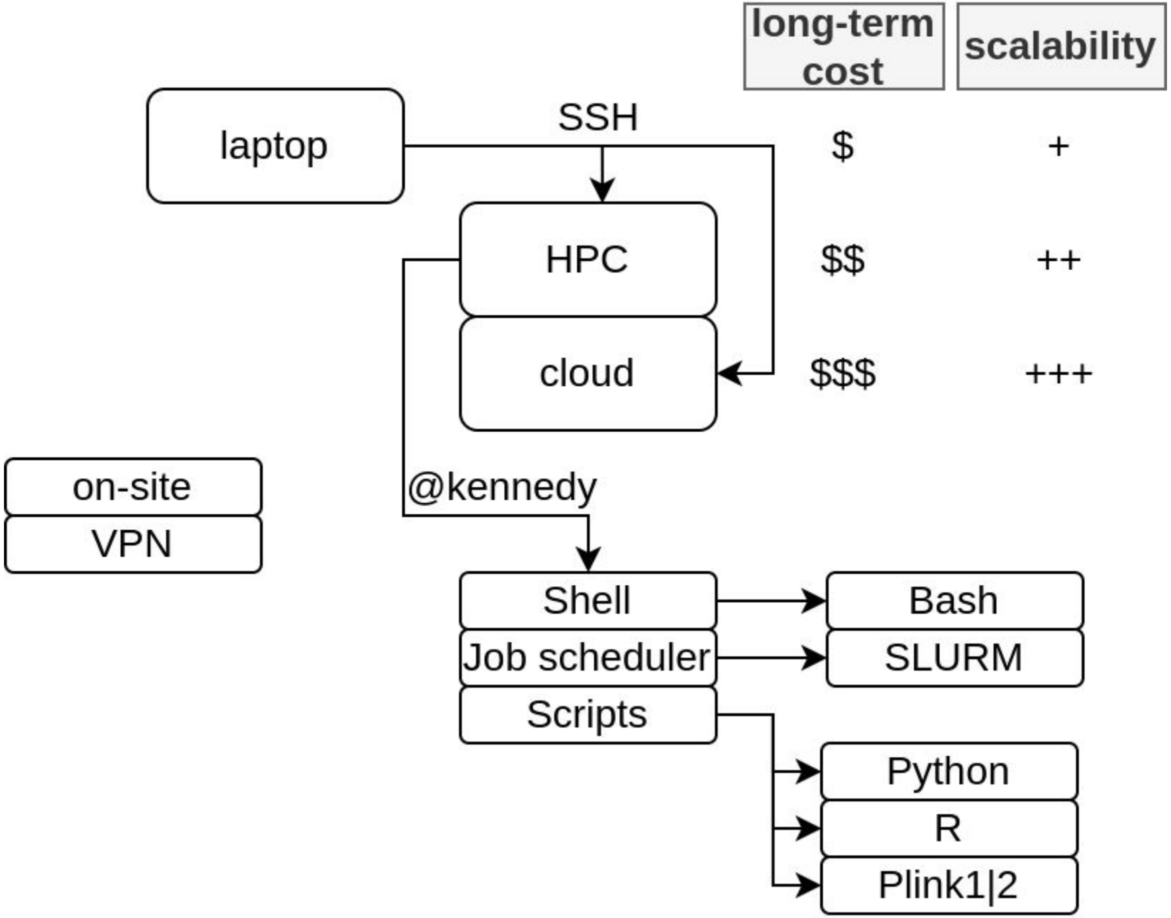
Lecture 1 - GWAS Statistics + Polygenic Risk Scores (PRS)

Wed, Mar 20, 9-10AM

Recap

HPC | GWAS

Recap: HPC + SSH



	long-term cost	scalability
laptop	$	+
HPC	$$	++
cloud	$$$	+++

SSH

on-site

VPN

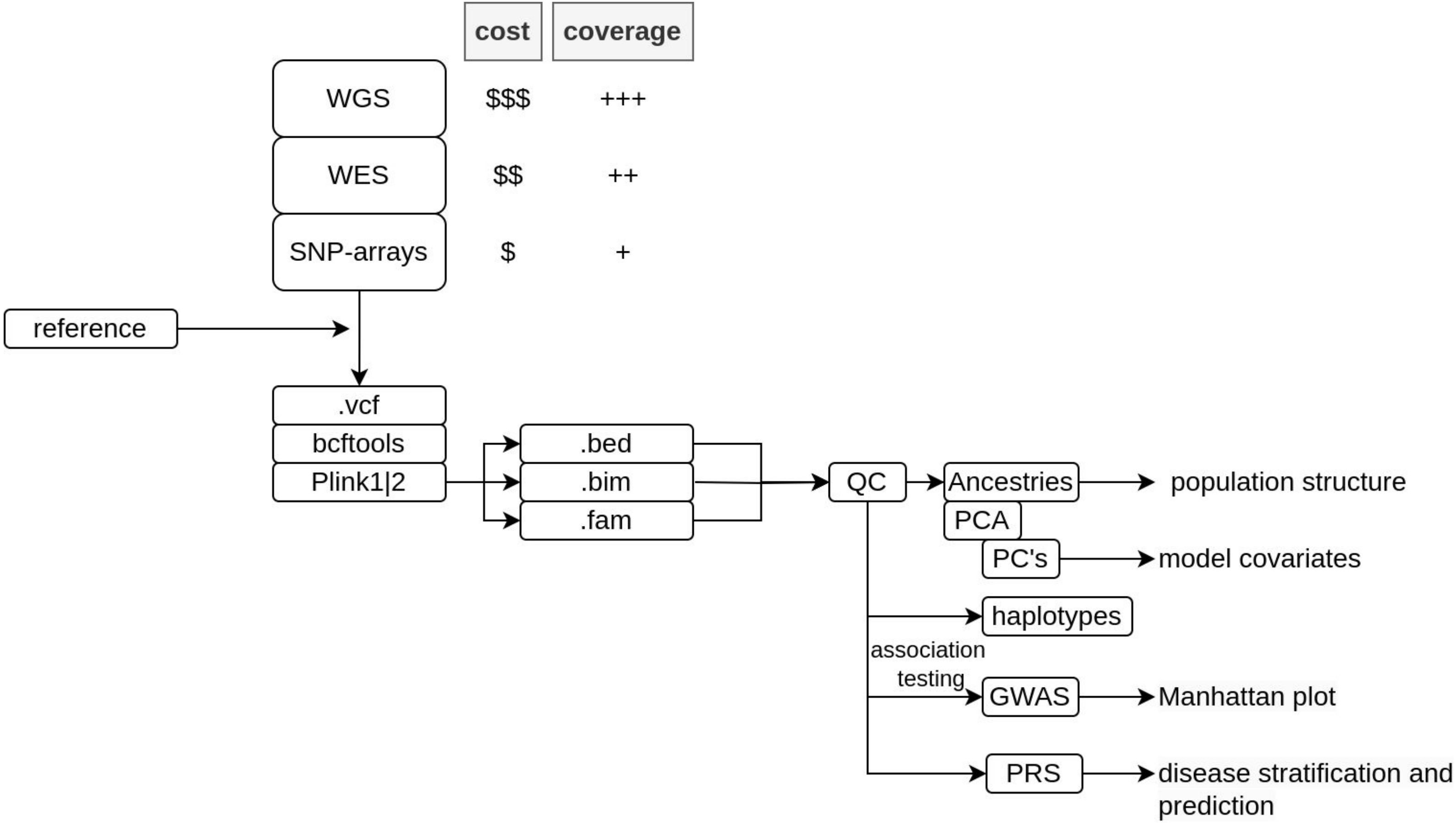
@kennedy

Shell → Bash

Job scheduler → SLURM

Scripts → Python, R, Plink1|2

Recap: GWAS



	cost	coverage
WGS	$$$	+++
WES	$$	++
SNP-arrays	$	+

reference

.vcf

bcftools

Plink1|2

.bed

.bim

.fam

QC

Ancestries

population structure

PCA

PC's

model covariates

haplotypes

association testing

GWAS

Manhattan plot

PRS

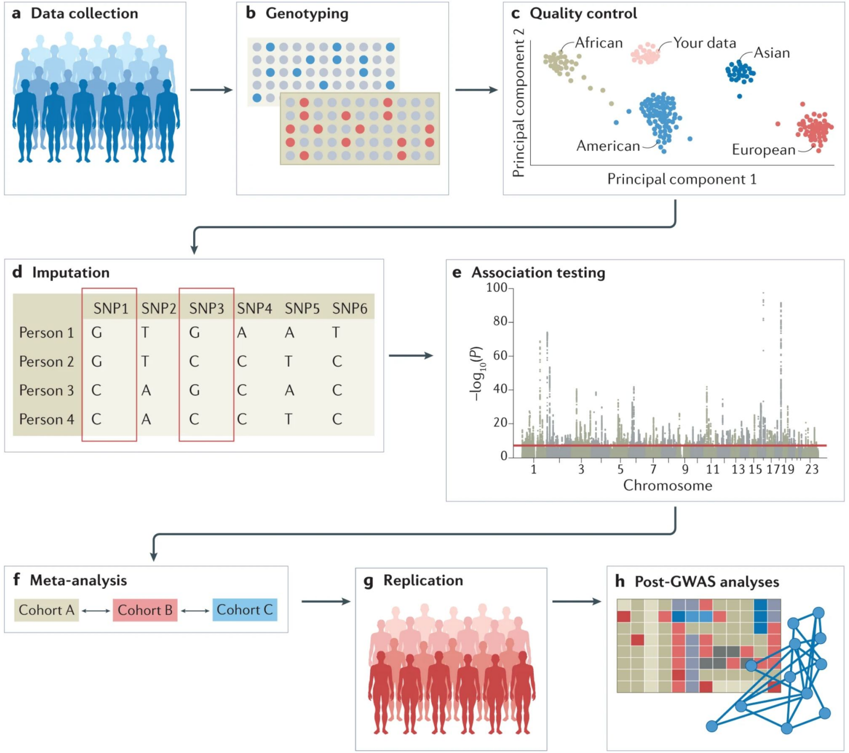
disease stratification and prediction

GWAS

Genome-wide Association Studies

Genome-wide Association studies (GWAS)

Single nucleotide polymorphism (SNP): This is a variation in a single nucleotide (i.e., **A**, **C**, **G**, or **T**) that occurs at a specific position in the genome. A SNP usually exists as two different forms (e.g., **A** vs. **T**). These different forms are called alleles. A SNP with two alleles has three different genotypes (e.g., **AA**, **AT**, and **TT**).



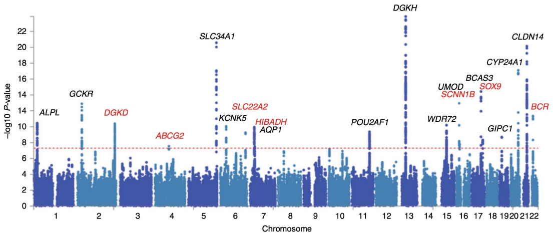
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

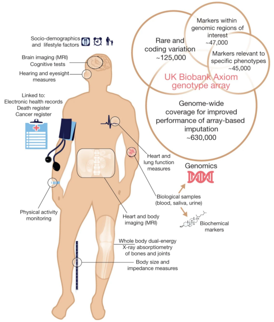
Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Data sources & repositories

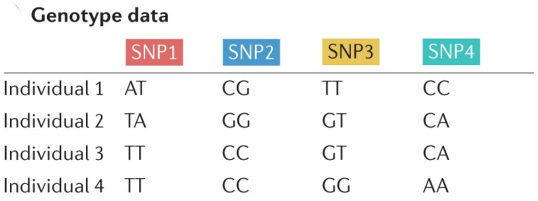
Summary statistics

Individual-level





Socio-demographics and lifestyle factors
Brain imaging (MRI) Cognitive tests
Hearing and eyesight measures
Linked to: Electronic health records Death register Cancer register
Physical activity monitoring
Rare and coding variation ~125,000
Markers within genomic regions of interest ~47,000
Markers relevant to specific phenotypes ~45,000
UK Biobank Axiom genotype array
Genome-wide coverage for improved performance of array-based imputation ~630,000
Heart and lung function measures
Genomics
Biological samples (blood, saliva, urine)
Heart and body imaging (MRI)
Biochemical markers
Whole body dual-energy X-ray absorptiometry of bones and joints
Body size and impedance measures



Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

GWAS Catalog ← general GWAS data repository

db GaP GENOTYPES and PHENOTYPES

CKD Gen CONSORTIUM ← specific for GWAS Chronic Kidney Disease

biobank uk

Bycroft *et al.* Nature (2018). doi:10.1038/s41586-018-0579-z.

Testing for associations

Genetic models

	AA	AG	GG
Additive model	0	1	2
Dominant model	0	1	1
Recessive model	0	0	1

- Additive model (ADD)
- Dominant model (DOM)
- Recessive model (REC)

biallelic SNP whose reference allele is **A** and the alternative allele is **G**.

Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Testing for associations

Contingency table

genotype	AA	AG	GG	Total
case	800	400	800	2000
control	1000	500	500	2000
Total	1800	900	1300	4000

Dominant model

genotype	**control:AA**	**case:AG/GG**	Total
case	800	1200	2000
control	1000	1000	2000
Total	1800	2200	4000

Recessive model

genotype	**control:AA/AG**	**case:GG**	Total
case	1200	800	2000
control	1500	500	2000
Total	2700	1300	4000

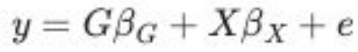
Additive model

Allele (G)	**0**	**1**	**2**	Total
case	800	400	800	2000
control	1000	500	500	2000
Total	1800	900	1300	4000

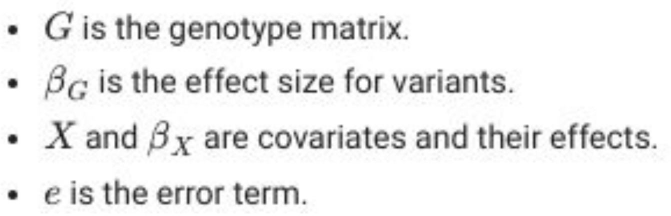
Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Testing for associations

Quantitative traits

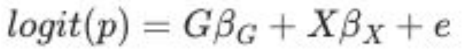


$$y = G\beta_G + X\beta_X + e$$



- G is the genotype matrix.
- β_G is the effect size for variants.
- X and β_X are covariates and their effects.
- e is the error term.

Binary traits

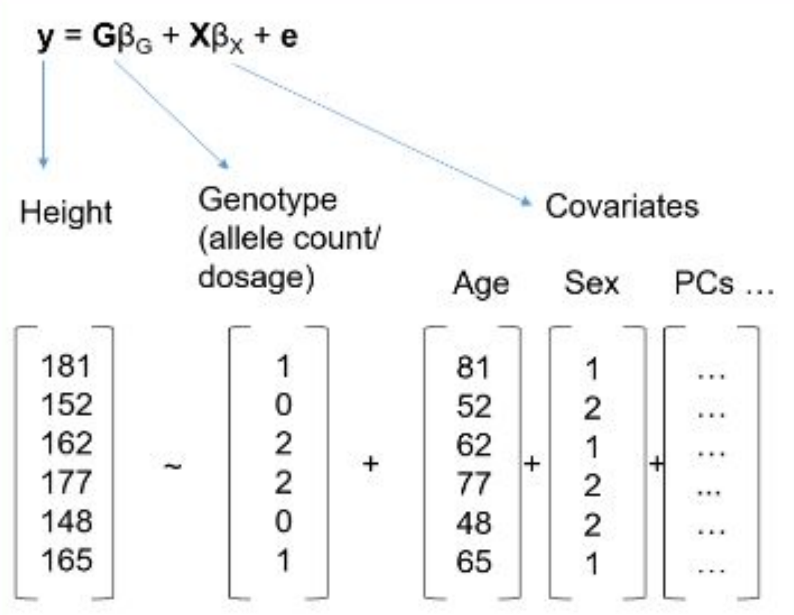


$$logit(p) = G\beta_G + X\beta_X + e$$

Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Testing for associations

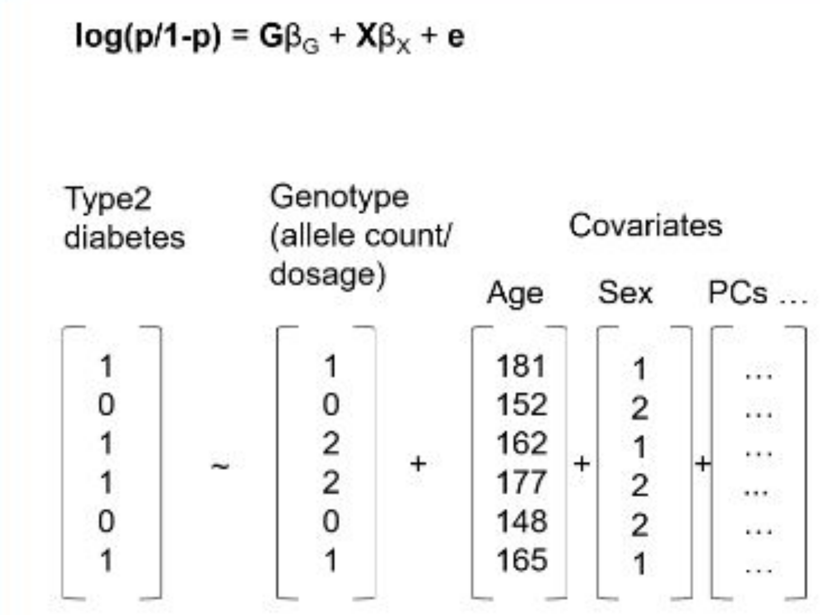
Quantitative trait



$$y = G\beta_G + X\beta_X + e$$

$$\begin{array}{ccccccccc}
\text{Height} & & \text{Genotype (allele count/ dosage)} & & \text{Covariates} & & & & \\
 & & & & \text{Age} & & \text{Sex} & & \text{PCs ...} \\
\begin{bmatrix}181\\152\\162\\177\\148\\165\end{bmatrix} & \sim & \begin{bmatrix}1\\0\\2\\2\\0\\1\end{bmatrix} & + & \begin{bmatrix}81\\52\\62\\77\\48\\65\end{bmatrix} & + & \begin{bmatrix}1\\2\\1\\2\\2\\1\end{bmatrix} & + & \begin{bmatrix}\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\end{bmatrix}
\end{array}$$

Binary trait

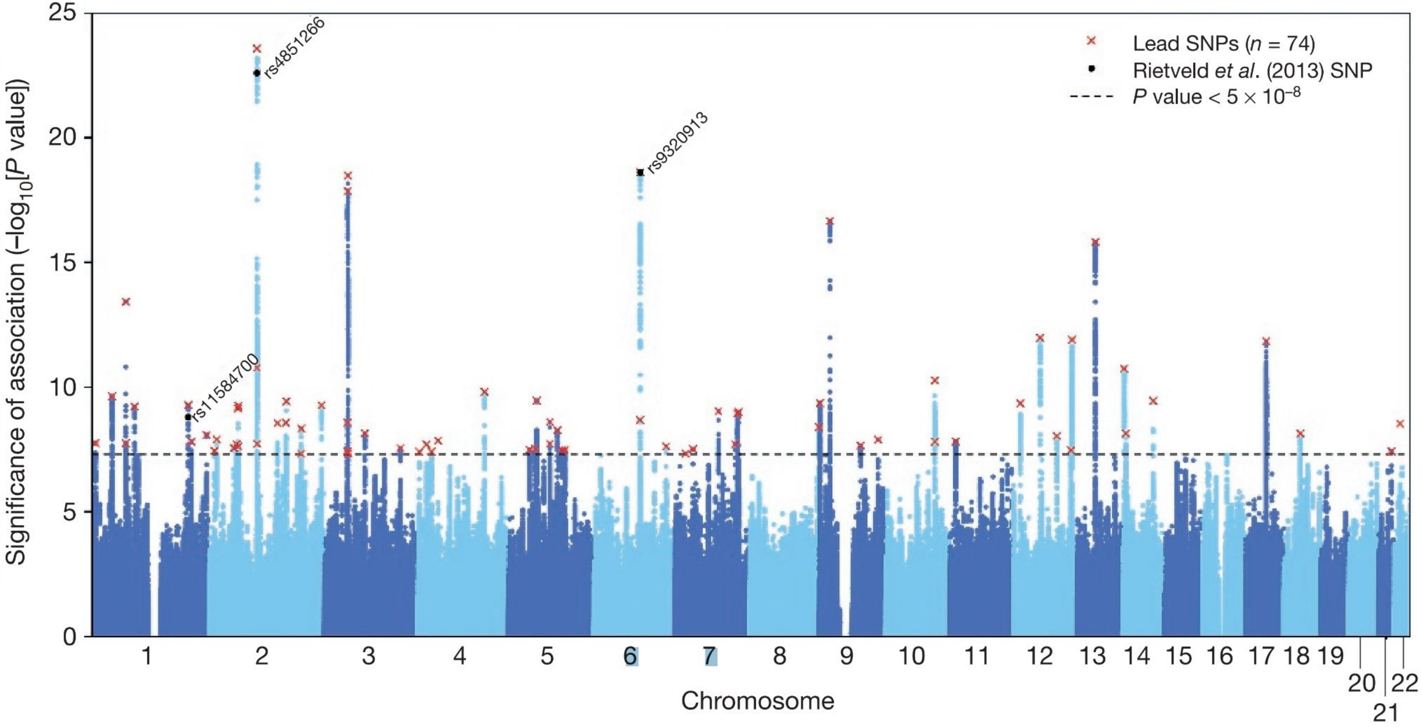


$$\log(p/1\text{-}p) = G\beta_G + X\beta_X + e$$

$$\begin{array}{ccccccccc}
\text{Type2 diabetes} & & \text{Genotype (allele count/ dosage)} & & \text{Covariates} & & & & \\
 & & & & \text{Age} & & \text{Sex} & & \text{PCs ...} \\
\begin{bmatrix}1\\0\\1\\1\\0\\1\end{bmatrix} & \sim & \begin{bmatrix}1\\0\\2\\2\\0\\1\end{bmatrix} & + & \begin{bmatrix}181\\152\\162\\177\\148\\165\end{bmatrix} & + & \begin{bmatrix}1\\2\\1\\2\\2\\1\end{bmatrix} & + & \begin{bmatrix}\cdots\\\cdots\\\cdots\\\cdots\\\cdots\\\cdots\end{bmatrix}
\end{array}$$

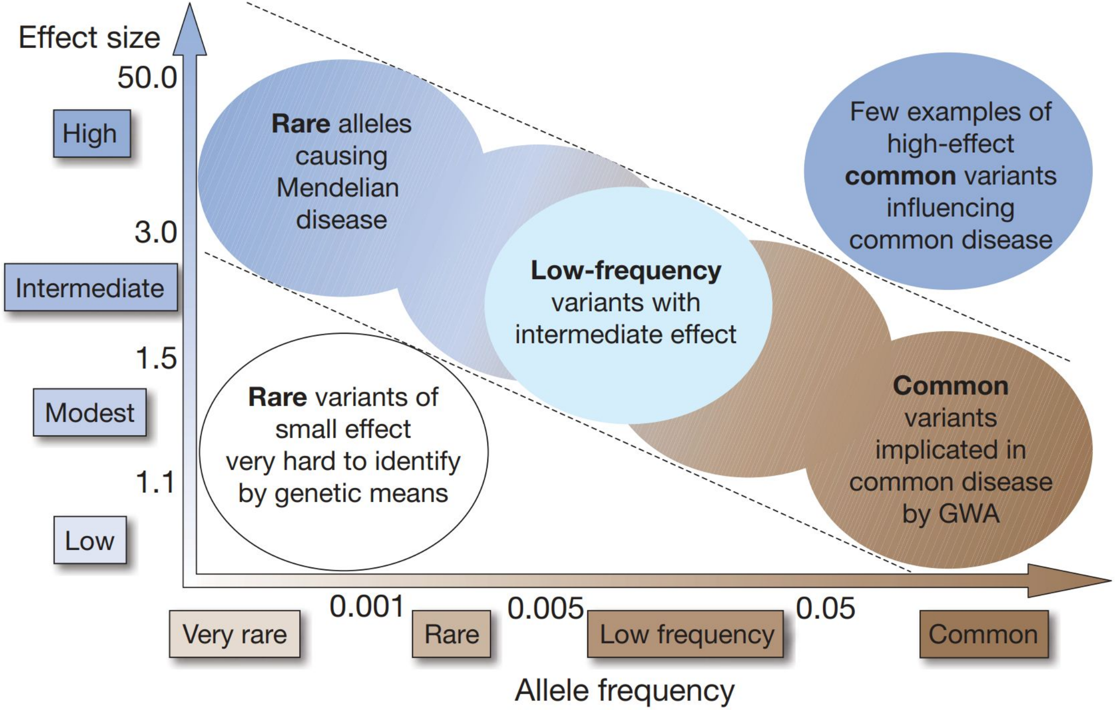
Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Manhattan plot for EduYears associations (n = 293,723)



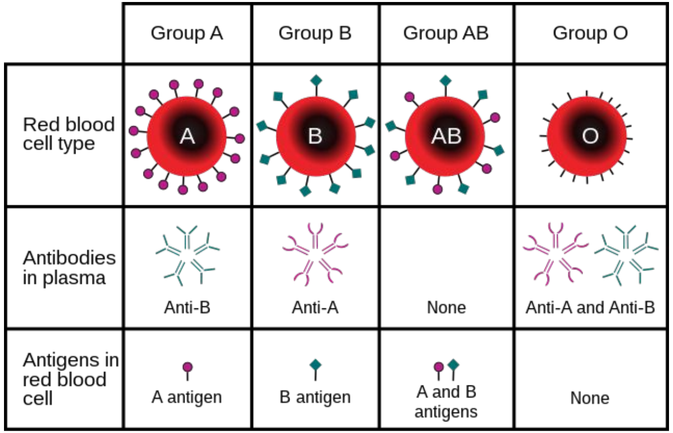
Okbay, A., Beauchamp, J., Fontana, M. et al. (2016). doi.org/10.1038/nature17671.

Rare and common variants

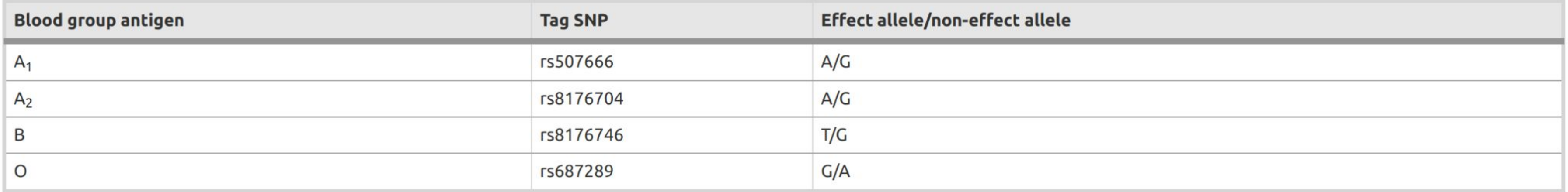

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect **common** variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Manolio, T., Collins, F., Cox, N. et al. (2009). doi.org/10.1038/nature08494.

Haplotypes: ABO serological groups



	Group A	Group B	Group AB	Group O
Red blood cell type	A	B	AB	O
Antibodies in plasma	Anti-B	Anti-A	None	Anti-A and Anti-B
Antigens in red blood cell	A antigen	B antigen	A and B antigens	None



Blood group antigen	Tag SNP	Effect allele/non-effect allele
A_1	rs507666	A/G
A_2	rs8176704	A/G
B	rs8176746	T/G
O	rs687289	G/A

Li S, Schooling CM (2020). doi: 10.1186/s12916-020-01795-4.

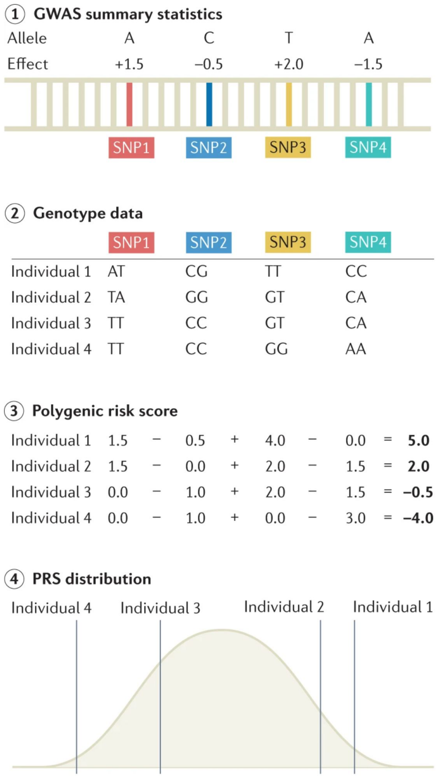
Polygenic Risk Scores (PRS)

stratification & disease trajectories

Common workflow - single-trait PRS

Polygenic risk scores (PRS)

score that summarizes the effect sizes of genetic variants on a certain disease or trait.



① GWAS summary statistics

Allele	A	C	T	A
Effect	+1.5	−0.5	+2.0	−1.5
	SNP1	SNP2	SNP3	SNP4

② Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

③ Polygenic risk score

Individual 1	1.5	−	0.5	+	4.0	−	0.0	=	**5.0**
Individual 2	1.5	−	0.0	+	2.0	−	1.5	=	**2.0**
Individual 3	0.0	−	1.0	+	2.0	−	1.5	=	**−0.5**
Individual 4	0.0	−	1.0	+	0.0	−	3.0	=	**−4.0**

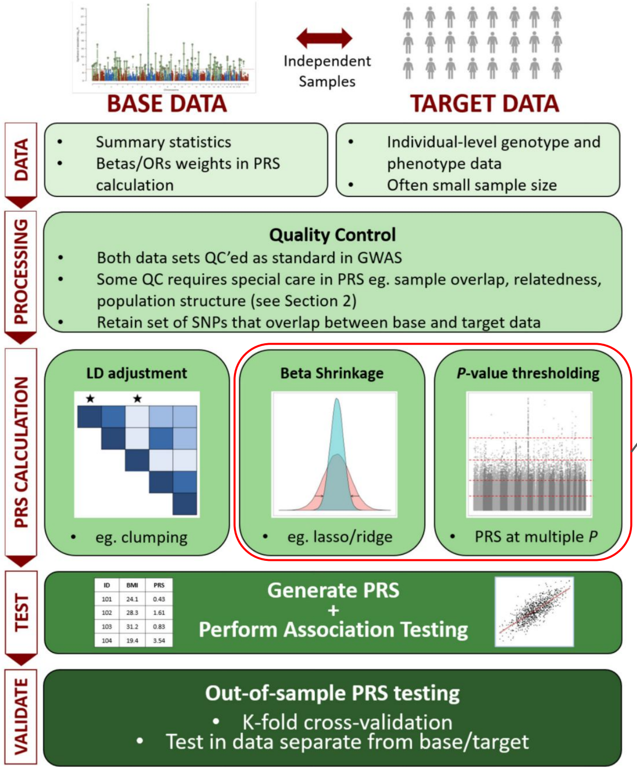
④ PRS distribution

$$\mathrm{PRS}_i = \sum_{j \in J} \beta_j \, G_{ij}$$

i-th individual
j-th variant
G: genotype
β: effect size

Choi *et al.* Nat Protoc (2020). doi:10.1038/s41596-020-0353-1. | Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Common workflow - single-trait PRS



BASE DATA

Independent Samples

TARGET DATA

DATA

- Summary statistics
- Betas/ORs weights in PRS calculation

- Individual-level genotype and phenotype data
- Often small sample size

PROCESSING

Quality Control

- Both data sets QC'ed as standard in GWAS
- Some QC requires special care in PRS eg. sample overlap, relatedness, population structure (see Section 2)
- Retain set of SNPs that overlap between base and target data

PRS CALCULATION

LD adjustment

- eg. clumping

Beta Shrinkage

- eg. lasso/ridge

***P*-value thresholding**

- PRS at multiple *P*

TEST

Generate PRS + Perform Association Testing

VALIDATE

Out-of-sample PRS testing

- K-fold cross-validation
- Test in data separate from base/target

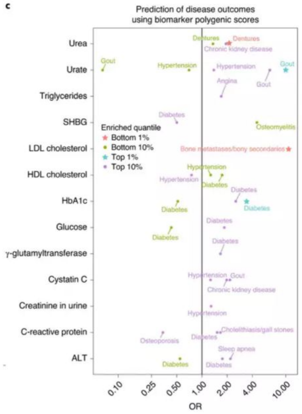
Category	**Description**	**Methods/ software**
P value thresholding	P + T	C+T, PRSice, Plink
Beta shrinkage	genome-wide PRS model	LDpred

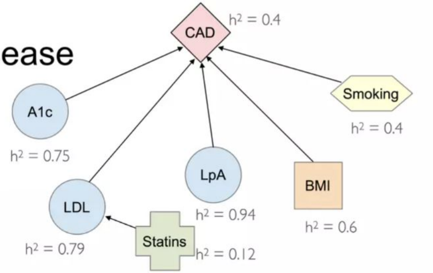
P+T stands for Pruning + Thresholding, also known as Clumping and Thresholding (**C+T)**

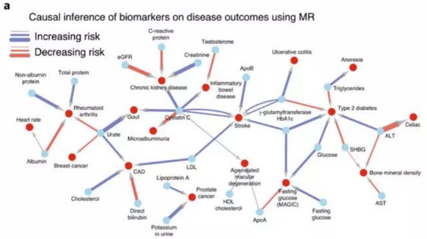
Choi *et al.* Nat Protoc (2020). doi:10.1038/s41596-020-0353-1. | Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Common workflow - multi-trait PRS

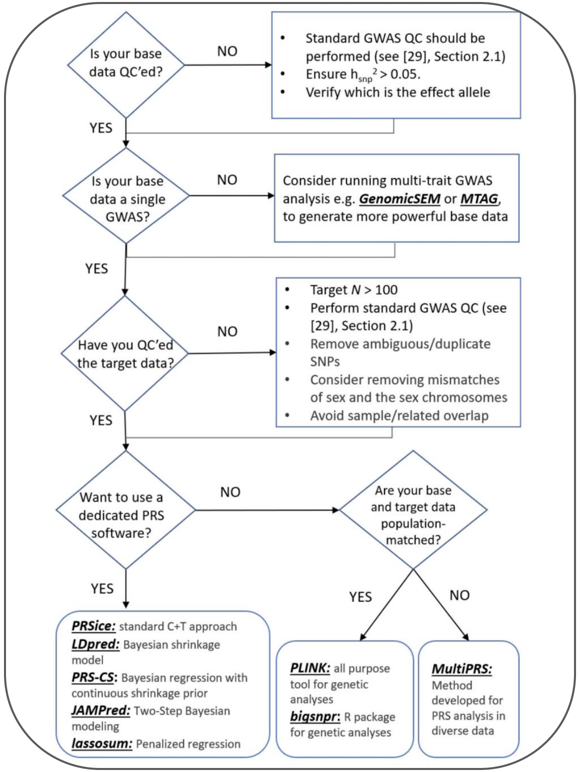
- Multiple observations suggest “biomarkers → disease” links
  - PRS-PheWAS analysis
  - Biomarkers are more heritable than disease
  - Mendelian Randomization







- Multi-PRS is a weighted sum of PRSs
  i.e. $w_1(PRS_1) + w_2(PRS_2) + w_3(PRS_3) + \ldots$



Is your base data QC’ed?

NO

- Standard GWAS QC should be performed (see [29], Section 2.1)
- Ensure $h_{snp}^2 > 0.05$.
- Verify which is the effect allele

YES

Is your base data a single GWAS?

NO

Consider running multi-trait GWAS analysis e.g. ***GenomicSEM*** or ***MTAG***, to generate more powerful base data

YES

Have you QC’ed the target data?

NO

- Target *N* > 100
- Perform standard GWAS QC (see [29], Section 2.1)
- Remove ambiguous/duplicate SNPs
- Consider removing mismatches of sex and the sex chromosomes
- Avoid sample/related overlap

YES

Want to use a dedicated PRS software?

NO

Are your base and target data population-matched?

YES

PRSice: standard C+T approach
LDpred: Bayesian shrinkage model
PRS-CS: Bayesian regression with continuous shrinkage prior
JAMPred: Two-Step Bayesian modeling
lassosum: Penalized regression

YES

PLINK: all purpose tool for genetic analyses
bigsnpr: R package for genetic analyses

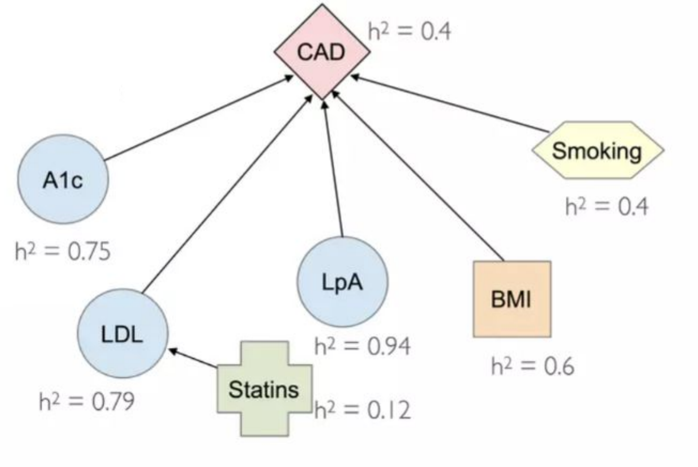
NO

MultiPRS: Method developed for PRS analysis in diverse data

Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z. | Choi *et al.* Nat Protoc. 2020. doi:10.1038/s41596-020-0353-1.

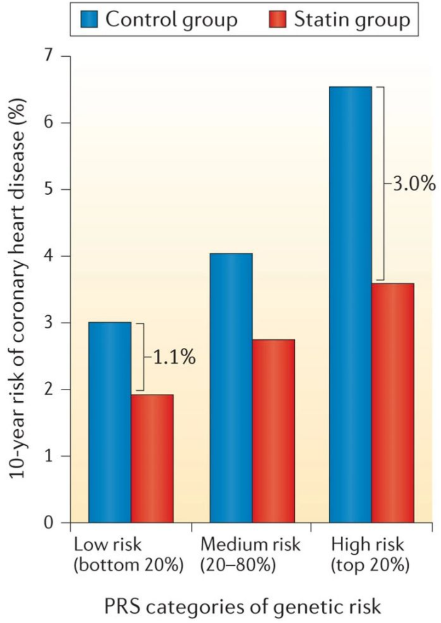
Common workflow - multi-trait PRS

- Multi-PRS is a weighted sum of PRSs
  i.e. $w_1(PRS_1) + w_2(PRS_2) + w_3(PRS_3) + \ldots$

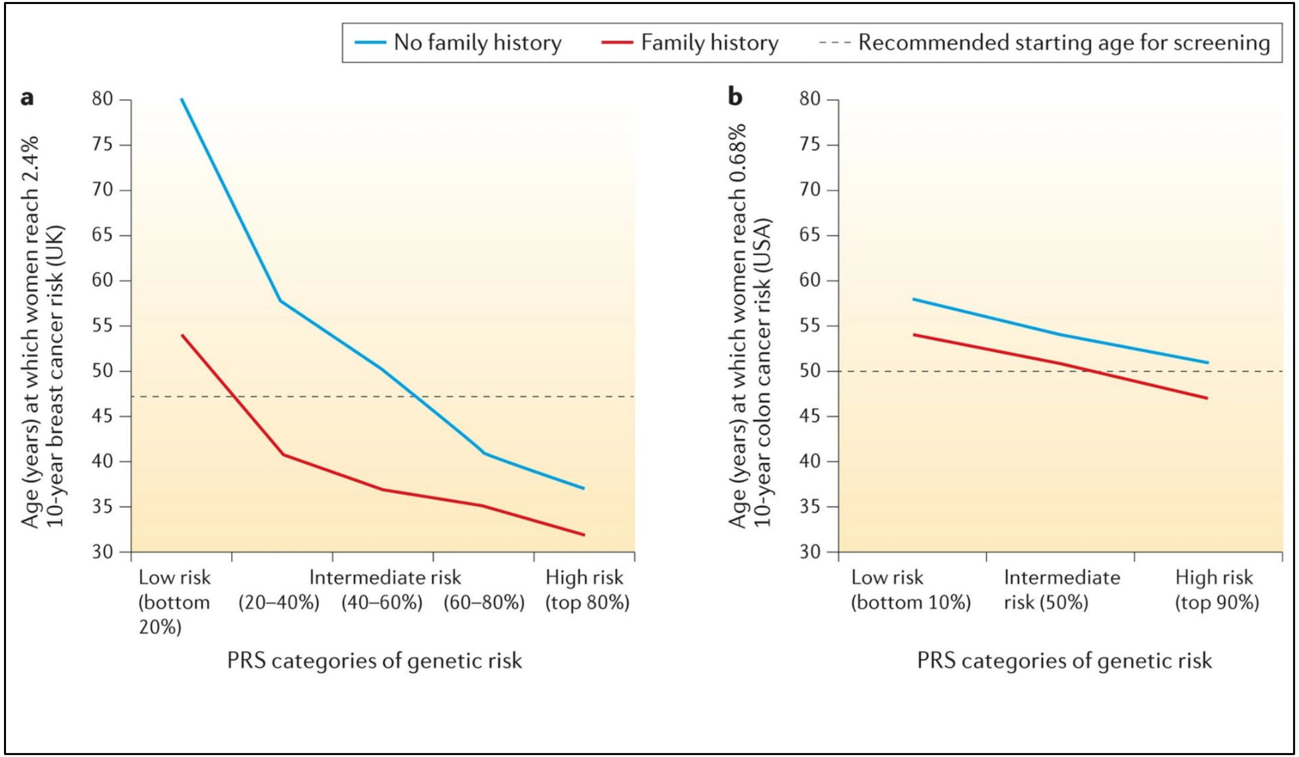


Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z. | Choi *et al.* Nat Protoc. 2020. doi:10.1038/s41596-020-0353-1.

Case study - disease stratification



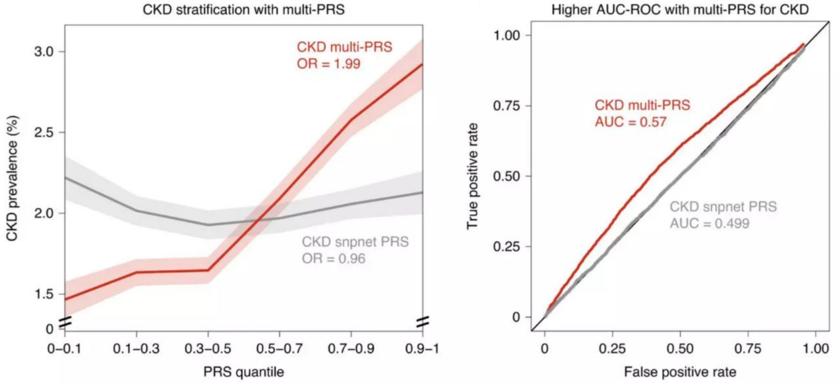
ARR (%)	1.1	1.3	3.0
RRR	0.36	0.32	0.46



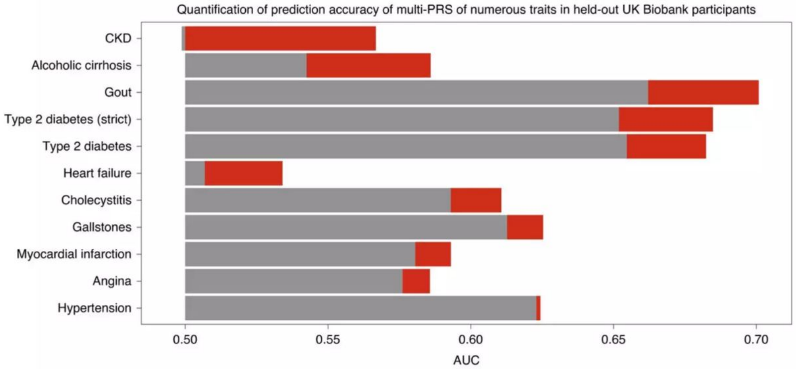
Marston NA *et al*. JAMA Cardiol. (2023). doi:10.1001/jamacardio.2022.4466. | Chatterjee N *et al.* Nat Rev Genet. (2016). doi: 10.1038/nrg.2016.27.

Case study - multi-trait PRS improves disease prevalence prediction

Chronic kidney disease (CKD)

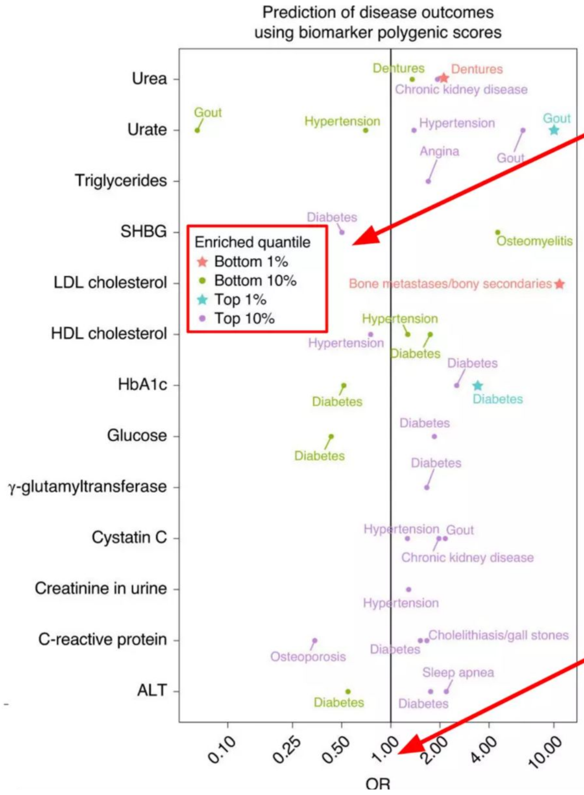


Other diseases in UK Biobank

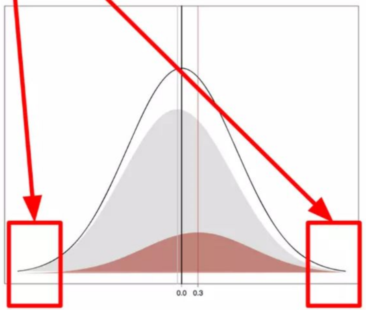


Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z.

Use of PRS for trait / disease prediction



Take extreme in PRS for biomarkers

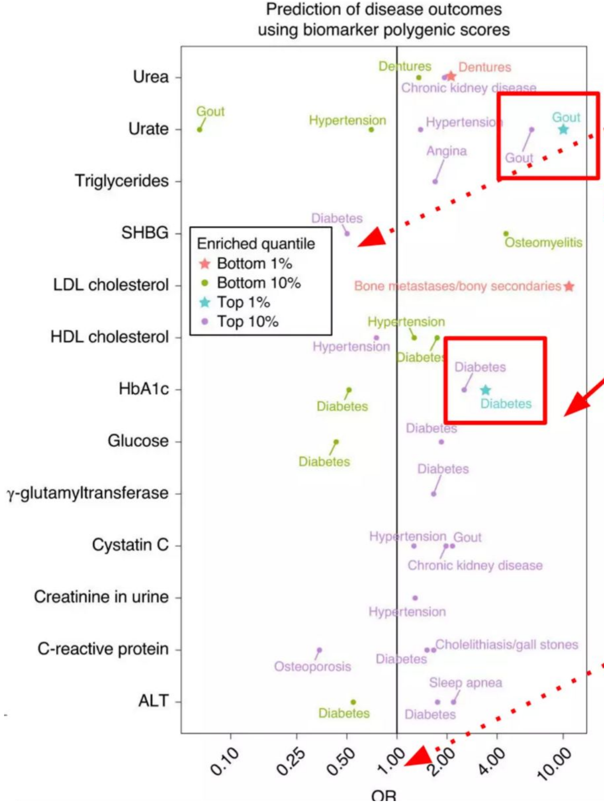


Compare odds ratio for disease outcome relative to 40-60%ile bin

Applied PheWAS for ~160 diseases

Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z.

Use of PRS for trait / disease prediction



Take extreme in PRS for biomarkers

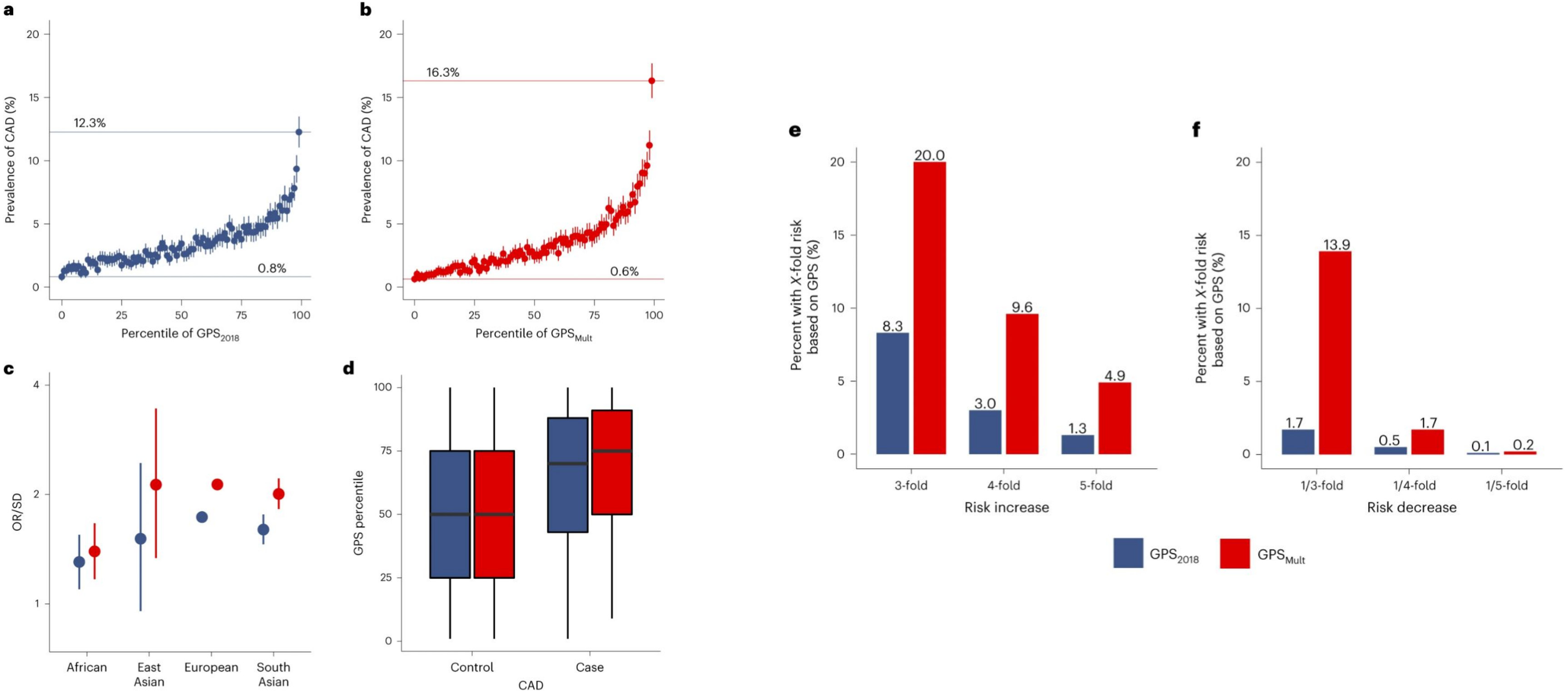
Identify diseases with biomarker PRS associations

Compare odds ratio for disease outcome relative to 40-60%ile bin

Applied PheWAS for ~160 diseases

Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z.

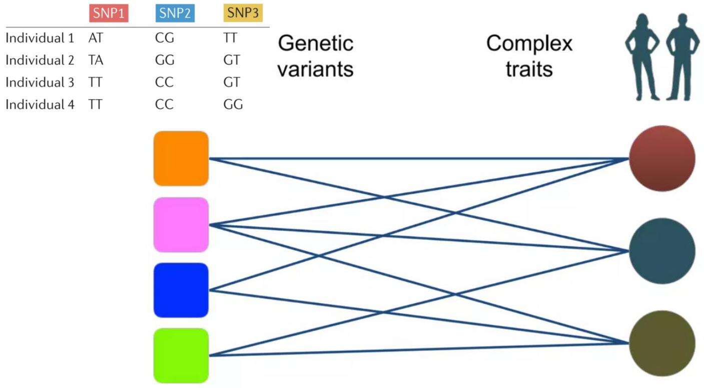
Limitations - ethnicity / ancestry



Patel *et al. Nat Med* (2023). doi:10.1038/s41591-023-02429-x.

Limitations - polygenicity & pleiotropy

- Polygenicity: many variants - one trait
- Pleiotropy: one variant - many traits

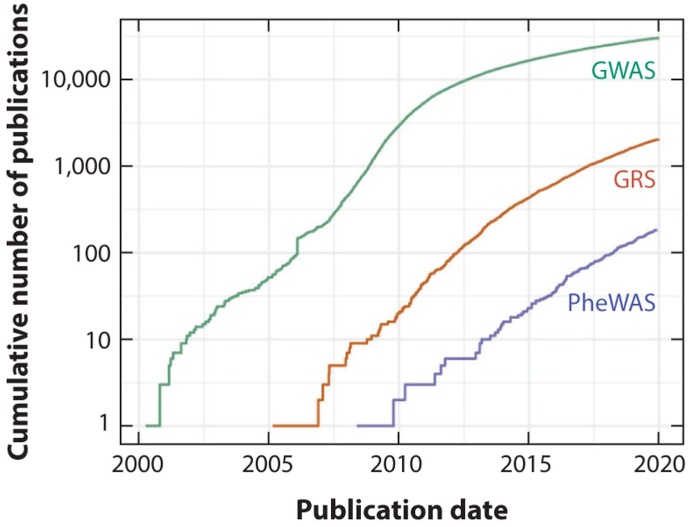


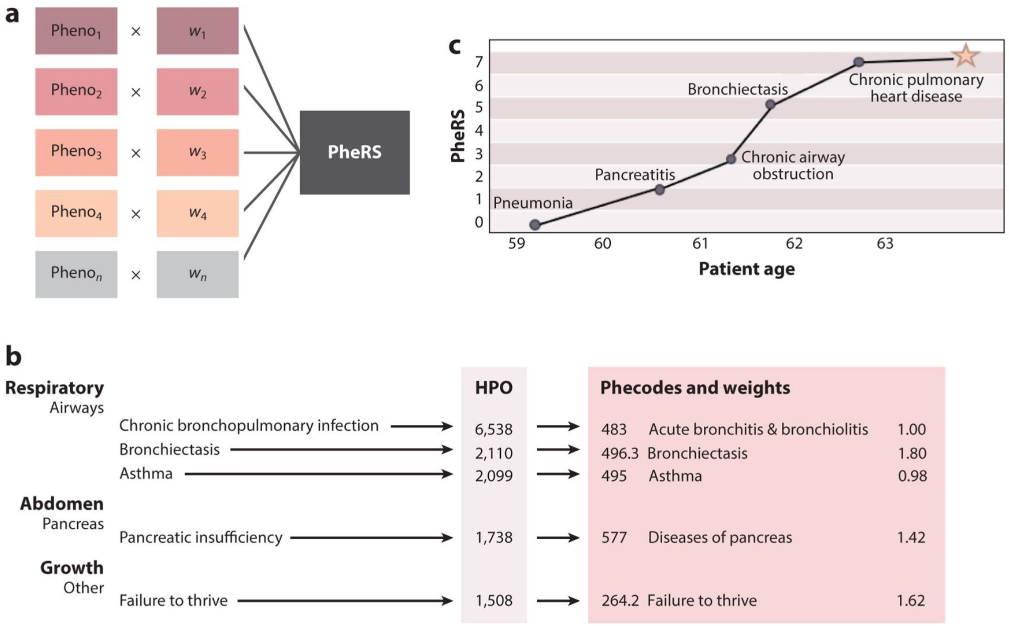
	SNP1	SNP2	SNP3
Individual 1	AT	CG	TT
Individual 2	TA	GG	GT
Individual 3	TT	CC	GT
Individual 4	TT	CC	GG

- Large number of associations in population-based cohorts
- Can we group them together for enhanced interpretation?

Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z.

Going further - EHR’s & PRS





		HPO	Phecodes and weights		
Respiratory Airways	Chronic bronchopulmonary infection →	6,538 →	483	Acute bronchitis & bronchiolitis	1.00
	Bronchiectasis →	2,110 →	496.3	Bronchiectasis	1.80
	Asthma →	2,099 →	495	Asthma	0.98
Abdomen Pancreas	Pancreatic insufficiency →	1,738 →	577	Diseases of pancreas	1.42
Growth Other	Failure to thrive →	1,508 →	264.2	Failure to thrive	1.62

Linder JE *et al.* Annu Rev Genomics Hum Genet. (2021). doi: 10.1146/annurev-genom-121120-125204.

Summary

Two complementary approaches to improve predictive performance:

- Sample size → increase in **statistical power**
- Multi-trait PRS analysis

Why does multi-PRS work?

- Quantitative traits have more power
- **Genetic correlation** between biomarkers and disease

The multi-trait PRS model:

Genetics →Biomarkers (molecular traits) →Disease