

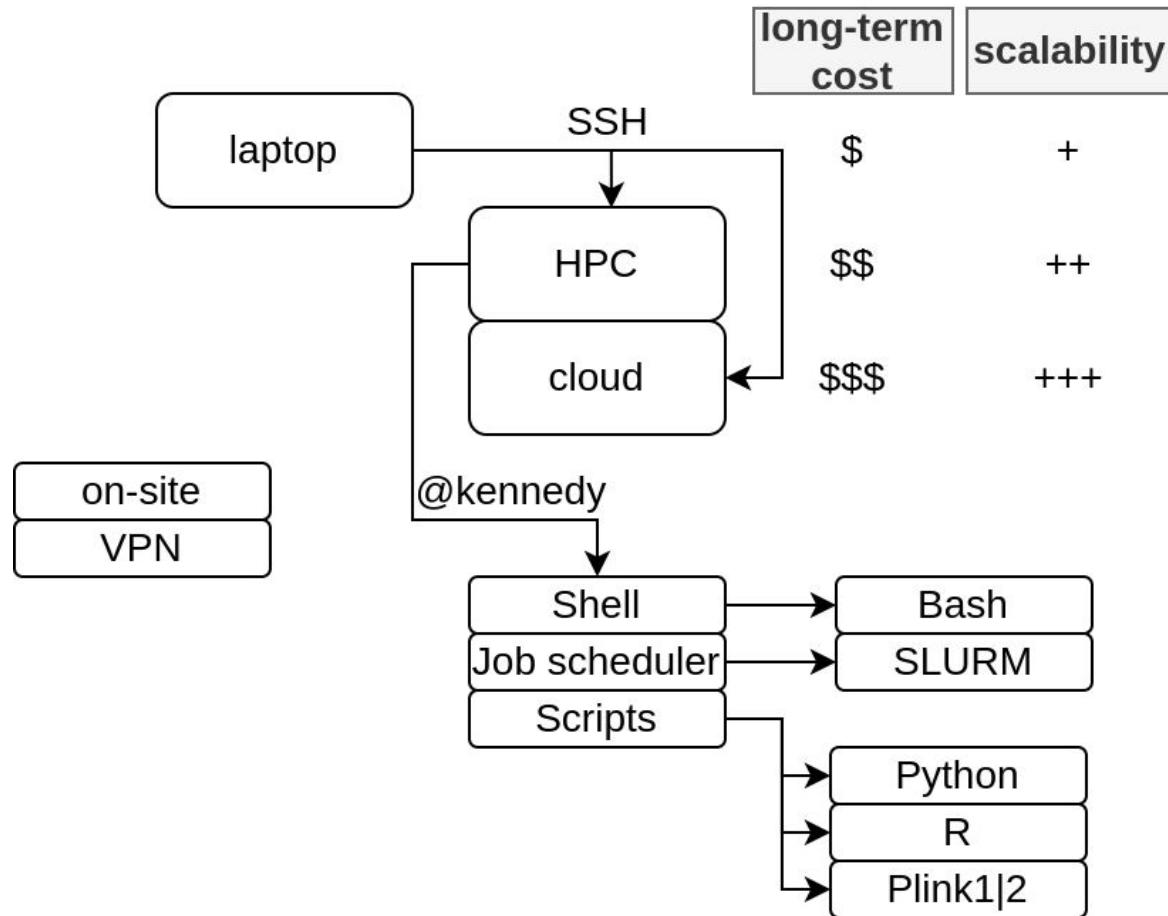
Lecture 1 - GWAS Statistics + Polygenic Risk Scores (PRS)

Wed, Mar 20, 9-10AM

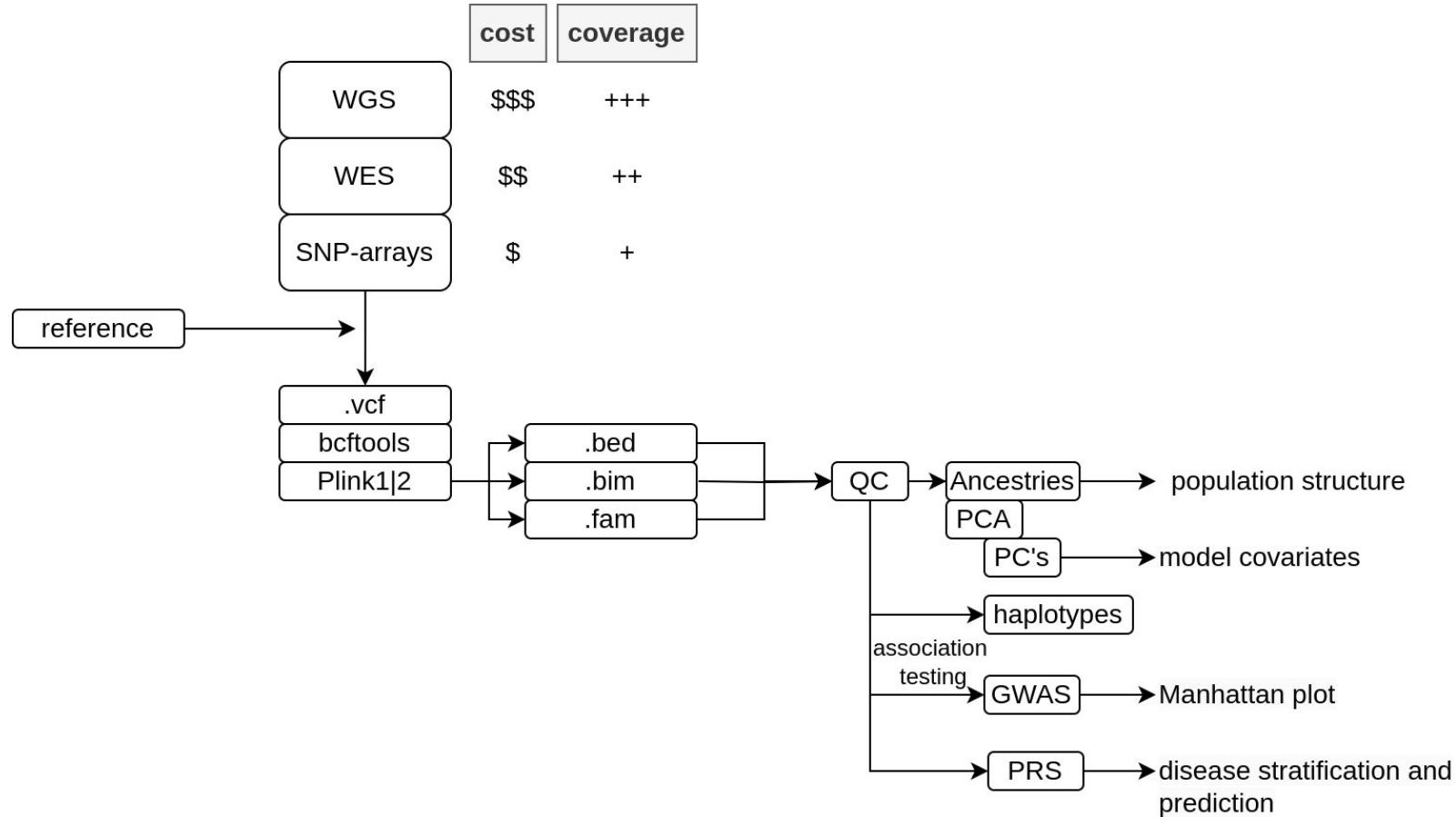
Recap

HPC | GWAS

Recap: HPC + SSH



Recap: GWAS

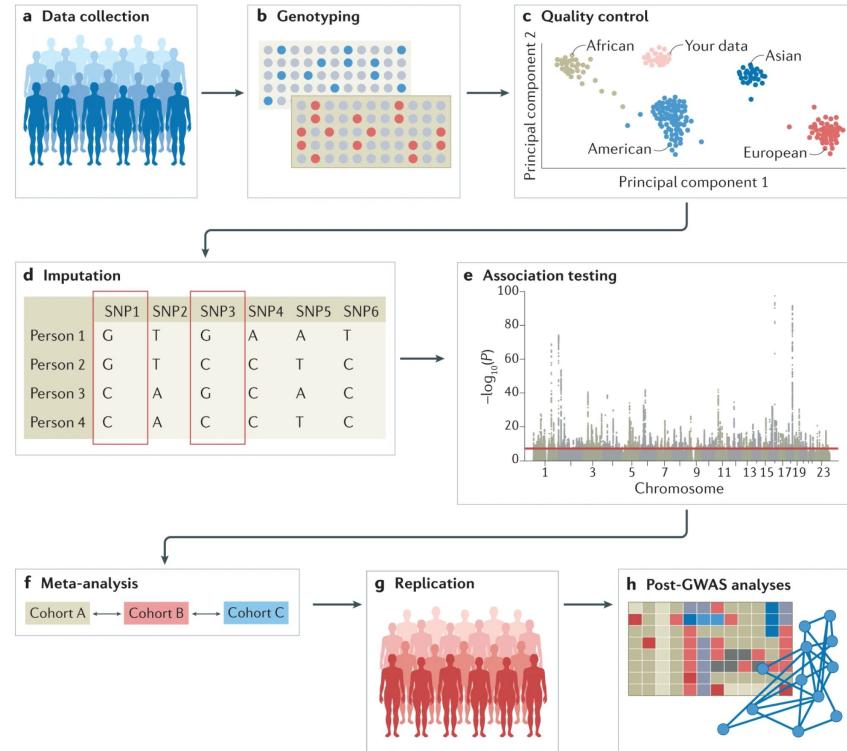


GWAS

Genome-wide Association Studies

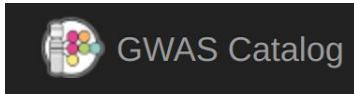
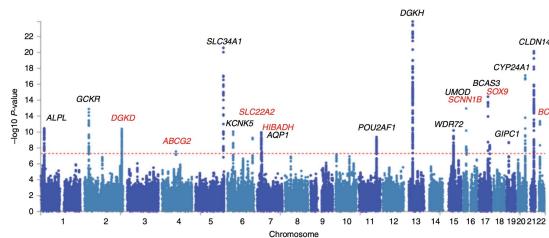
Genome-wide Association studies (GWAS)

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**).



Data sources & repositories

Summary statistics

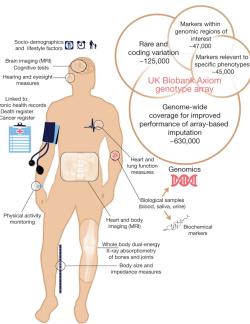


general GWAS data repository



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



biobank^{uk}

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**.

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

Testing for associations

Quantitative traits

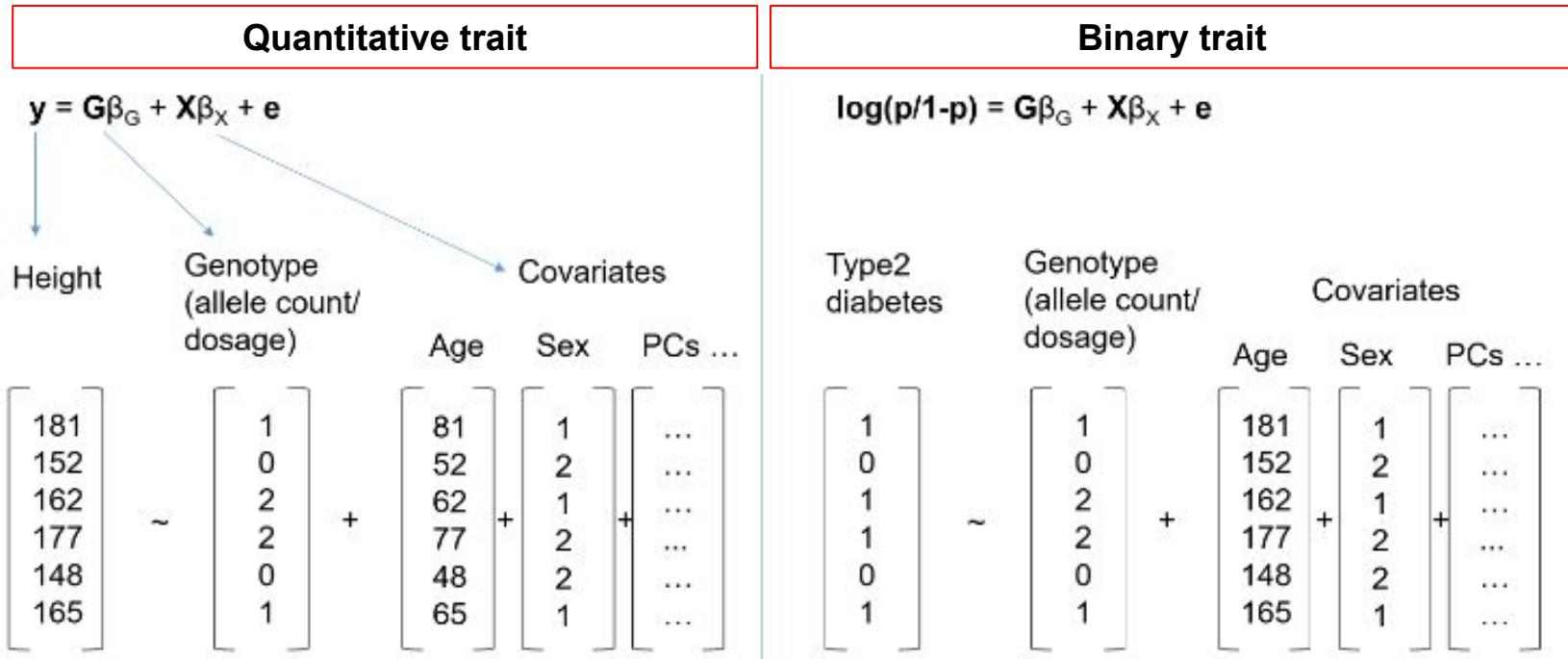
$$y = G\beta_G + X\beta_X + e$$

- 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.

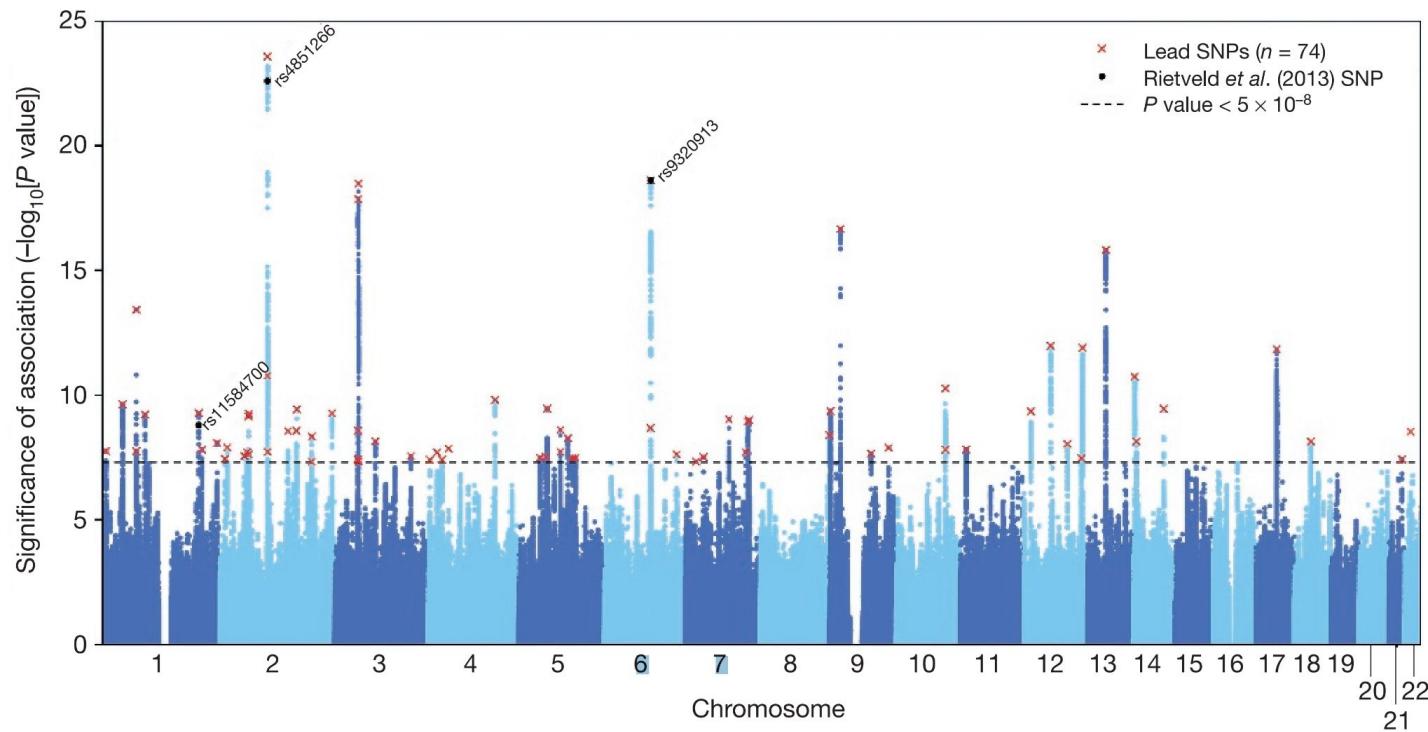
Binary traits

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

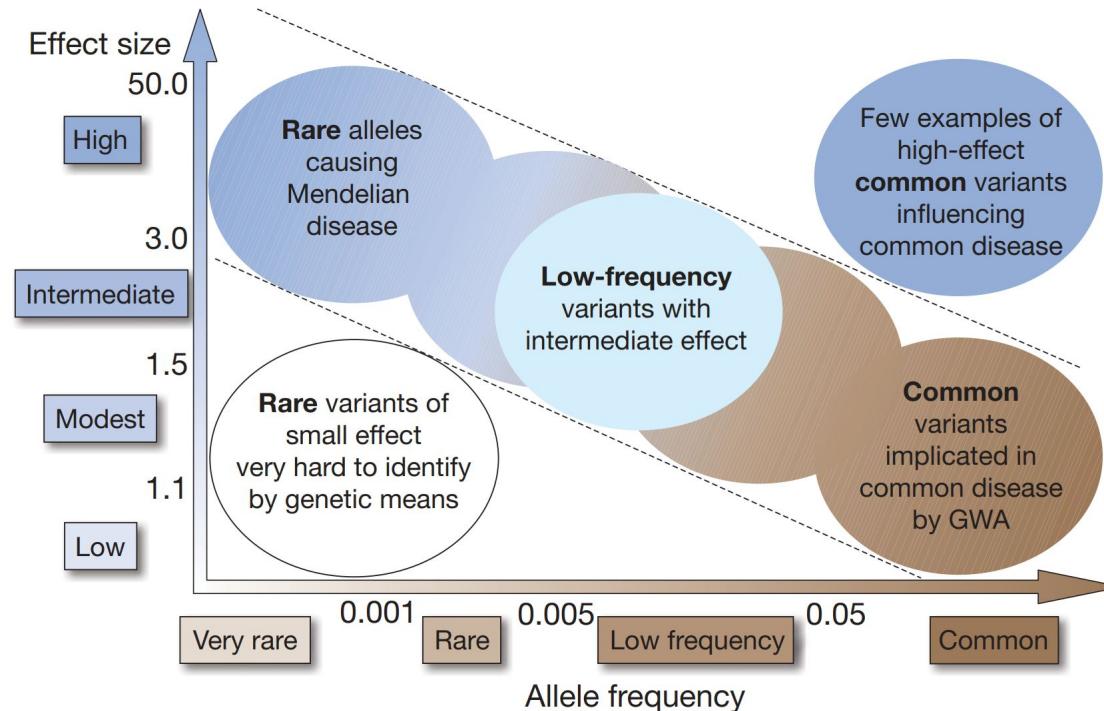
Testing for associations



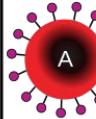
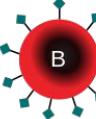
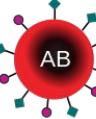
Manhattan plot for EduYears associations (n = 293,723)



Rare and common variants



Haplotypes: ABO serological groups

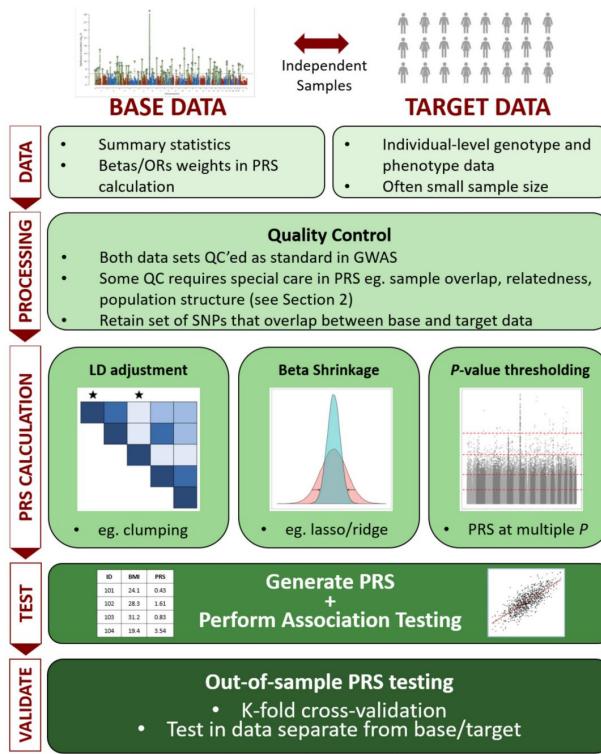
	Group A	Group B	Group AB	Group O
Red blood cell type				
Antibodies in plasma			None	
Antigens in red blood cell	A antigen	B antigen	A and B antigens	None

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

Polygenic Risk Scores

stratification & disease trajectories

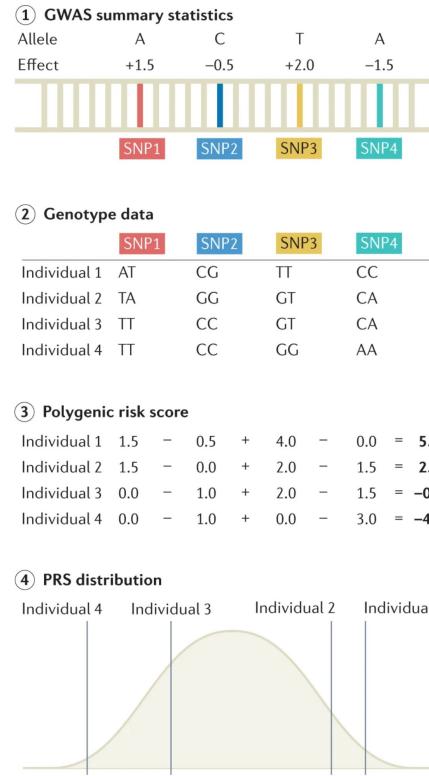
Common workflow - single-trait PRS



Polygenic risk scores (PRS)

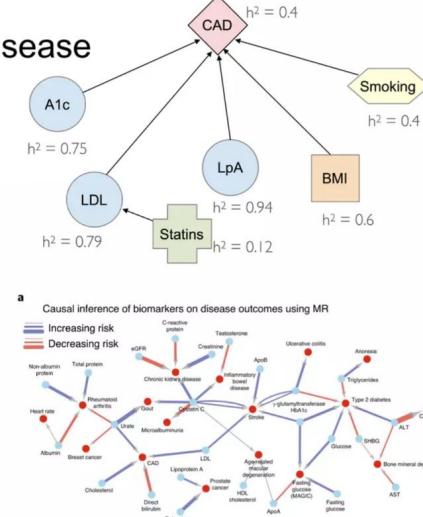
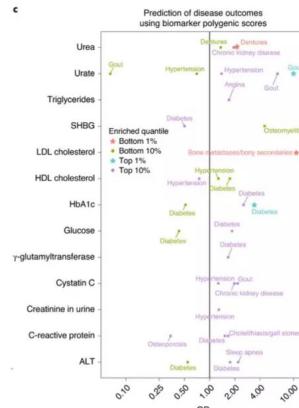
$$PRS_i = \sum_{j \in J} \beta_j G_{ij}$$

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

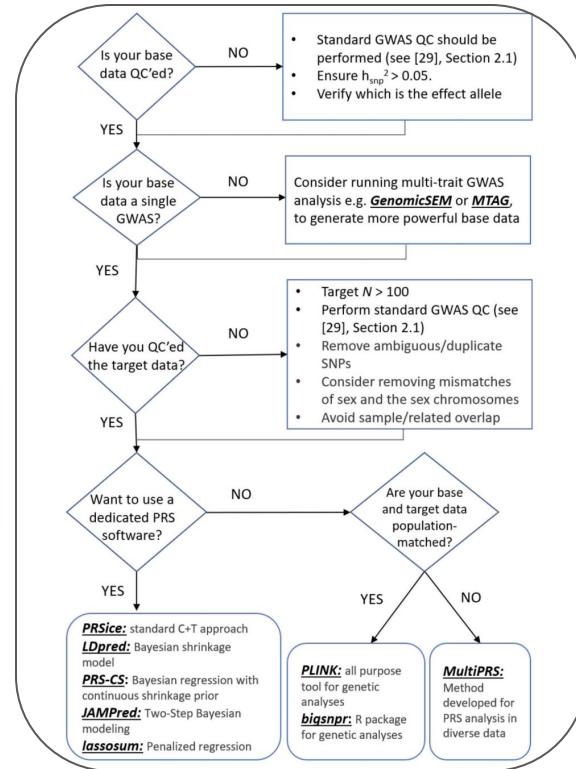


Common workflow - multi-trait PRS

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

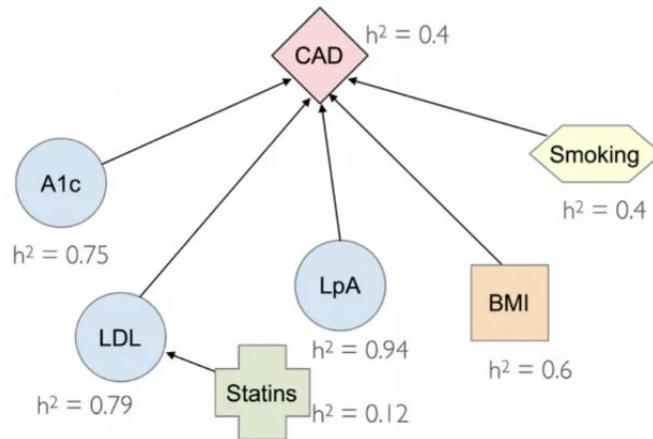


- Multi-PRS is a weighted sum of PRSs
i.e. $w_1(\text{PRS}_1) + w_2(\text{PRS}_2) + w_3(\text{PRS}_3) + \dots$

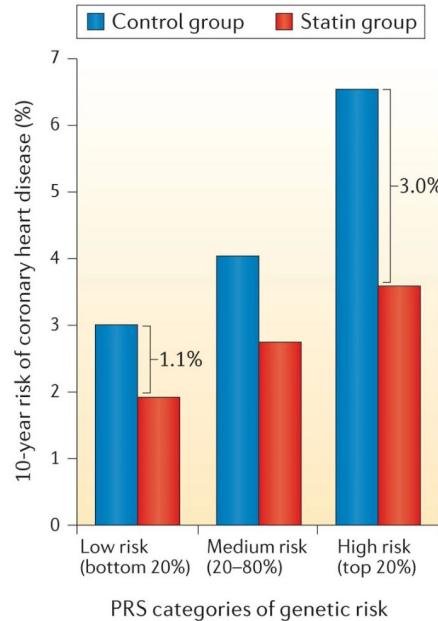


Common workflow - multi-trait PRS

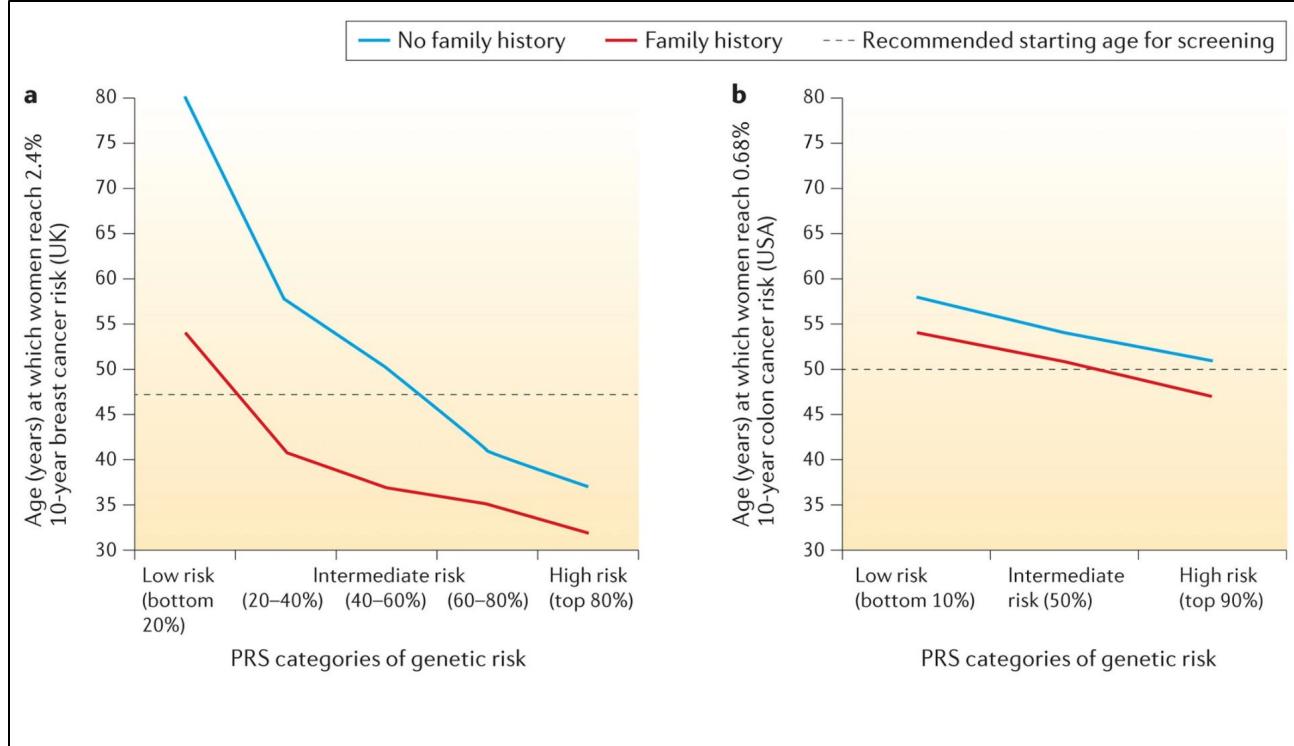
- Multi-PRS is a weighted sum of PRSs
i.e. $w_1(\text{PRS}_1) + w_2(\text{PRS}_2) + w_3(\text{PRS}_3) + \dots$



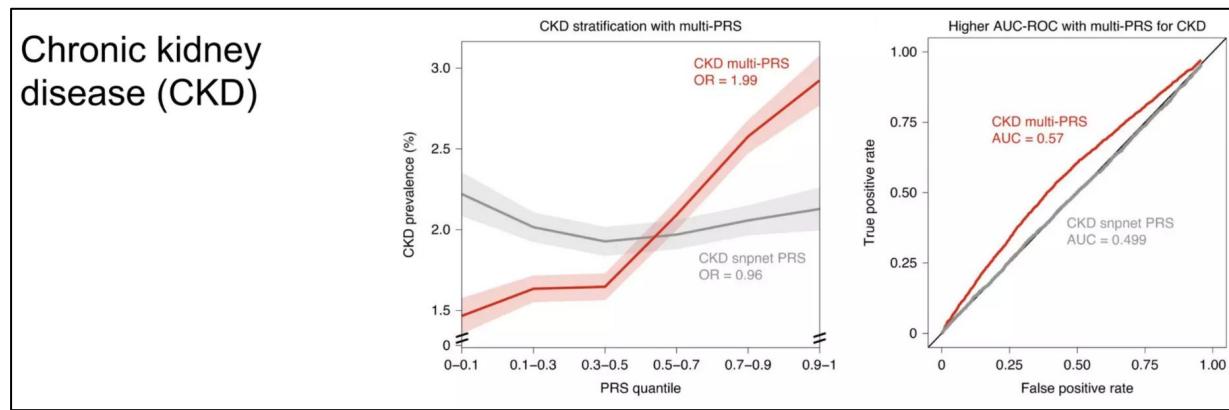
Case study - disease stratification



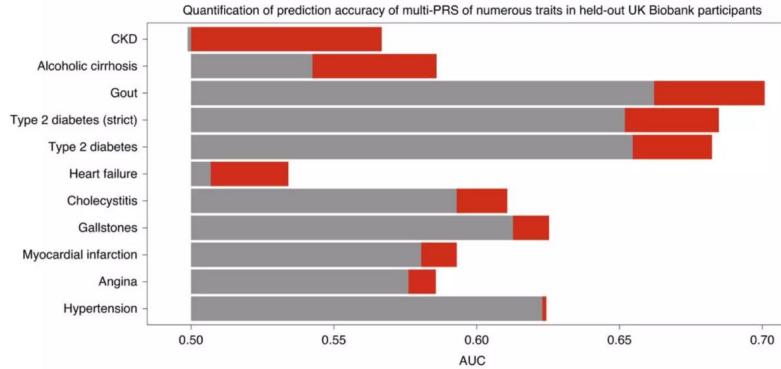
	ARR (%)	1.1	1.3	3.0
	RRR	0.36	0.32	0.46



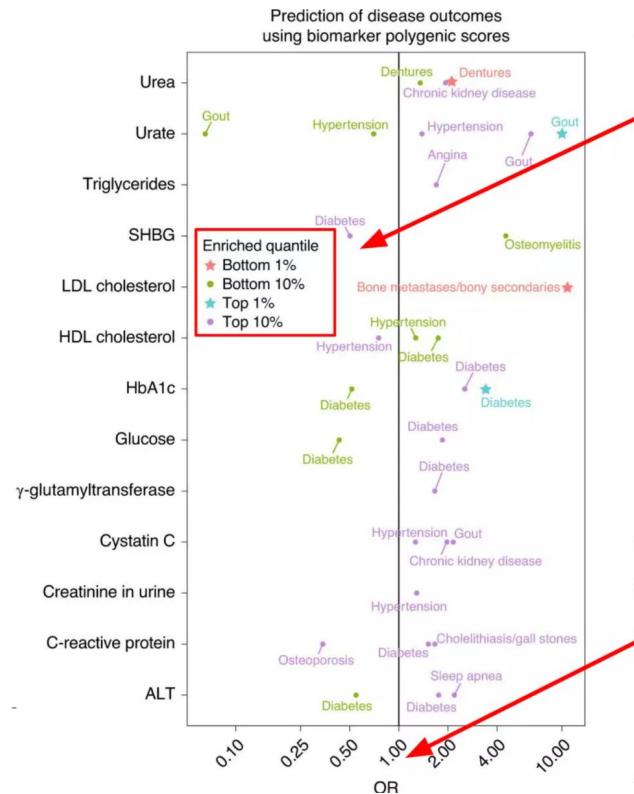
Case study - multi-trait PRS improves disease prevalence prediction



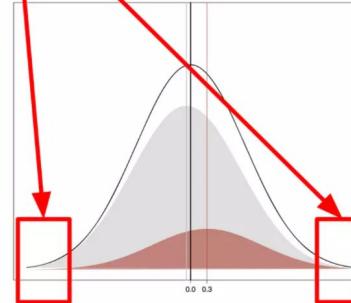
Other diseases in UK Biobank



Use of PRS for trait / disease prediction



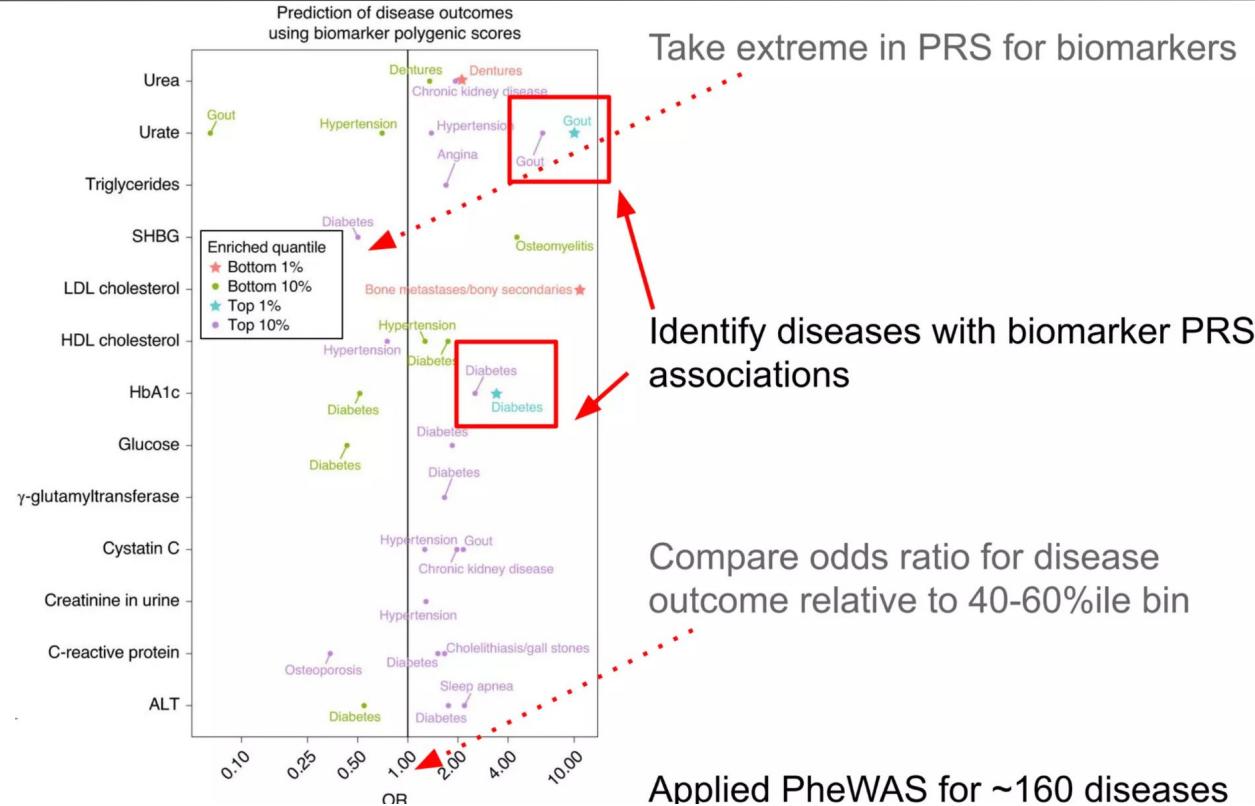
Take extreme in PRS for biomarkers



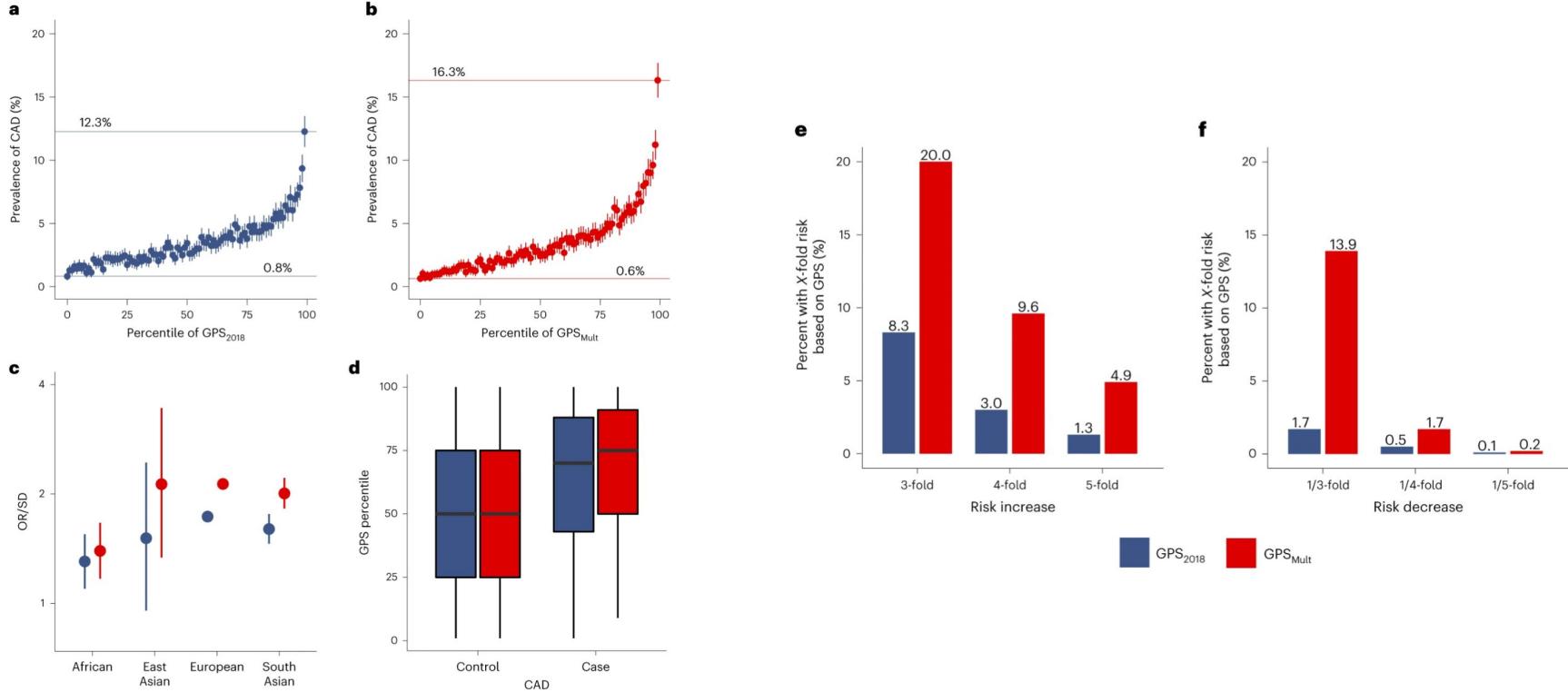
Compare odds ratio for disease outcome relative to 40-60%ile bin

Applied PheWAS for ~160 diseases

Use of PRS for trait / disease prediction

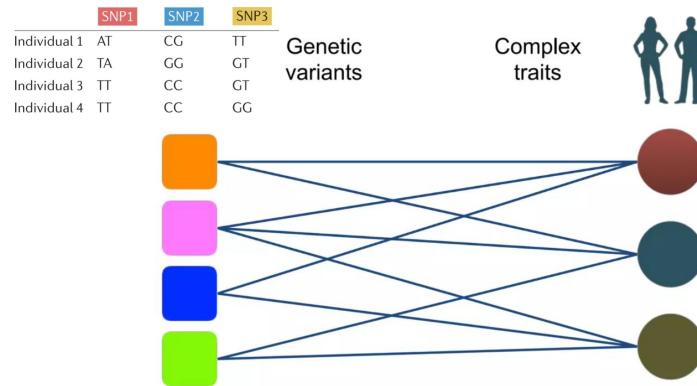


Limitations - ethnicity / ancestry



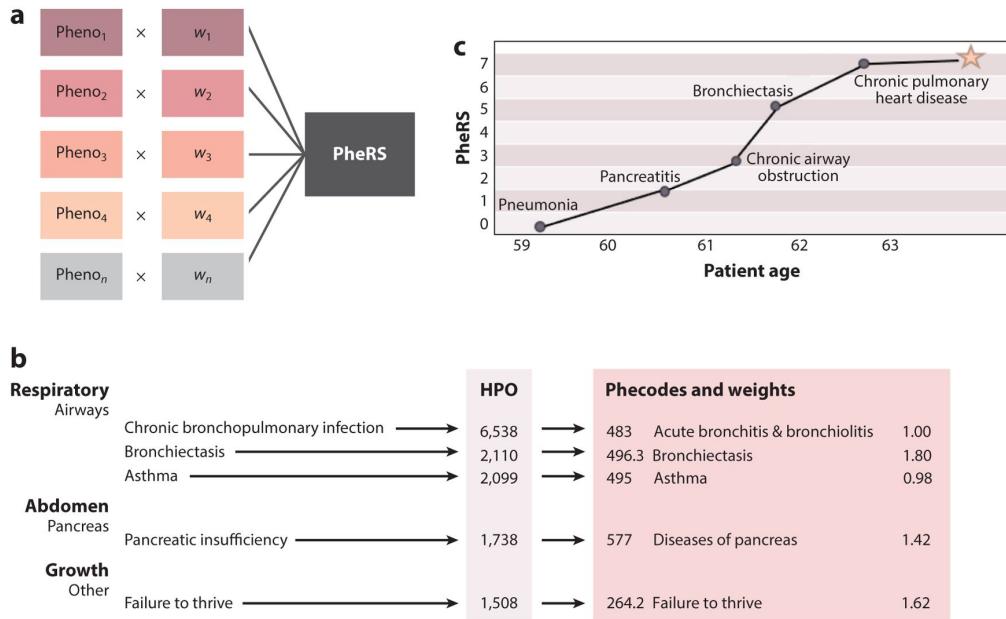
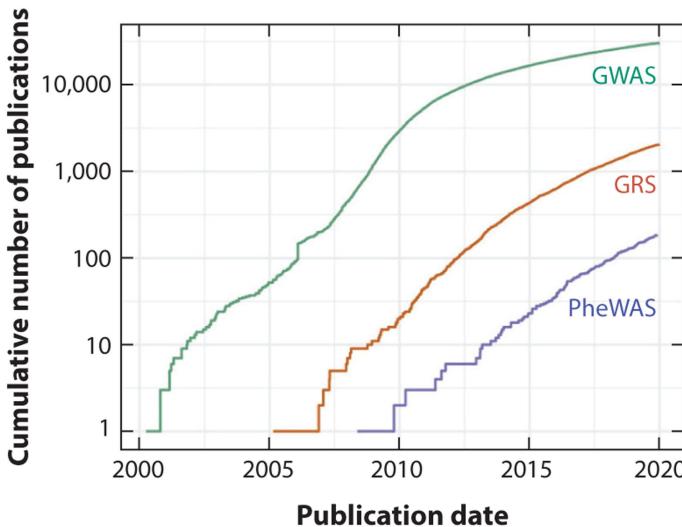
Limitations - polygenicity & pleiotropy

- 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?

Going further - EHR's & PRS



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