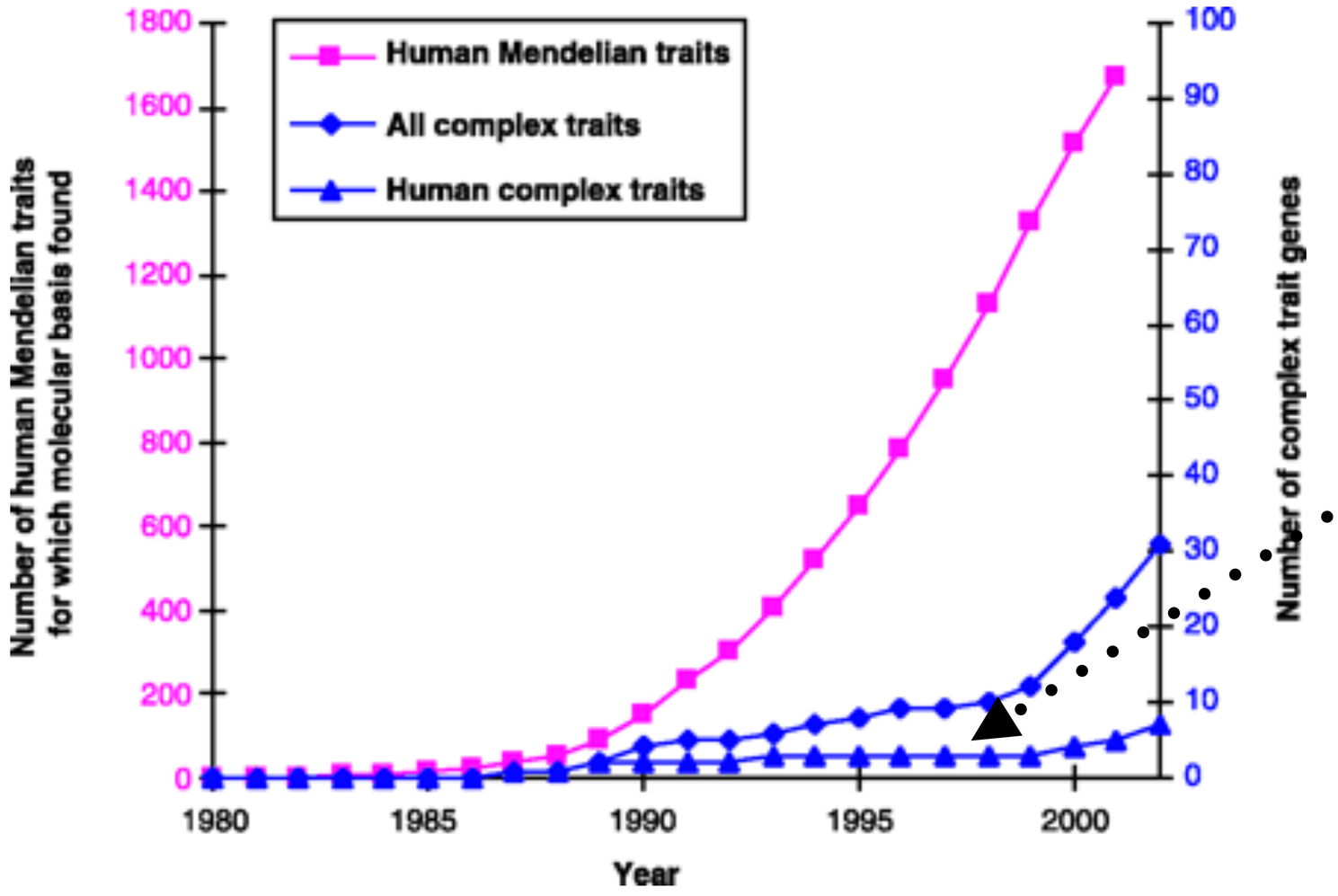
Dark ages of human genetics of complex traits



Mendelian traits (rare diseases)

RIVASLAB

Dark ages of human genetics of complex traits

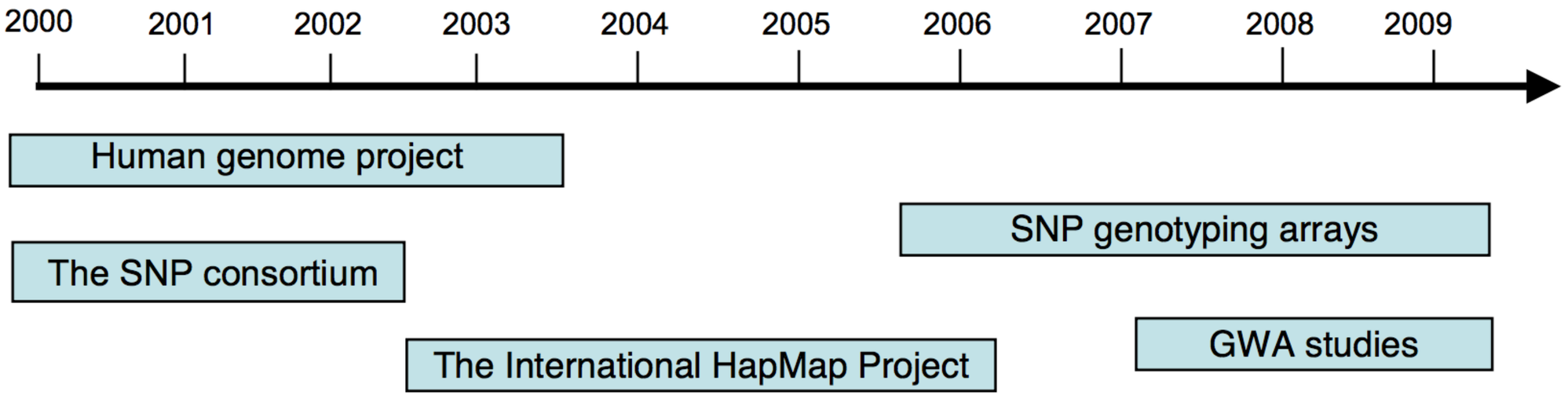


Comlex traits
(common diseases)

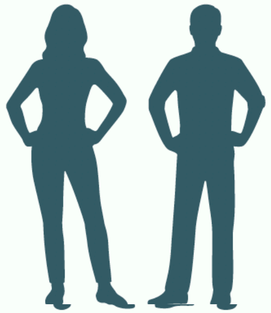
Human Genome Variation

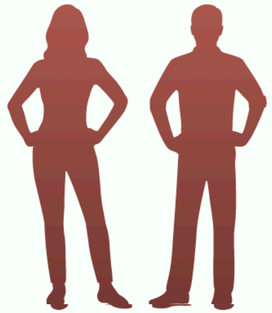
- Hailed as “Breakthrough Of The Year” in 2007

Genome wide association studies have been very successful*



Genetic association study

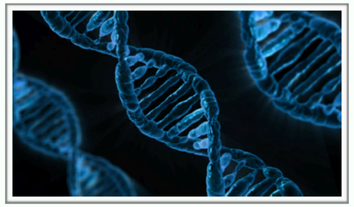




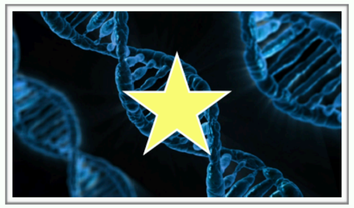
Control

Patients

For a given variant..

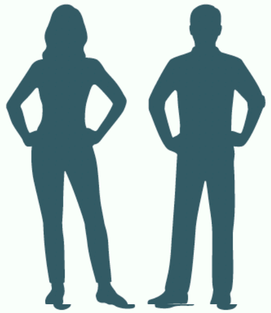


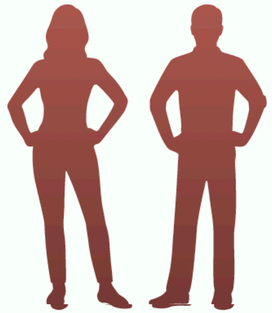
A



C

Genetic association study

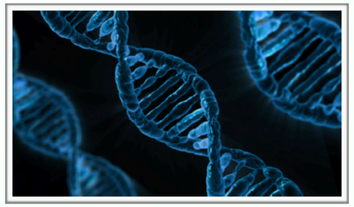




Control

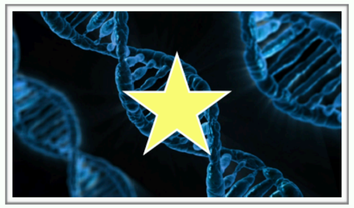
Patients

For a given variant..



A

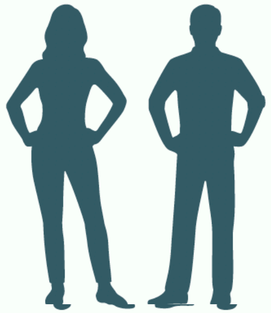
80%

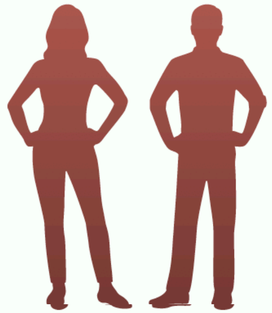


C

20%

Genetic association study

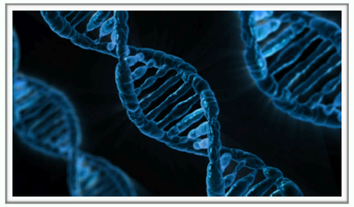




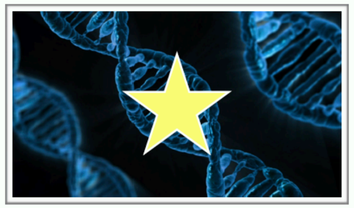
For a given variant..

Control

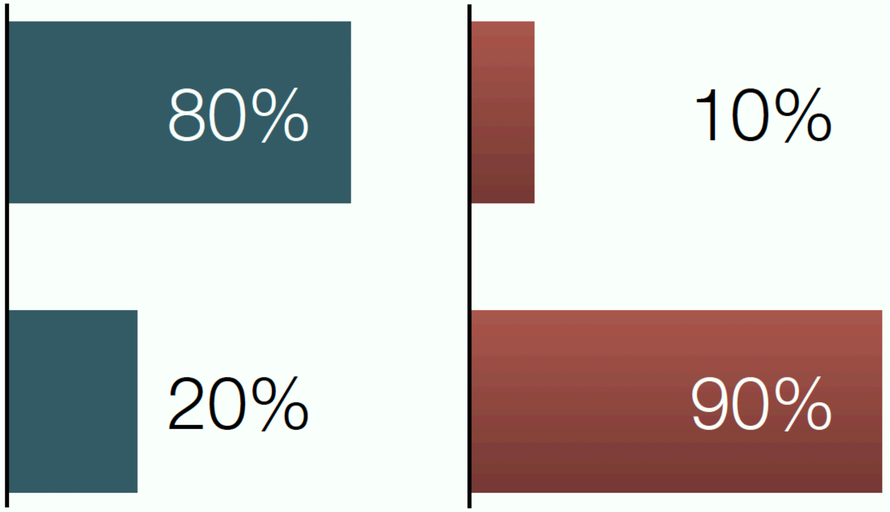
Patients



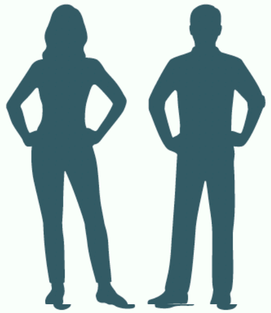
A

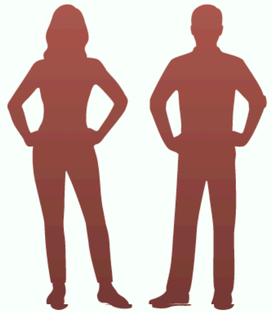


C



Genetic association study





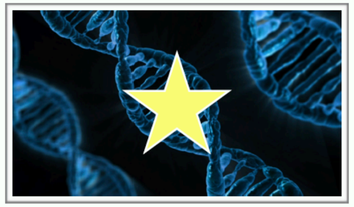
Control

Patients

For a given variant..

Variation on DNA is associated with the trait of interest (disease)

10%



C

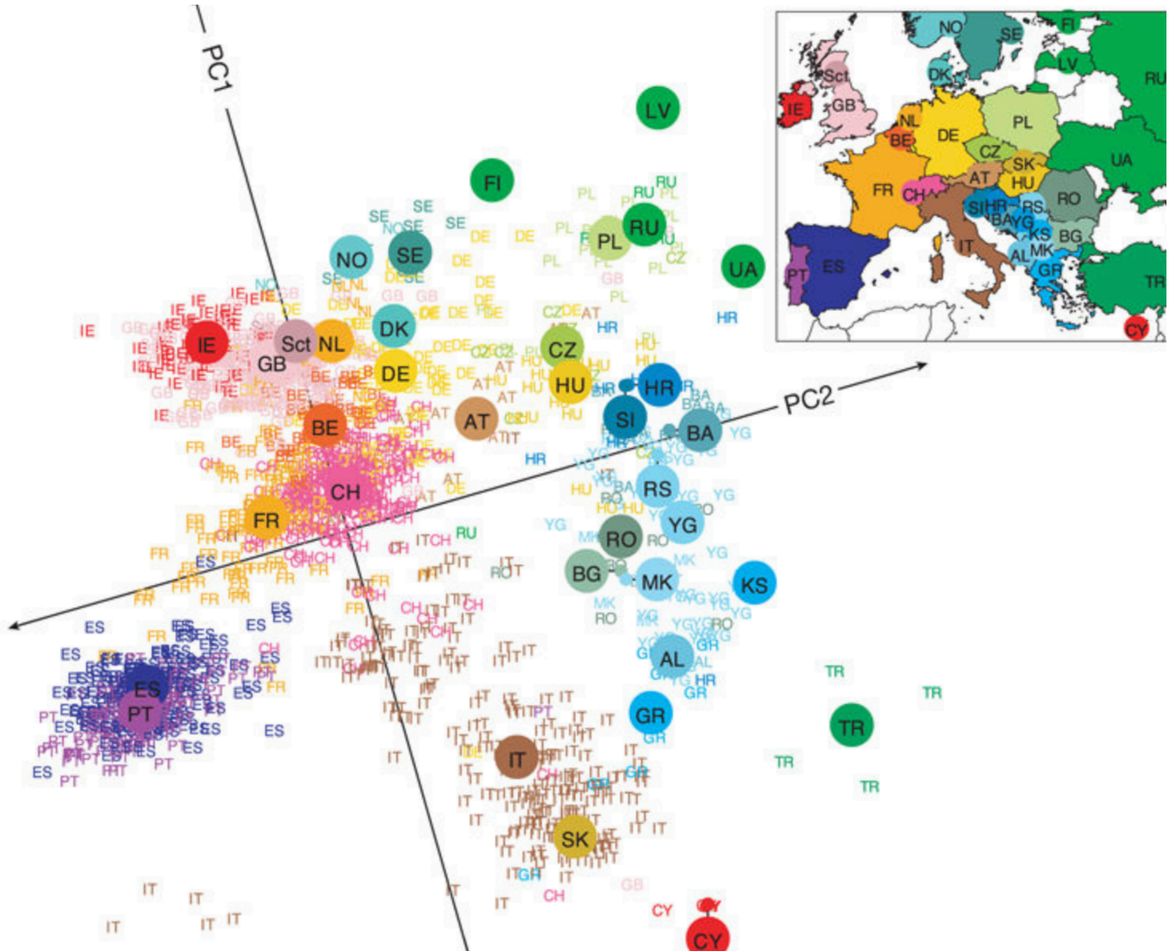
20%

90%

Typical GWAS analysis

Logistic regression analysis with covariates

Principal components to adjust for population structure

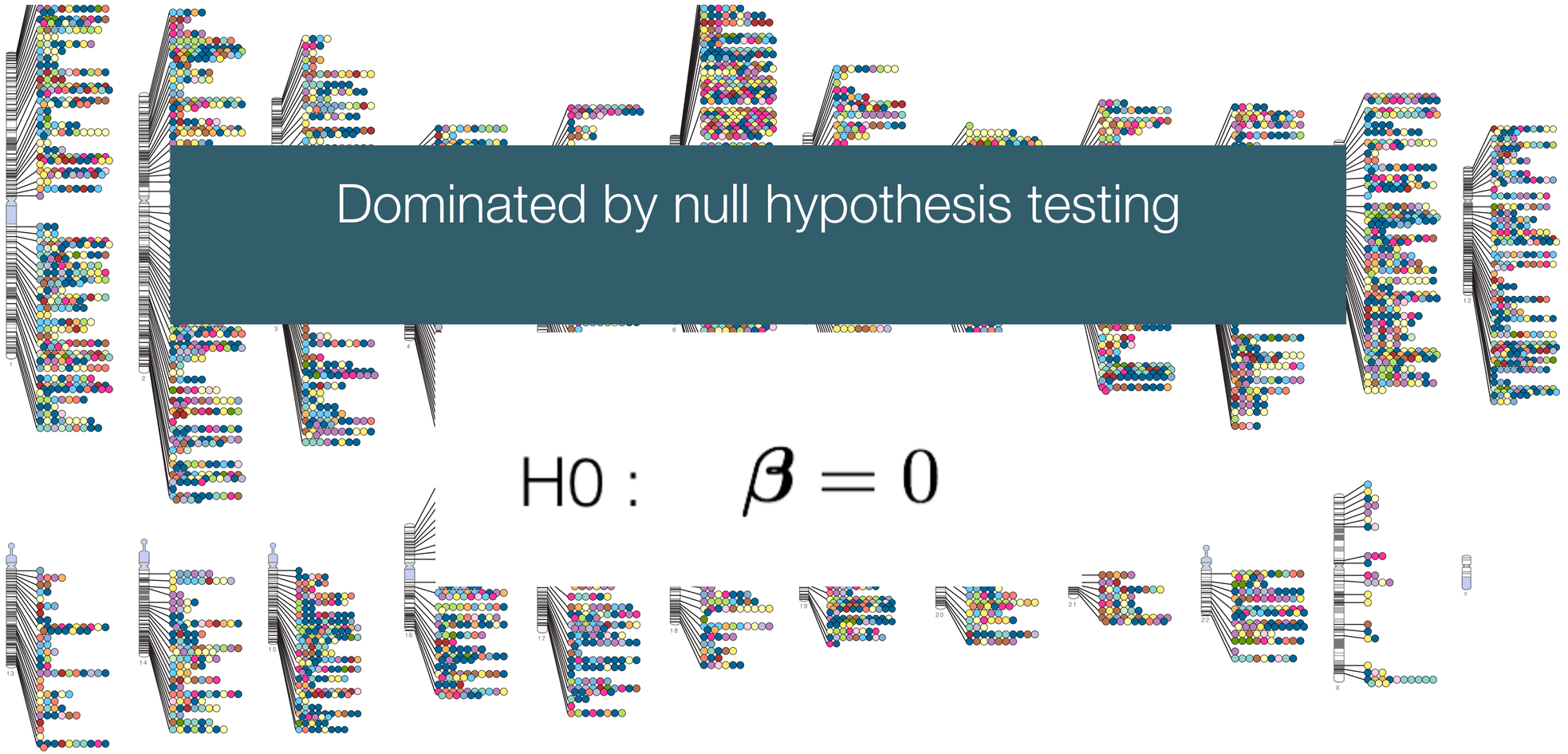


RIVASLAB

Genome wide association studies have been very successful*

Over 1000 genetic associations

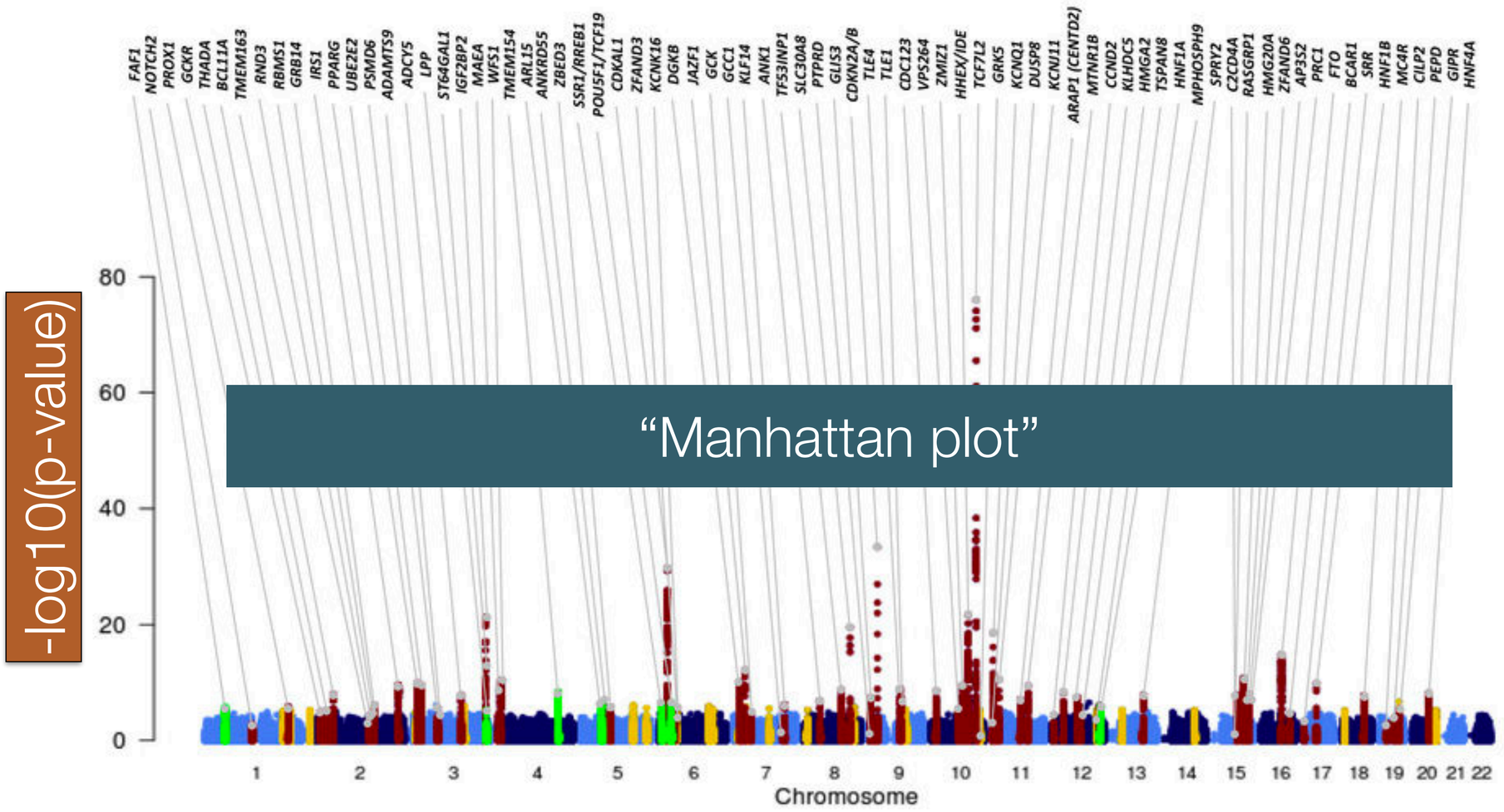
Genome wide association studies have been very successful*



i.e. the effect of the genetic variant is 0.

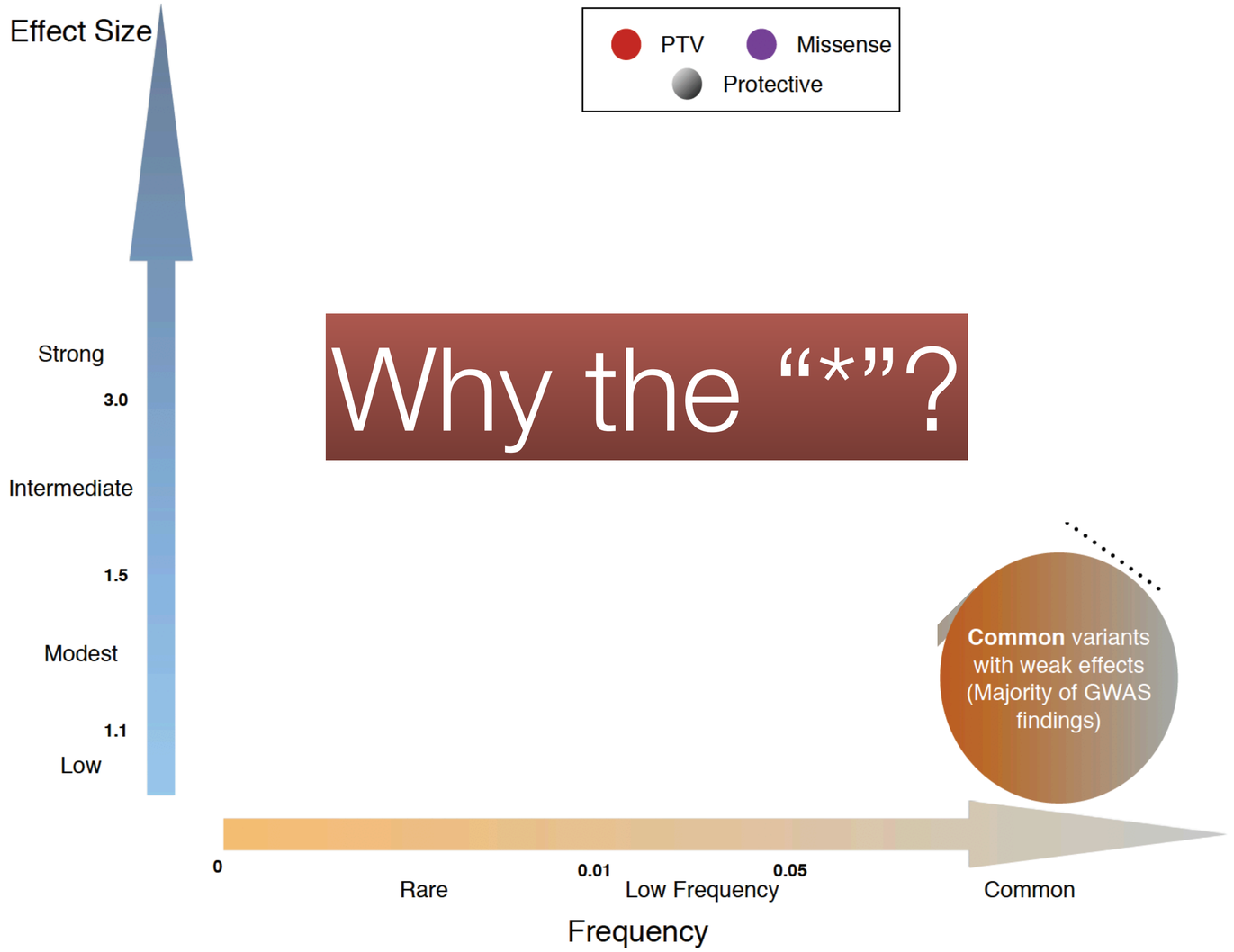
RIVASLAB

Genome wide association studies have been very successful*



RIVASLAB

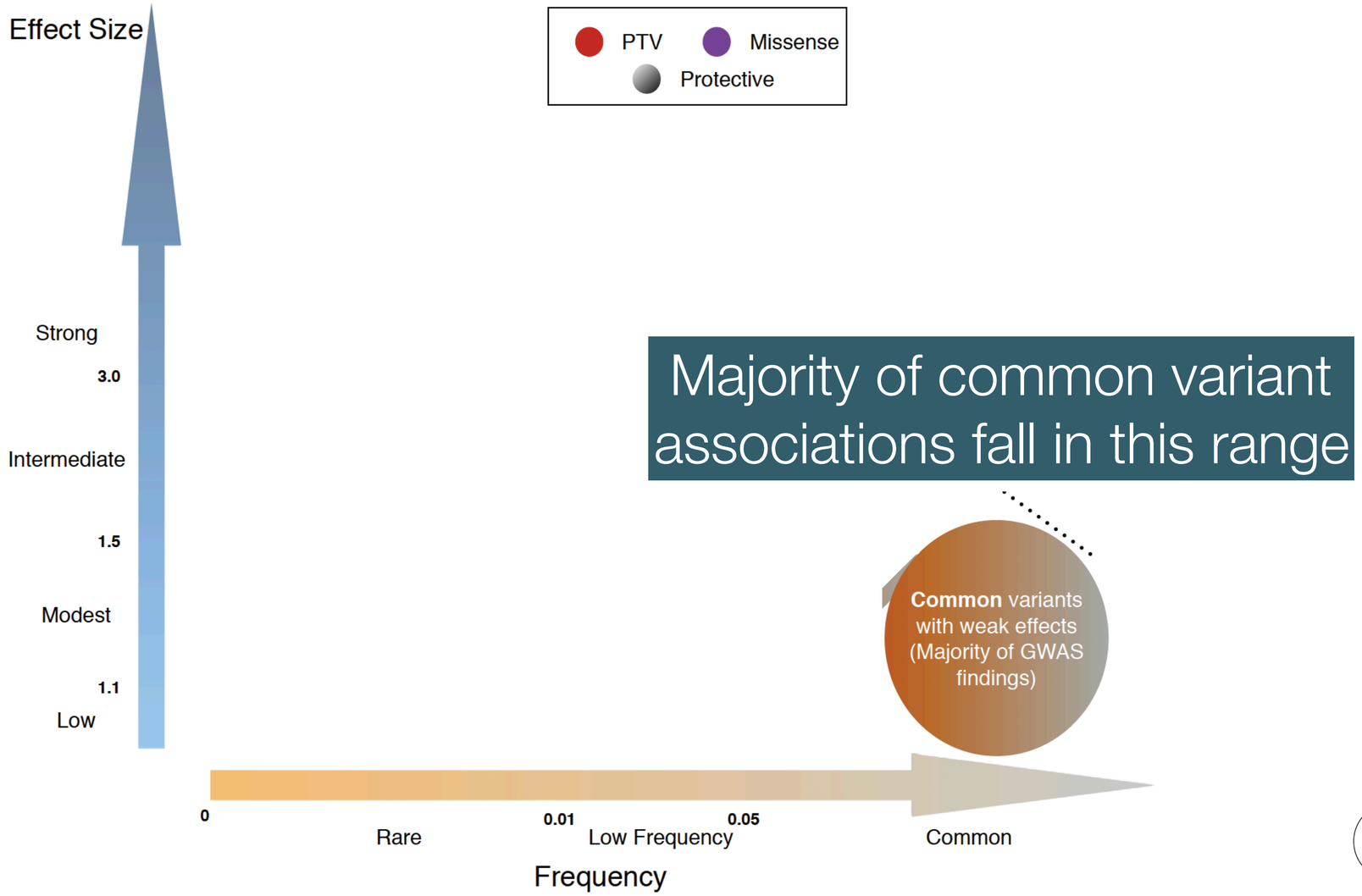
Genome wide association studies have been very successful*



RIVASLAB

Genome wide association studies have been very successful*

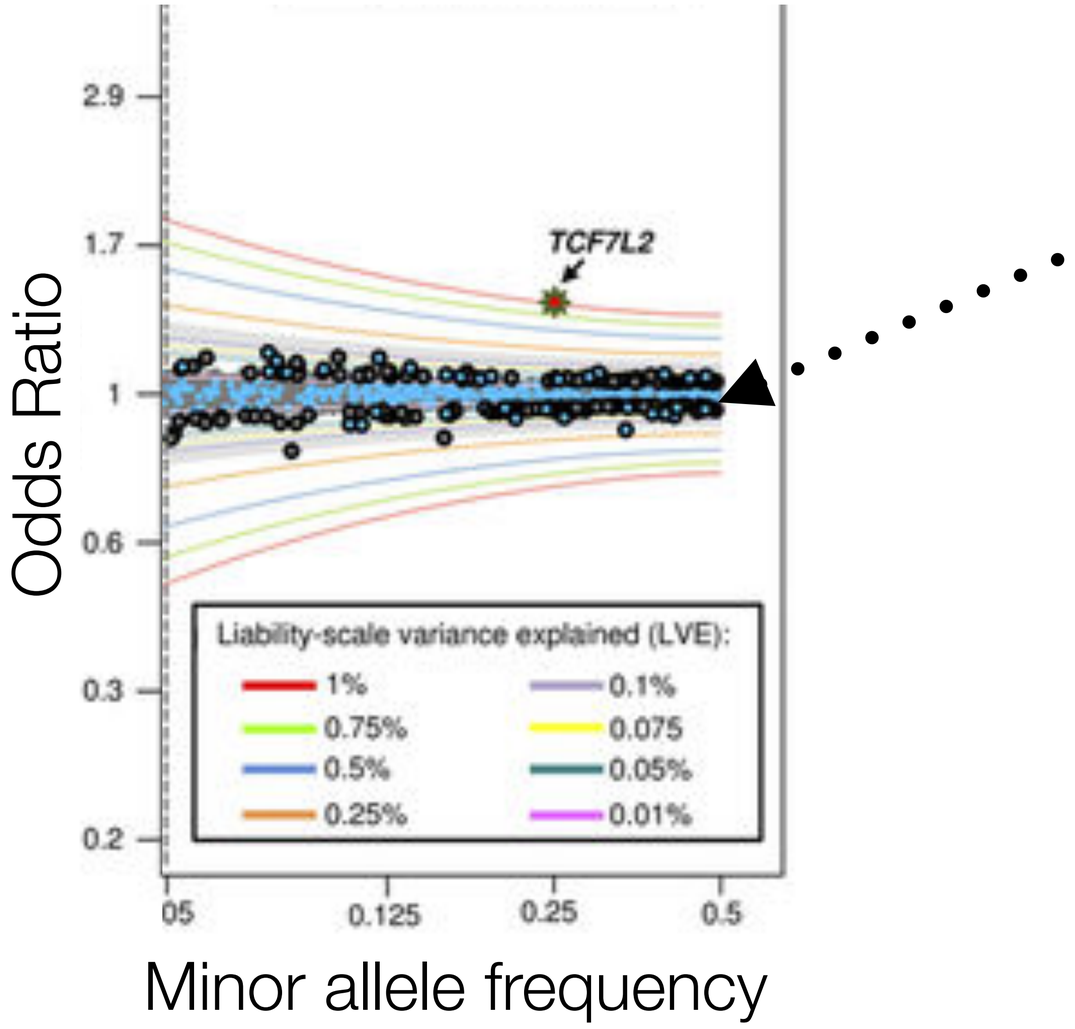
Why the "*"?



RIVASLAB

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

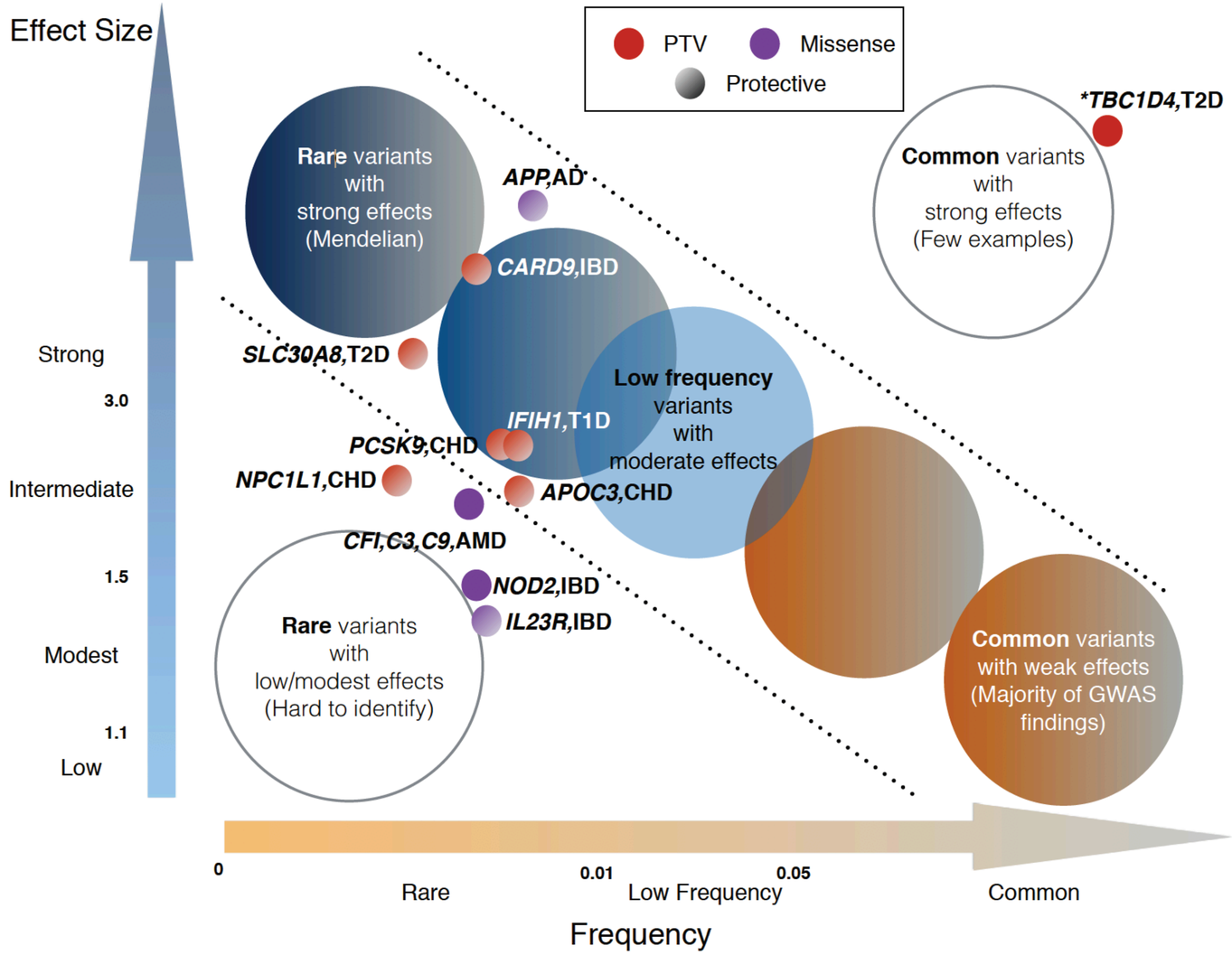
Why the "*"?



Tiny effect sizes for all associated variants

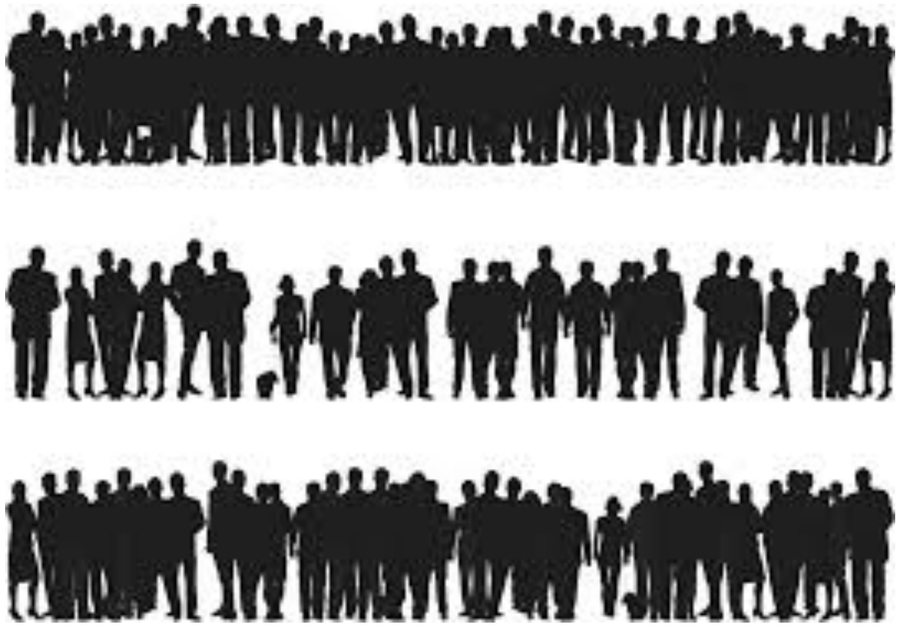
RIVASLAB

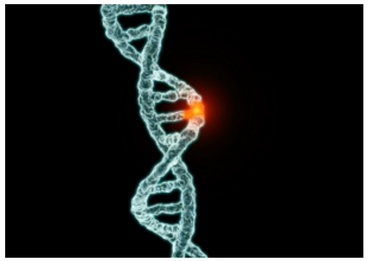
Additional signals started emerging from rare variant studies



RIVASLAB

“Experiments of nature” that protect can guide selection of drug targets

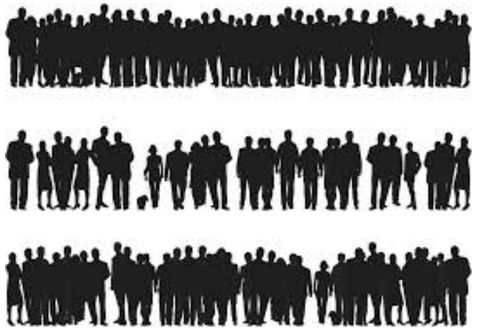


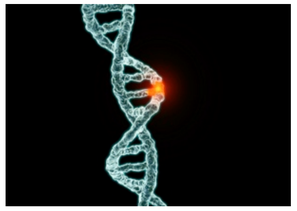


Lower risk for disease

Examples of protective mutations

PCSK9 for LDL and MI
Nav 1.7 for pain
CARD9 for Crohn's disease and ulcerative colitis
RNF186 for ulcerative colitis
CCR5 for HIV

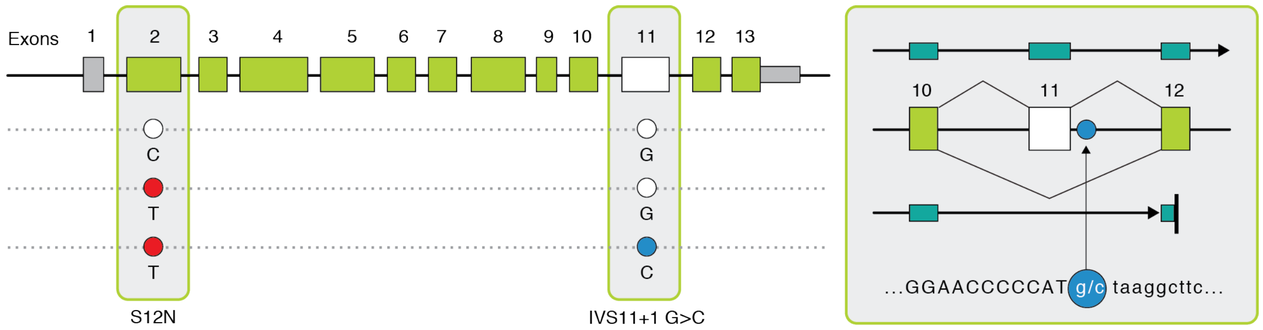




Lower risk for disease

Rare, strong acting alleles provide interpretation of the GWAS results

- Splice variant in *CARD9* cause premature truncating of protein and ***strongly protects*** against the development of Crohn's disease and ulcerative colitis ($p < 10^{-16}$).

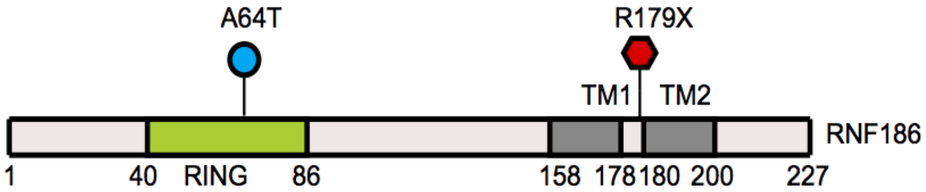


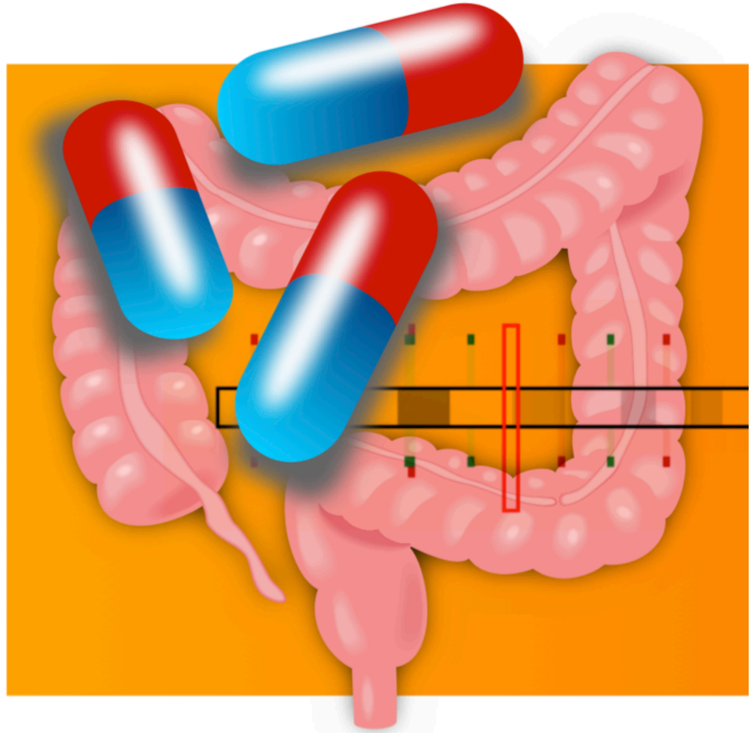
Rivas et al., *Nature Genetics* 2011.

Rare, strong acting alleles provide interpretation of the GWAS results

- Protective loss-of-function variant in *RNF186* found to confer protection against ulcerative colitis (3-fold protective effect).
- Protective genetic variants reveal process that is:
  - **safe** (naturally occurs in healthy adults)
  - **effective** (proven to reduce risk of disease).

Found in Iceland and Finland





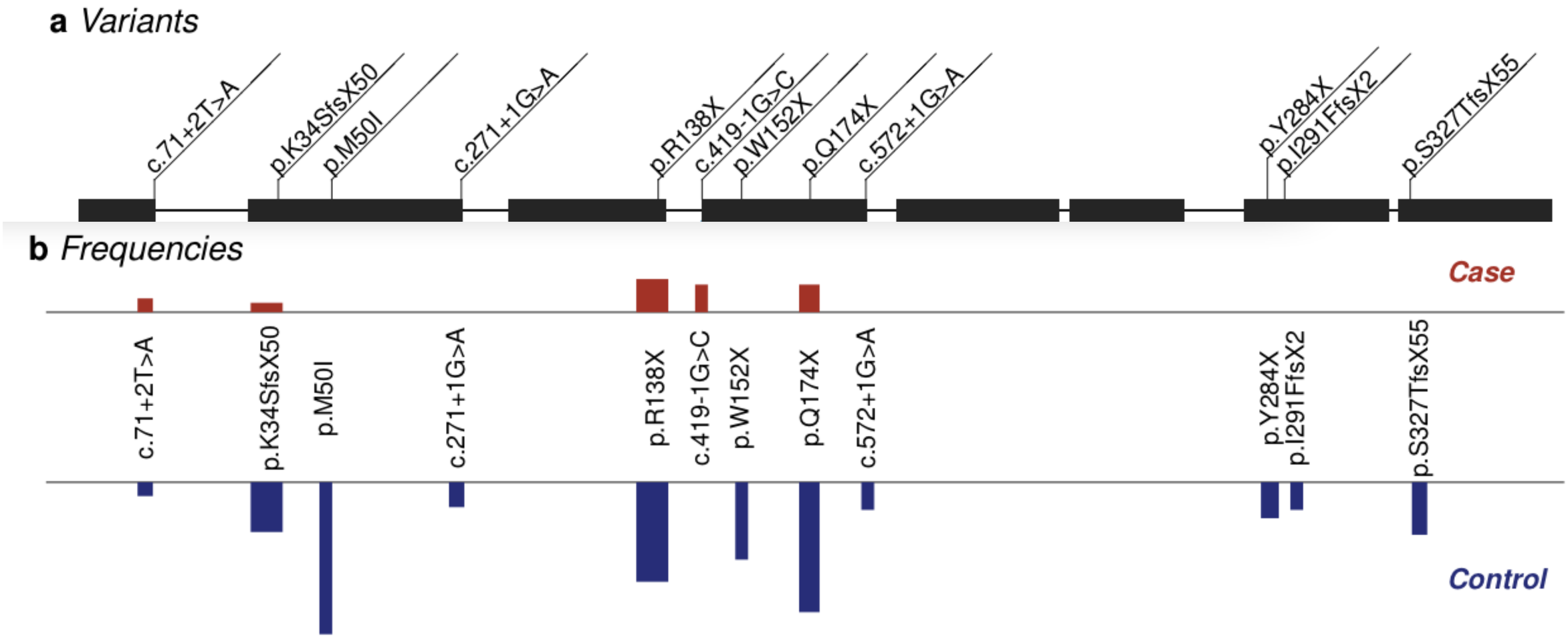
Rivas et al., *Nature Communications* 2016.

nature genetics

Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes

Jason Flannick[1–3], Gudmar Thorleifsson[4], Nicola L Beer[1,5], Suzanne B R Jacobs[1], Niels Grarup[6], Noël P Burtt[1], Anubha Mahajan[7], Christian Fuchsberger[8], Gil Atzmon[9,10], Rafn Benediktsson[11], John Blangero[12], Don W Bowden[13–16], Ivan Brandslund[17,18], Julia Brosnan[19], Frank Burslem[20], John Chambers[21–23], Yoon Shin Cho[24], Cramer Christensen[25], Desirée A Douglas[26], Ravindranath Duggirala[12], Zachary Dymek[1], Yossi Farjoun[1], Timothy Fennell[1], Pierre Fontanillas[1], Tom Forsén[27,28], Stacey Gabriel[1], Benjamin Glaser[29,30], Daniel F Gudbjartsson[4], Craig Hanis[31], Torben Hansen[6,32], Astradur B Hreidarsson[11], Kristian Hveem[33], Erik Ingelsson[7,34], Bo Isomaa[35,36], Stefan Johansson[37–39], Torben Jørgensen[40–42], Marit Eika Jørgensen[43], Sekar Kathiresan[1,44–46], Augustine Kong[4], Jaspal Kooner[22,23,47], Jasmina Kravic[48], Markku Laakso[49], Jong-Young Lee[50], Lars Lind[51], Cecilia M Lindgren[1,7], Allan Linneberg[40,41,52], Gisli Masson[4], Thomas Meitinger[53], Karen L Mohlke[54], Anders Molven[37,55,56], Andrew P Morris[7,57], Shobha Potluri[58], Rainer Rauramaa[59,60], Rasmus Ribel-Madsen[6], Ann-Marie Richard[19], Tim Rolph[19], Veikko Salomaa[61], Ayellet V Segrè[1,2], Hanna Skärstrand[26], Valgerdur Steinthorsdottir[4], Heather M Stringham[8], Patrick Sulem[4], E Shyong Tai[62–64], Yik Ying Teo[62,65–68], Tanya Teslovich[8], Unnur Thorsteinsdottir[4,69], Jeff K Trimmer[19], Tiinamaija Tuomi[27,35], Jaakko Tuomilehto[70–72], Fariba Vaziri-Sani[26], Benjamin F Voight[1,73,74], James G Wilson[75], Michael Boehnke[8], Mark I McCarthy[5,7,76], Pål R Njølstad[1,37,77], Oluf Pedersen[6], Go-T2D Consortium[78], T2D-GENES Consortium[78], Leif Groop[48,79], David R Cox[58], Kari Stefansson[4,69] & David Altshuler[1–3,44,45,80,81]

In aggregate, **12 loss of function** mutations associated with **65% decreased risk** of **T2D**



RIVASLAB

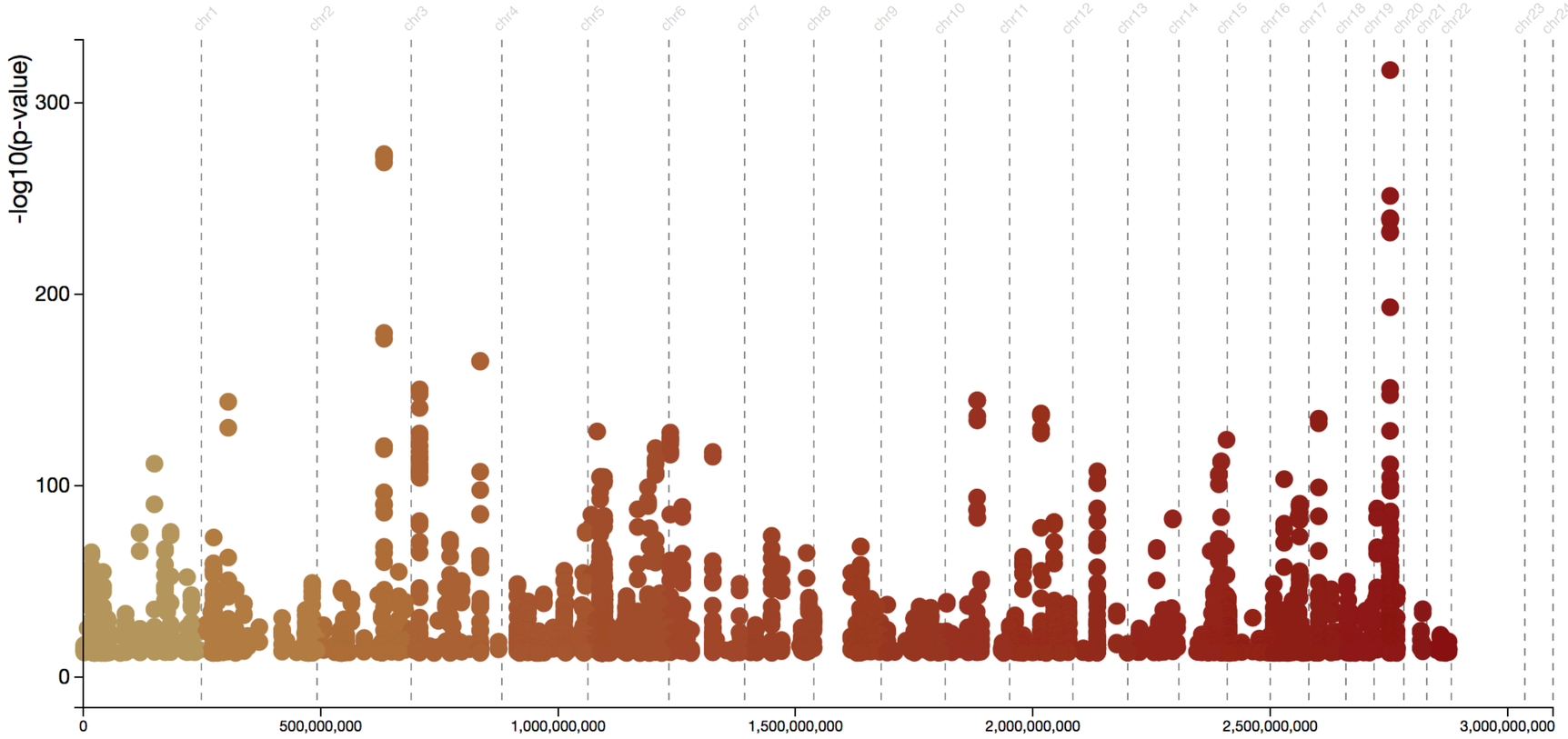
Why the need for **Stan** in **Human Genetics**?

Parameters became more exciting (and informative)…

Location, scale, correlation, effect size

p-value is not informative especially…

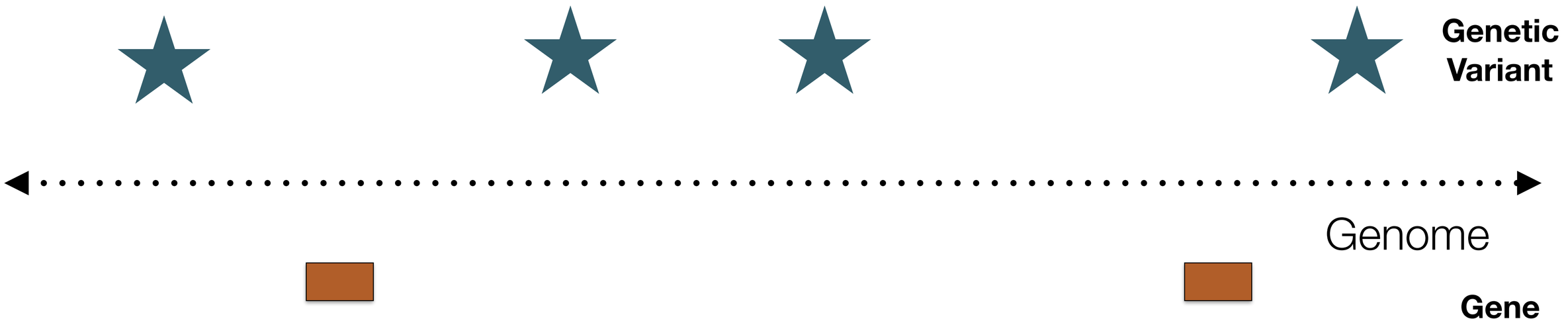
when you realize the entire genome is “associated” to disease



Note: only variants with a p-value less than 0.001 are included in the manhattan plot.

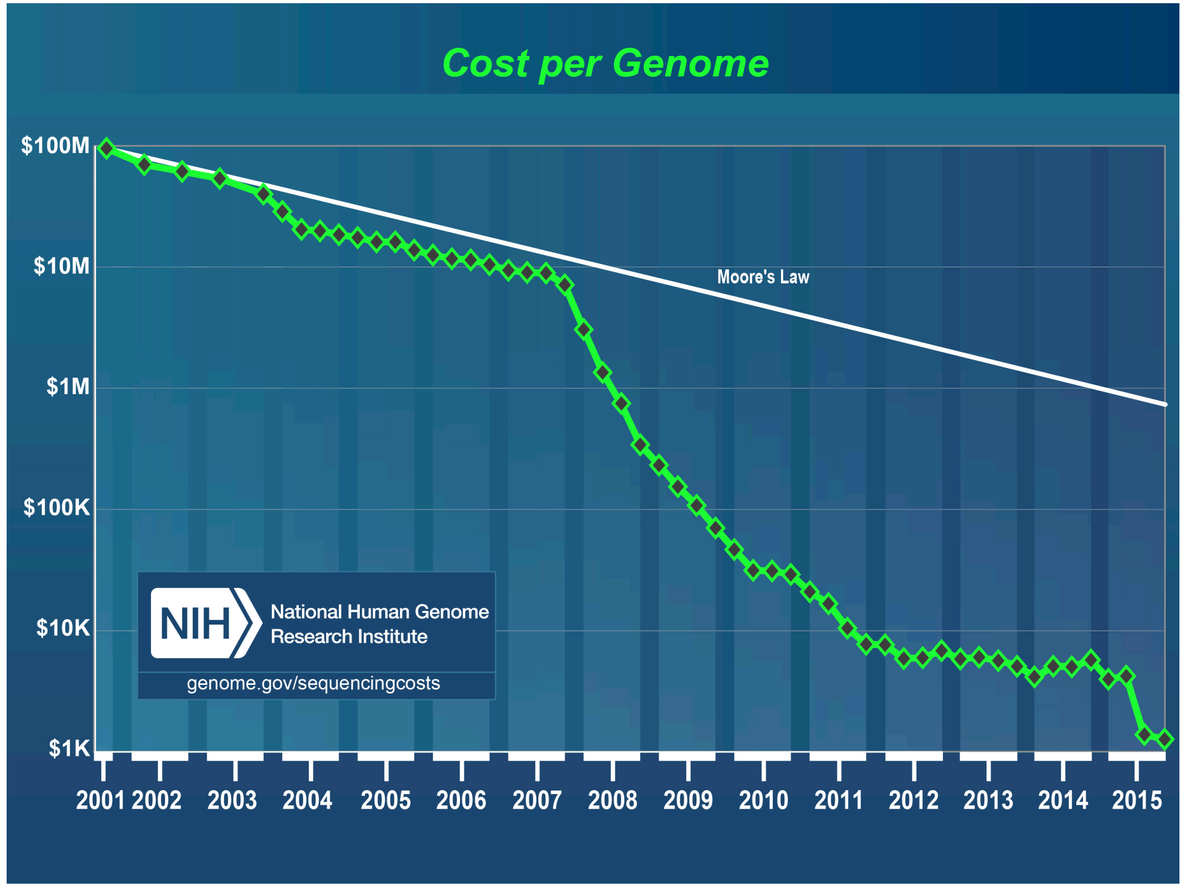
RIVASLAB

Age of common variants

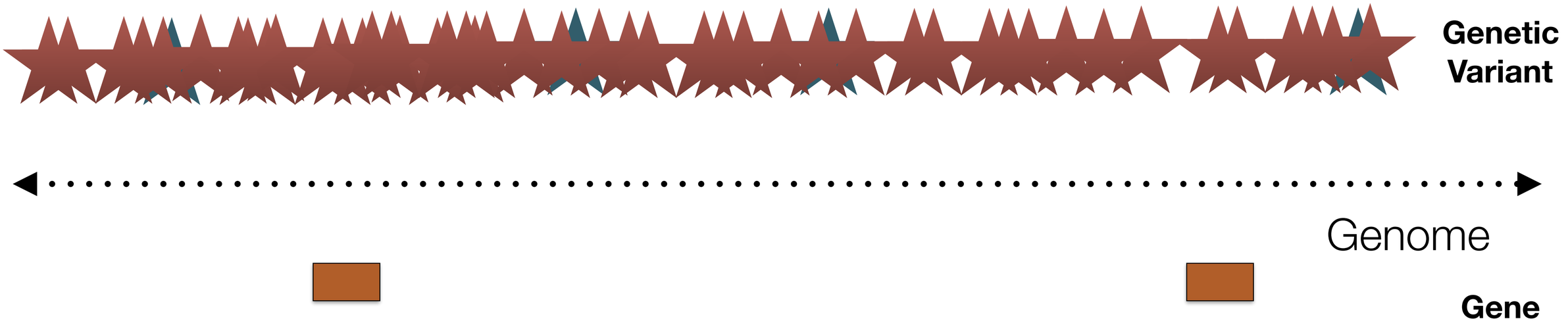


Technologies transforming biomedicine

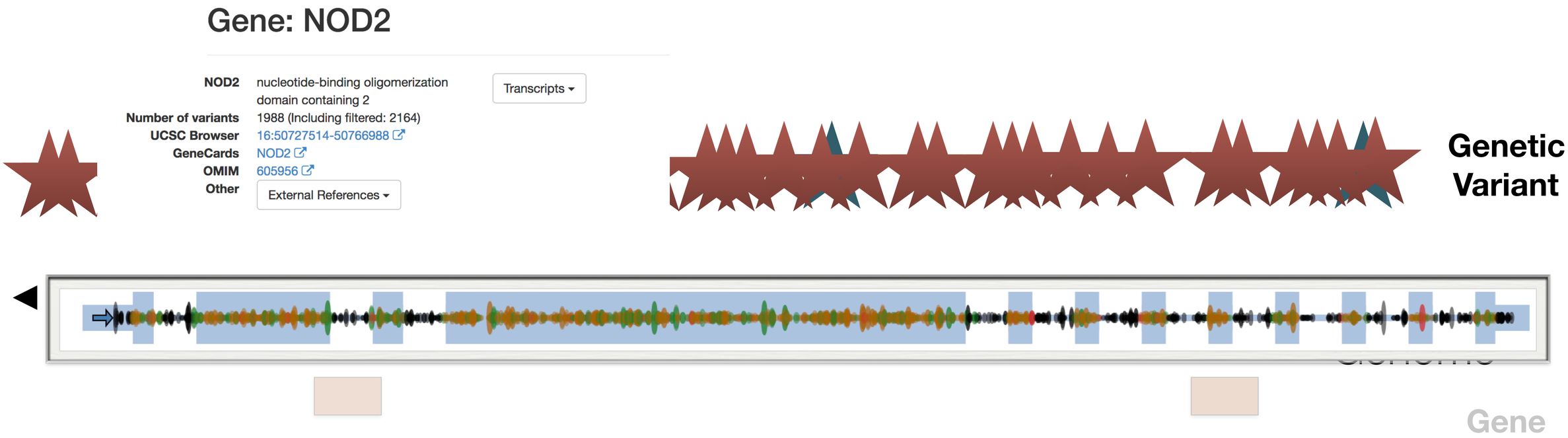
Cost of sequencing has plummeted over the past 15 years



Age of rare variants



Age of rare variants

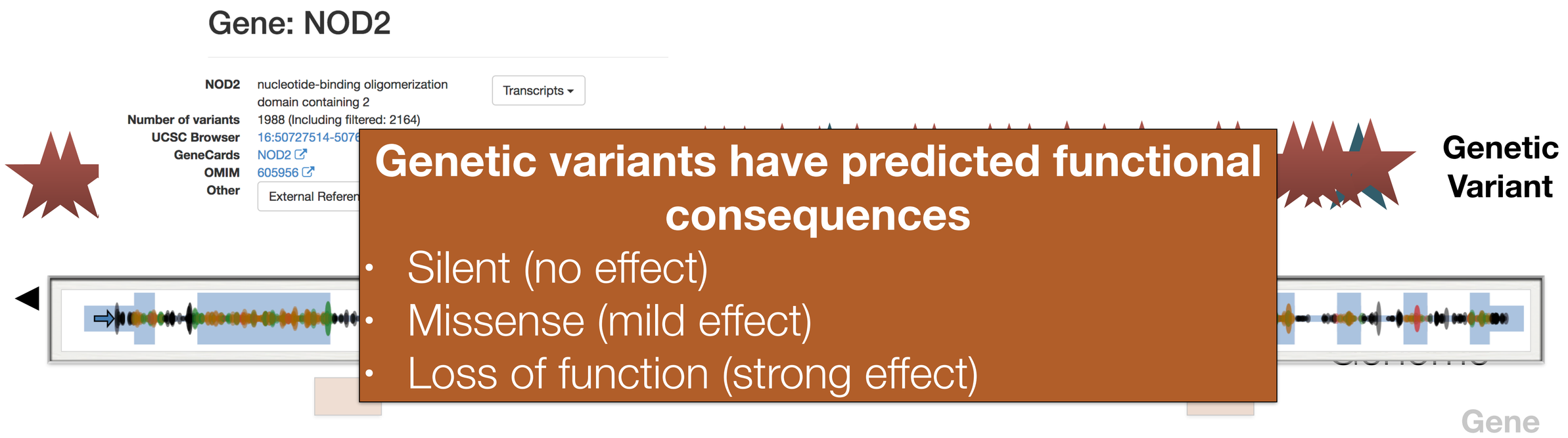


https://ibd.broadinstitute.org/gene/ENSG00000167207

Age of rare variants

https://ibd.broadinstitute.org/gene/ENSG00000167207

Age of rare variants

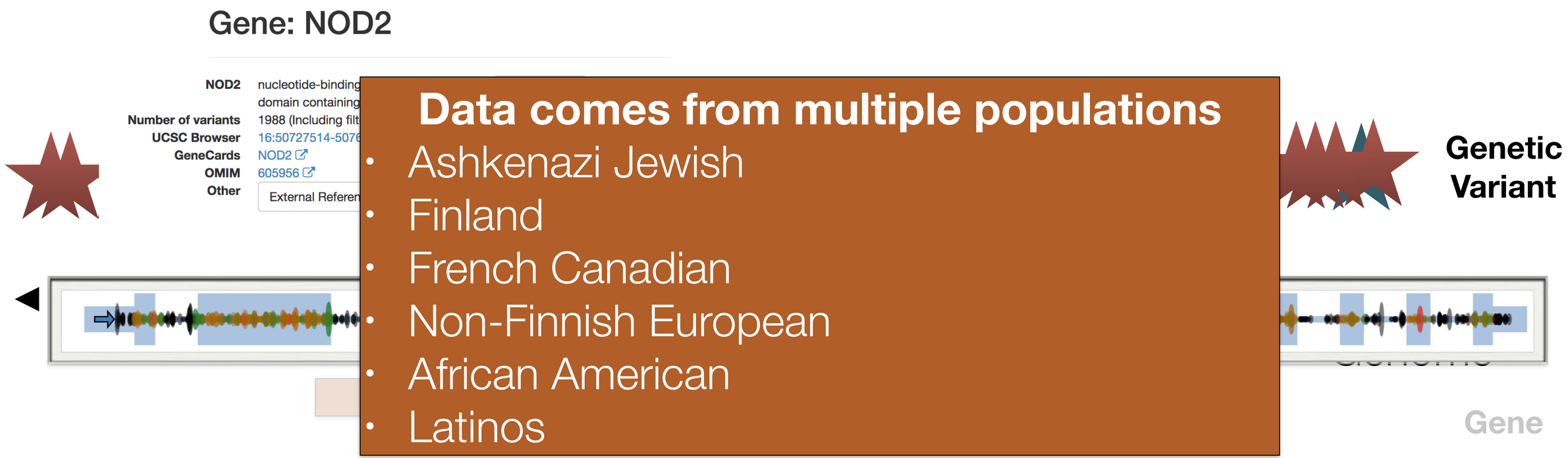


Genetic variants have predicted functional consequences

- Silent (no effect)
- Missense (mild effect)
- Loss of function (strong effect)

https://ibd.broadinstitute.org/gene/ENSG00000167207

Age of rare variants



Data comes from multiple populations

- Ashkenazi Jewish
- Finland
- French Canadian
- Non-Finnish European
- African American
- Latinos

https://ibd.broadinstitute.org/gene/ENSG00000167207

RIVASLAB

Data from genetic study

Crohn's Disease				
BETA	**SE**	**Annotation**	**Gene**	**Population**
1.2	0.03	lof	NOD2	AJ
1.5	0.9	mis	IL23R	AJ
-1.1	0.006	syn	NOD2	NFE
…	…	…	…	…

Single population

~50 genetic variants

Modeling of rare variants in *NOD2*, *IL23R*, and *CARD9*

brms facilitates multilevel modeling with “meta-analysis” type data

\hat{B}| se(SE) ~ Annotation + Gene + Annotation | Gene, prior = gaussian, prior = “horseshoe(3)”, algorithm = “meanfield, fullrank, sampling”)

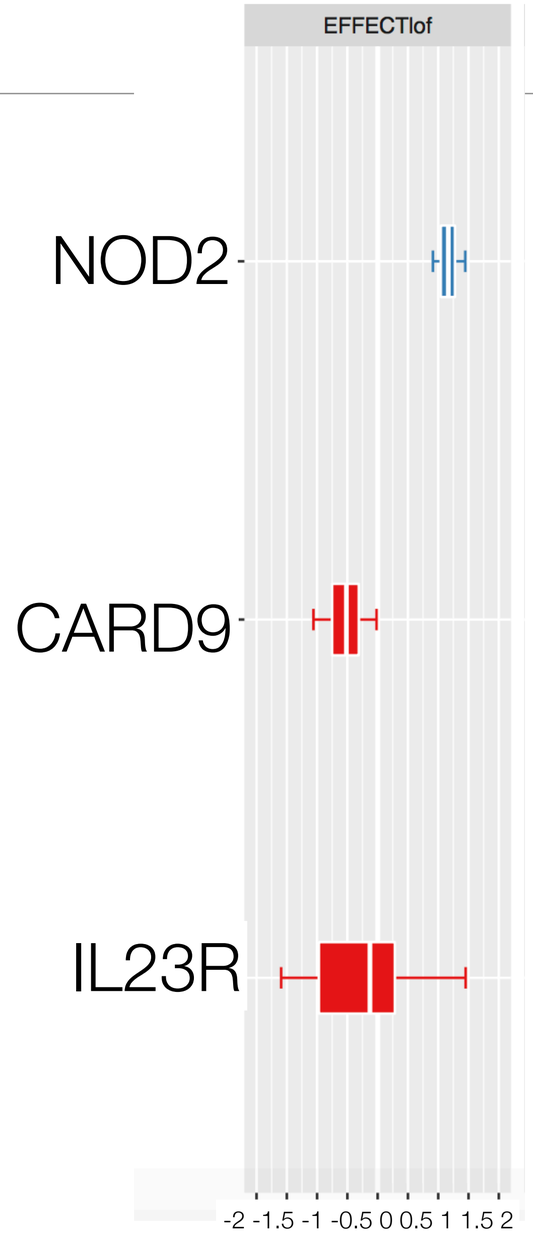
Population level effects — assumes to be the same across population

Group level effects — varying across grouping variables

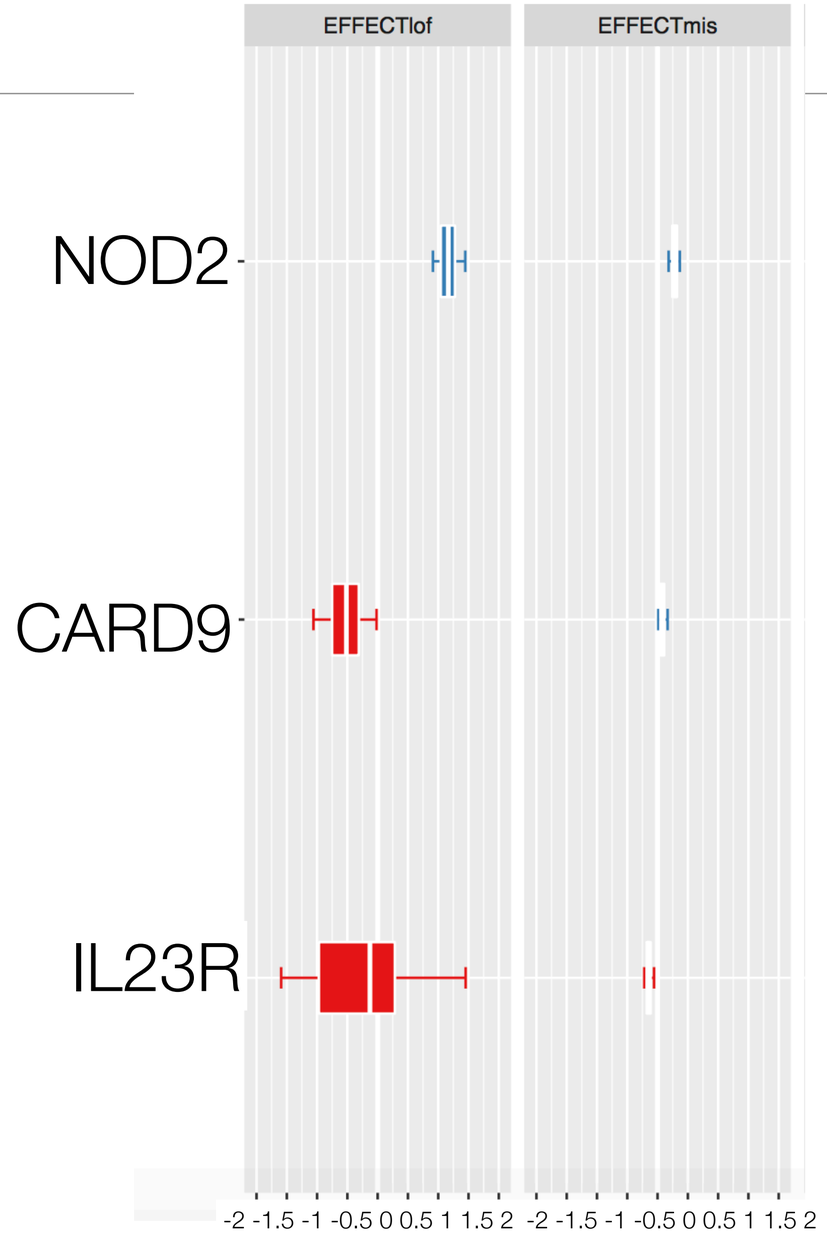
Single population

~50 genetic variants

RIVASLAB



RIVASLAB



Group-Level Effects:
~GENE (Number of levels 3

	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample
sd(EFFECTlof)	1.38	0.14	1.11	1.70	1000
sd(EFFECTmis)	0.20	0.02	0.16	0.25	875
sd(EFFECTsplice)	0.38	0.14	0.18	0.71	1000
sd(EFFECTsyn)	0.08	0.02	0.05	0.12	1000

RIVASLAB

Population cohorts enables the study of **shared genetics** with other diseases and phenotypes

First large-scale population biobank cohort widely available to researchers is **UK Biobank** with approximately **500,000 participants**

- Prospective cohort study
- 500k individuals (337k European British)
- Large number of phenotypes + Array genotypes

- Hospital in-patient record
- Verbal questionnaire
- Cancer and death registry
- Imaging (MRI) derived features
- Biomarker measurements
- Lifestyle measurements

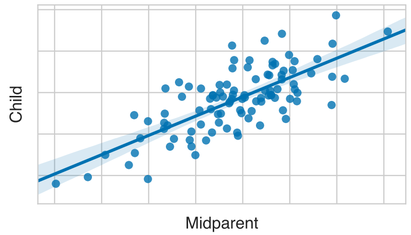
Bayesian Mixture Model to estimate genetic parameters

Estimate genetic parameters using GWAS summary statistics

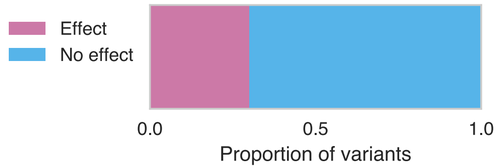
Phenotype 1		Phenotype 2	
BETA	SE	BETA	SE
1.2	0.03	1.15	0.04
1.5	0.9	1.3	0.1
-1.1	0.006	-1.2	0.05
…	…	…	…

MVPMM

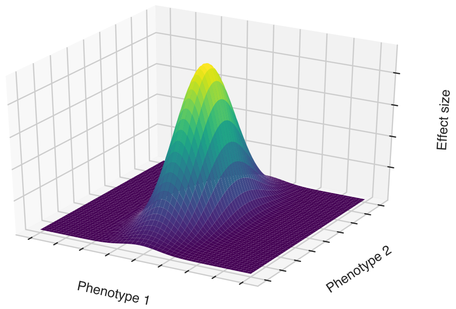
- h^2 : heritability



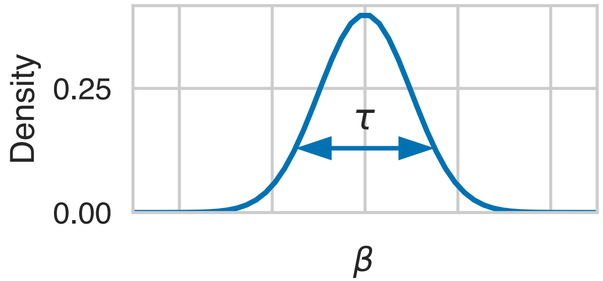
- π : membership/ polygenicity



- Ω : genetic correlation



- τ : spread/scale of effects



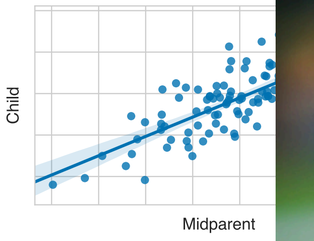
RIVASLAB

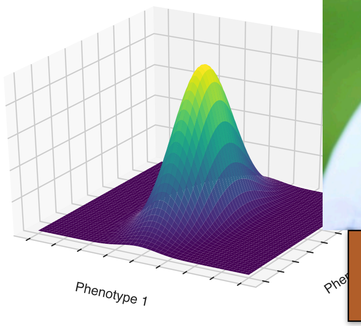
Bayesian Mixture Model to estimate genetic parameters

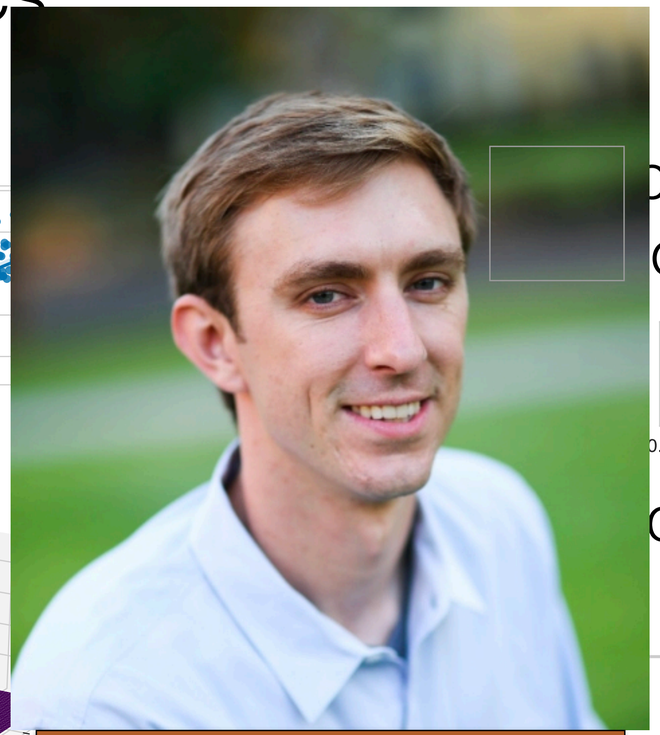
Estimate genetic parameters using GWAS summary statistics

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
…	…	…	…

- h^2 : heritability
- Ω : genetic correlation

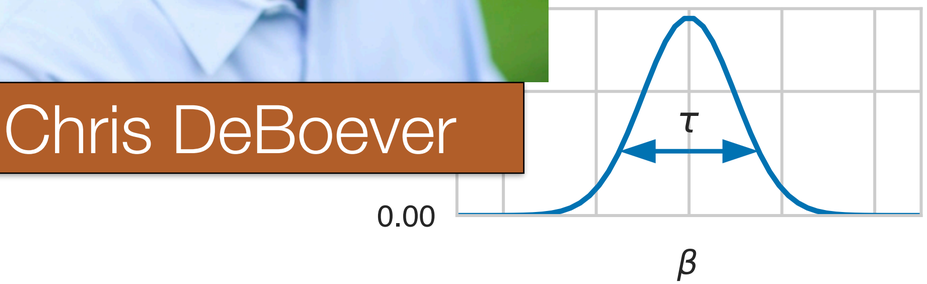






Chris DeBoever

Genetic parameter estimates across diseases, biomarkers, and lifestyle measures in the UK Biobank



RIVASLAB

Multivariate Polygenic Mixture Model (MVPMM)

- 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

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

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

Null component

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

Correlated errors

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

Non-null component

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

Correlated genetic effects

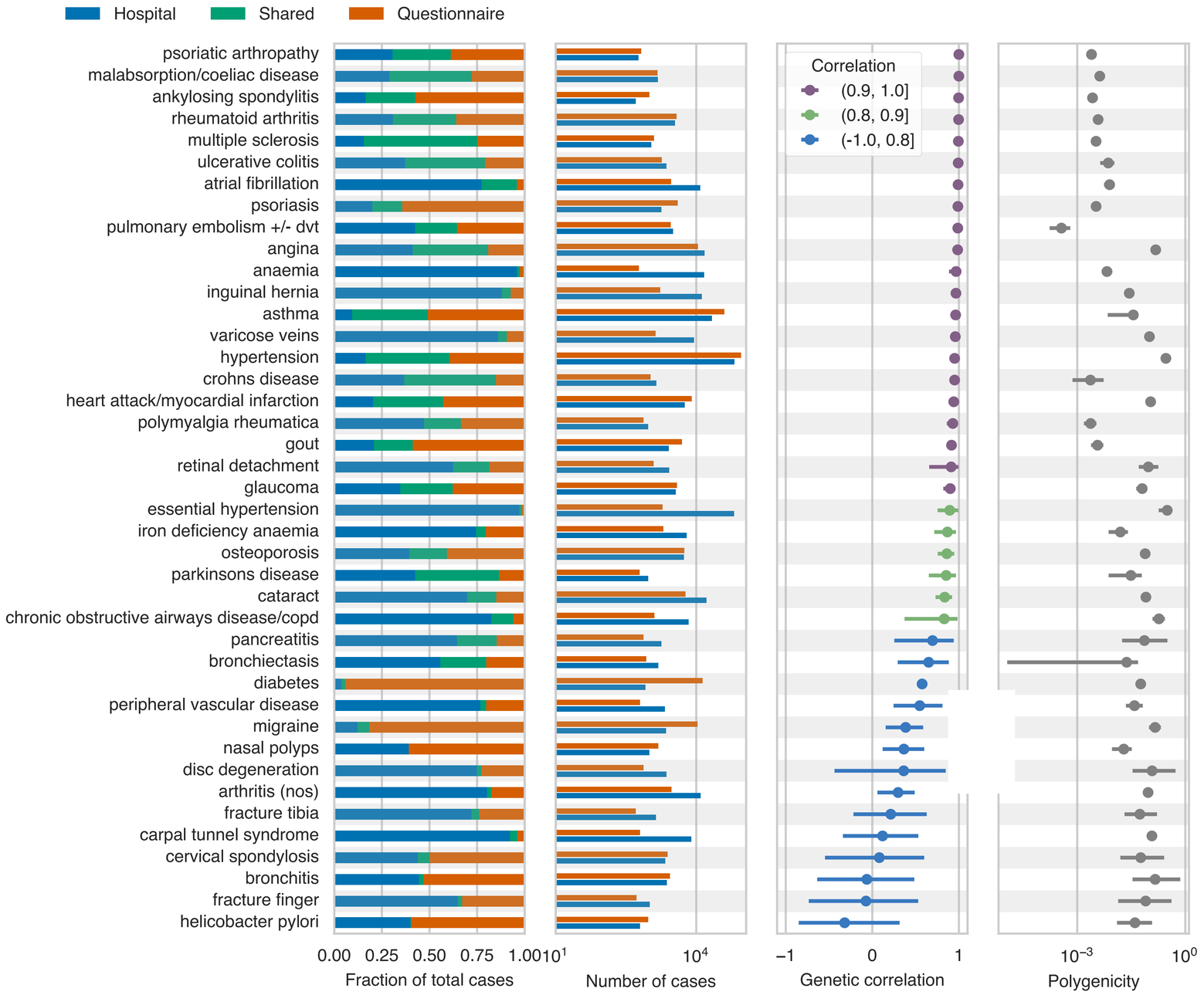
$$\Sigma_{\Omega} = \text{diag}\left(\tau\right) \cdot \Omega \cdot \text{diag}\left(\tau\right)$$

$$\pi : \text{mixture proportion}$$

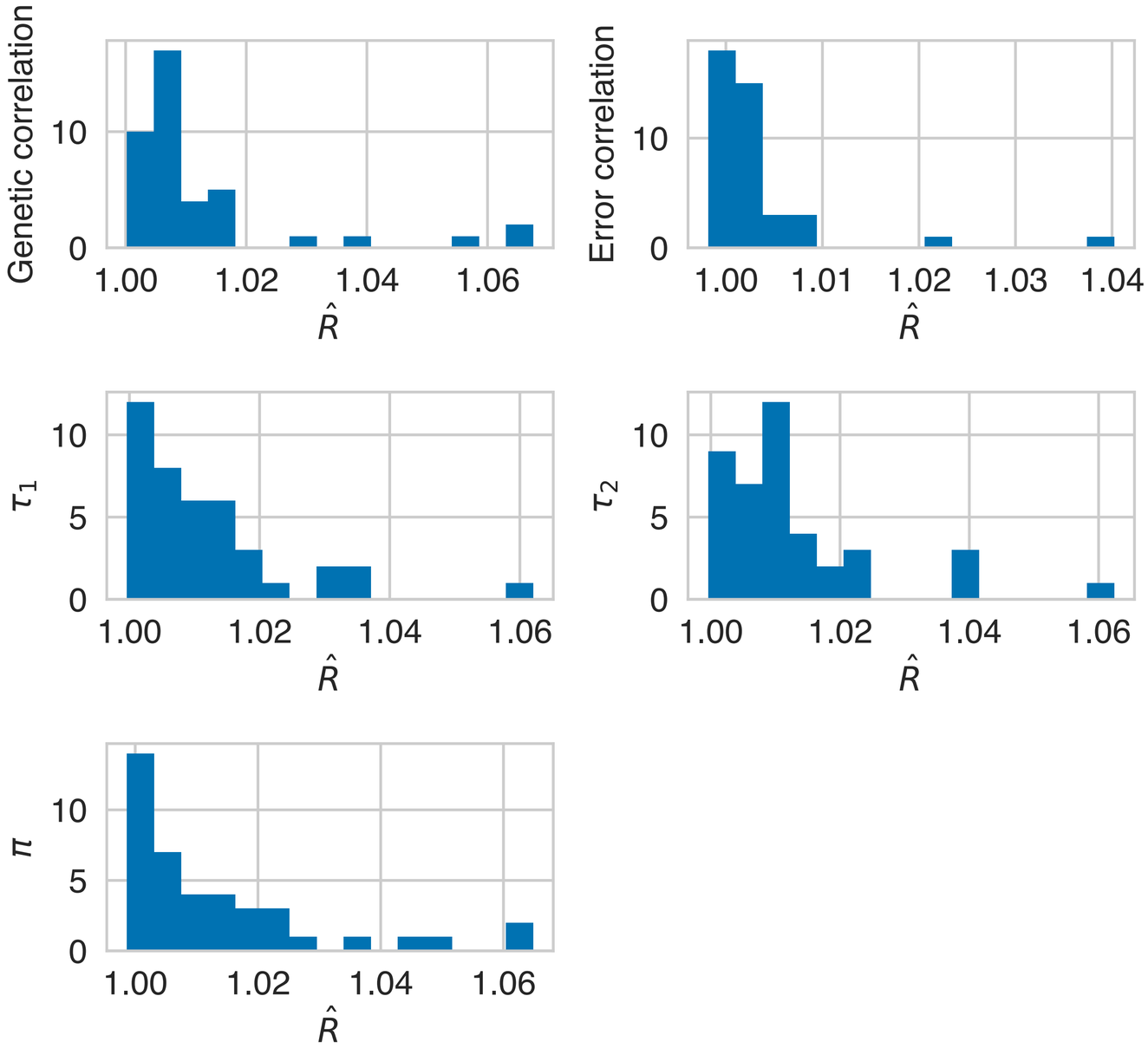
RIVASLAB

Hospital record vs. verbal questionnaire digital phenotyping

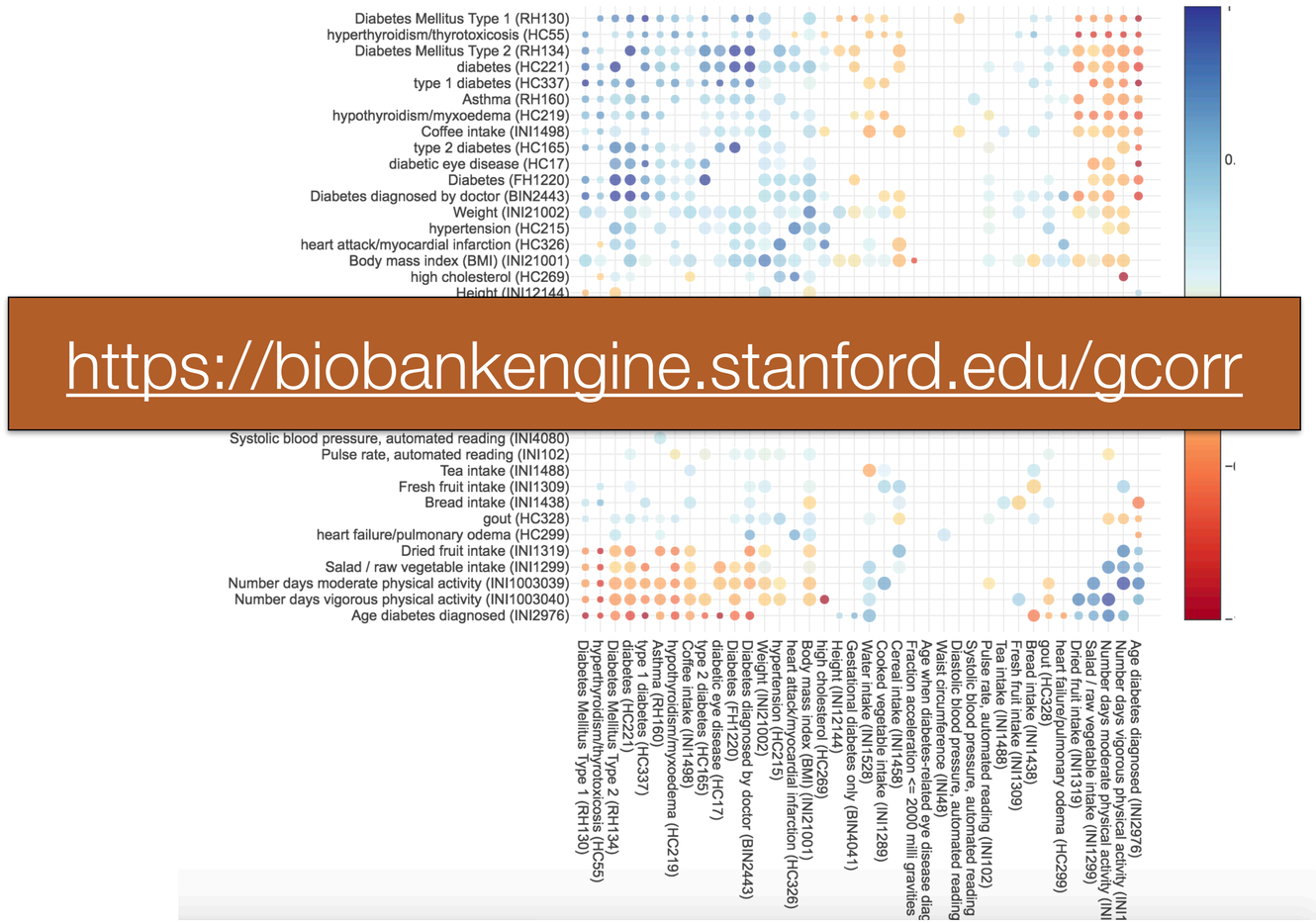
- We have defined binary disease traits using a combination of hospital records and verbal questionnaire answers
- Question: to what extent do these two phenotyping methods agree?
- Approach
  - Run GWAS using only cases from hospital records or only cases from questionnaire results
  - Calculate genetic correlation between GWAS results for the “same” traits


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

RIVASLAB


Genetic correlation
Error correlation
τ_1
τ_2
π
$\hat{R}$
0
5
10
1.00
1.01
1.02
1.03
1.04
1.06

Estimate sharing of genetic effects across diseases



RIVASLAB

Bayesian Mixture Models for estimating sex-related contributions to variance

Previous studies have identified sex differences in heritability (Rawlik et al. 2017, Ober et al. 2008)

We developed a new approach to estimate sex-related contributions to variance from GWAS summary statistics using a Bayesian Mixture Model (BMM), implemented in Stan

Applied to quantitative traits in UK Biobank

Bayesian Mixture Models for estimating sex-related contributions to variance

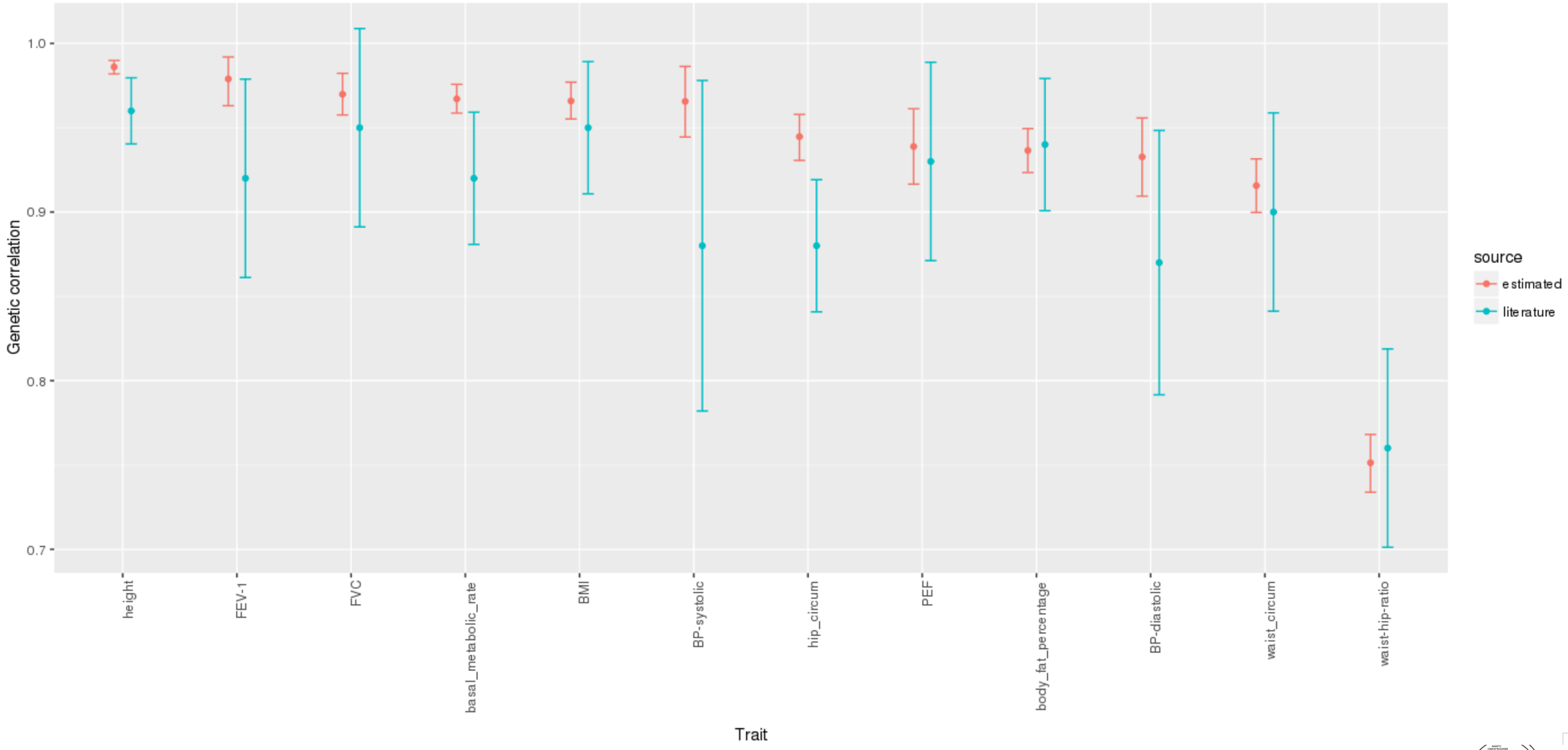
Estimated component fractions and sex-specific variance-covariance matrix, and genetic correlation

Results similar to literature values

Emily Flynn

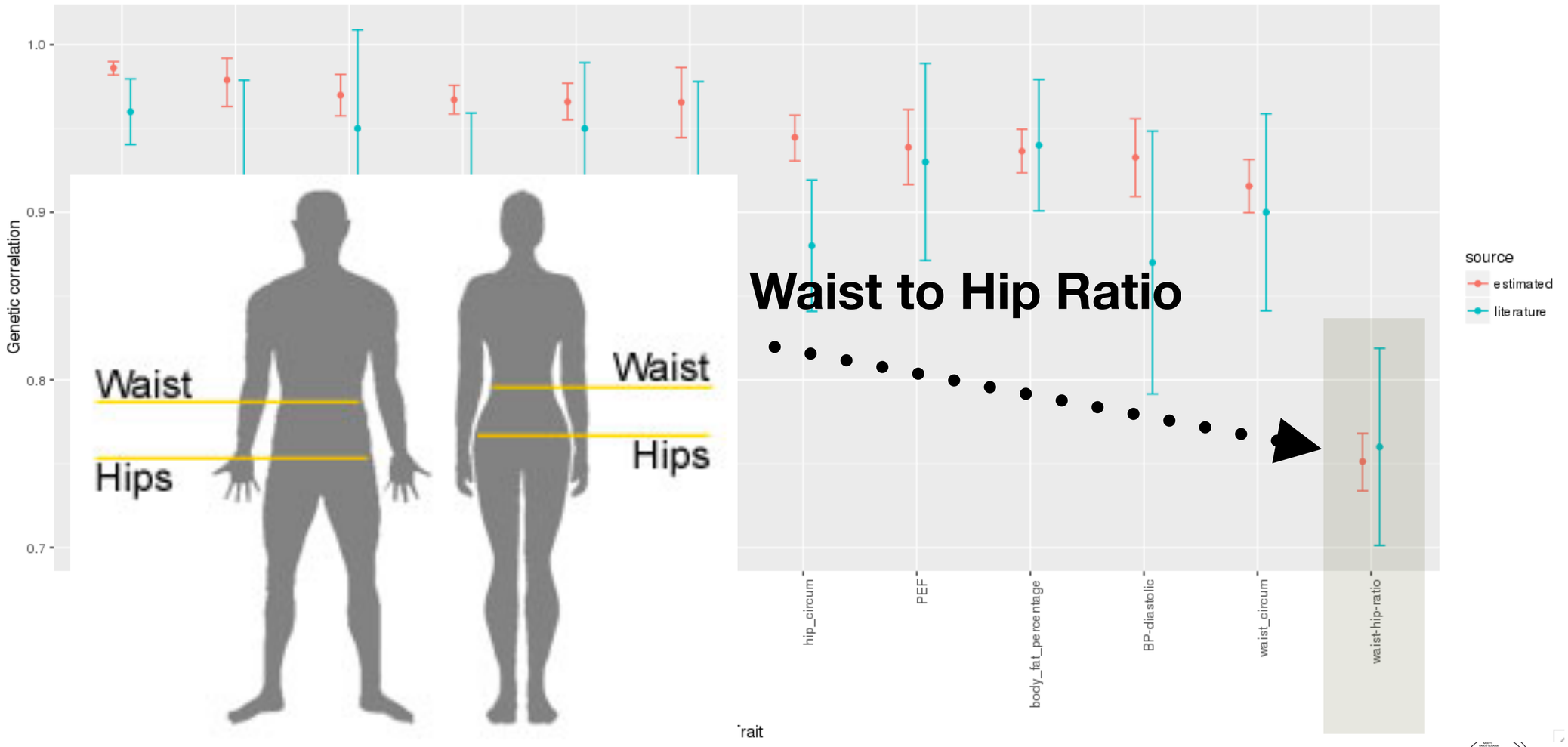
RIVASLAB

Genetic correlation estimates are similar to the literature



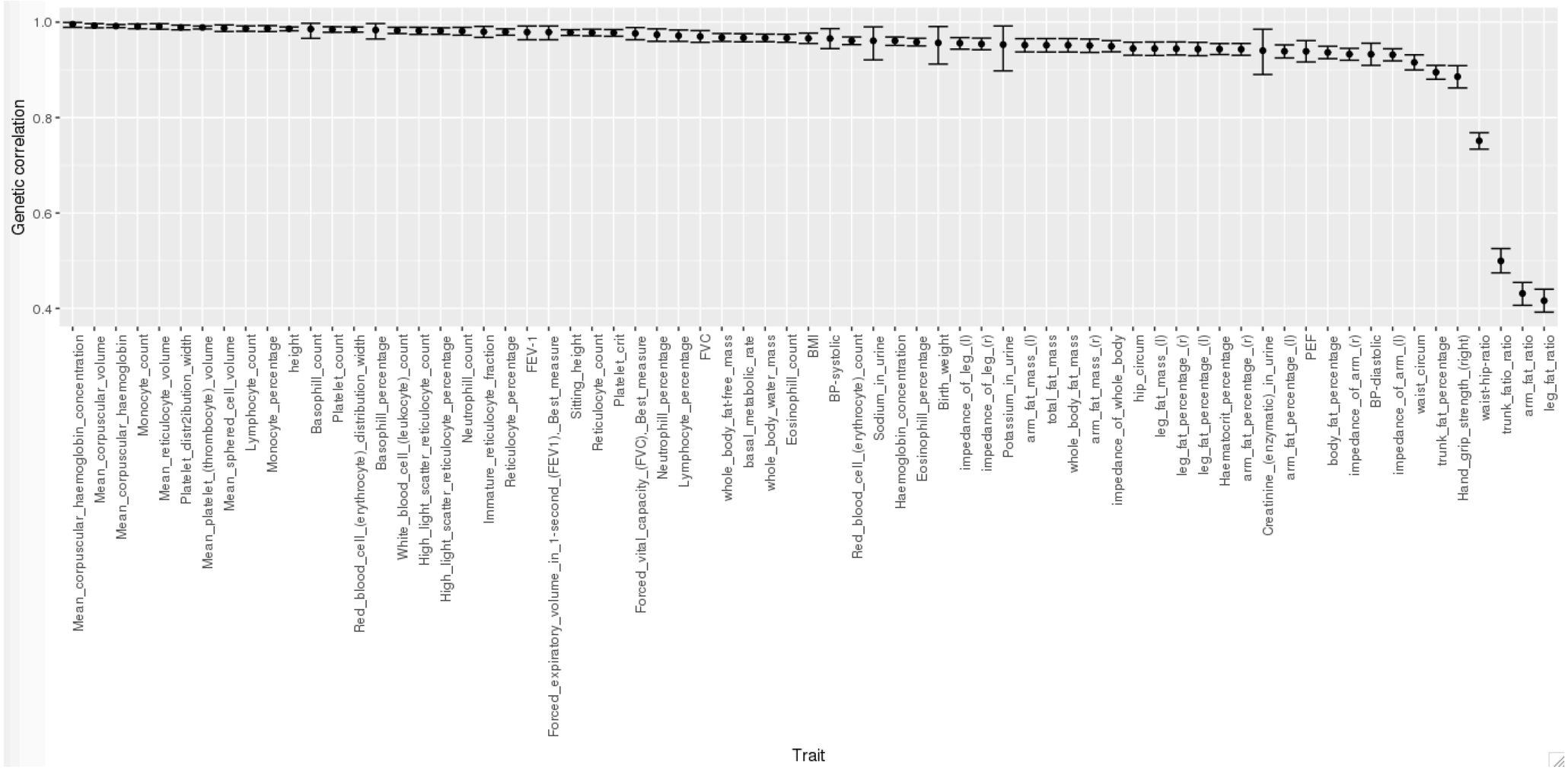
RIVASLAB

Genetic architecture of WHR differs across men and women



RIVASLAB

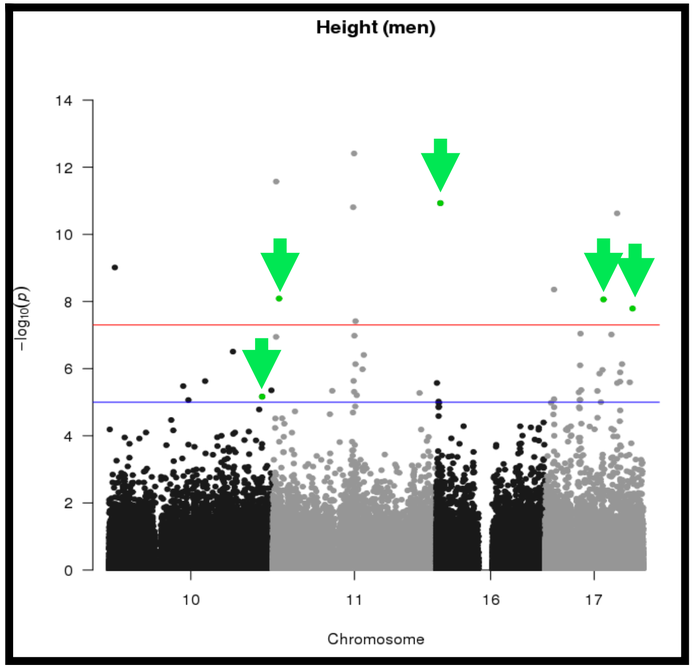
Sex affects many genetic correlation across many traits

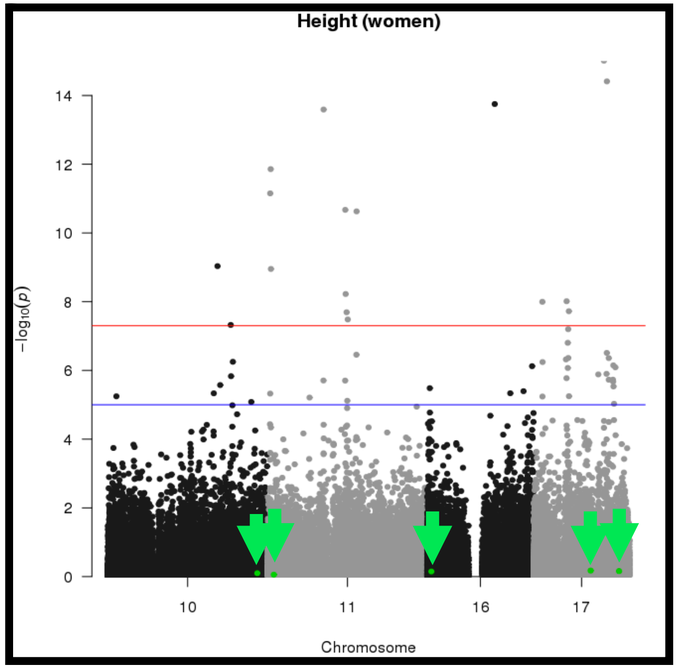


RIVASLAB

Identification of sex-specific effects

- Mixture model with four components: (1) null/no effect, (2) female-specific effect, (3) male-specific effect, and (4) effects in both sexes
- Estimated component fractions and parameters, assigned SNPs based on posterior probability
- Identified five male-specific SNPs in height





Male-specific SNPs for height

SNP	posterior	gene	B_f	B_m	p_f	p_m
Affx-58847809	1.000	CTBP2	-0.0296	-0.476	0.795	6.84E-06
Affx-35292772	0.960	HBB	-0.0465	-2.14	0.881	8.17E-09
Affx-80252875	1.000	NLRC3	-0.128	-2.52	0.710	1.17E-11
Affx-80298599	1.000	SGCA	0.121	-1.57	0.668	8.68E-09
Affx-80286332	1.000	DNAI2	-0.123	-1.96	0.692	1.62E-08

RIVASLAB

Identification of sex-specific effects

	Female-spec.	Male-specific	Literature genes[1,2]
Waist-hip ratio	20	0	VEGFA, COBLL1, ADAMTS9
Leg fat ratio	28	0	ADAMTS3, ID4
Arm fat ratio	11	4	SLC12A2
Trunk fat ratio	44	0	PCSK5, ADAMTS14, ADAMTS17

Conclusions

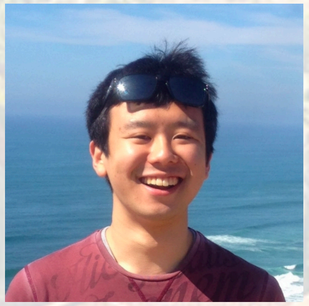
Ample opportunities in linking

Medical population data  Human genetic data  Data science technologies

STAN *has been a powerful statistical modeling language for our applications*

Oliver Bear Don't Walk IV
Natural Language Processing (NLP)
Clinical Note Tagger

Chris DeBoever
Gene discovery, statistical methods development

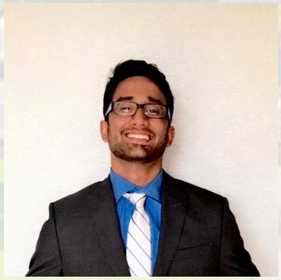


Yosuke Tanigawa
Pathway identification, Risk modeling, and methods development

Matthew Aguirre
Ancestry inference, Global Biobank Engine

Johanne Justesen
Disease progression
Temporal data

Julia Olivieri
Inference of HLA allelotypes to human diseases



Guhan Venkataraman
Integrating genomic data with clinical notes

Anna Cichonska
Machine Learning, Drug-protein-genetics target repositioning and interaction inference

The Lab

RIVASLAB