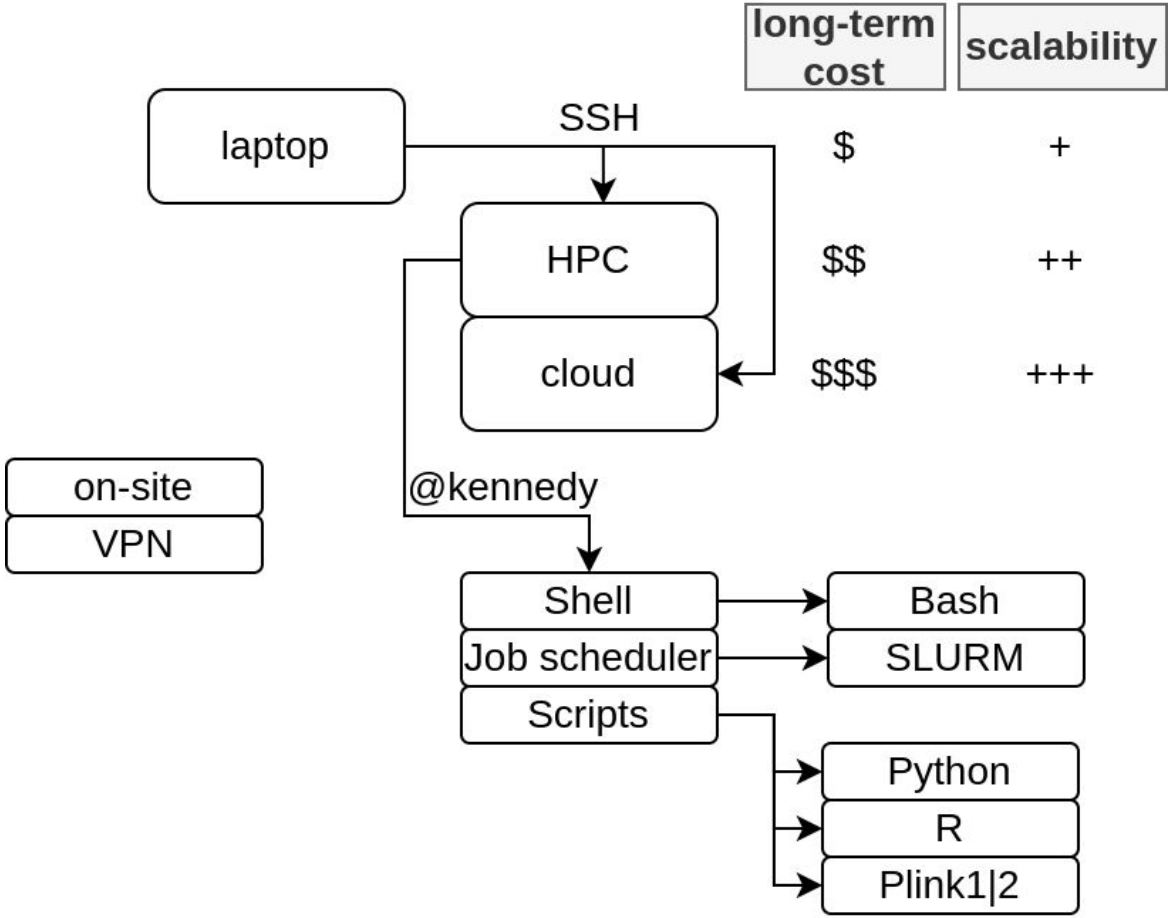
Lecture 1 - GWAS Statistics + Polygenic Risk Scores (PRS)

Wed, Mar 20, 9-10AM

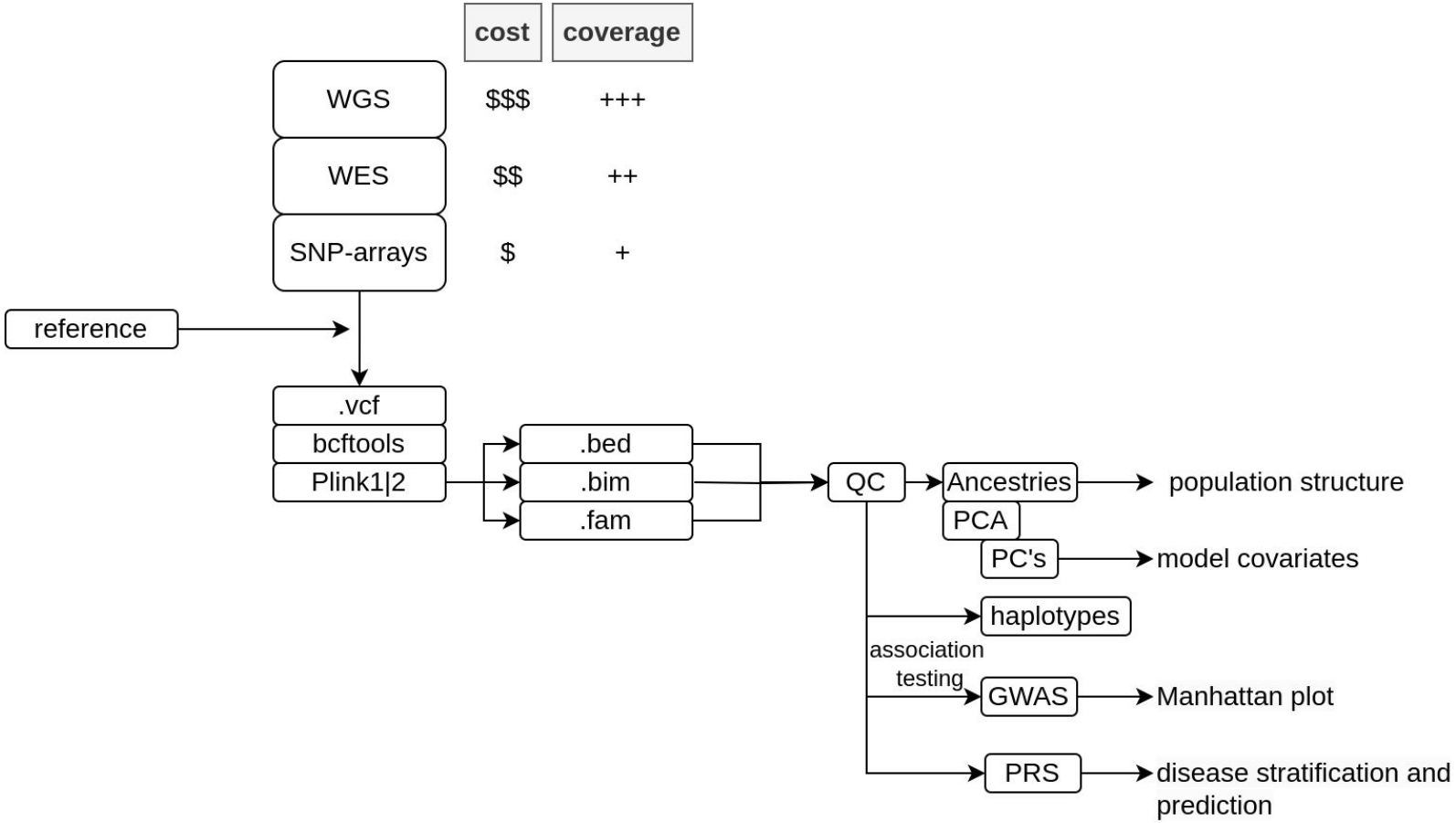
Recap

HPC | GWAS

Recap: HPC + SSH



Recap: GWAS

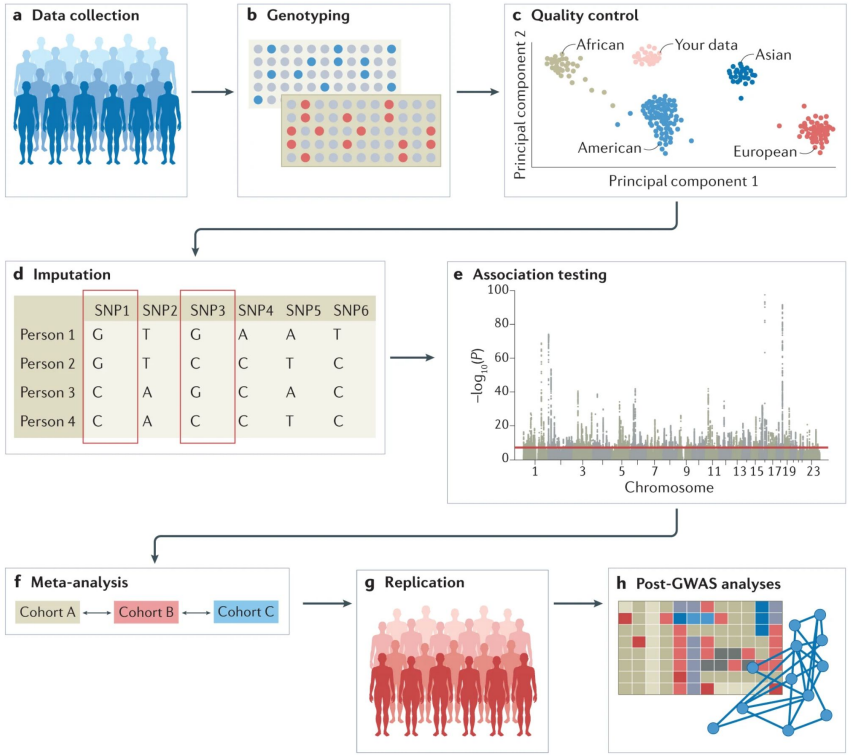

cost
coverage
WGS
$$$
+++
WES
$$
++
SNP-arrays
$
+
reference
.vcf
bcftools
Plink1|2
.bed
.bim
.fam
QC
Ancestries
population structure
PCA
PC's
model covariates
haplotypes
association testing
GWAS
Manhattan plot
PRS
disease stratification and prediction

GWAS

Genome-wide Association Studies

Genome-wide Association studies (GWAS)

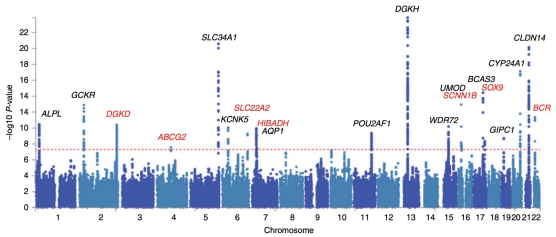
Single nucleotide polymorphism (SNP): This is a variation in a single nucleotide (i.e., **A**, **C**, **G**, or **T**) that occurs at a specific position in the genome. A SNP usually exists as two different forms (e.g., **A** vs. **T**). These different forms are called alleles. A SNP with two alleles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



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

Data sources & repositories

Summary statistics



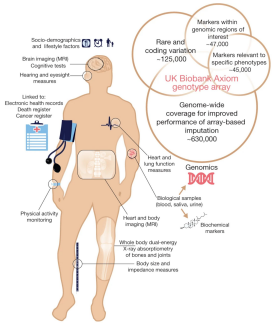
GWAS Catalog

general GWAS data repository

CKD Gen CONSORTIUM

specific for GWAS Chronic Kidney Disease

Individual-level



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

dbGaP GENOTYPES and PHENOTYPES

biobank uk

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

Testing for associations

Genetic models

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)

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
Controls	1000	500	500	2000
Total	1800	900	1300	4000

Dominant model

Genotype	Control: AA	Case: GG/AG	Total
Case	800	1200	2000
Controls	1000	1000	2000
Total	1800	2200	4000

Chi-square test

Recessive model

Genotype	Control: AA/AG	Case: GG	Total
Case	1200	800	2000
Controls	1500	500	2000
Total	1800	2200	4000

Chi-square test

Additive model

Alt allele (G) count	0	1	2	Total
Case	800	400	800	2000
Controls	1000	500	500	2000
Total	1800	900	1300	4000

Cochran–Armitage trend test

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

Testing for associations

Quantitative traits

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

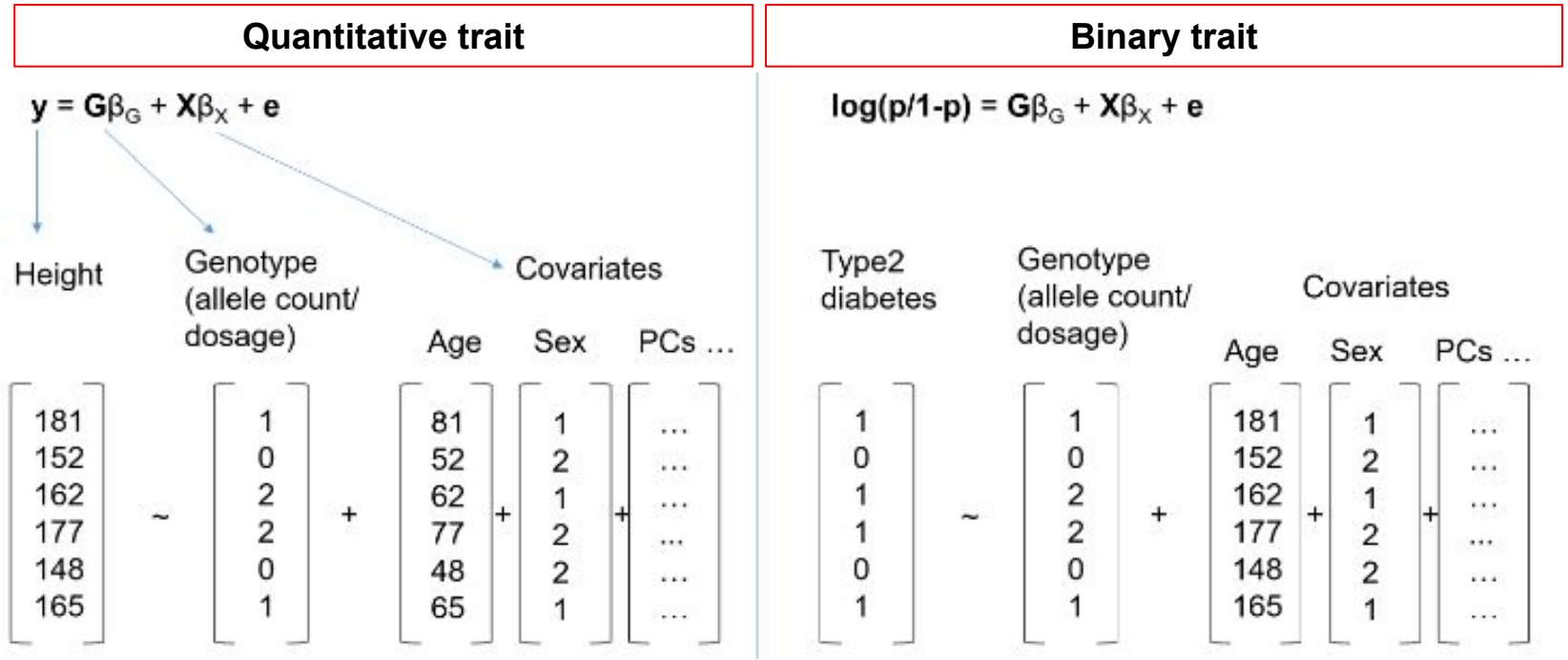
- G is the genotype matrix.
- β_G is the effect size for variants.
- X and β_X are covariates and their effects.
- e is the error term.

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

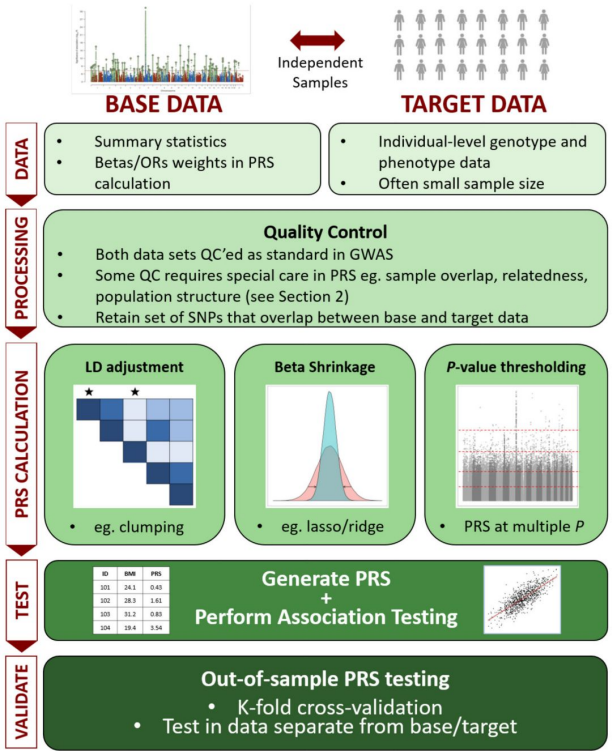


Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

Polygenic Risk Scores

stratification & disease trajectories

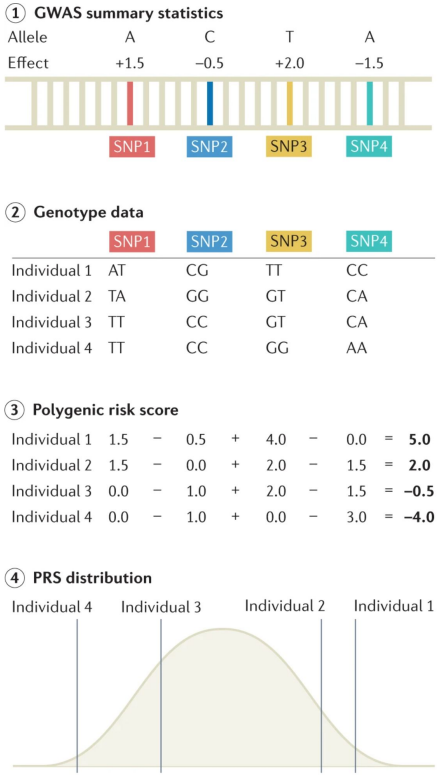
Common workflow - single-trait PRS



Polygenic risk scores (PRS)

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

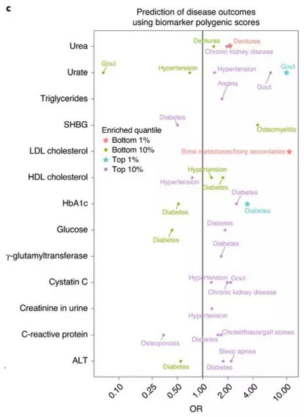
i-th individual
j-th variant
G: genotype
β: effect size

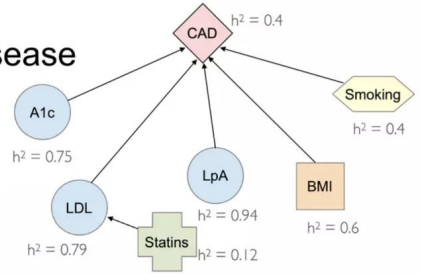


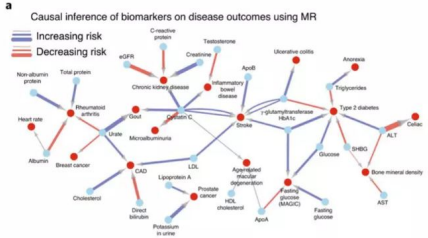
Choi *et al.* Nat Protoc (2020). doi:10.1038/s41596-020-0353-1. | Uffelmann *et al.* Nat Rev Methods Primers (2021). doi:10.1038/s43586-021-00056-9.

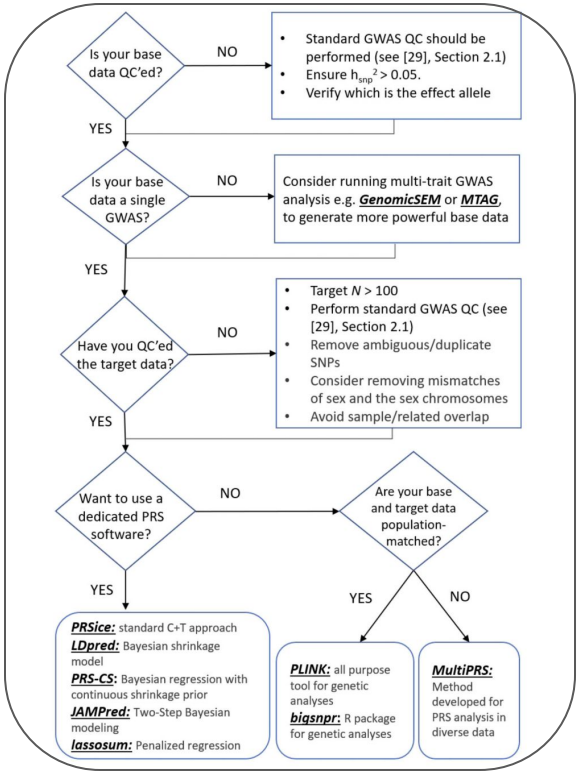
Common workflow - multi-trait PRS

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







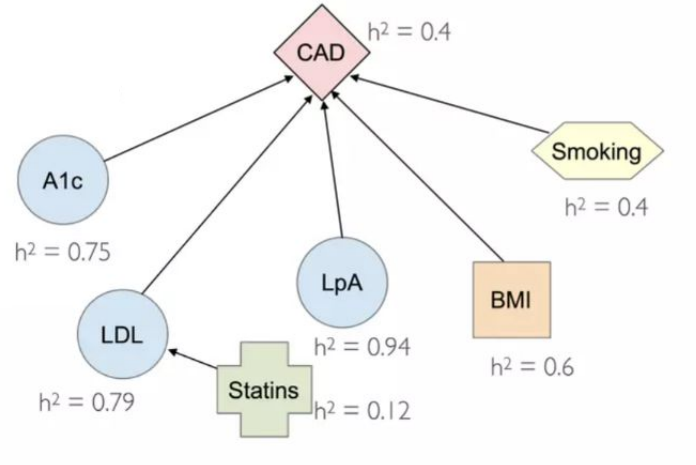


- Multi-PRS is a weighted sum of PRSs
  i.e. $w_1(PRS_1) + w_2(PRS_2) + w_3(PRS_3) + \ldots$

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.

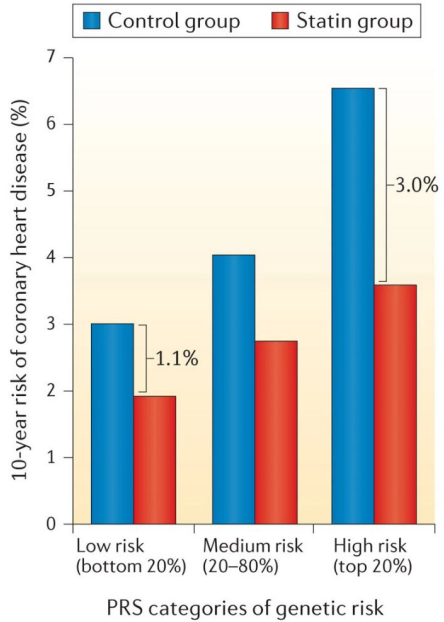
Common workflow - multi-trait PRS

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

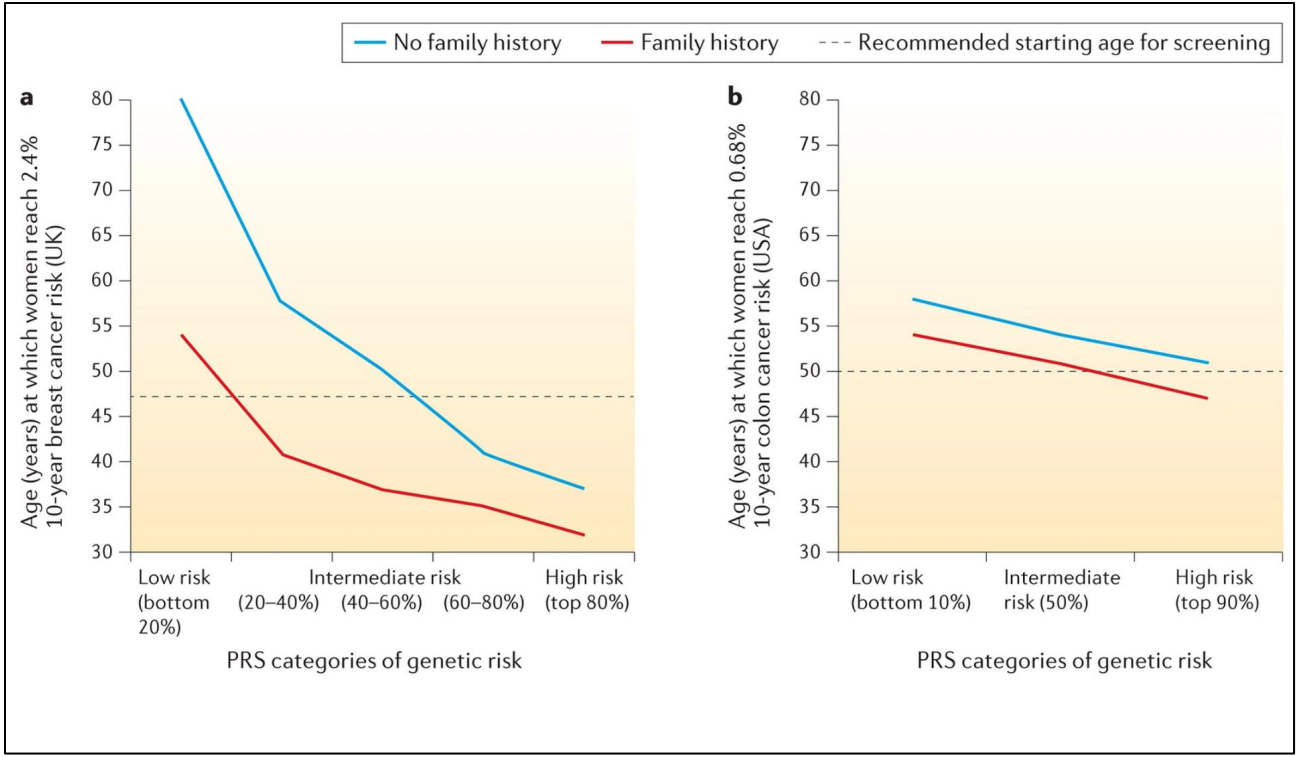


Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z. | Choi *et al*. Nat Protoc. 2020. doi:10.1038/s41596-020-0353-1.

Case study - disease stratification



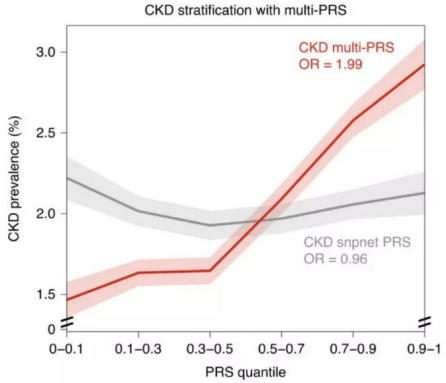
ARR (%)	1.1	1.3	3.0
RRR	0.36	0.32	0.46

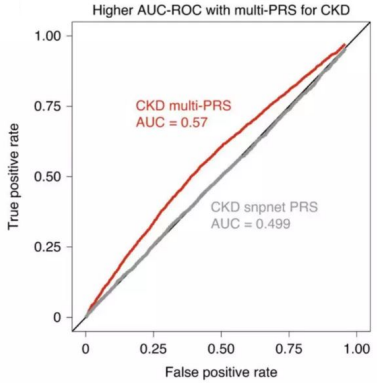


Marston NA *et al*. JAMA Cardiol. (2023). doi:10.1001/jamacardio.2022.4466. | Chatterjee N *et al.* Nat Rev Genet. (2016). doi: 10.1038/nrg.2016.27.

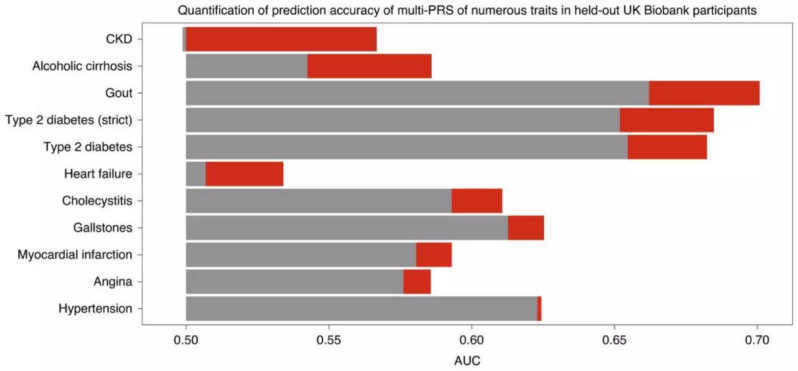
Case study - multi-trait PRS improves disease prevalence prediction

Chronic kidney disease (CKD)



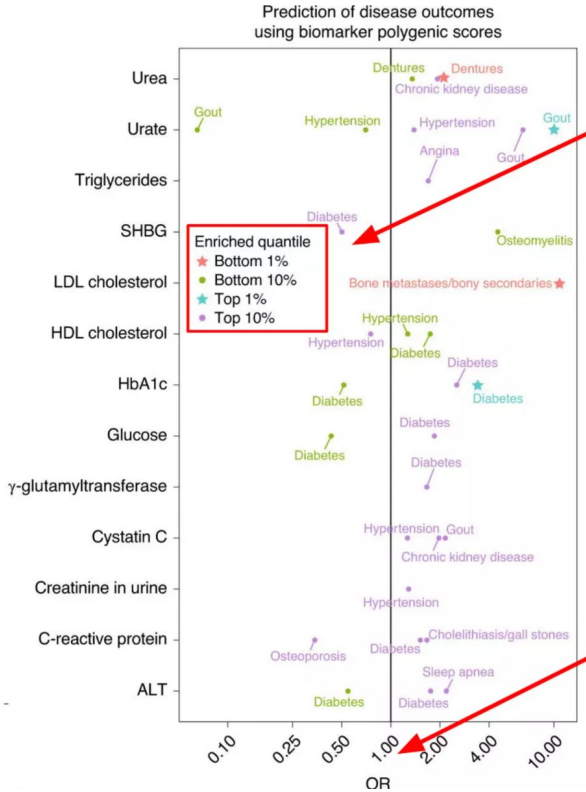


Other diseases in UK Biobank

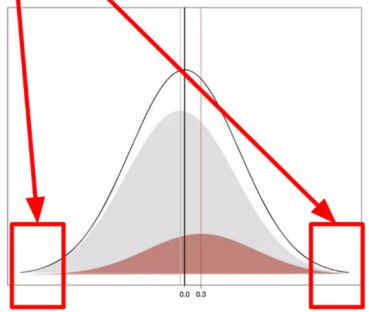


Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z.

Use of PRS for trait / disease prediction



Take extreme in PRS for biomarkers

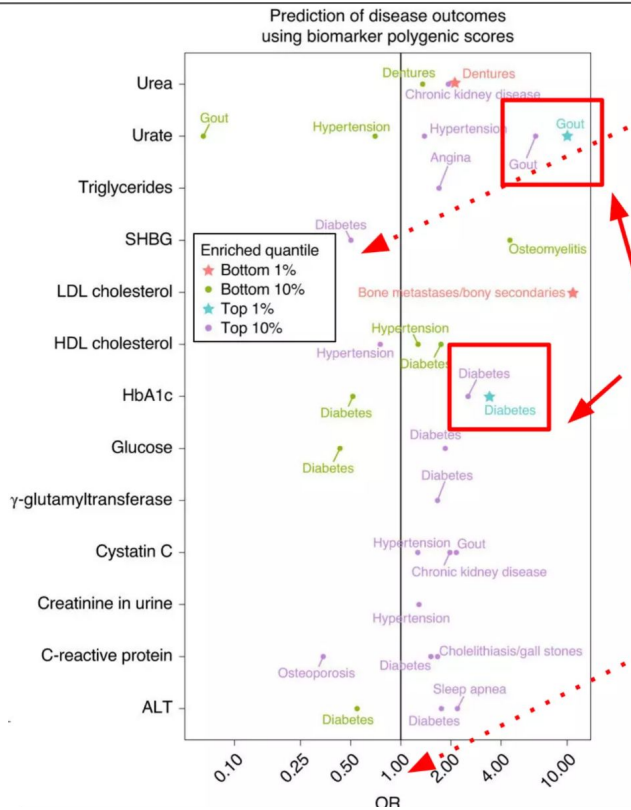


Compare odds ratio for disease outcome relative to 40-60%ile bin

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

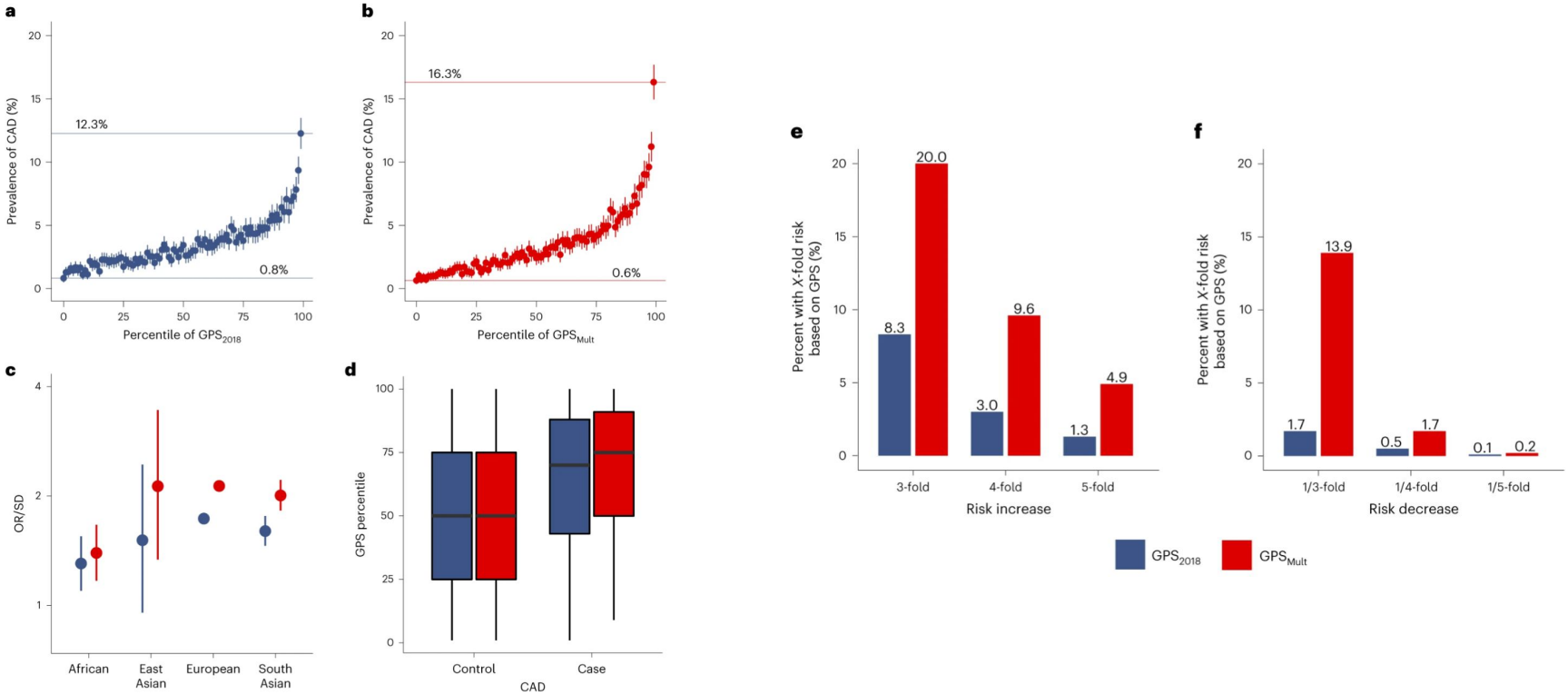
Identify diseases with biomarker PRS associations

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.

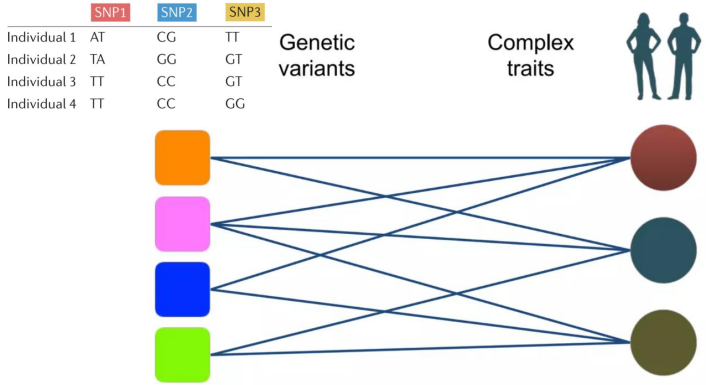
Limitations - ethnicity / ancestry



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

Limitations - polygenicity & pleiotropy

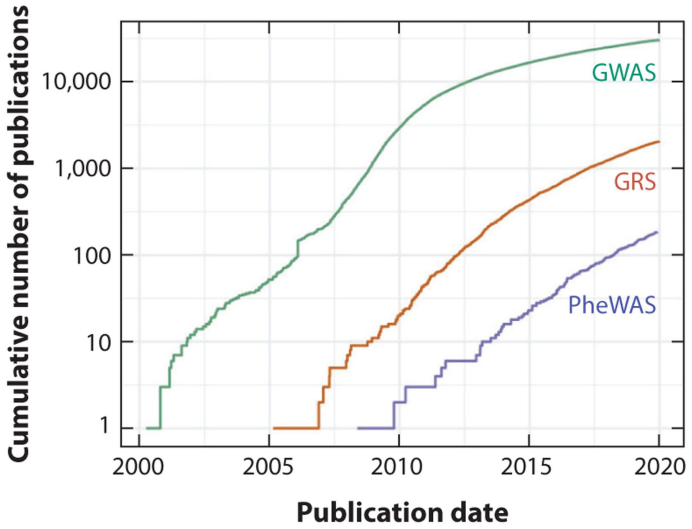
- Polygenicity: many variants - one trait
- Pleiotropy: one variant - many traits

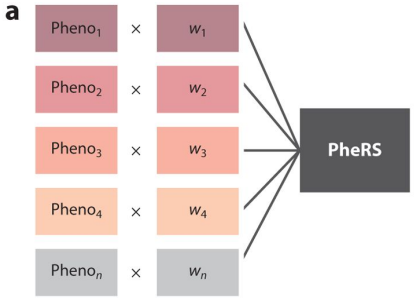


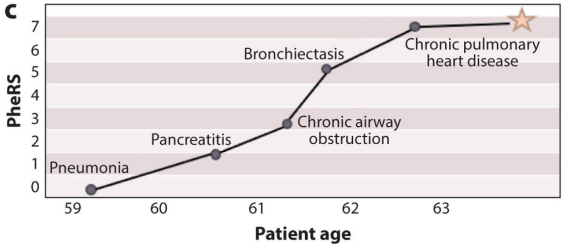
- Large number of associations in population-based cohorts
- Can we group them together for enhanced interpretation?

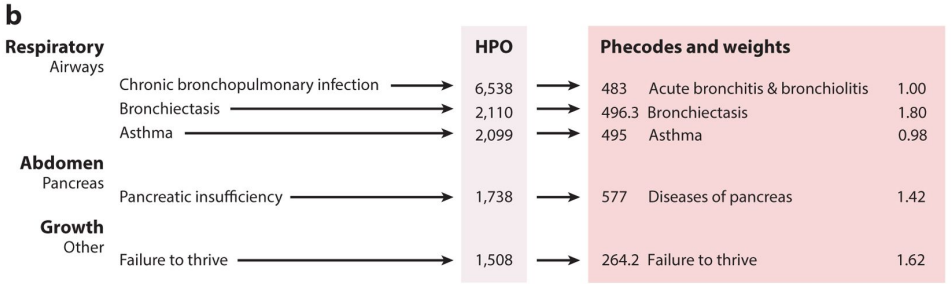
Sinnott-Armstrong N *et al*. Nat Genet. (2021). doi:10.1038/s41588-020-00757-z.

Going further - EHR's & PRS









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