



GWAS Applications

from the
Health Data Science
Sandbox



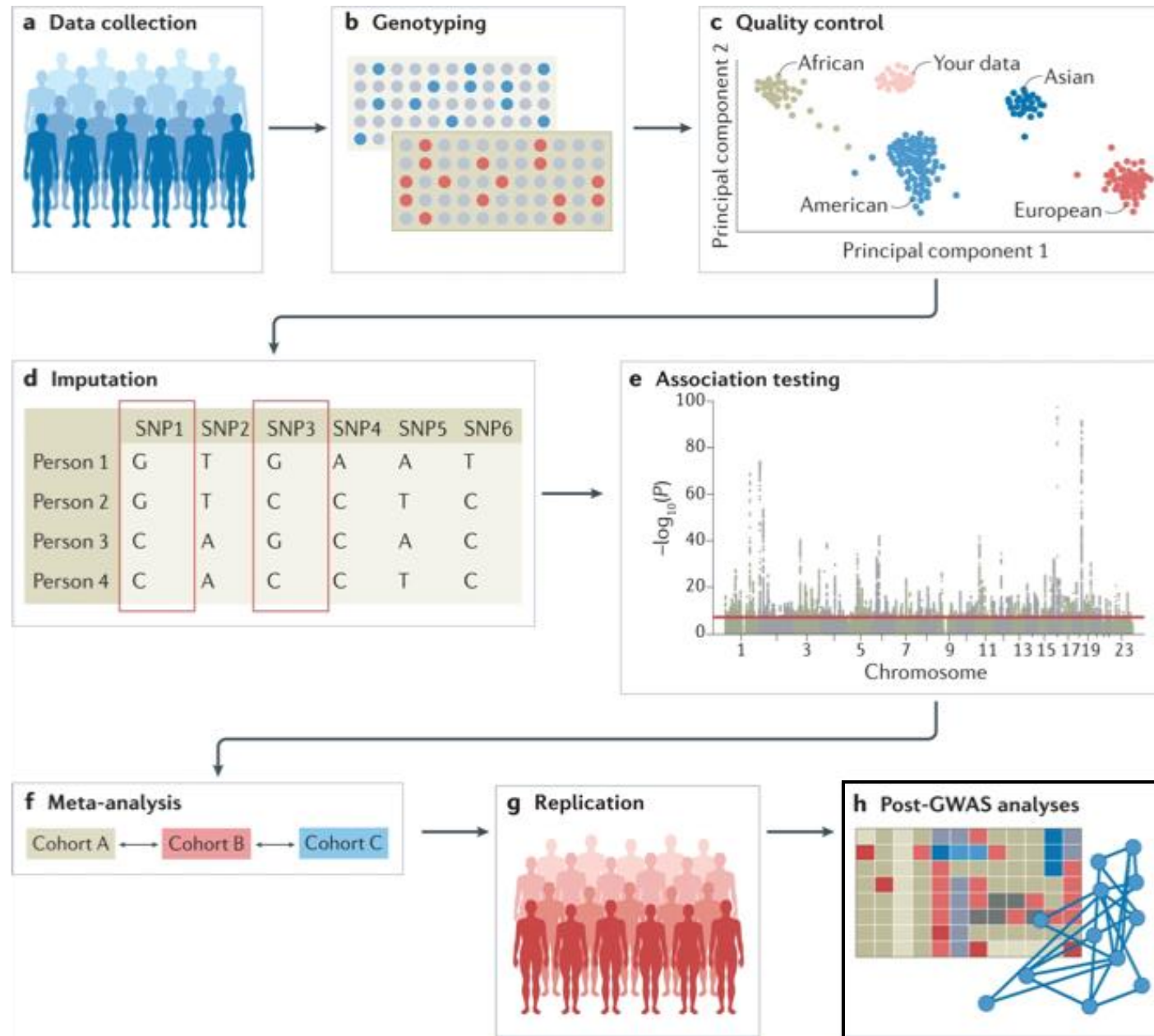
Alba Refoyo Martinez, PhD

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF
COPENHAGEN





Post-GWAS analysis

After the GWAS (and GWAS meta-analysis), the challenging work begins:

- Interpretation of the results
- Finding the causal variant (linkage disequilibrium)
- Assessing the causal gene or functional mechanism
- Pathway enrichment analyses, pleiotropic effects, risk prediction,
....



GWAS with the Genomics Sandbox



Today's topics

GWAS catalog

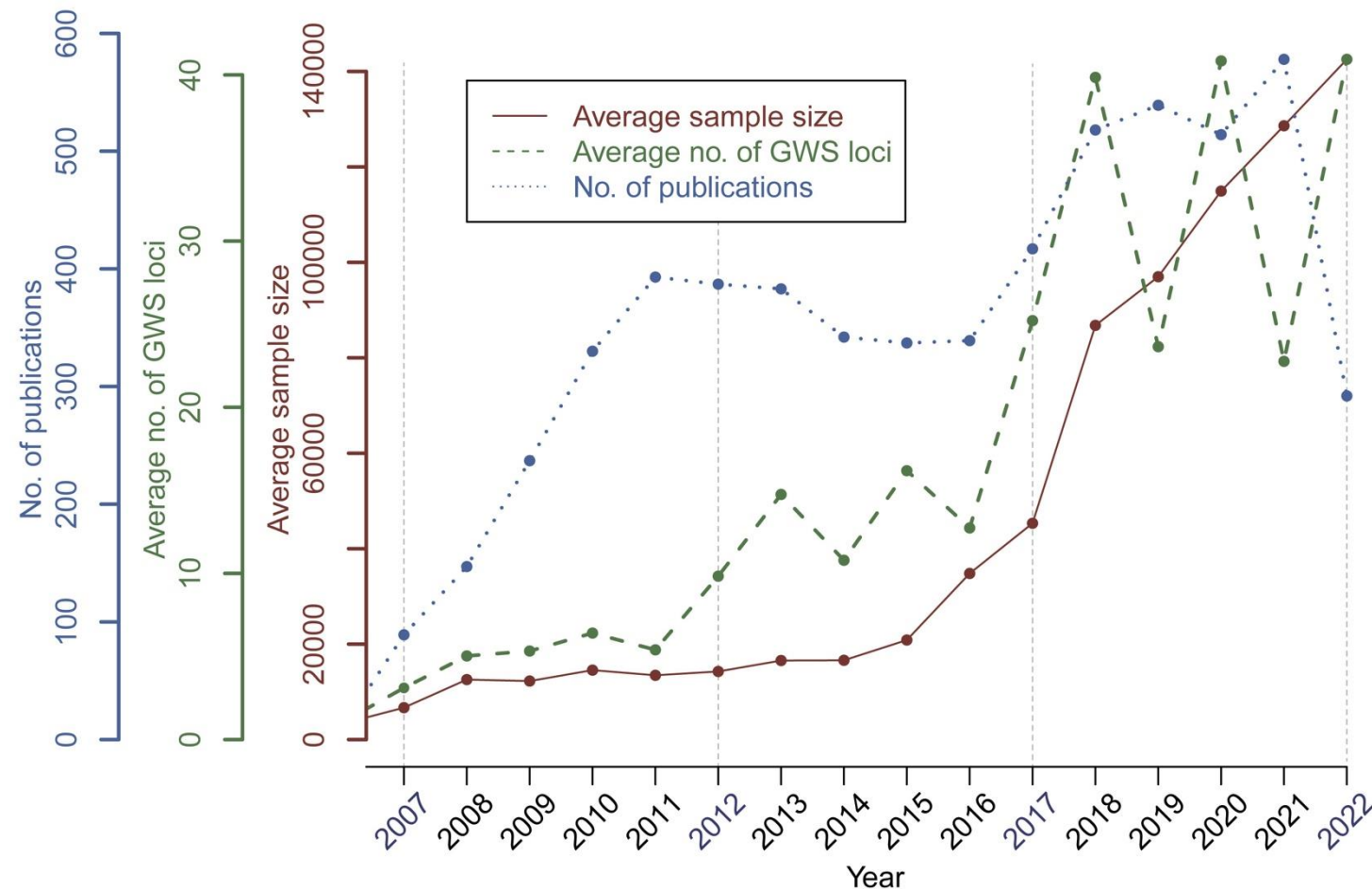
- Caveats and pitfalls

Polygenic scores

- What are PGS and PRS?
- How to calculate PGS
- Interpreting PGS
- Portability of PGS
- Caveats and pitfalls



Over the past 5 years, the average sample size per publication is >x3, increasing the number of significant associations



Over 650k GWAS-significant variants!!

85 000 full genome-wide summary statistics datasets available for downstream analysis (e.g. meta-analysis, PRS...).




GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000





GWAS Catalog Submission

Available data: **Associations** Studies Full summary statistics LocusZoom

GWAS-SSF standard format

chromosome	base_pair_location	effect_allele	other_allele	beta	standard_error	effect_allele_frequency	p_value
1	869388	A	G	-0.016619	0.00806496	0.997221	0.1
1	205813916	G	C	-0.008959	0.00331941	0.983589	9.70E-03
2	70478797	T	TG	0.0187528	0.00167685	0.934121	3.50E-30
7	8458030	TC	T	-0.0184	0.00101051	0.78451	5.70E-76
23	24173186	A	C	0.0038776	0.08757958	0.627178	2.30E-08

Variant and risk allele	P-value	RAF	OR	CI	Mapped gene	Reported trait	Trait(s)
rs2075650-G	1 x 10 ⁻²⁹⁵	0.14	2.53	[2.41-2.66]	TOMM40	Alzheimer's disease	Alzheimer disease

- S** 110,000 Studies
- P** 7,000 Publications
- V** 650,000 Associations
- T** 15,000 Mapped trait ontology terms

Parent Directory

- GCST90104534_buildGRCh37.tsv.gz
- GCST90104534_buildGRCh37.tsv.gz-meta
- README.txt
- harmonised/
- md5sum.txt

Name	Last modified	Size	Description
GCST90104534.h.tsv.gz	2024-08-11 02:40	311M	
GCST90104534.h.tsv.gz-meta.yaml	2024-08-11 02:40	1.4K	
GCST90104534.h.tsv.gz.tbi	2024-08-11 02:40	1.4M	
GCST90104534.running_log	2024-08-11 02:40	2.1K	
md5sum.txt	2024-08-11 02:40	116	

Home / Documentation

Curation of population descriptors

A description of our data extraction and standardisation process.

Diversity analysis

Distribution of ancestry labels in the GWAS Catalog.



Author recommendations

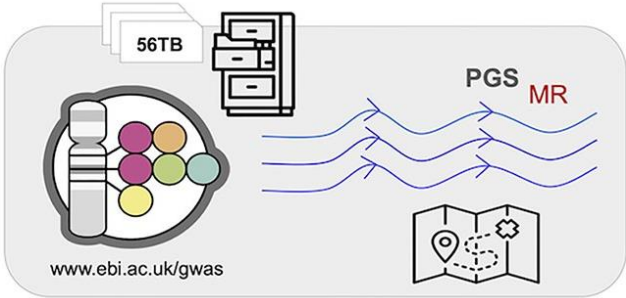
Recommendations for the reporting of GWAS sample metadata.



Select study type:

- ☒ GxE
- ☐ seqGWAS

Study information	
Reported trait	Age at adiposity rebound x sex interaction
Genotyping technology	Genome-wide genotyping array
Discovery sample description	904 Admixed Chileans with European and Native American ancestry individuals
Discovery ancestry label (country of recruitment)	904 Hispanic or Latin American (Chile)
PubMed ID	36844456
First author	Vicuña L
Full Summary Statistics	FTP Download
GxE	<input checked="" type="checkbox"/>



Example of GWAS Catalog for lung cancer

A

(1) New tabbed layout, loads only the current table

Available data: Associations 1154 Studies 118 **Full summary statistics 29** LocusZoom Download Associations

☐ Include background traits data
☒ Include child trait data

Studies with summary statistics 29

Show 5 entries

First author	Study accession	Pub. date	Reported trait	Trait(s)	Association count	Summary statistics
Wei X	GCST90277434	2023-10-17	Lung cancer in never smokers	lung carcinoma	9	FTP Download
Walters RG	GCST90246027	2023-07-20	ICD10 C34: Malignant neoplasm of bronchus and lung	lung carcinoma	1	FTP Download
Lebrecht MB	GCST90315948	2023-05-05	Non-small cell lung cancer	non-small cell lung carcinoma	17	FTP Download
Byun J	GCST90134661	2022-08-01	Lung cancer	lung carcinoma	16	FTP Download
Jiang L	GCST90043863	2021-11-04	ICD10 C34.1: Malignant neoplasm of upper lobe, bronchus or lung	lung carcinoma	0	FTP Download

Showing 1 to 5 of 29 entries

« 1 2 3 4 5 6 »

(2) Improved pagination allows rows to be loaded as needed

B

Select study type:
☒ GxE
☐ seqGWAS

Studies 880

Show 5 entries

First author	Study accession	Pub. date	Reported trait
Lelievre R	GCST90399837	2024-07-31	Serum ascorbic acid 2-sulfate levels x sex interaction

C

PubMed ID 37659414

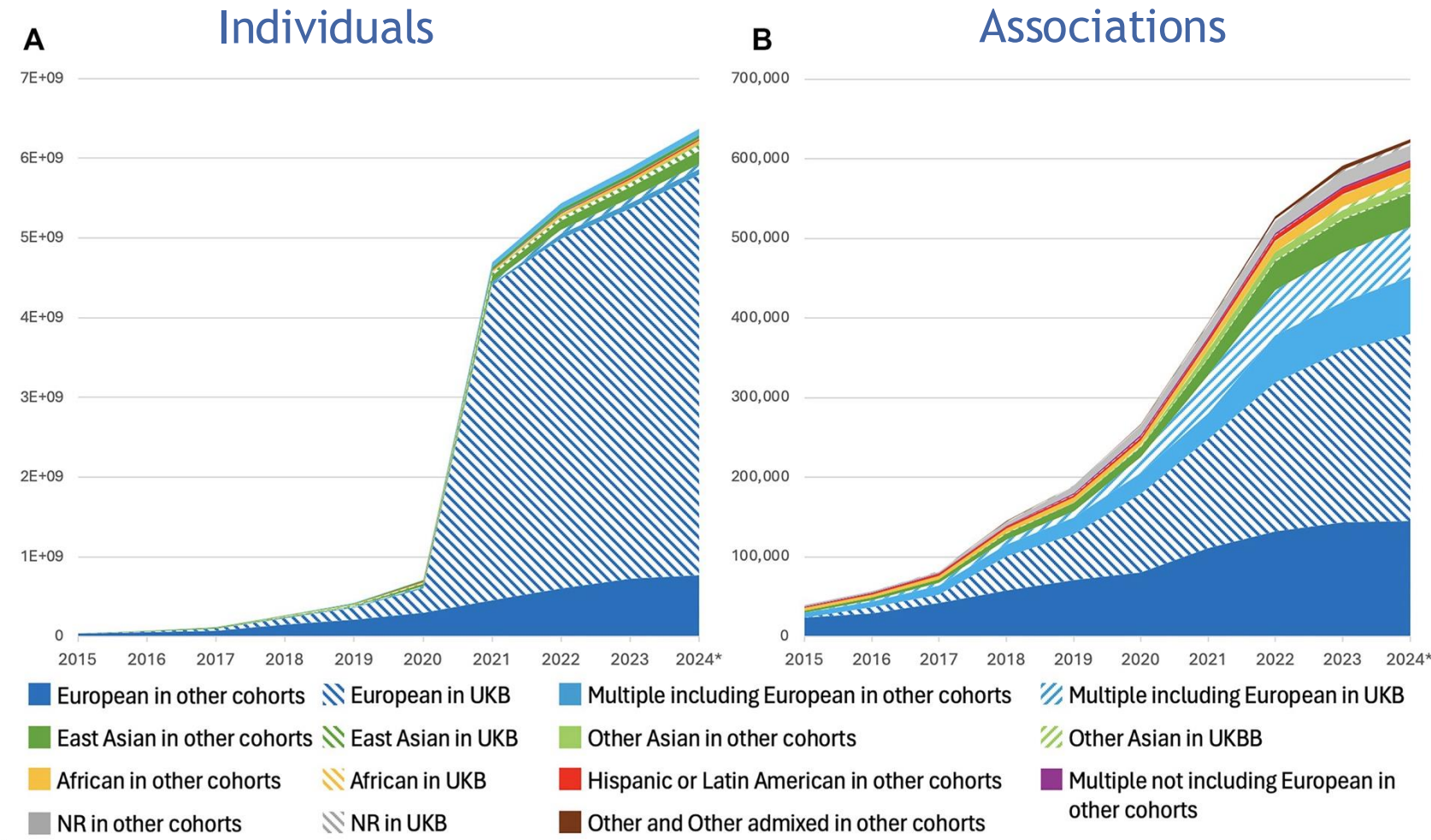
First author Zhang C

Full Summary Statistics FTP Download

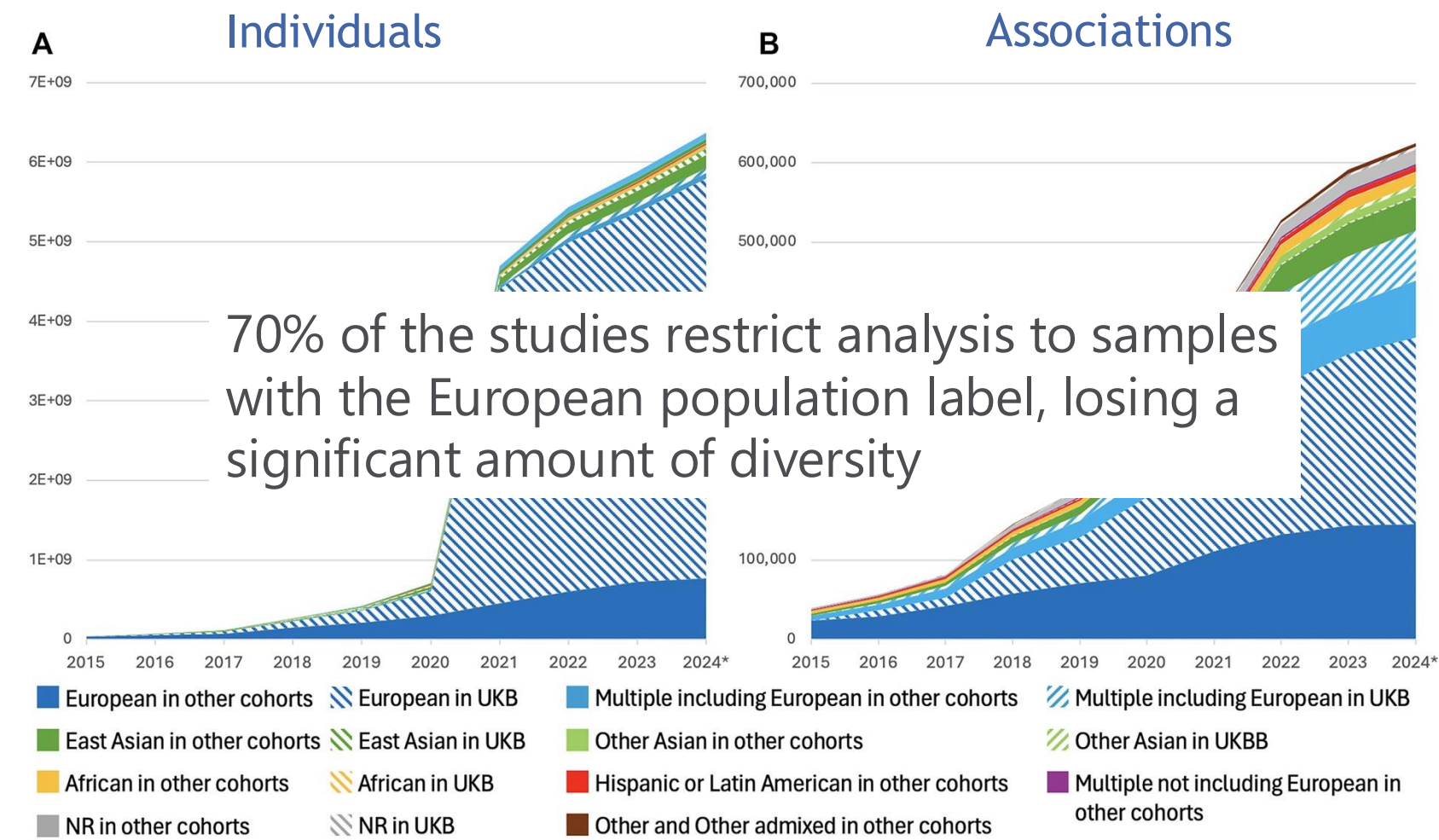
GxE ✓



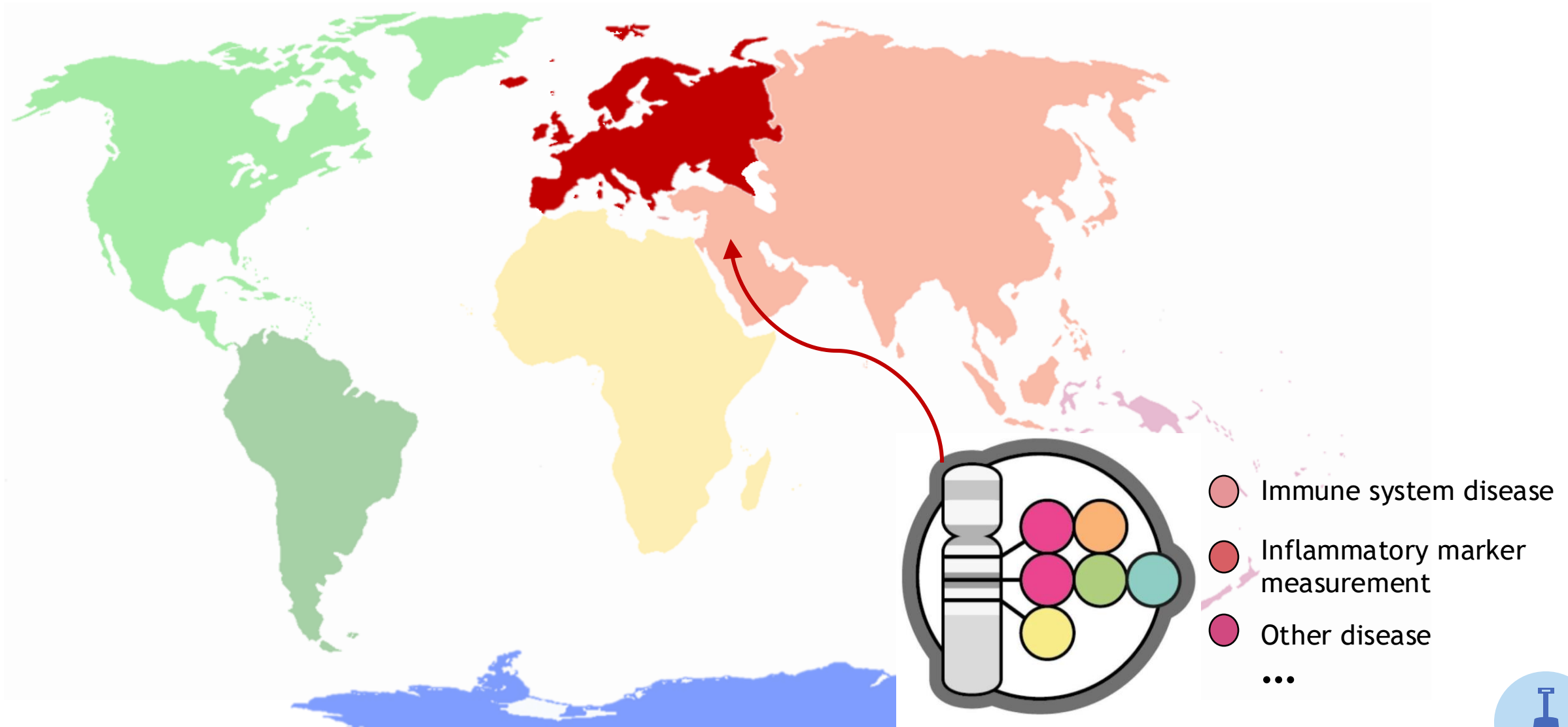
UK Biobank contribution to the ancestry in the GWAS catalog



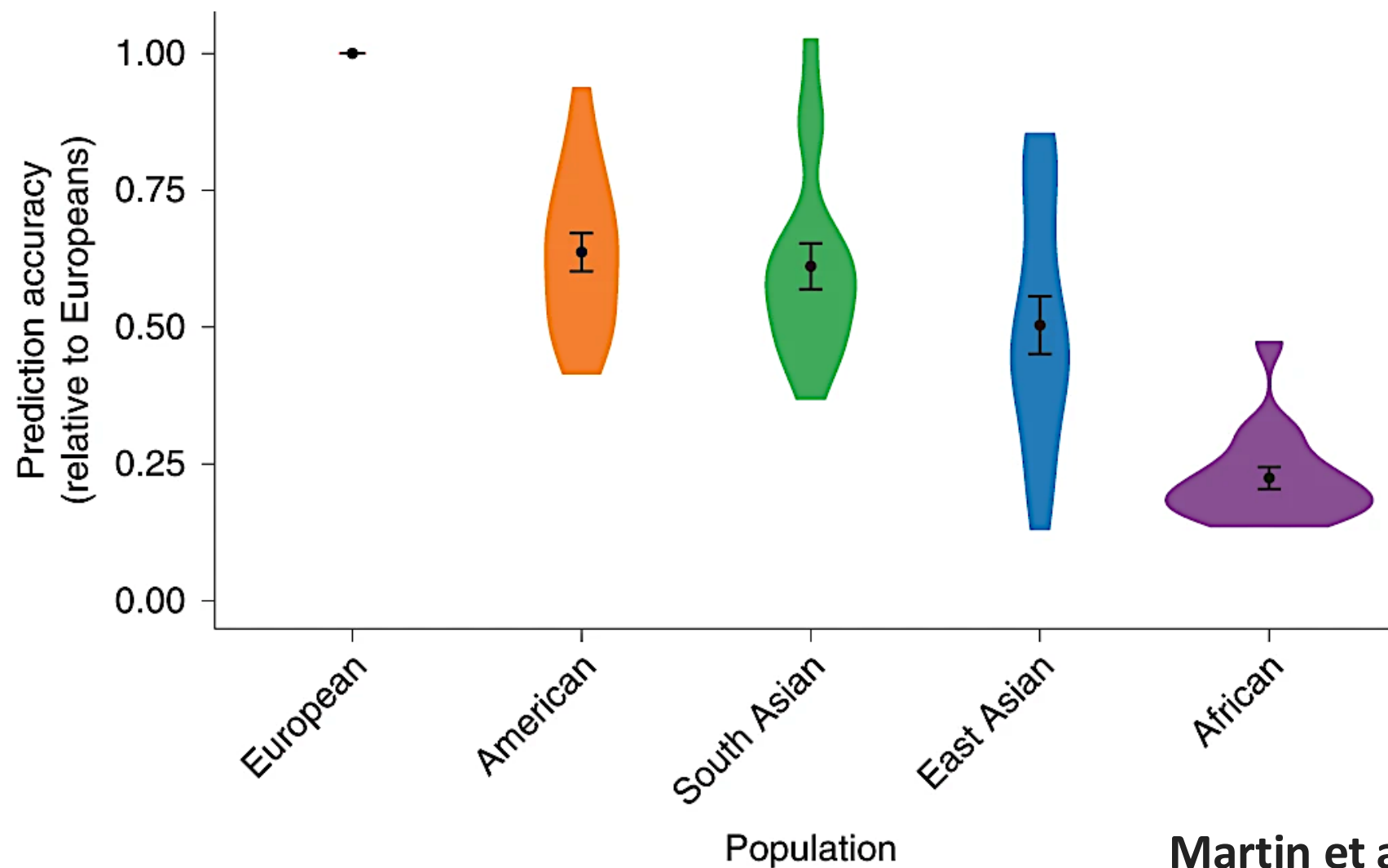
UK Biobank contribution to the ancestry in the GWAS catalog



European genomic data won't represent all human genomic differences



Prediction accuracy relative to European-ancestry individuals



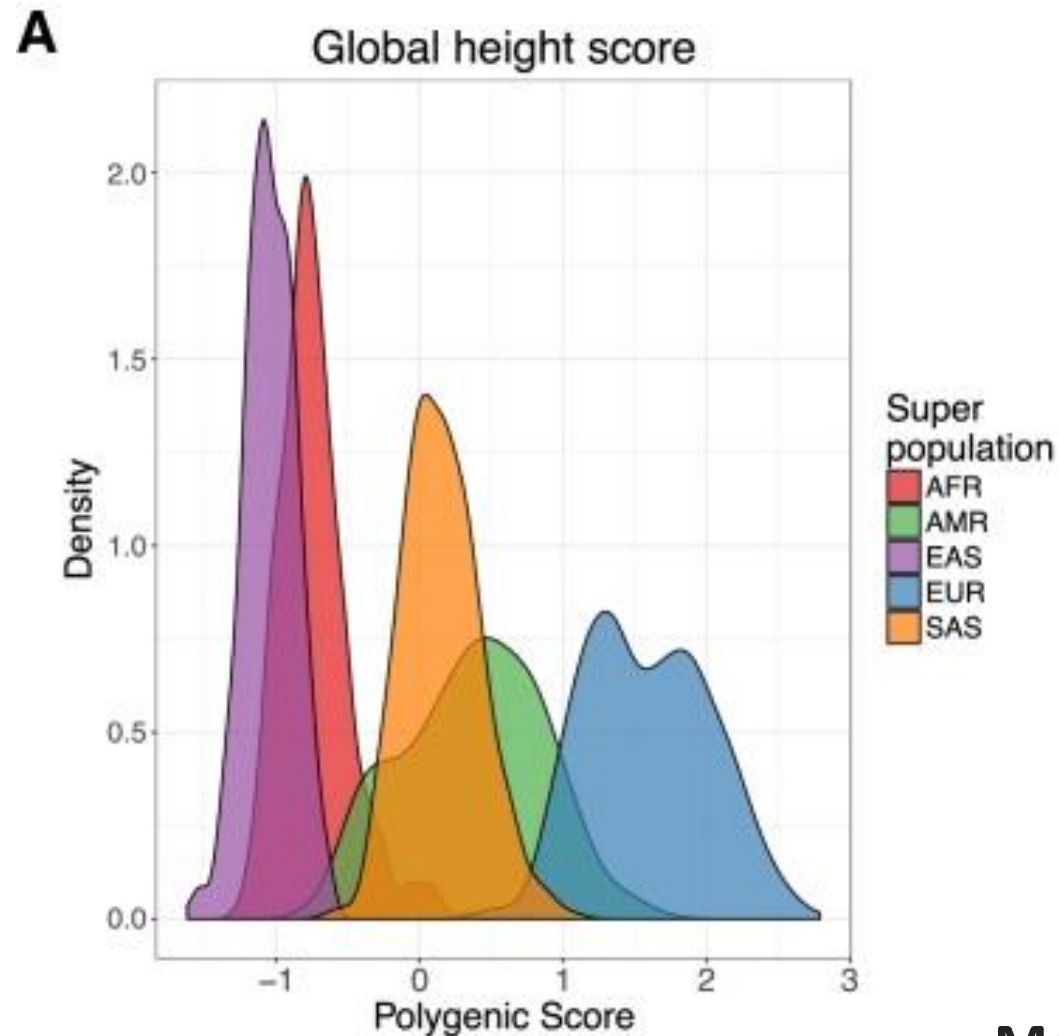
Martin et al. 2019



PRS are not portable across global populations

Base GWAS:
Europeans

How would
Africans look
compared to
Europeans if
these scores
were accurate?

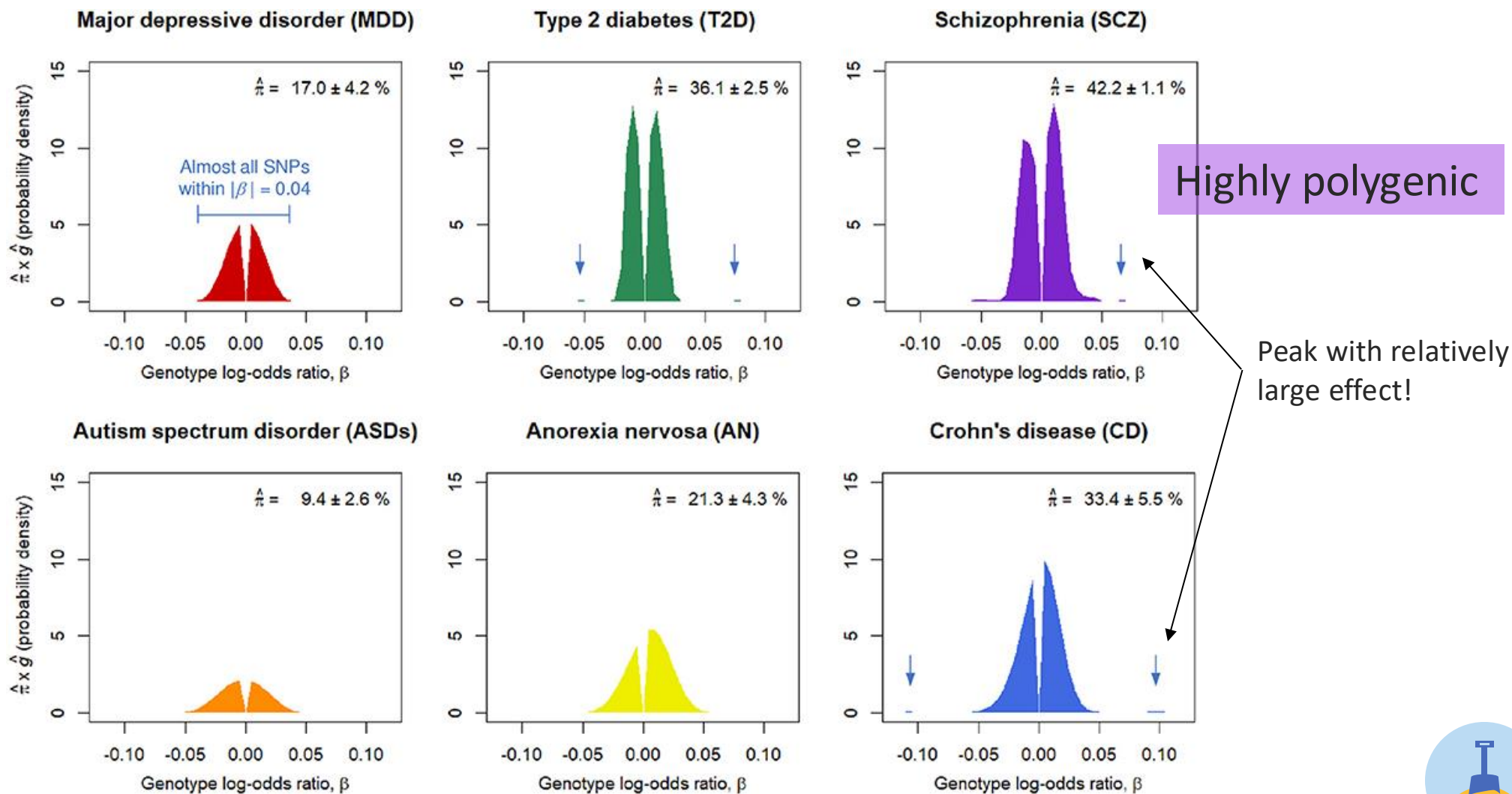


Martin et al. 2017



Effect sizes of GWAS variants are (mostly) small

Carrying a single disease variant increases your risk by less than 5%!

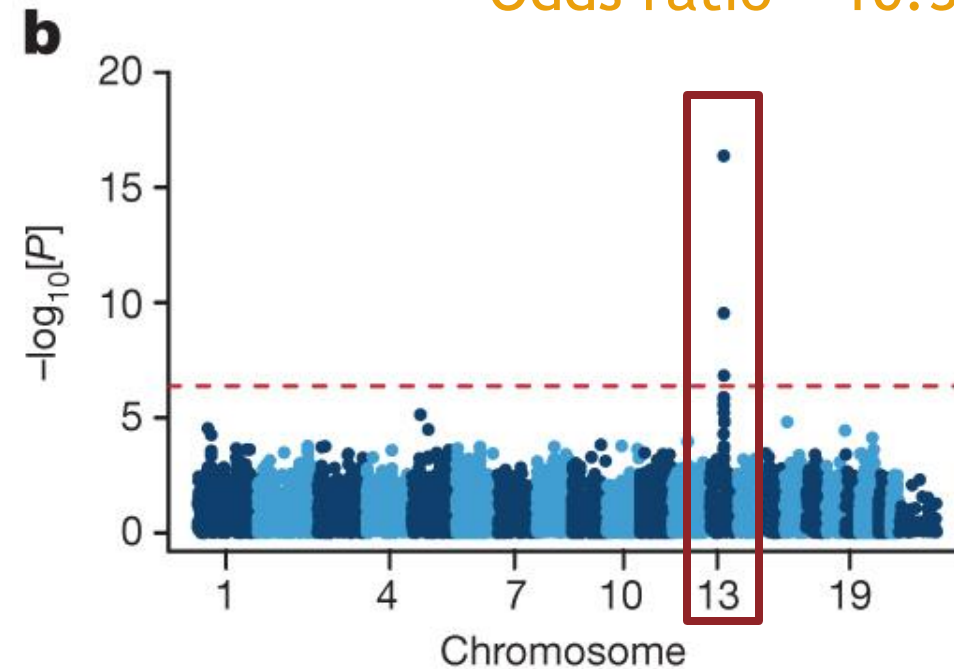
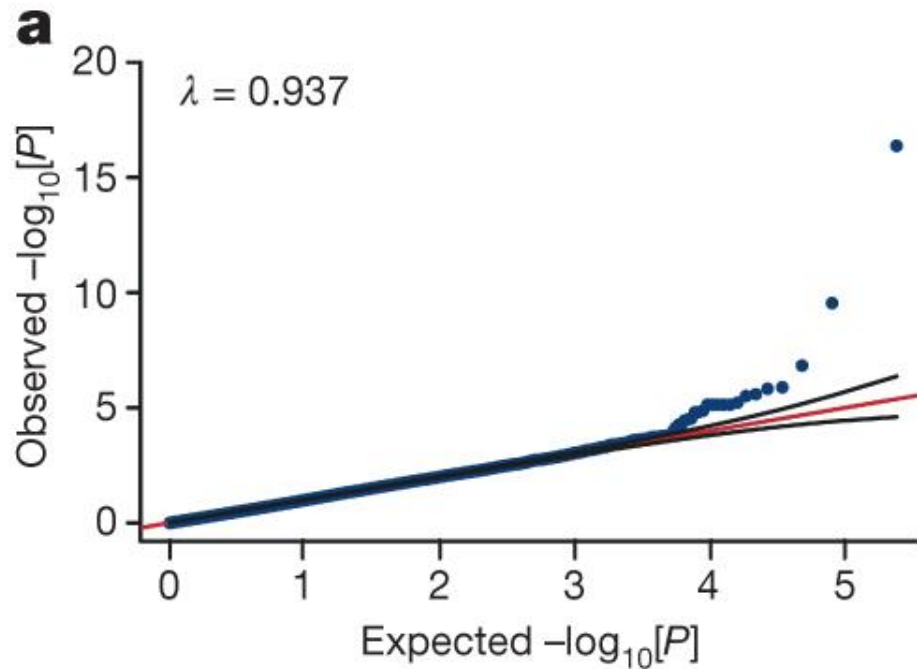


Proportion of Disease-Associated SNPs and Their Effect-Size Distributions



Some exceptions...

A common Greenlandic TBC1D4 variant confers insulin resistance and type 2 diabetes

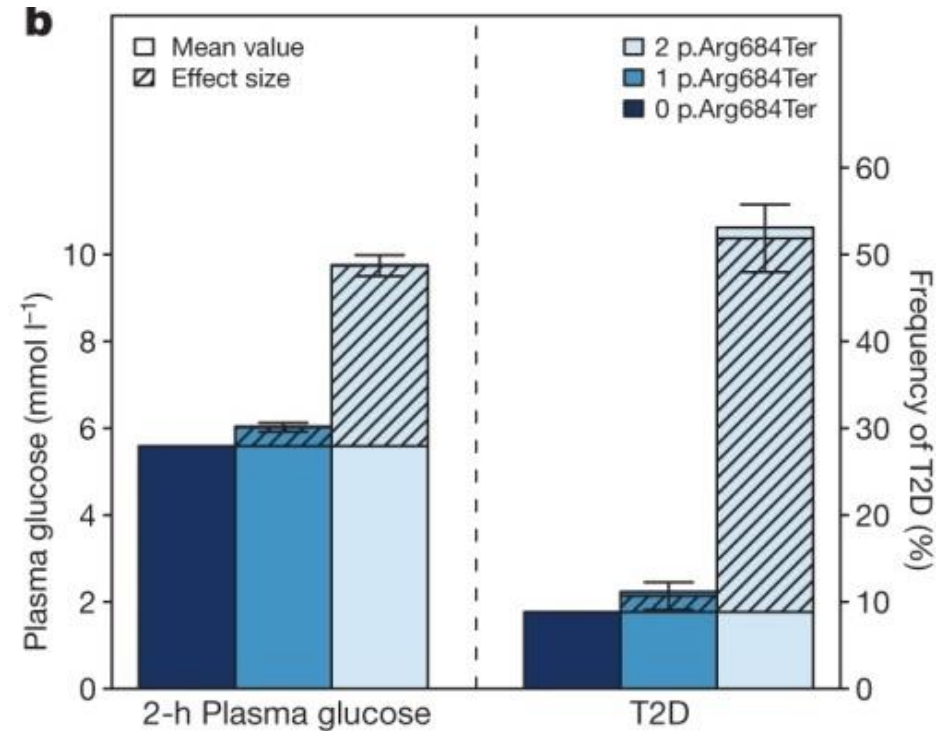
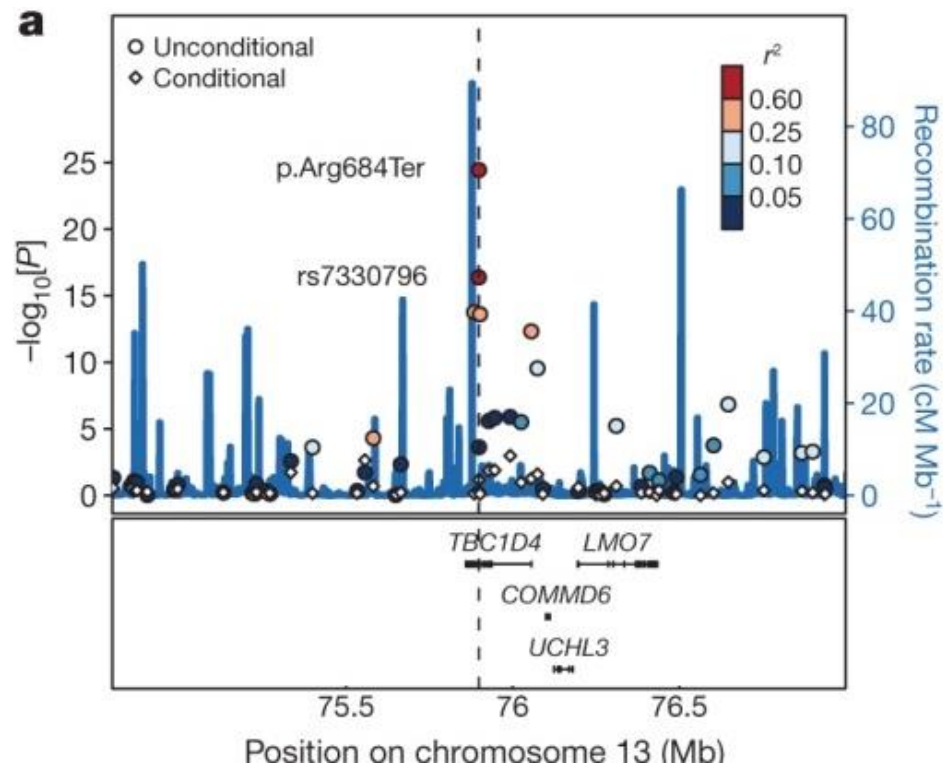


Some exceptions...

A common Greenlandic TBC1D4 variant confers insulin resistance and type 2 diabetes

zoom-in

Odds ratio = 10.3



What is a polygenic risk score (PGS)?

Many variants across the genome affecting a trait, each with a **small effect**
Pooling information across all significant variants to derive a composite predictor

Genome-based predictor about the overall risk of having a disease, or the genetic value for continuous traits.

Genotype at associated-SNP
“j” in an individual

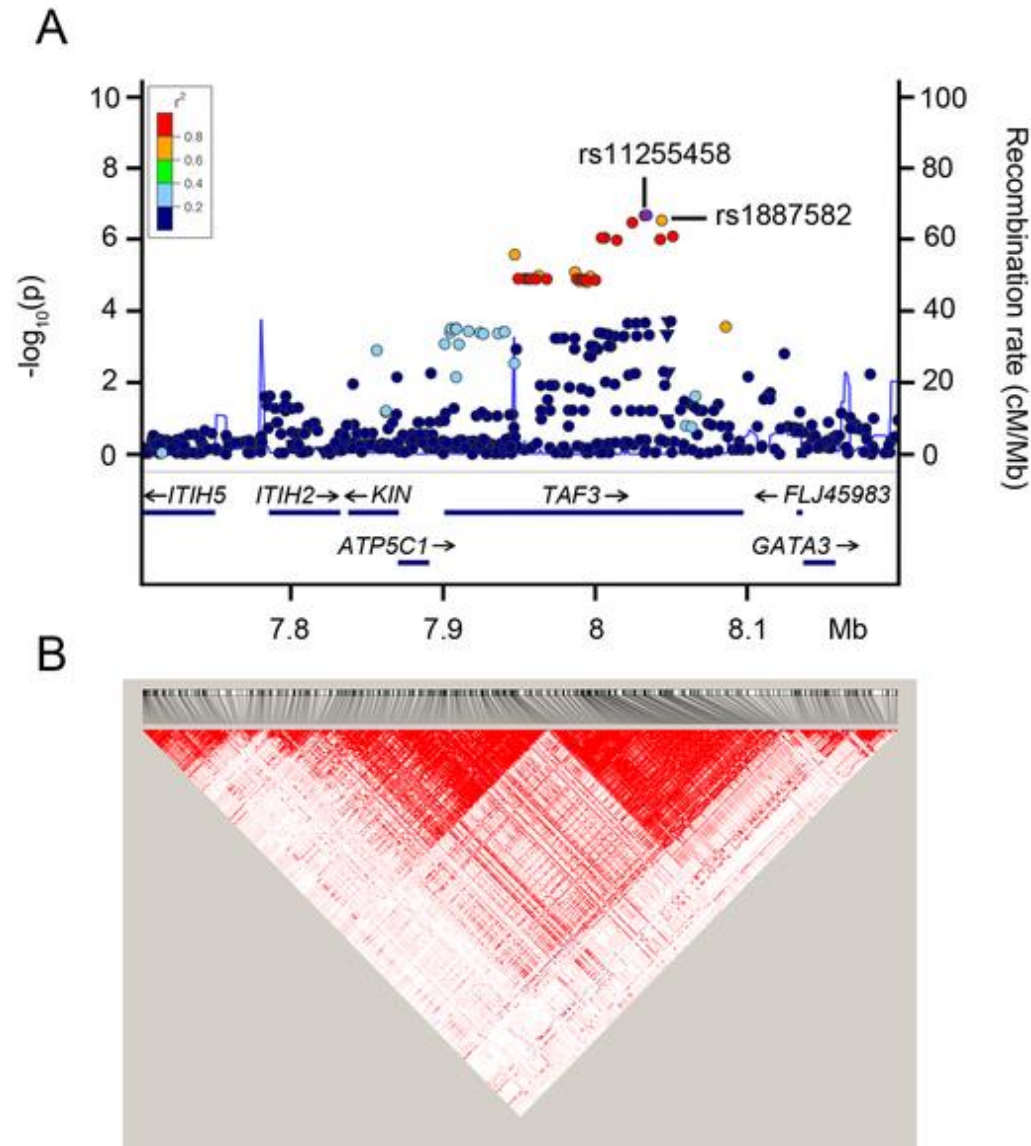
Effect size of
SNP “j” on trait

Polygenic score
for an individual

$$\hat{S} = \sum_{j=1}^m X_j \hat{\beta}_j$$



Problem: SNPs are not independent



Regional association plot

<http://locuszoom.org>

- rs1887582 association is due to LD to rs11255458
- 1 recombination hot spot separating 2 LD-blocks

Linkage disequilibrium pattern



How are polygenic scores calculated?

Naïve methods: **a priori filter** SNPs so that the ones included in the model are approximately independent while only using significantly associated (e.g. genome-wide significant)

Bayesian methods: **explicitly account for** the linkage disequilibrium (LD) across the genome by using a **prior on the effect sizes** that depends on the LD surrounding a SNP. They take into account the underlying genetic structure.

Penalized regression methods: use all SNPs in the genome, but it **penalize large regression coefficients** for many SNPs; learning a "sparse" model where only some SNPs contribute to the trait



Standard method

Clumping and thresholding

- Consider only SNPs with $P\text{-value} < \text{cutoff}$
- Among SNPs in LD ($LD > r^2$), choose the one with the smaller $P\text{-value}$
 - This process “clumps” significant SNPs with each other and picks the most significant
 - R^2 alone determines whether 2 SNPs have independent signals
- Use marginal allelic effect estimated in PRS calculation
- To optimise performance you can tune the cutoff and the r^2



Adjusting effect sizes can increase PRS accuracy using LD information from a external reference panel

LDpred

(Vilhjalmsson et al. AJHG 97:576-592)



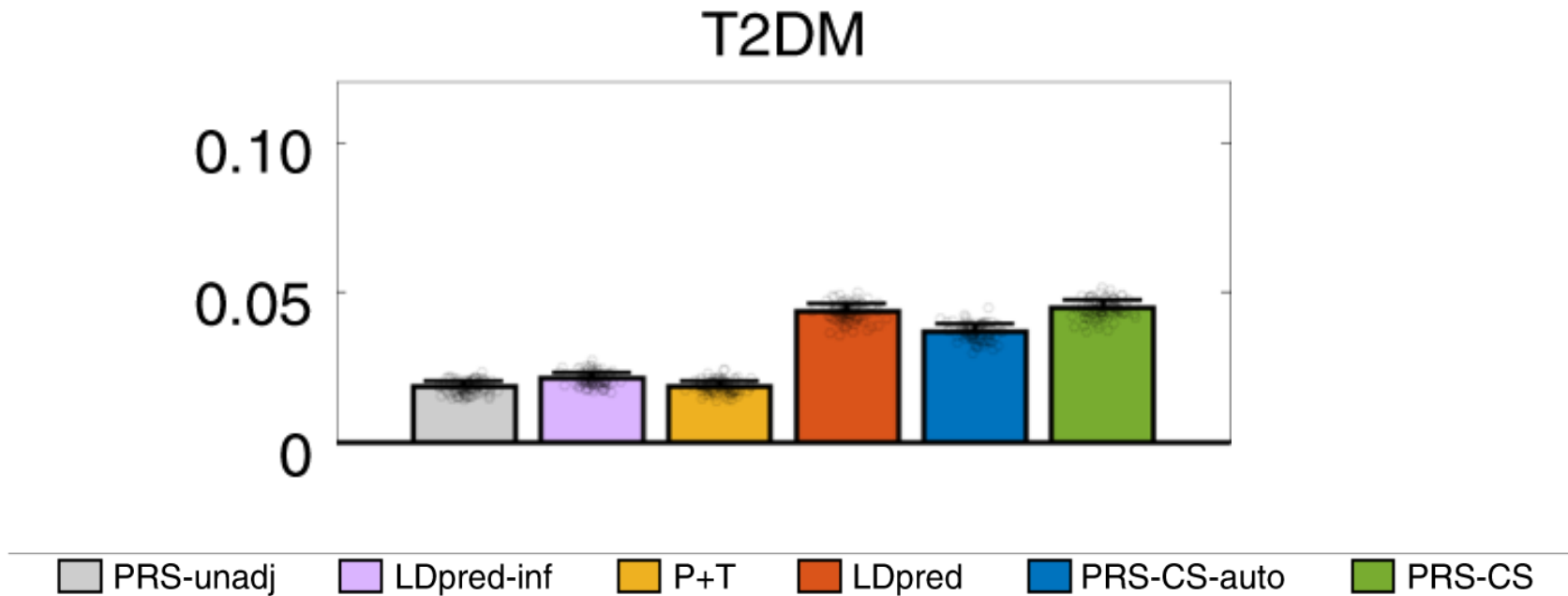
LDpred

(Vilhjalmsson et al. AJHG 97:576-592)

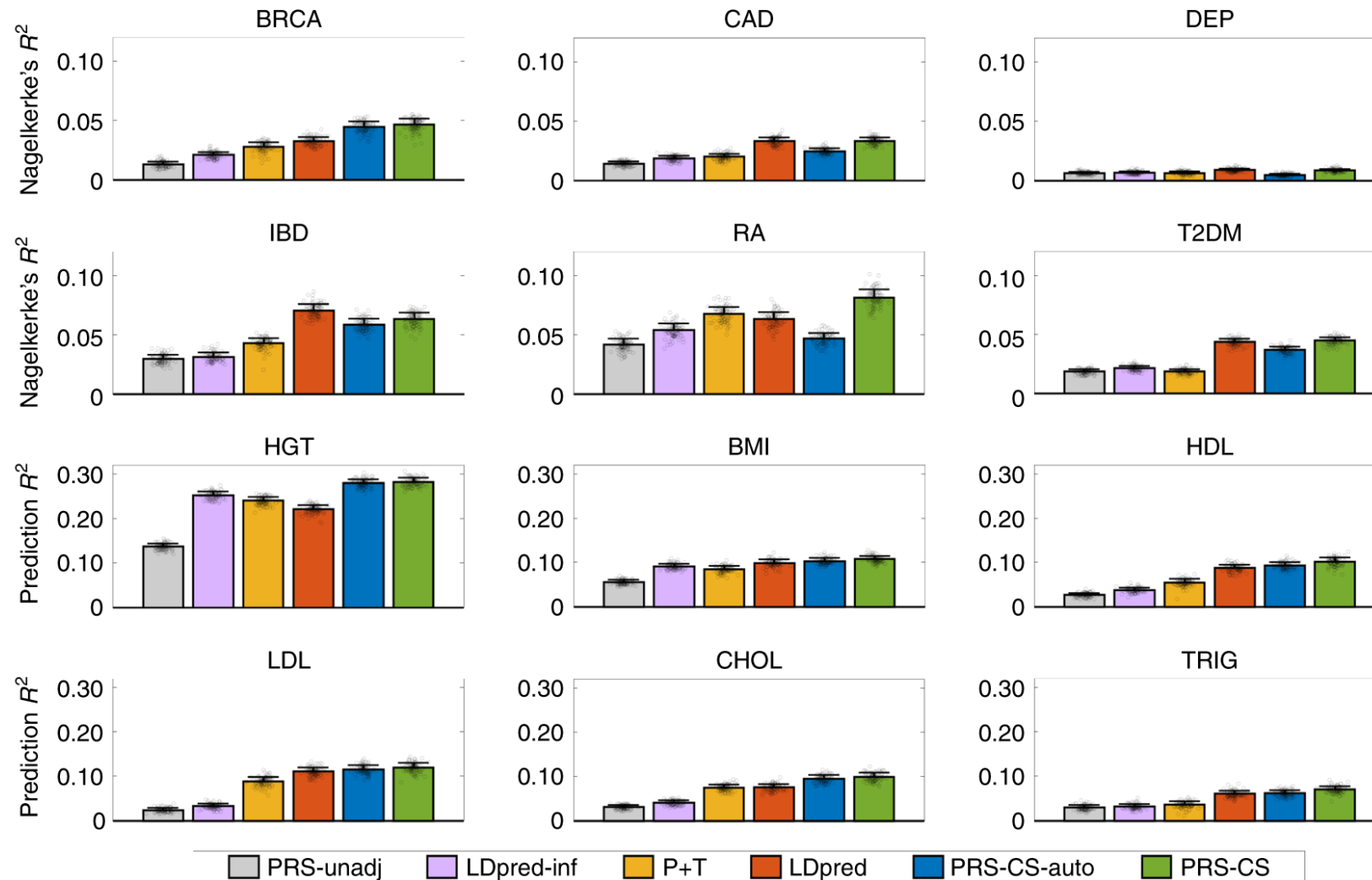
- Assume a prior $\lambda_l \sim \begin{cases} N\left(0, \frac{h^2}{p\theta}\right) & \text{with prob. } \theta \\ 0, & \text{with prob. } 1 - \theta \end{cases}$
- Given marginal GWAS effect estimates $\hat{\beta} = (\hat{\beta}_l)$ and their SEs, LDpred computes the posterior expectation of the causal effects $E(\lambda \mid \hat{\beta}, \mathbf{R}, h^2, \theta)$, where \mathbf{R} is the LD matrix
 - In practice, the LD matrix is only considered within a predefined window
 - The heritability estimate (h^2) can be obtained externally using methods such as linear mixed models (LMM) or linkage disequilibrium score regression (LDSC)
- A grid θ values is evaluated to identify the best-performing model
- The estimated causal effects are then used as weights in PRS



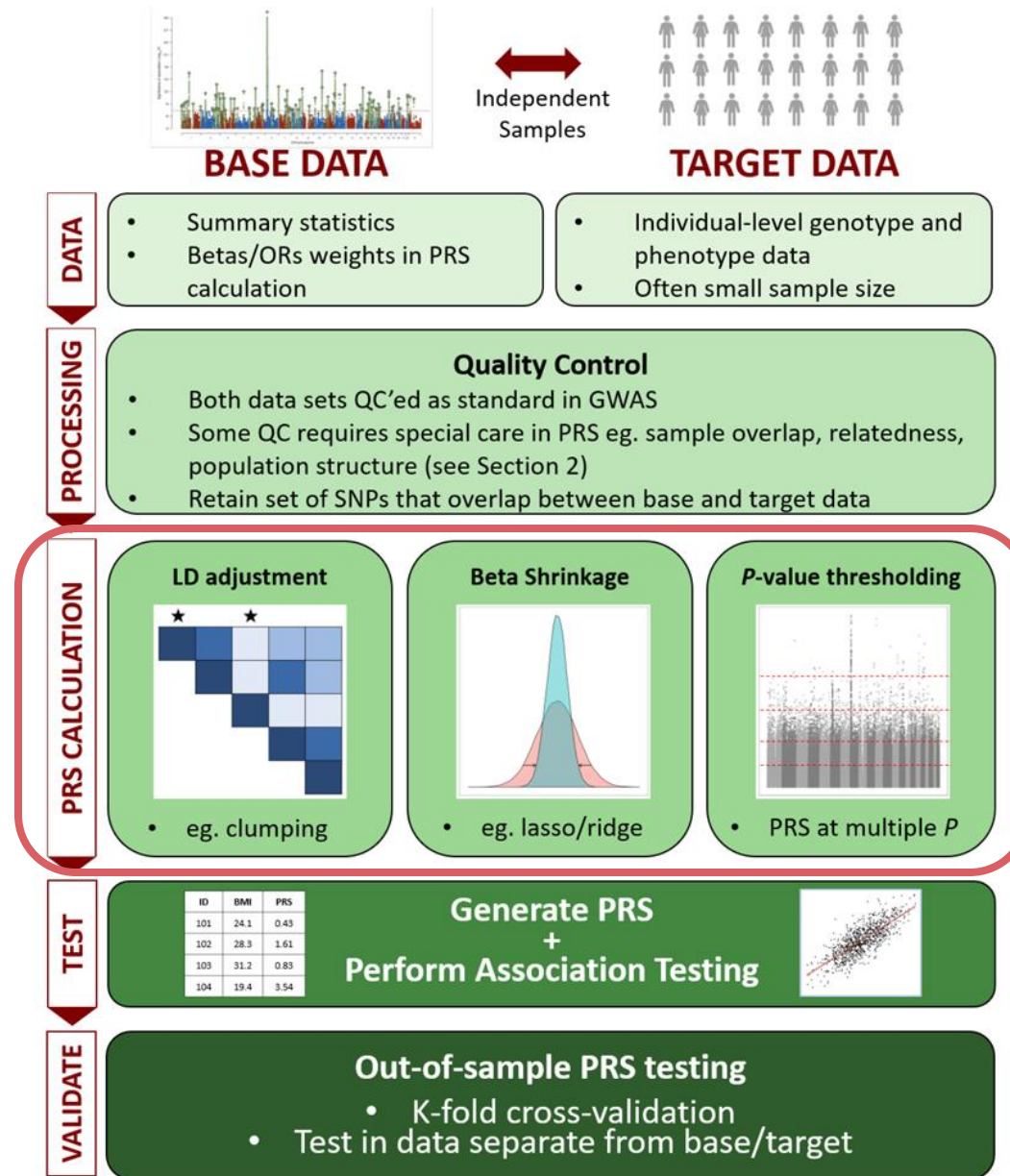
Prediction accuracy of six polygenic prediction methods in the Partners HealthCare Biobank



Prediction accuracy of six polygenic prediction methods in the Partners HealthCare Biobank



Recap



How accurate are polygenic risk scores?

- **Heritability of the trait** (variance attributed to genetic differences)
- **What aspect of the trait are we trying to predict?** (onset, different subtypes, severity...)
- **Who are we predicting the trait on?** (differences in genetic architecture such AF or LD patterns)
- **The power of the base GWAS** (quality + size)
- **The power of the method used to build the score** (how well it accounts for the complexity of the trait)



What if we applied PRS to “populations”?

Effect size of SNP “l” on trait

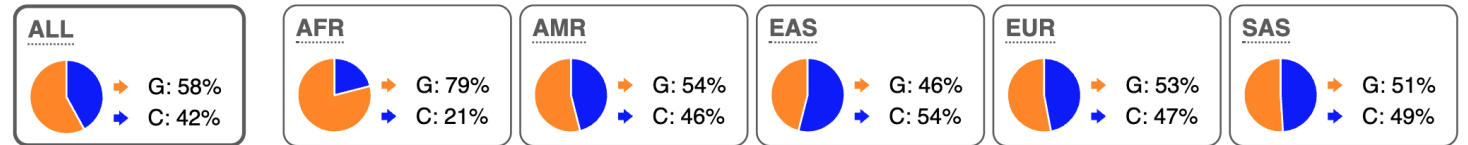
Allele **frequency** at
associated-SNP “l”

$$Z_m = \sum_{l=1}^L 2\alpha_l p_{l,m}$$



How are polygenic scores constructed?

1000 Genomes Project Phase 3 allele frequencies



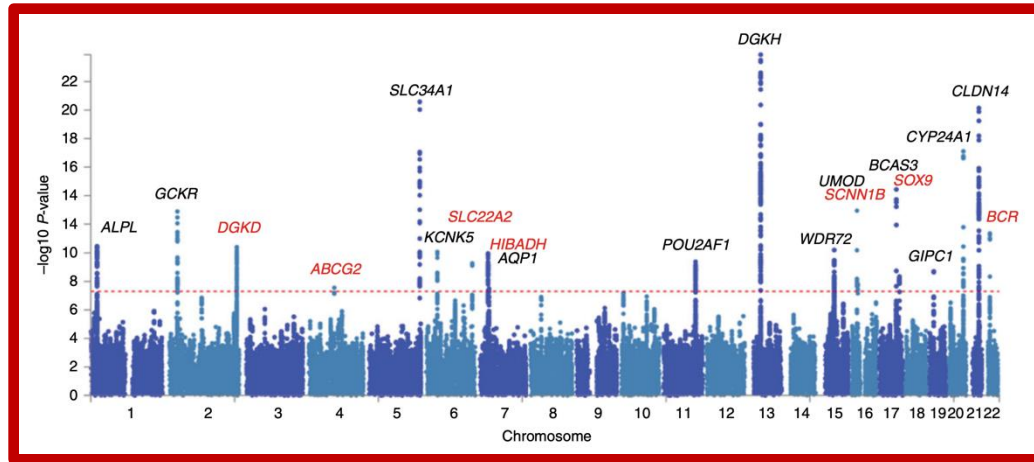
Allele frequency from 26 populations
across 5 super-populations

$$PGS = Z_m = \sum_{l=1}^L 2\alpha_l p_{l,m}$$

A dashed arrow points from the text "Allele frequency" to the term $p_{l,m}$ in the equation.



How are polygenic scores constructed?

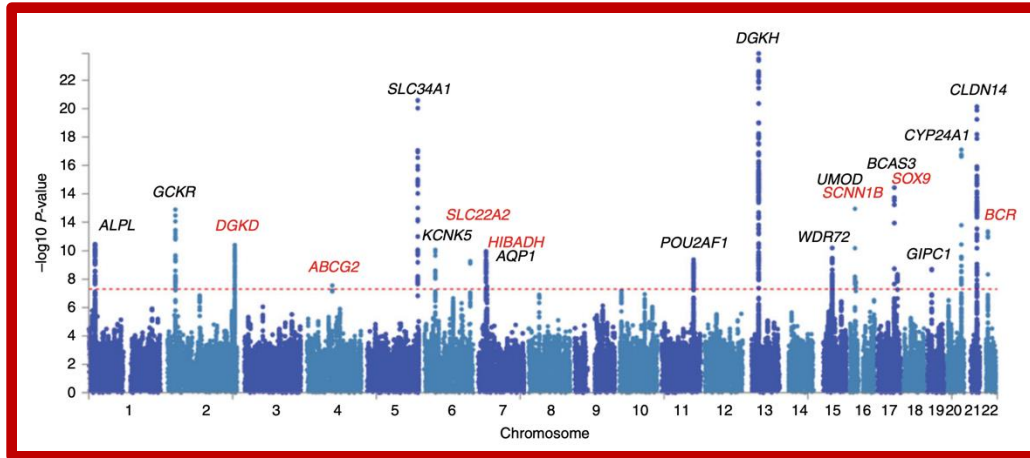


Effect size from trait-associated
SNPs -estimates from **UK Biobank**

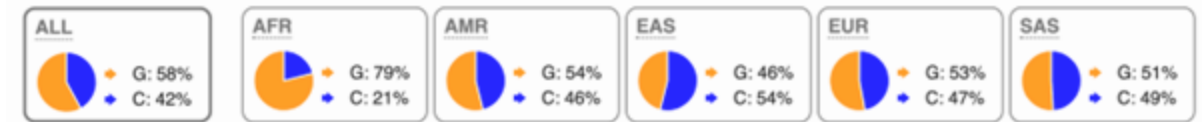
$$PGS = Z_m = \sum_{l=1}^L 2\alpha_l p_{l,m}$$



How are polygenic scores constructed?

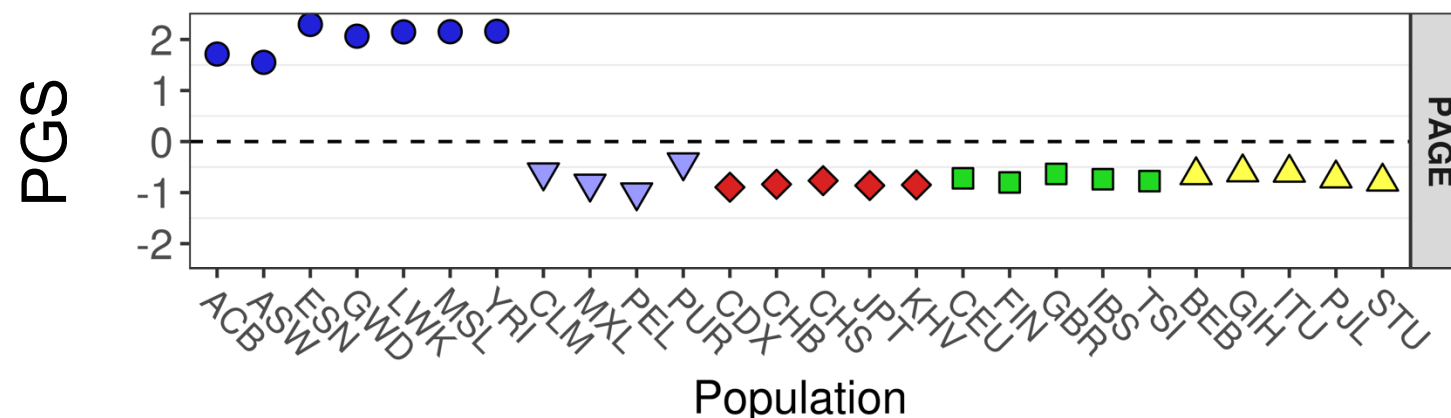


1000 Genomes Project Phase 3 allele frequencies



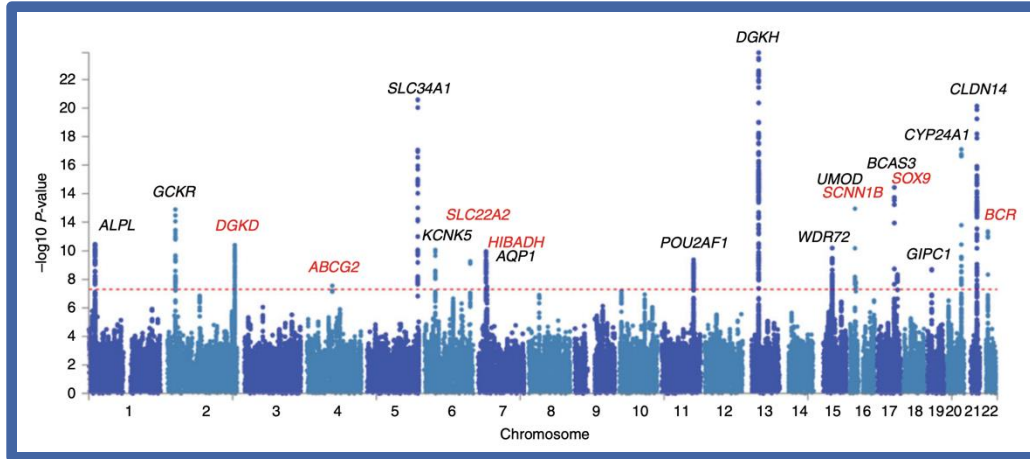
Effect size from trait-associated SNPs -estimates from **UK Biobank**

Allele frequency from 26 populations across 5 super-populations

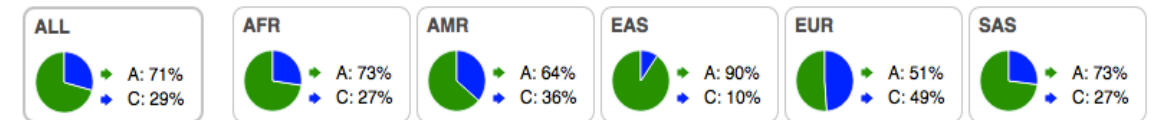


- AFR : African
- ▼ AMR : American
- ◆ EAS : East Asian
- EUR : European
- ▲ SAS : South Asian

Height PGS using another GWAS cohort

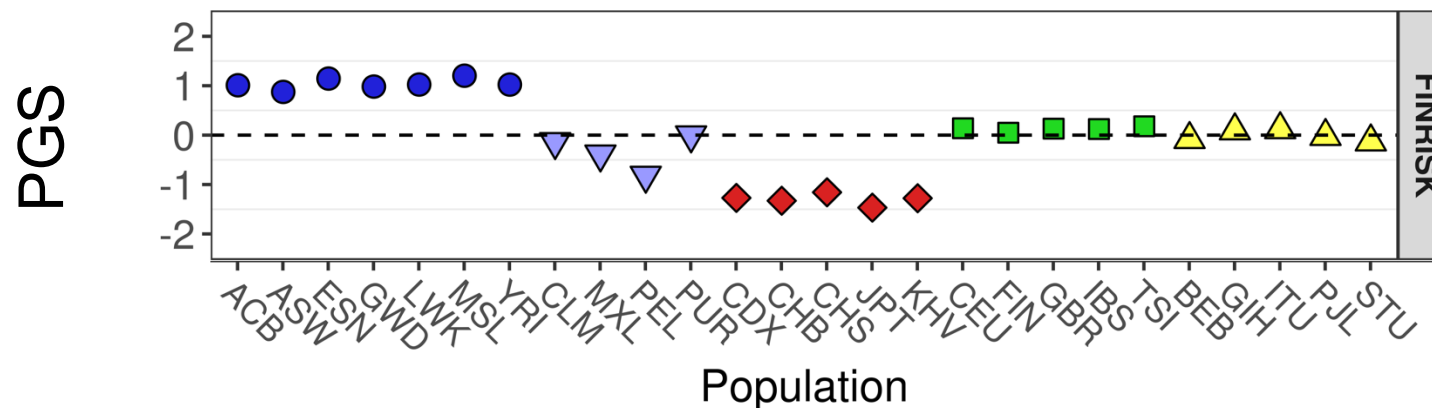


1000 Genomes Project Phase 3 allele frequencies



Effect size from trait-associated SNPs -estimates from **National FINRISK study**

Allele frequency from 26 populations across 5 super-populations



- AFR : African
- ▼ AMR : American
- ◆ EAS : East Asian
- EUR : European
- ▲ SAS : South Asian

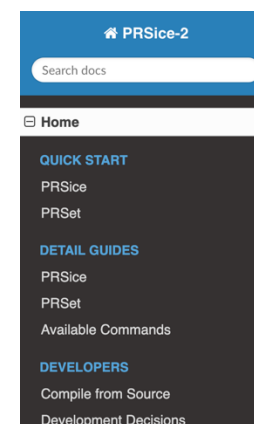


~ 25 min

GWAS6-PRSAnalysis.ipynb



- Compute PRS scores using PRSice for binary trait.



PRSice-2: Polygenic Risk Score software

PRSice (pronounced 'precise') is a Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses. Some of the features include:

1. High-resolution scoring (PRS calculated across a large number of P-value thresholds)
2. Identify Most predictive PRS
3. Empirical P-values output (not subject to over-fitting)
4. Genotyped (PLINK binary) and imputed (Oxford bgen v1.2) data input
5. Biobank-scale genotyped data can be analysed within hours
6. Incorporation of covariates
7. Application across multiple target traits simultaneously
8. Results plotted in several formats (bar plots, high-res plots, quantile plots)
9. PRSet: function for calculating PRS across user-defined pathways / gene sets

Executable downloads DOI: [10.5281/zenodo.3703335](https://doi.org/10.5281/zenodo.3703335) coverage: 68%



Choose the Bash kernel



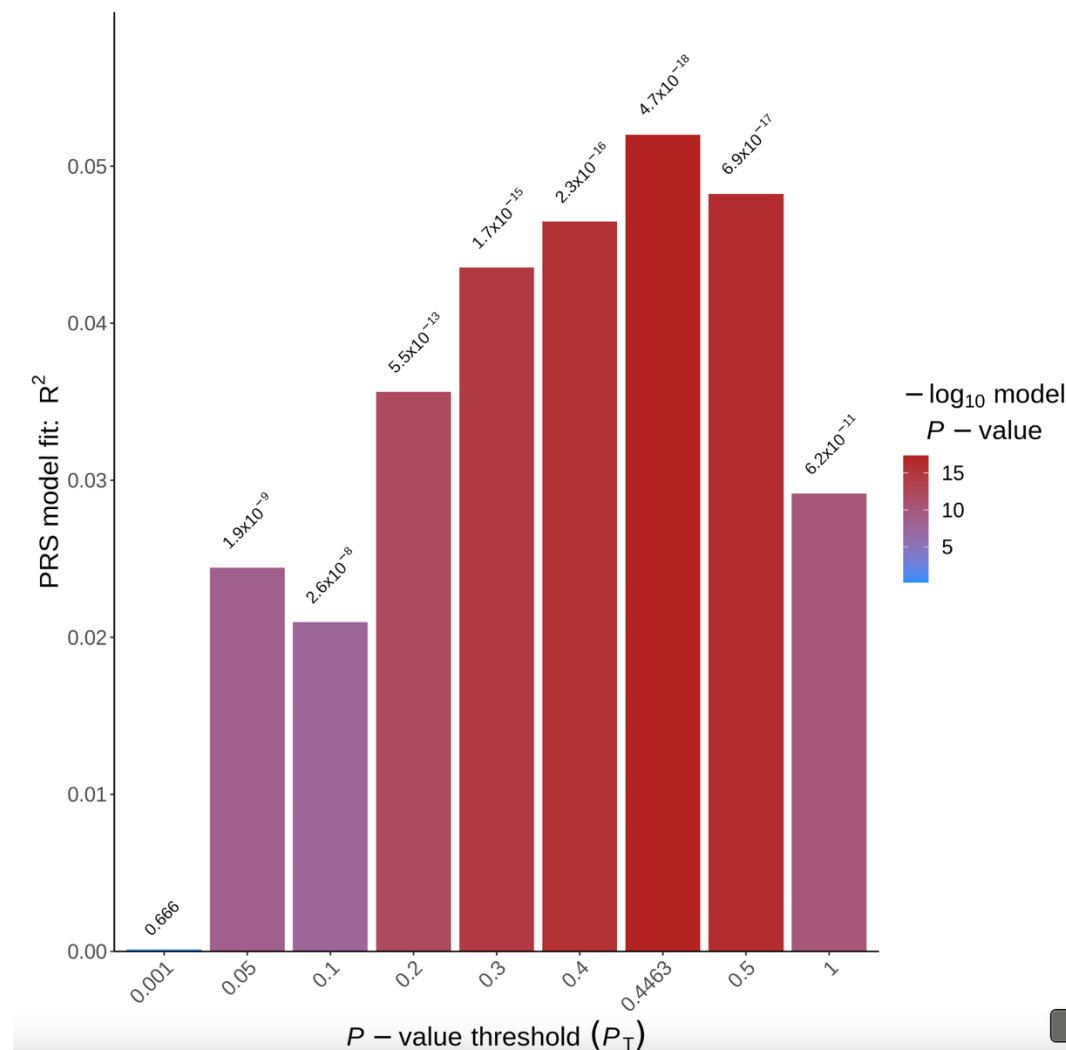
Choose the R-GWAS kernel

Solutions

- Problems/Issues/Comments?

Choosing threshold

Choosing the optimal threshold will really influence the performance of the PRs



(Future) Application of PRS in precision medicine

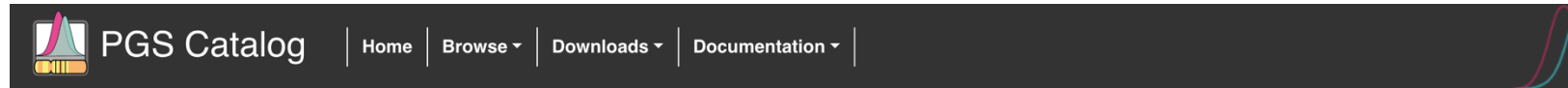


Prevention:

- Lifestyle change
- Screening programs

To best treat X person?

Provides the SNP weights of thousands of published PGS in a standardized format



Latest release: Feb. 4, 2025

The Polygenic Score (PGS) Catalog

An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.

Search the PGS Catalog



Examples: [breast cancer](#), [glaucoma](#), [BMI](#), [EFO_0001645](#)

Available tool: **pgsc_calc**

A reproducible workflow to calculate both PGS Catalog and custom polygenic scores. [See more information](#)

Explore the Data

In the current PGS Catalog you can **browse** the scores and metadata through the following categories:

Polygenic Scores

⌘ 5,053

Traits

🧑 656

Publications

📖 692

[Submit a PGS](#)

Applications of PRS in population genomics

How did evolution shape **genetic variation**?

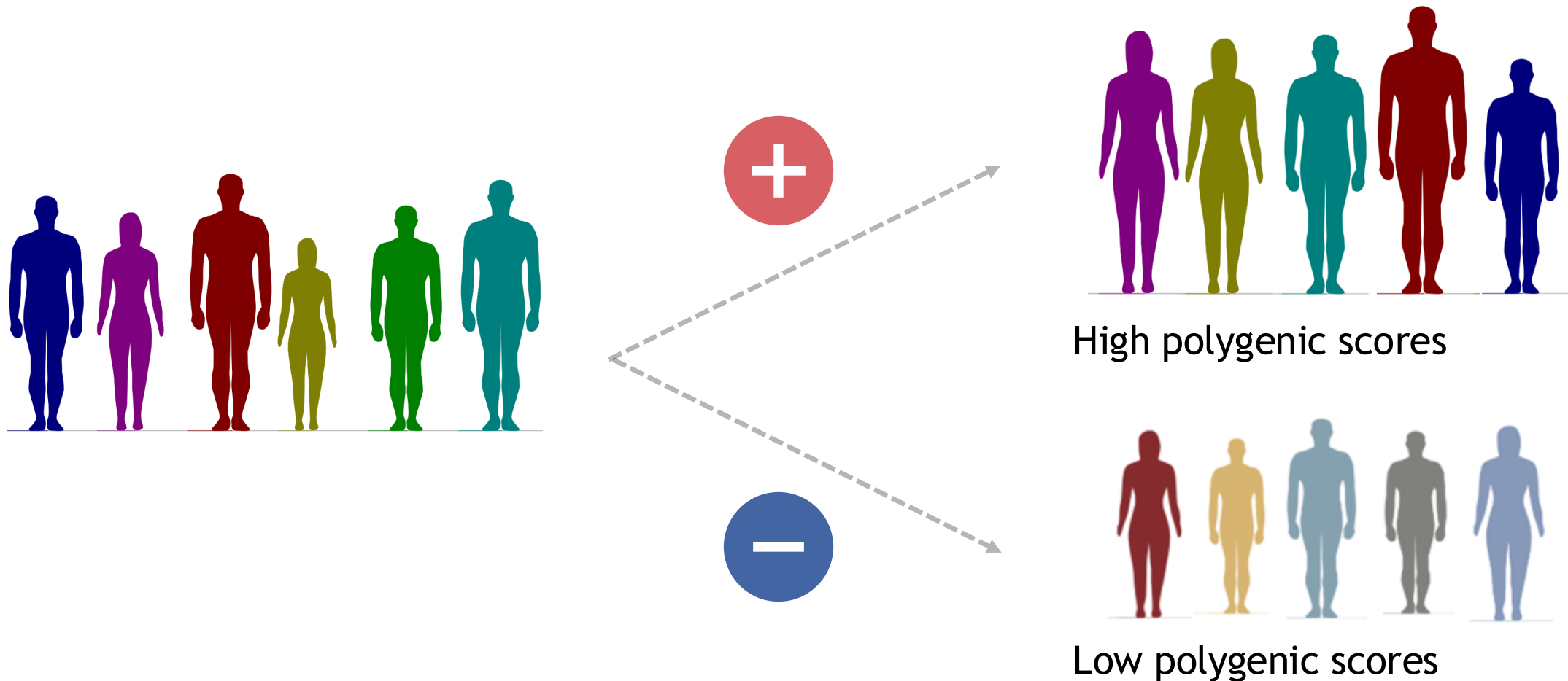
To what extent are **phenotypic differences** among human populations driven by natural selection?



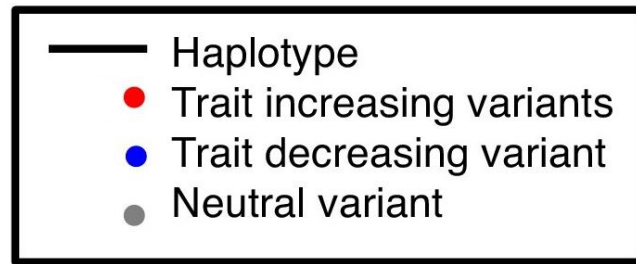
Image credit: [humanae.tumblr.com](https://www.tumblr.com/humanae)

Applications of PRS in population genomics

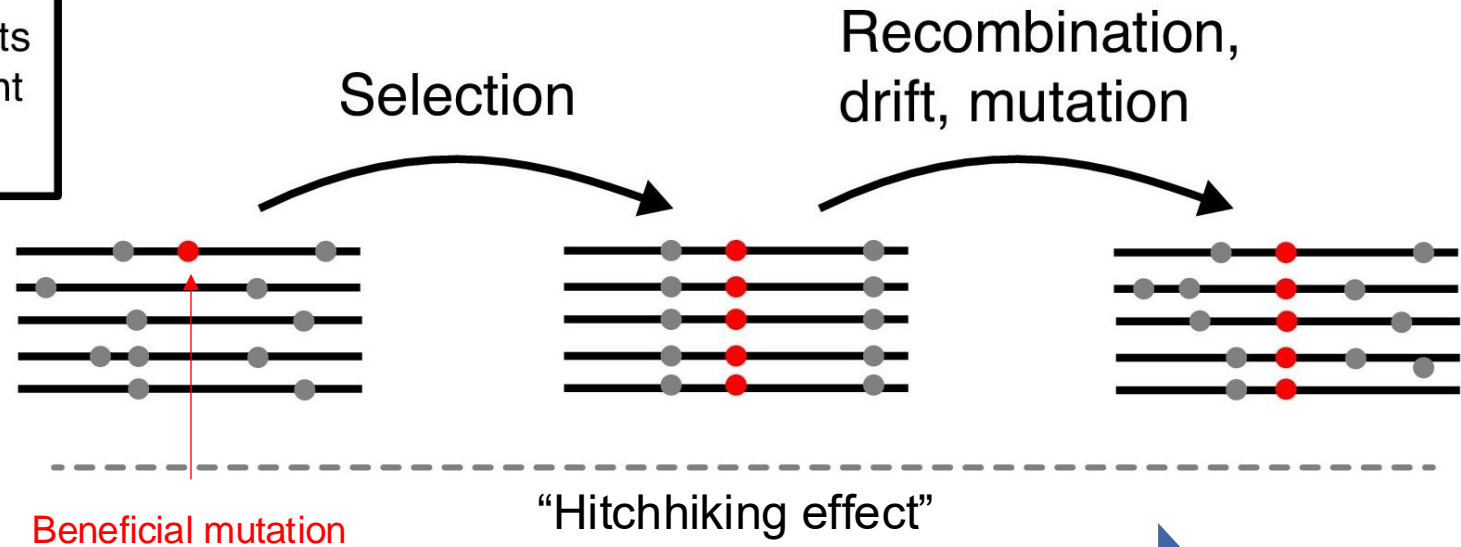
What if we used GWAS variants to test for natural selection?



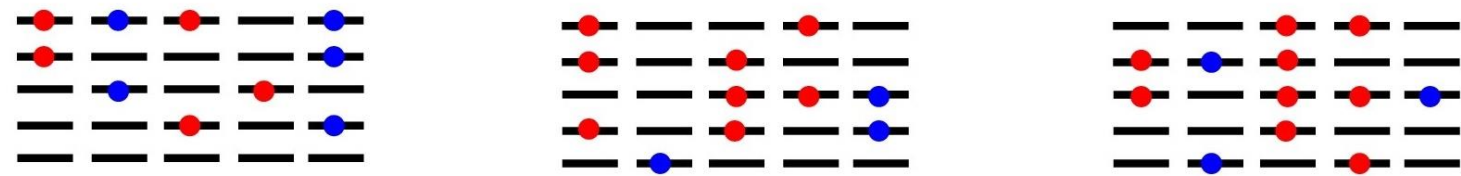
Polygenic adaptation



Hard sweep



Polygenic selection



Polygenic adaptation

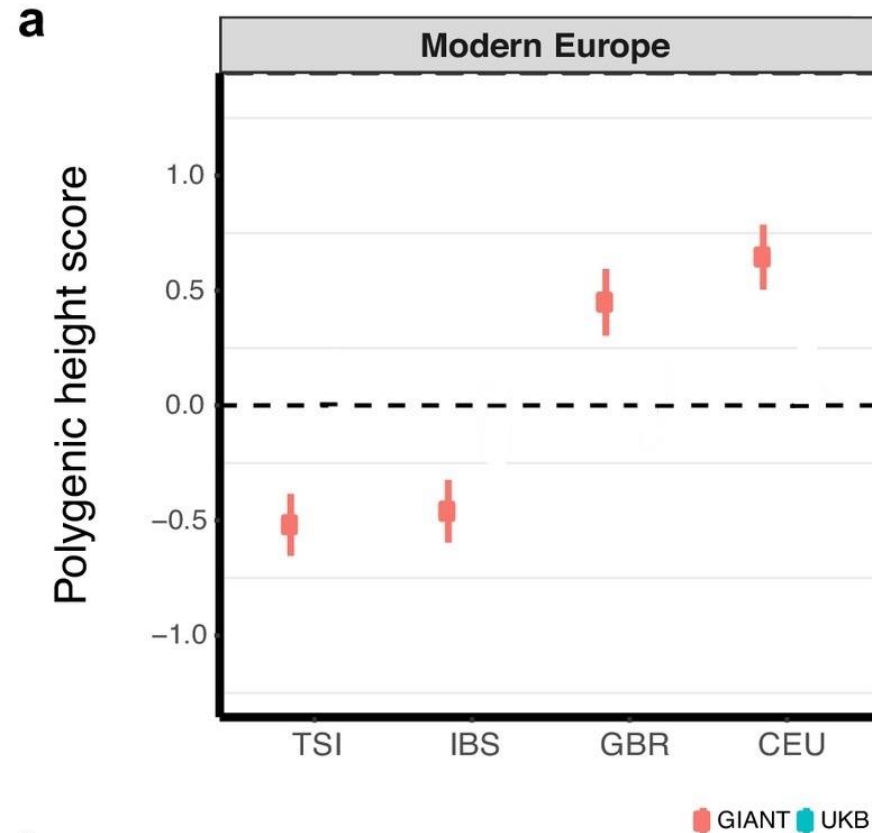
We consider the vector of allele frequencies for GWAS-significant SNPs as a **test statistic** and evaluate **whether these frequencies show greater divergence** in other populations than expected under genetic drift alone

Key assumption: The GWAS used to construct the polygenic score must have adequately accounted for **population stratification** (we will revisit this point later).



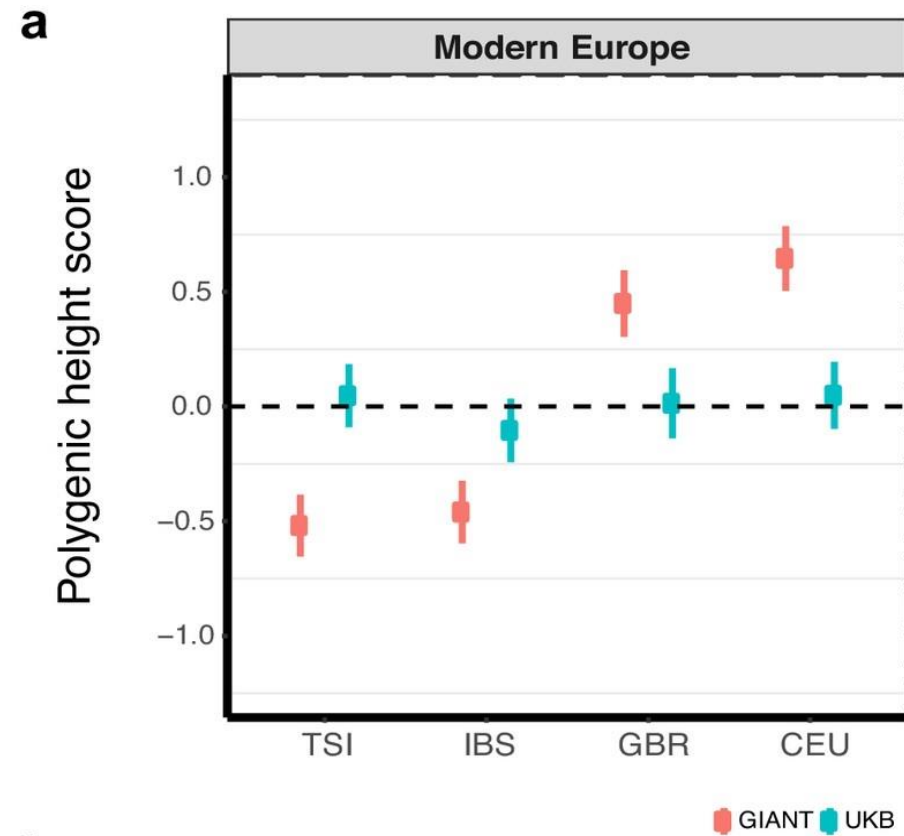
Can we measure the scores overdispersion among populations?

Significant measure of overdispersion -> evidence for polygenic adaption on height

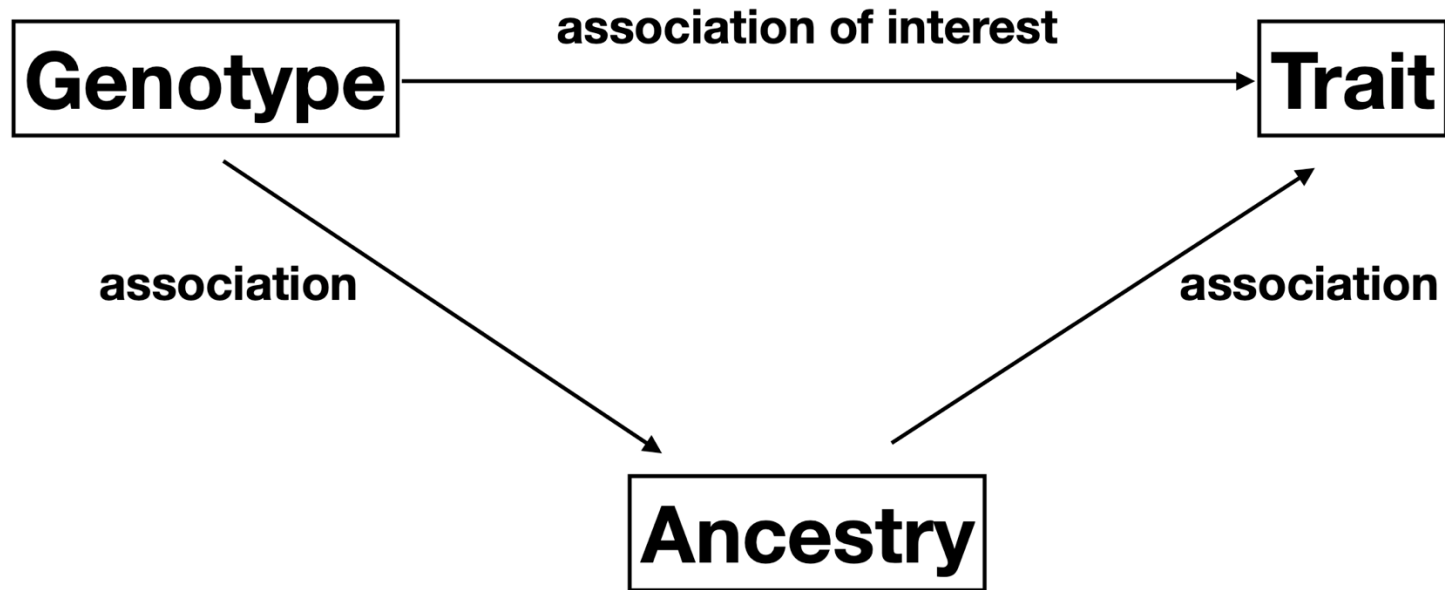


Evidence was not so strong

Not significant
measure of
overdispersion

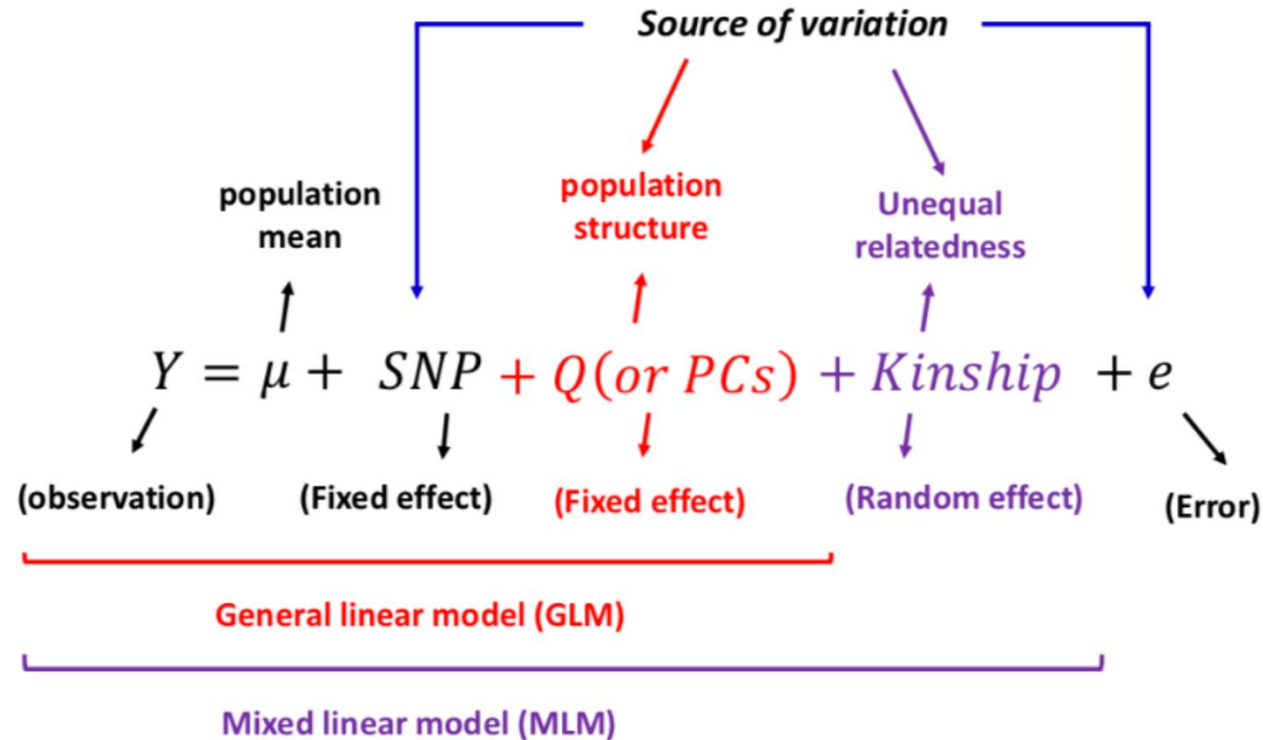


What went wrong?



Example: propensity to drink Carlsberg beer and alleles that happen to be at high frequency in Danes





On an individual SNP basis, this is likely OK. However, this was not enough to correct for **subtle biases that accumulate** in polygenic scores



Other caveats

- Even if we find evidence for selection at trait-associated SNPs, it doesn't mean we have found the **true trait under selection**.
- Even if we find evidence for selection, it doesn't necessarily mean there are **phenotypic differences** between populations in that trait: genetic compensation, environmental effects, etc.
- Unclear how **differences in effect sizes and LD** across populations may affect inference
- Major effect alleles may be different across populations



Post-GWAS analysis

Recommendation for human studies:

- Have a look at this tutorial: <https://github.com/AngelaMinaVargas/eMAGMA-tutorial> to conduct eQTL informed gene-based tests by assigning SNPs to tissue-specific eGenes

JOURNAL ARTICLE

E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics FREE

Zachary F Gerring ✉, Angela Mina-Vargas, Eric R Gamazon, Eske M Derks

Bioinformatics, Volume 37, Issue 16, August 2021, Pages 2245–2249, <https://doi.org/10.1093/bioinformatics/btab115>

Published: 24 February 2021 **Article history** ▼



PDF

Split View

Cite

Permissions

Share ▼



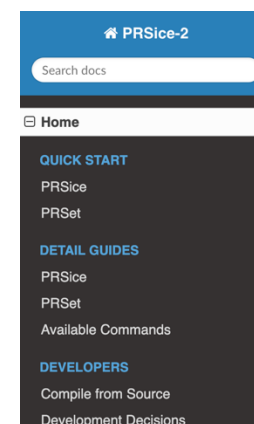


~ 25 min

GWAS7-PRSIII.ipynb



- Compute PRS scores using PRSice for quantitative trait.



PRSice-2: Polygenic Risk Score software

PRSice (pronounced 'precise') is a Polygenic Risk Score software for calculating, applying, evaluating and plotting the results of polygenic risk scores (PRS) analyses. Some of the features include:

1. High-resolution scoring (PRS calculated across a large number of P-value thresholds)
2. Identify Most predictive PRS
3. Empirical P-values output (not subject to over-fitting)
4. Genotyped (PLINK binary) and imputed (Oxford bgen v1.2) data input
5. Biobank-scale genotyped data can be analysed within hours
6. Incorporation of covariates
7. Application across multiple target traits simultaneously
8. Results plotted in several formats (bar plots, high-res plots, quantile plots)
9. PRSet: function for calculating PRS across user-defined pathways / gene sets

Executable downloads DOI: [10.5281/zenodo.3703335](https://doi.org/10.5281/zenodo.3703335) coverage 68%



Choose the Bash kernel

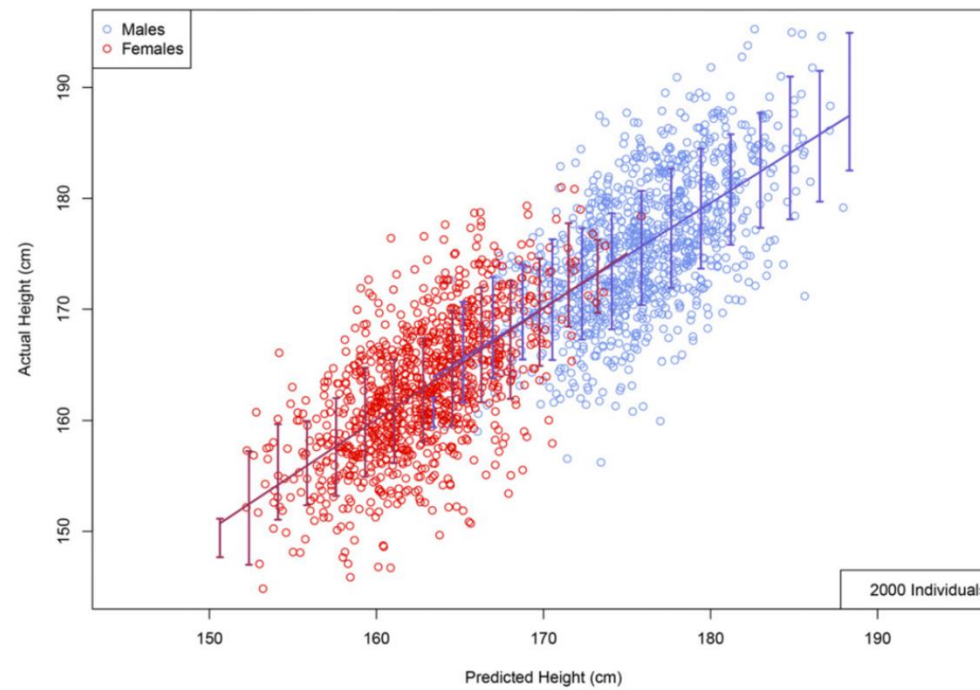


Choose the R-GWAS kernel

Solutions

- Problems/Issues/Comments?

65% correlation between actual and predicted height



Wrap-up

