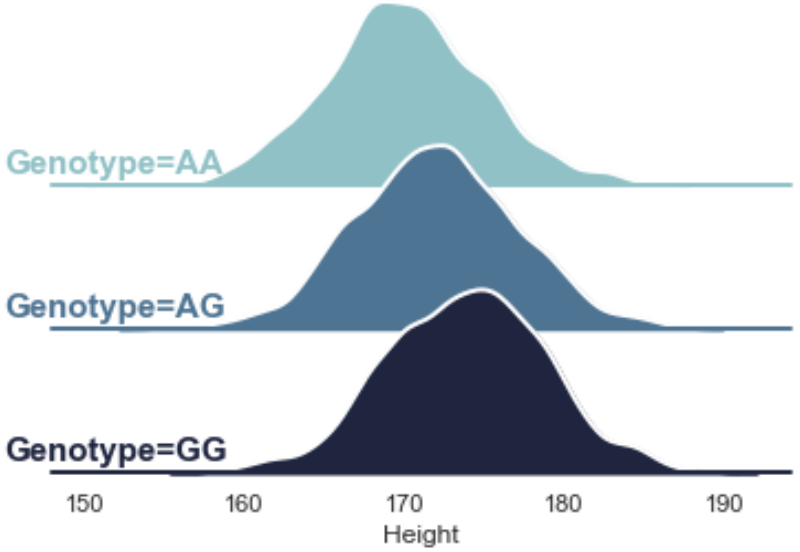
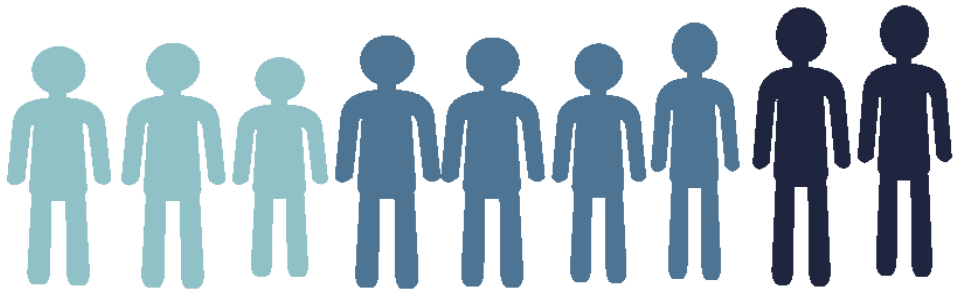
Polygenic Prediction

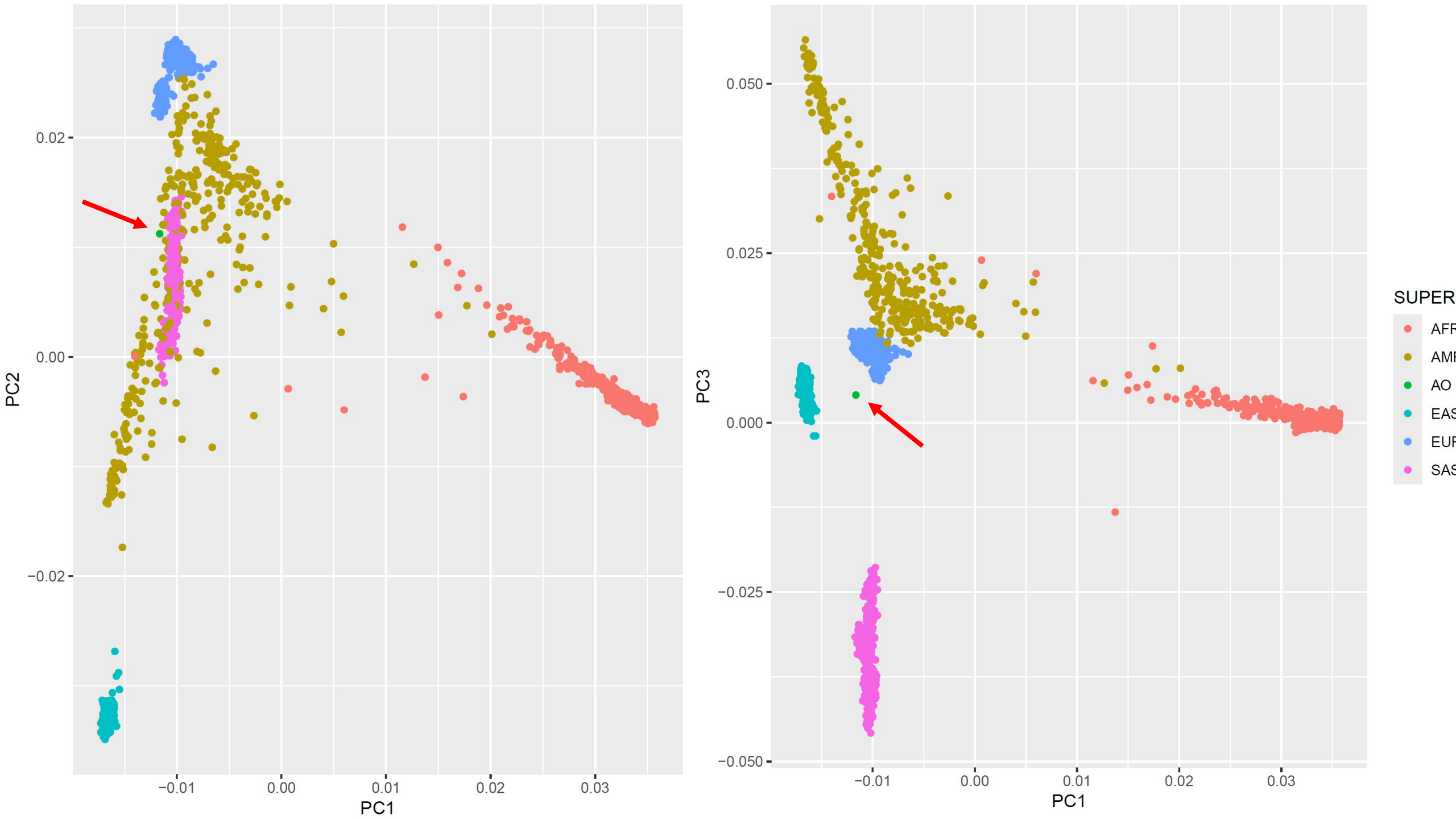
Aysu Okbay

Amsterdam UMC

a.okbay@amsterdamumc.nl






0.02
0.00
−0.02
PC2
−0.01
0.00
0.01
0.02
0.03
PC1
0.050
0.025
0.000
−0.025
−0.050
PC3
−0.01
0.00
0.01
0.02
0.03
PC1
SUPER
AFR
AMR
AO
EAS
EUR
SAS

Outline

- A literal history of PGIs
- What is a polygenic index?
- Theoretical framework
- Constructing polygenic indices
- Applications
- Limitations & pitfalls

Polygenic score (PGS)

Polygenic index (PGI)

Polygenic risk score (PRS)

Genome-wide score (GWS)

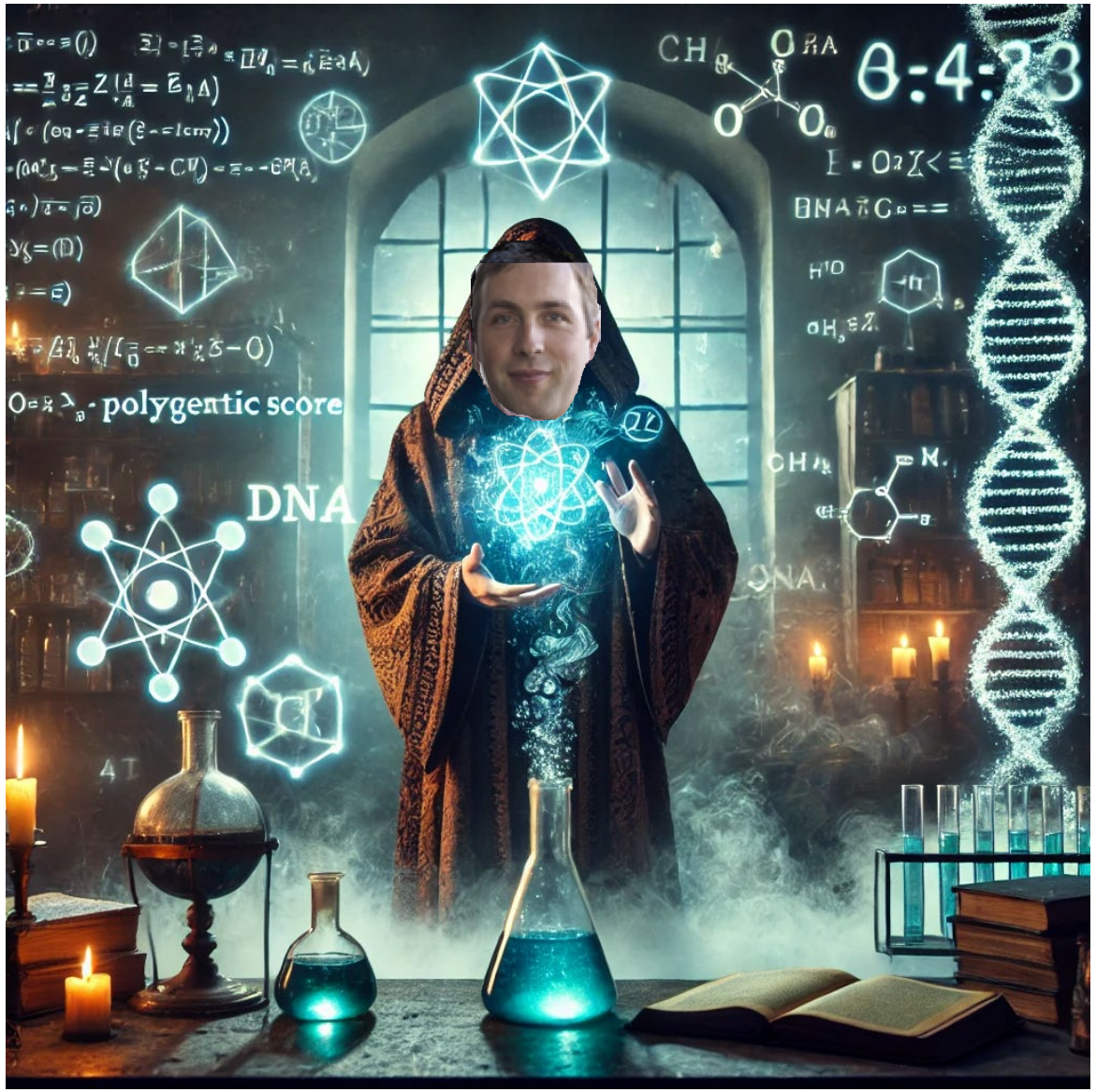
Genetic risk score (GRS)

A literal history of polygenic indices

- Once upon a time, people thought all human traits were influenced by a handful of genes.
- In their folly, they came up with candidate genes chosen out of a few whose function were known

Thus, started the era of candidate gene studies, and the following replication crisis.

- In the silence that followed the wake of these studies, tiny little GWAS started to emerge from the ashes
- They identified a few tiny genetic breadcrumbs that barely explained a fraction of the traits they were chasing. The top hits were underwhelming, the heritability gaps yawning…



It was then (2009) that a statistical alchemist came to the rescue..
Today, we know him as…

doi:10.1038/nature08185 nature

Sean Purcell

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

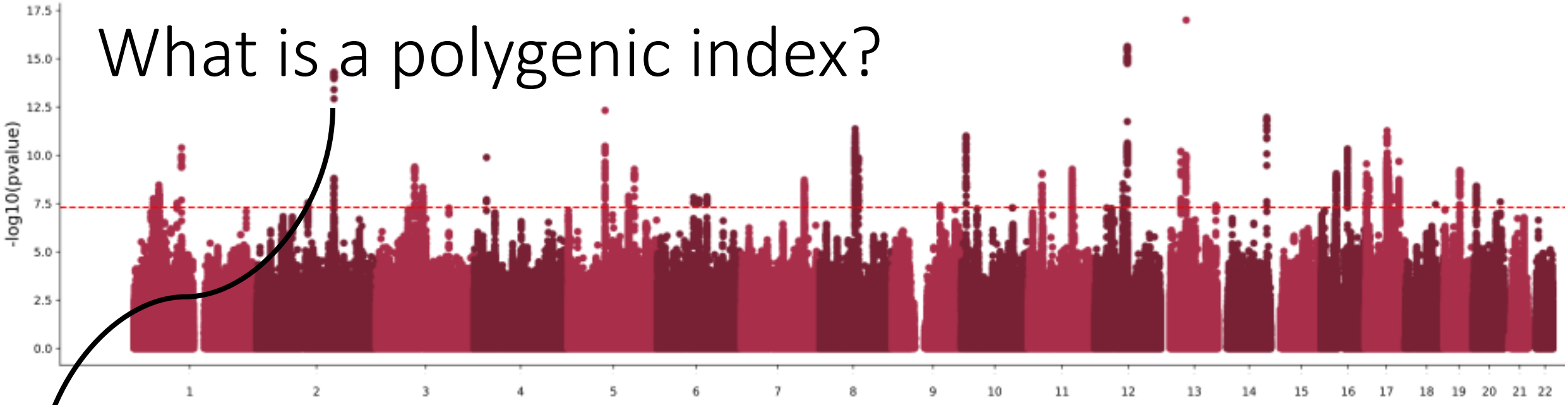
The International Schizophrenia Consortium*

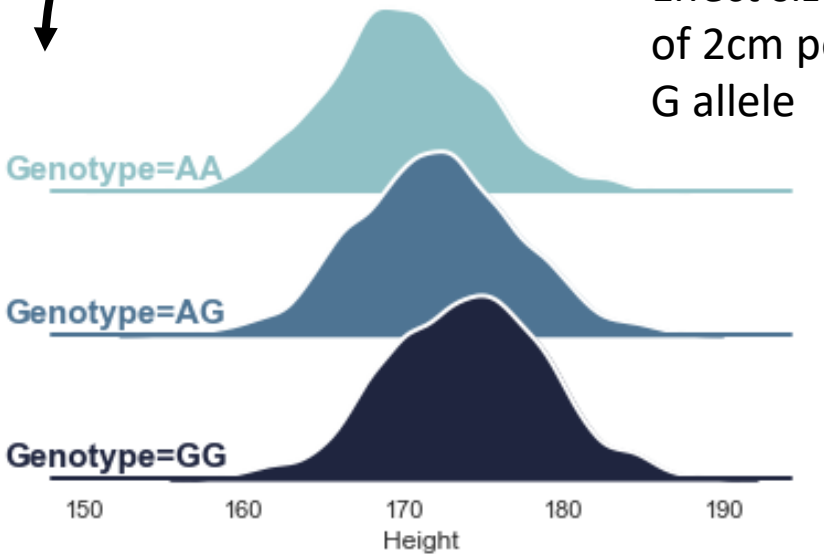
It is said that around the same time, in the land down under, a lesser wizard came up with the same idea, but there is absolutely no written record of this.

Outline

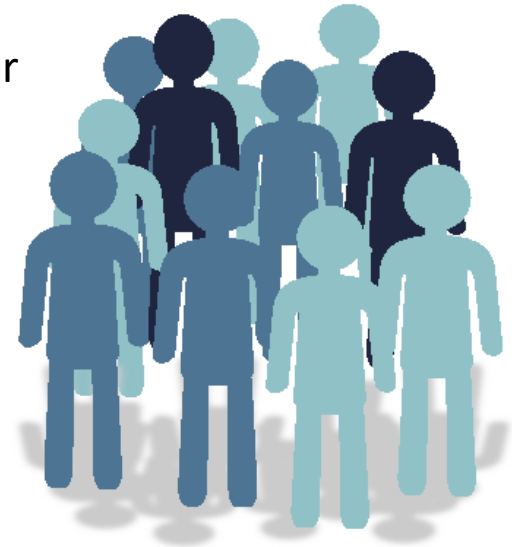
- A literal history of PGIs
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations & pitfalls

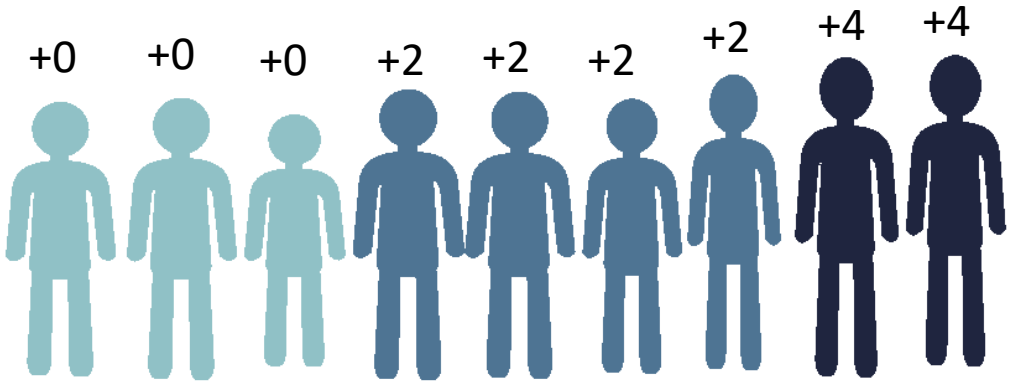
What is a polygenic index?

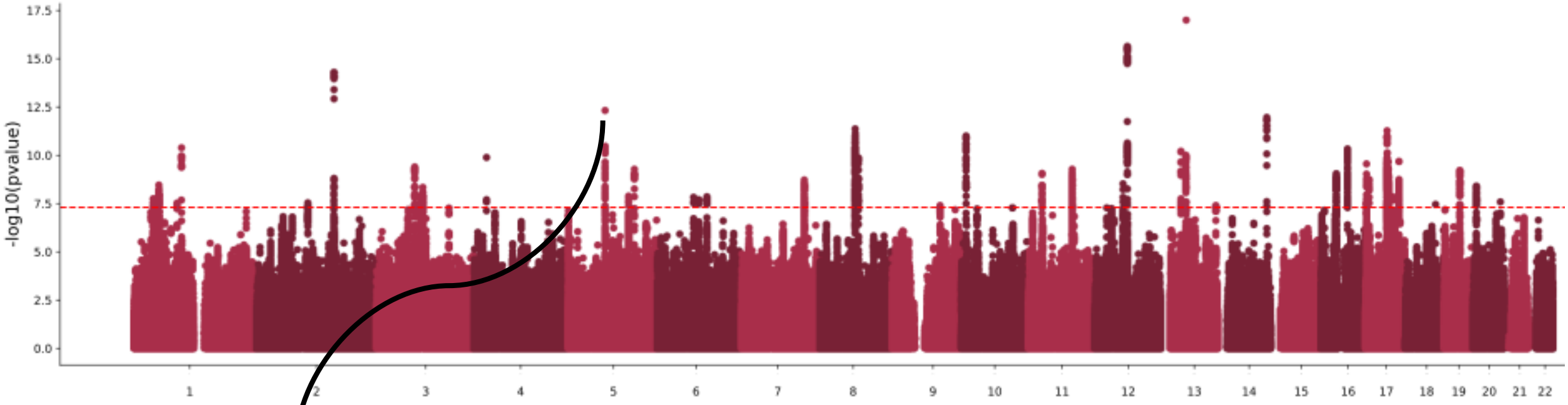




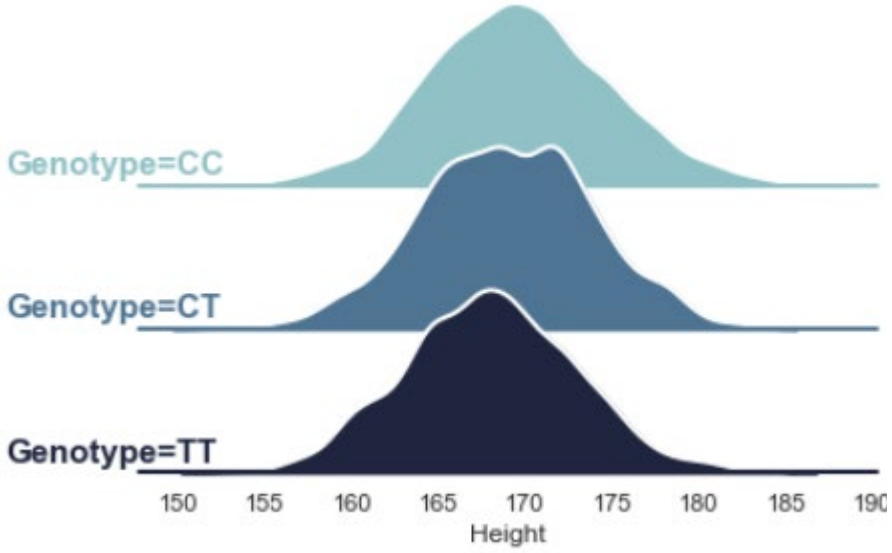
Effect size of 2cm per G allele

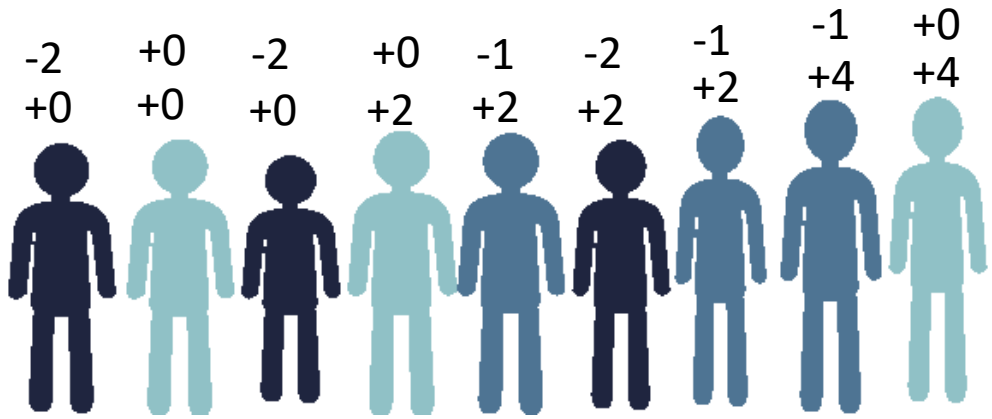






Effect size of -1 per T allele





What is a polygenic index?

- An index that linearly aggregates the estimated effects of individual SNPs on the trait of interest.
- Can be considered a measure of an individual's **genetic propensity** towards a trait.
- Defined as a **weighted sum of a persons genotypes at K loci**.
- Start with additive model using measured SNPs:

$$y_i = A_{SNP,i}(x_i) + \epsilon_{i,SNP} = \sum_{j=1}^{K} \beta_j x_{ij} + \epsilon_{i,SNP}$$

additive SNP factor

What is a polygenic index?

Additive SNP factor:

$$A_{SNP,i}(x_i) \equiv \sum_{j=1}^{K} \beta_j x_{ij}$$

True effect size of SNP j

PGI:

$$\hat{A}_{SNP,i}(x_i) \equiv \sum_{j=1}^{K} \hat{\beta}_j x_{ij}$$

Estimated effect size of SNP j

$$\hat{\beta}_j = \beta_j + u_j \Rightarrow \hat{A}_{SNP,i} = \sum_{j=1}^{K} (\beta_j + u_j) x_{ij} = A_{SNP,i} + U_i \text{ where } U_i = \sum_{j=1}^{K} u_j x_{ij}$$

If u is mean-zero estimation error uncorrelated with β_j → U is mean-zero measurement error →

$$E(\hat{A}_i | A_i) = A_i$$

Outline

- A literal history of PGIs
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations & pitfalls

Predictive power of a polygenic index

If we regress y on $\hat{A}_{SNP}$ we get an OLS coefficient of

$$b = \frac{Cov(\hat{A}_{SNP}, y)}{Var(\hat{A}_{SNP})}$$

$$= \frac{Cov(A_{SNP} + U_i, A_{SNP} + \epsilon_{SNP})}{Var(A_{SNP} + U)}$$

$$= \frac{Var(A_{SNP})}{Var(A_{SNP}) + Var(U)}$$

And the expected predictive power is:

$$R^2 \approx \frac{b^2 Var(\hat{A}_{SNP})}{Var(y)}$$

$$= \left(\frac{Var(A_{SNP})}{Var(A_{SNP}) + Var(U)}\right)^2 \frac{Var(\hat{A}_{SNP})}{Var(y)}$$

$$\vdots$$

$$\approx \frac{{h^2_{SNP}}^2}{h^2_{SNP} + \frac{M_e}{N}}$$

Sometimes called the Daetwyler formula (Daetwyler et al. 2008)

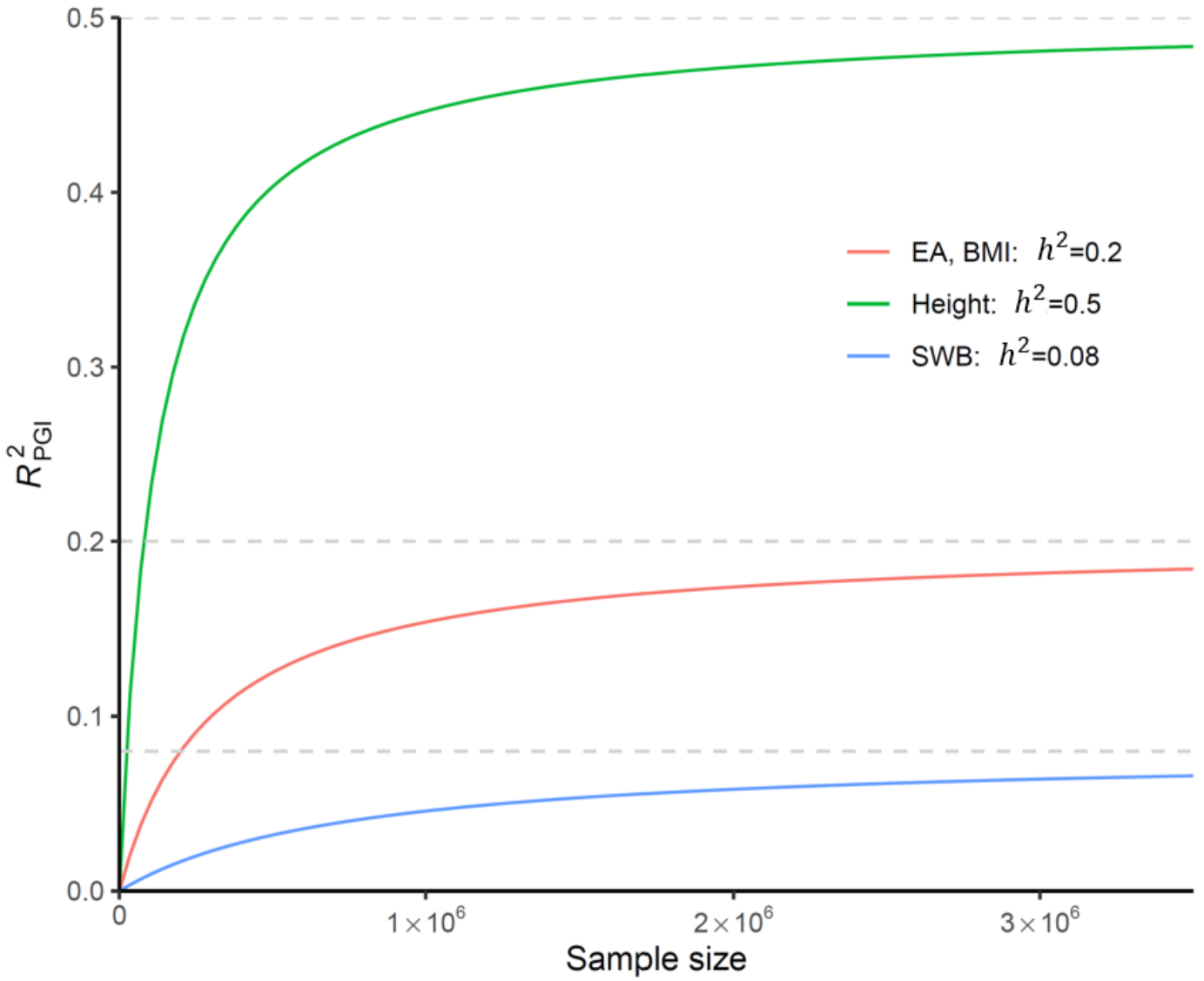
Effective number of SNPs in the PGI, estimated to be between 50k-70k in genome-wide data for EUR ancestry (Wray et al. 2013)

OLS:

$$y_i = a + bx_i + \epsilon_i$$

$$b = \frac{Cov(x,y)}{Var(x)}, R^2 = \frac{b^2 Var(x)}{Var(y)}$$

Theoretical projections for R^2_{PGI}



Predictive power and heterogeneity

What if there is heterogeneity between GWAS and validation samples?

$$A^*_{SNP,i} \neq A_{SNP,i} \rightarrow {h^2}^*_{SNP} \neq h^2_{SNP}$$

Define the genetic correlation to be

$$r_g = Corr(A^*_{SNP,i}, A_{SNP,i})$$

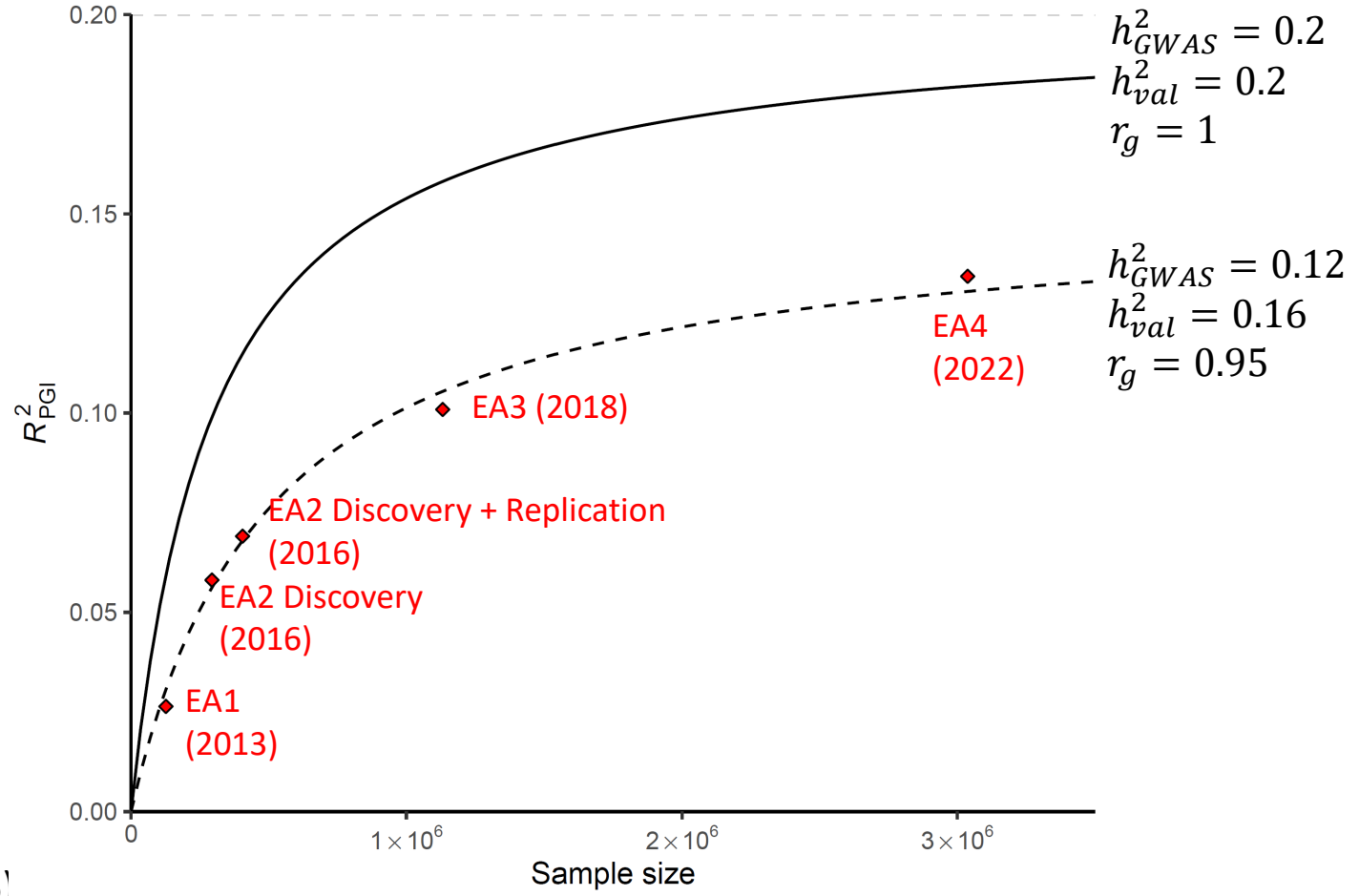
The expected predictive power

$$R^2 \approx \frac{{h^2_{SNP}}^2}{h^2_{SNP} + \frac{M_e}{N}}$$

now becomes

$$R^2 \approx \frac{r_g h^2_{SNP} {h^2}^*_{SNP}}{h^2_{SNP} + M_e/N}$$

(De Vlaming et al. 2016)



Measuring observed predictive power

Most commonly used measure for continuous phenotypes is **incremental-R^2**

- Regress phenotype on basic covariates (e.g. sex, age, age^2, $sex \times age$, $sex \times age^2$) and PCs
- Obtain the R^2 of the regression
- Add the PGI to the RHS, obtain the new R^2
- The difference between the two R^2's is the incremental-R^2 of the PGI

For binary phenotypes analyzed using logistic regression, various other measures are used, such as

- **Nagelkerke's pseudo-R^2** : $R^2 = \frac{\left(1-\left\{\frac{L_0}{L_1}\right\}^{\frac{2}{n}}\right)}{1-L_0^{\frac{2}{n}}}$ where L_0 and L_1 are the likelihoods of the null model (with only covariates) and of the model with PGI, respectively
- **Area under the ROC curve (AUC):** The probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative. Ranges between 0.5 and 1.

Outline

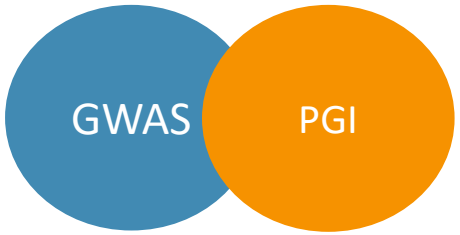
- A literal history of PGIs
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations & pitfalls

Constructing polygenic indices

What is needed?

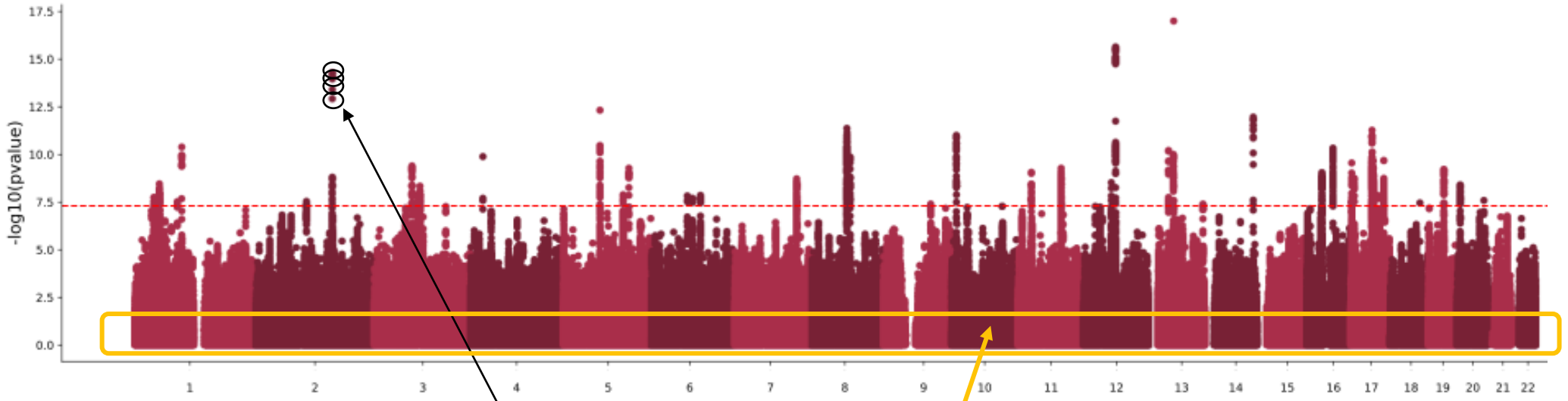
- Individual-level genotype data from a prediction sample.
- Weights: GWAS summary statistics from a discovery sample
- Reference genotypes to estimate LD

Caution: The prediction sample should not overlap with the discovery sample!



Overfitting!
The R^2 of the PGI will be biased upwards!

Weights



GWAS results give us $\hat{\beta}_j^{GWAS}$, not β_j. Two issues to consider when constructing $\sum_{j=1}^{K} \hat{\beta}_j^{GWAS} x_{ij}$:

1. For some SNPs, $\hat{\beta}_j^{GWAS}$ may be a very noisy estimate of β_j and/or β_j may be close to 0, so adding those SNPs will add more noise than signal
2. If we include all SNPs, we will overweight ("double-count") SNPs with high LD scores

Two solutions

$$\sum_{j=1}^{K} \hat{\beta}_j^{GWAS} x_{ij}$$

Clumping and thresholding

Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

Weights: Set equal to GWAS coefficients.

Loci: Selected by

1. using a **clumping** algorithm that ensures the included markers are all approximately independent of each other
2. omitting SNPs whose *P* value for association with the phenotype is above a certain **threshold**

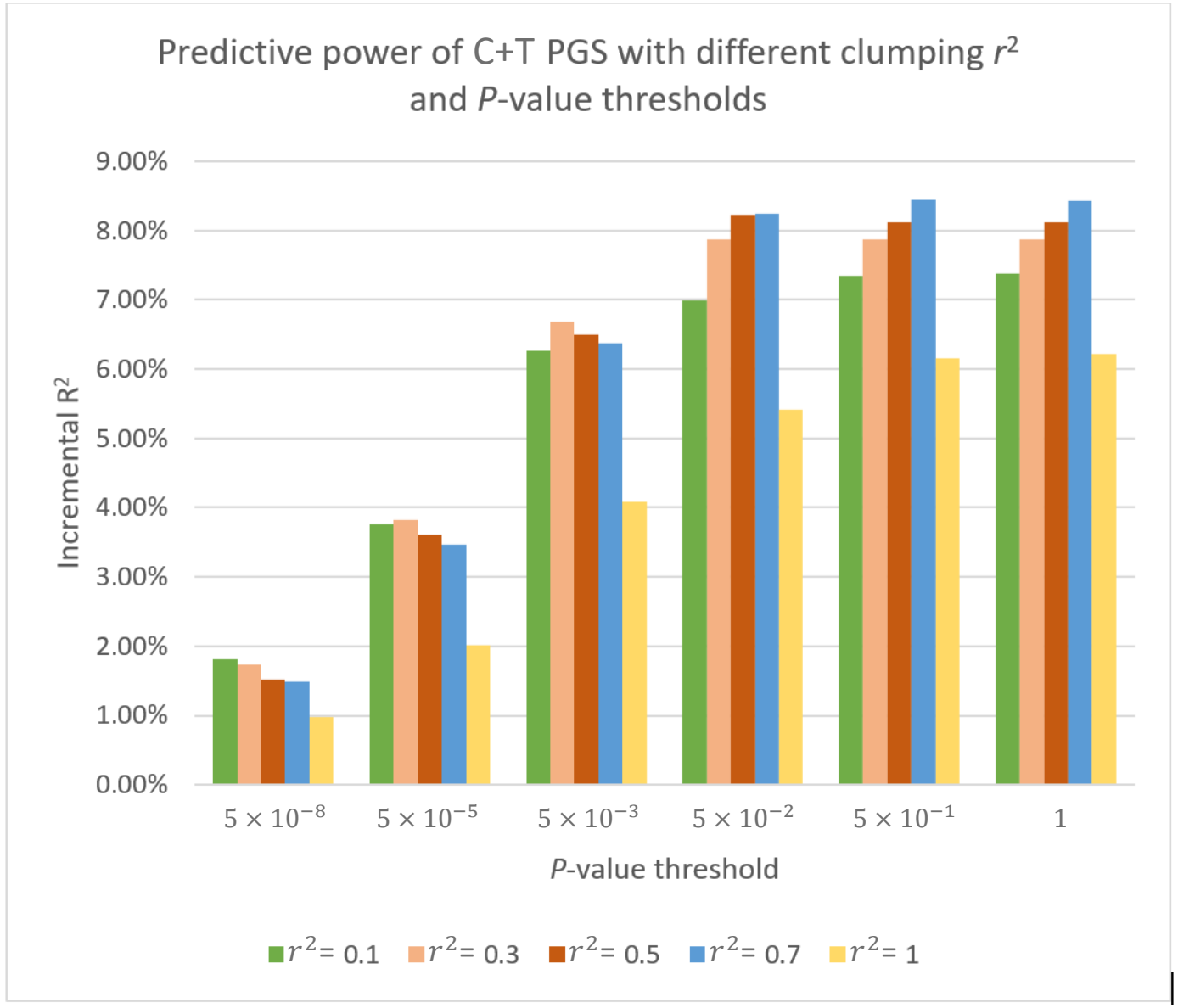
Bayesian approaches

Include all SNPs but adjust the effect sizes for LD

Weights: Set to GWAS coefficients **adjusted for LD** → approximate results from a theoretical multiple regression of the phenotype on all SNPs

Loci: Include **all SNPs**, no LD-based pruning

Examples: LDpred (Vilhjalmsson et al. 2015, Prive et al. 2020), PRS-CS (Ge et al. 2019), SBayesR (Lloyd-Jones et al. 2019)



- **Cohort:** Health and Retirement Study
- **Phenotype:** Educational attainment

Bayesian approaches

Uses as weights

$$E\left(\beta_j \middle| \hat{\beta}_j^{GWAS}, D\right)$$

LD matrix

By Bayes's rule,

$$f\left(\beta_j \middle| \hat{\beta}_j^{GWAS}, D\right) = \frac{f\left(\hat{\beta}_j^{GWAS} \middle| \beta, D\right) f\left(\beta_j \middle| D\right)}{f\left(\hat{\beta}_j^{GWAS} \middle| D\right)}$$

Shrinkage depends on the prior!

PRS-CS: "Continuous shrinkage"

$$\left(\beta_j \middle| D\right) \sim N(0, \phi\psi_j)$$
$$\psi_j \sim N\left(a, \delta_j\right)$$
$$\delta_j \sim N(b, 1)$$

Parameters a and b determine how aggressively to shrink small estimates and how much you don't shrink large ones

LDpred2: Gaussian or Spike-and-Slab

$$\left(\beta_j \middle| D\right) \sim \begin{cases} N\left(0, \tau^2\right), & with\ probability\ \pi \\ 0 & with\ probability\ 1-\pi \end{cases}$$

π can be estimated from data, sparsity allowed (if $\bar{\pi}_j < \pi$, b_j set to 0), $\tau^2 = h^2/M\pi$

SBayesR: flexible finite mixture of normal distributions, sparsity allowed

$$\left(\beta_j \middle| D\right) \sim \begin{cases} 0, & with\ probability\ \pi_1 \\ N\left(0, \gamma_2\sigma_b^2\right), & with\ probability\ \pi_2 \\ \ldots & \\ N(0, \gamma_C\sigma_b^2) & with\ probability\ 1-\sum_{c=1}^{C-1}\pi_c \end{cases}$$

Practical considerations – LD reference data

How to choose LD reference data?

- Some software tools (e.g. SBayesR, PRS-CS) make available previously calculated LD estimates
- These are usually limited to a set of good quality SNPs (e.g. HapMap3) to reduce errors in LD estimation and computational burden while ensuring sufficient coverage
- But you can also estimate your own! Why would you?
  - You may want to include more SNPs
  - The available LD reference data may not be a good match for the ancestry of your GWAS
- If you decide to obtain your own LD estimates, you should make sure that the quality of your data is good. Bayesian approaches are very sensitive to errors in LD estimates!

Practical considerations – LD reference data

Points to consider when estimating LD

- The sample is large and **representative of the GWAS sample**
- **Sequenced genotypes are best.** Imputation inaccuracy introduces noise into LD estimates. If you're using imputed data, apply a strict imputation accuracy filter
- Data are cleaned
  - sample-level filters: related individuals, ancestry outliers, individuals with low genotyping rate
  - SNP-level filters: low SNP call rate, MAF, HWE P-value (genotyped SNPs), imputation accuracy (imputed SNPs)
- There are no genotyping or imputation batch effects.
  - May lead to errors in LD estimation if genotyping is done with multiple arrays or imputation in multiple batches.

Practical considerations - GWAS

Restricting set of SNPs

- The SNPs included in the PGI will be limited to the intersection of GWAS, LD reference sample and validation data.
- It is a good idea to **limit the SNPs in the GWAS to those available in the validation data prior to adjusting for LD**, especially if the overlap is rather poor (e.g. if you only have array SNPs in the validation data)
- The Bayesian software will assume all SNPs in the GWAS will be included in the PGI and make LD adjustments to maximize prediction accuracy. If some SNPs cannot be included because they are not in the validation data, the adjustments will be suboptimal.
- As an additional QC step, you can exclude SNPs whose MAF is very different from the LD reference data

Practical considerations – Validation data

- Applying some QC to validation data to minimize noise and genotyping errors is recommended:
  - sample-level filters: limit to a single genetic ancestry, drop individuals with low genotyping rate
  - SNP-level filters: drop SNPs with low call rate, MAF, HWE P-value (genotyped SNPs), imputation accuracy (imputed SNPs)
- Restrict the GWAS to SNPs available in the validation data after the above QC steps are applied
- If you are using imputed data, **use dosages rather than hard calls**. Hard calls don't account for imputation uncertainty!

C+T vs Bayesian approaches

Clumping and thresholding

Faster and easier, but too black & white

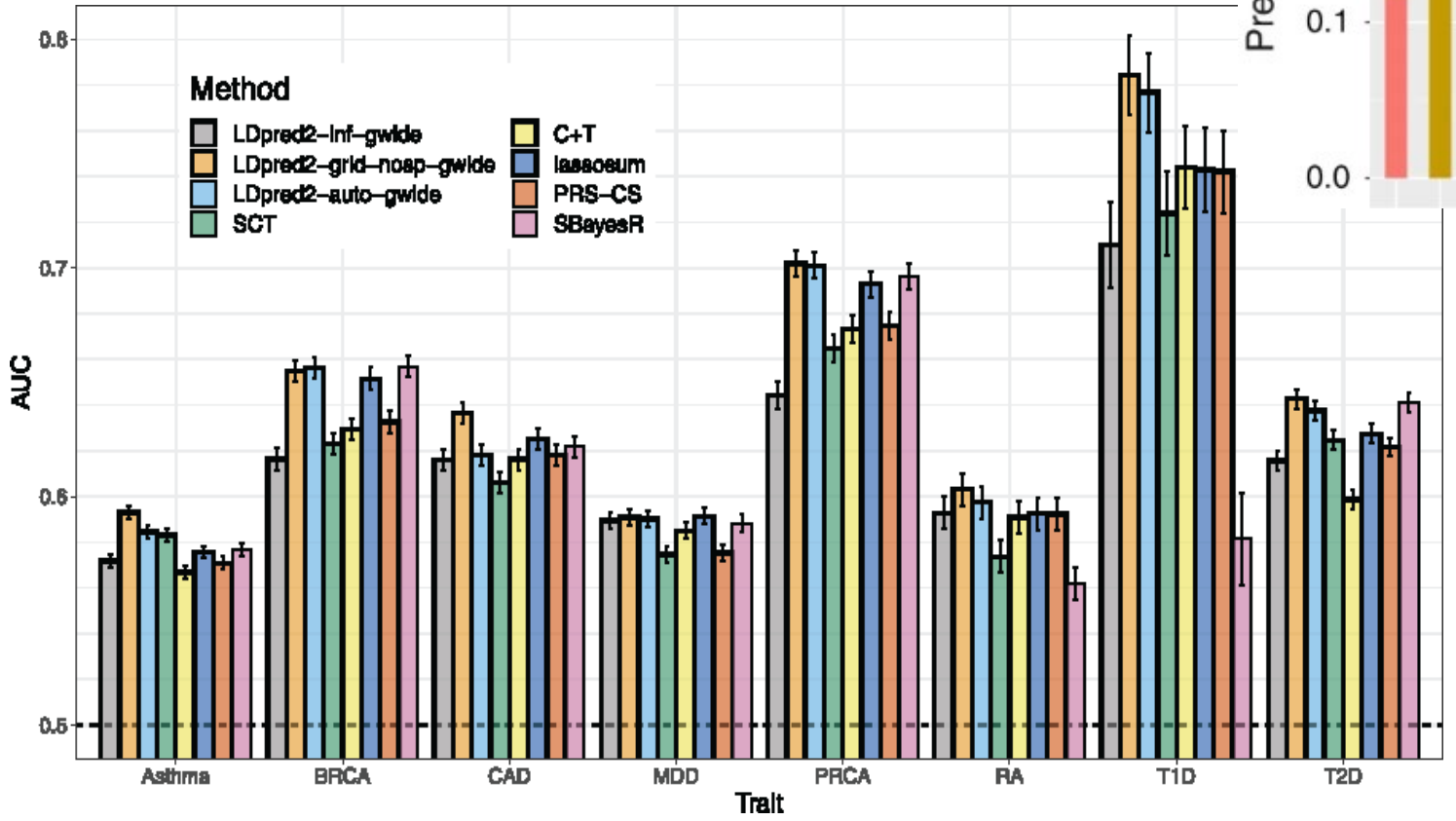
- If clumping r^2 or *P*-value cutoffs too strict, it drops potentially causal SNPs.
- If clumping r^2 and *P*-value cutoffs too relaxed, there is a lot of double-counting and noise

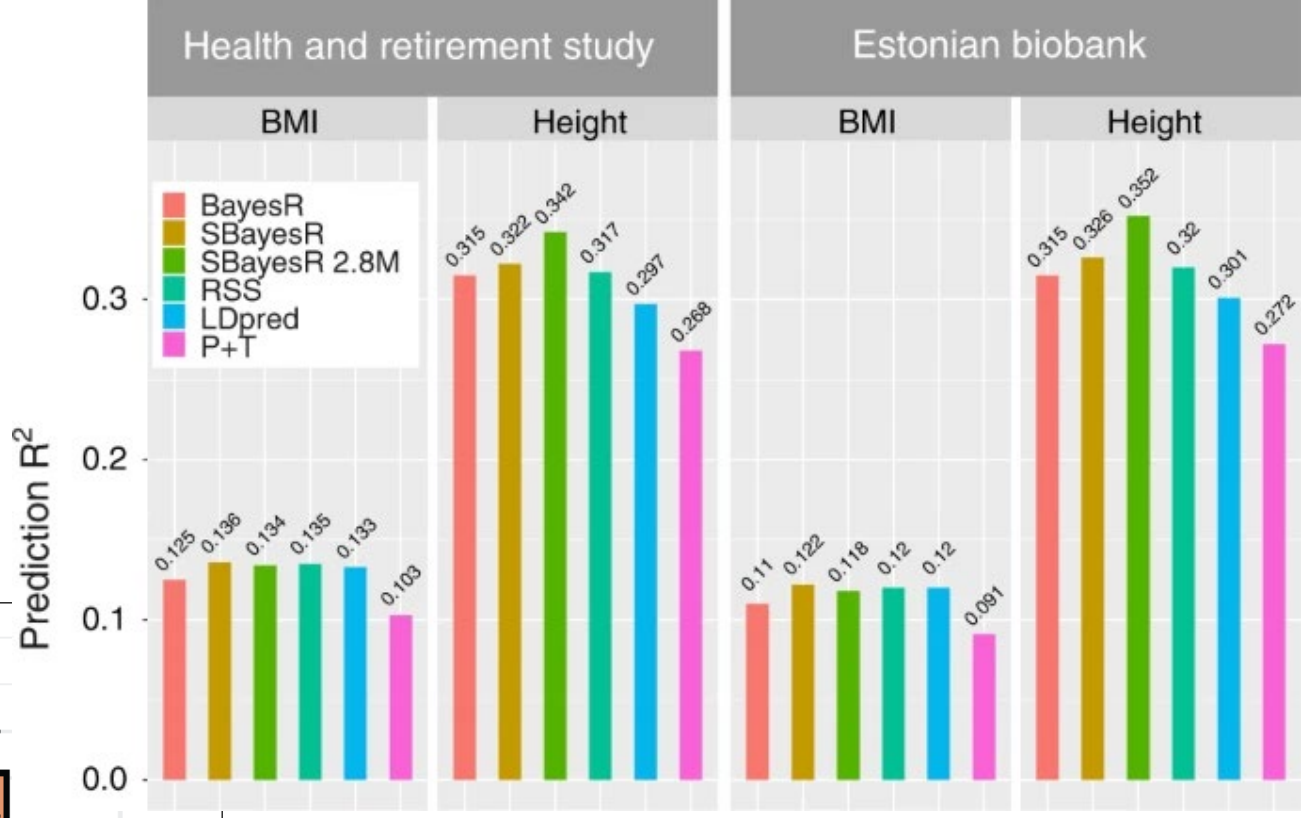
Bayesian approaches

- utilize information from all SNPs by adjusting SNP weights for LD, but
  - if the reference panel is not a good match for the population from which summary statistics were obtained, prediction accuracy might be compromised
  - the assumed prior distribution might not accurately model the true genetic architecture

If the purpose is to maximize predictive power, then Bayesian approaches clearly do better

Source: Privé, Arbel, Vilhjálmsson (2020)





Source: Lloyd-Jones et al (2019)

There may still be uses for C+T, for example when you want to include SNPs most likely to be associated (e.g. in mendelian randomization)

Outline

- A literal history of PGIs
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations and pitfalls

Applications

Major advantage of PGI over specific genetic variants: can have much greater predictive power

e.g., if $R^2_{PGI} = 0.07$, then 80% to detect its effect in a sample of size ~ 110 individuals. If $R^2_{PGI} = 0.09$, then ~ 85 individuals.

→ Can study PGI in datasets containing high quality measures of outcomes, mediators, and covariates.

→ **Identify correlates of genetic factors**

e.g. Educational attainment PGI predicts early speech acquisition and is mediated by cognitive ability (Belsky et al., 2016).

→ **Identify causal effects of genetic factors**

Sibling data and family fixed effects → causal effect of PGI

→ **Study GxE**

e.g. Increase of compulsory schooling age in U.K. reduces BMI only among those with a high-BMI PGI (Barcellos, Carvalho, and Turley 2016)

→ **Use as control variable**

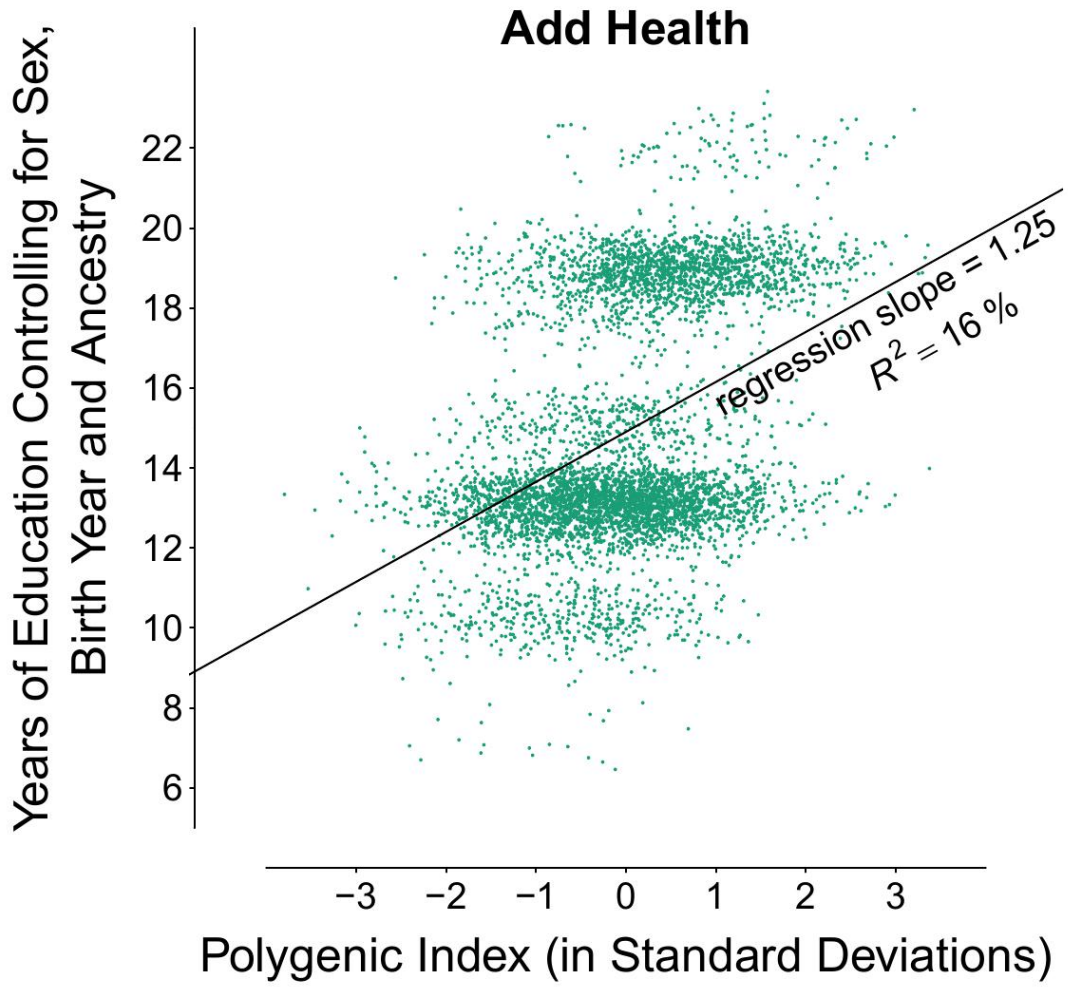
To control for confounding genetic factors or to increase statistical power for estimating the effect of a randomized treatment. If incremental R^2_{PGI} is 15%, then power increase is equivalent to 17% increase in sample size (Rietveld, 2013)

→ **Identify at-risk individuals**

→ **Personalized treatment**

⋮

Individual-level prediction is not accurate enough for most complex phenotypes!



Source: Okbay et al. (2022)

Prediction with related samples

If you are interested in incremental-R^2, no need to do anything special, R^2 is still valid, but

- the standard error for the coefficient of the PGI is going to be wrong!

What to do?

- If you have family IDs, cluster standard errors at the family level
- Otherwise, can control for the relatedness using the GRM and a linear mixed model
- Possible to do in GCTA

Outline

- A literal history of PGIs
- What is a polygenic index?
- Predictive power of polygenic indices
- Constructing polygenic indices
- Applications
- Limitations and pitfalls

LIMITATIONS & PITFALLS

Mechanisms are poorly understood.

- Including many genetic variants
  - increases predictive power
  - requires including genetic variants with unknown function

→makes it hard to specify what is captured by PGI.

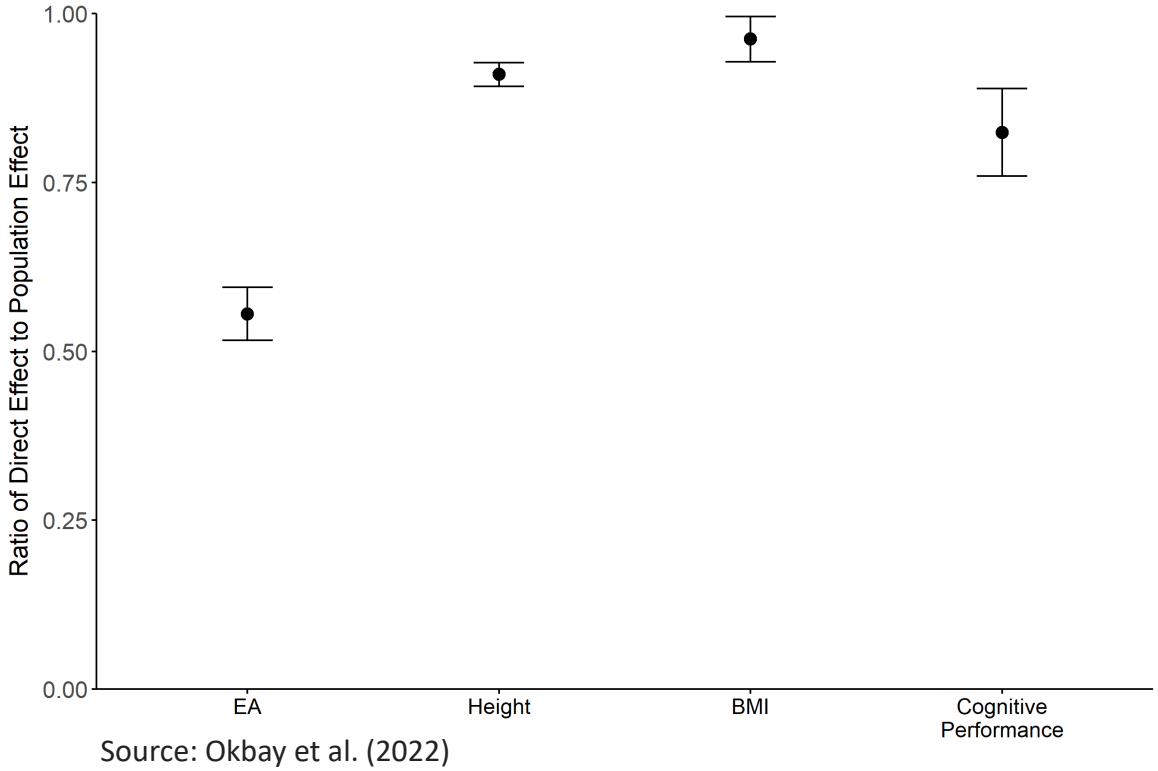
LIMITATIONS & PITFALLS

The predictive power of the PGI may not be solely due to the causal effect of genetic variants included in the PGI!

- Gene-environment correlation
  - Population stratification
  - Indirect genetic effects
- Assortative mating

Solutions:

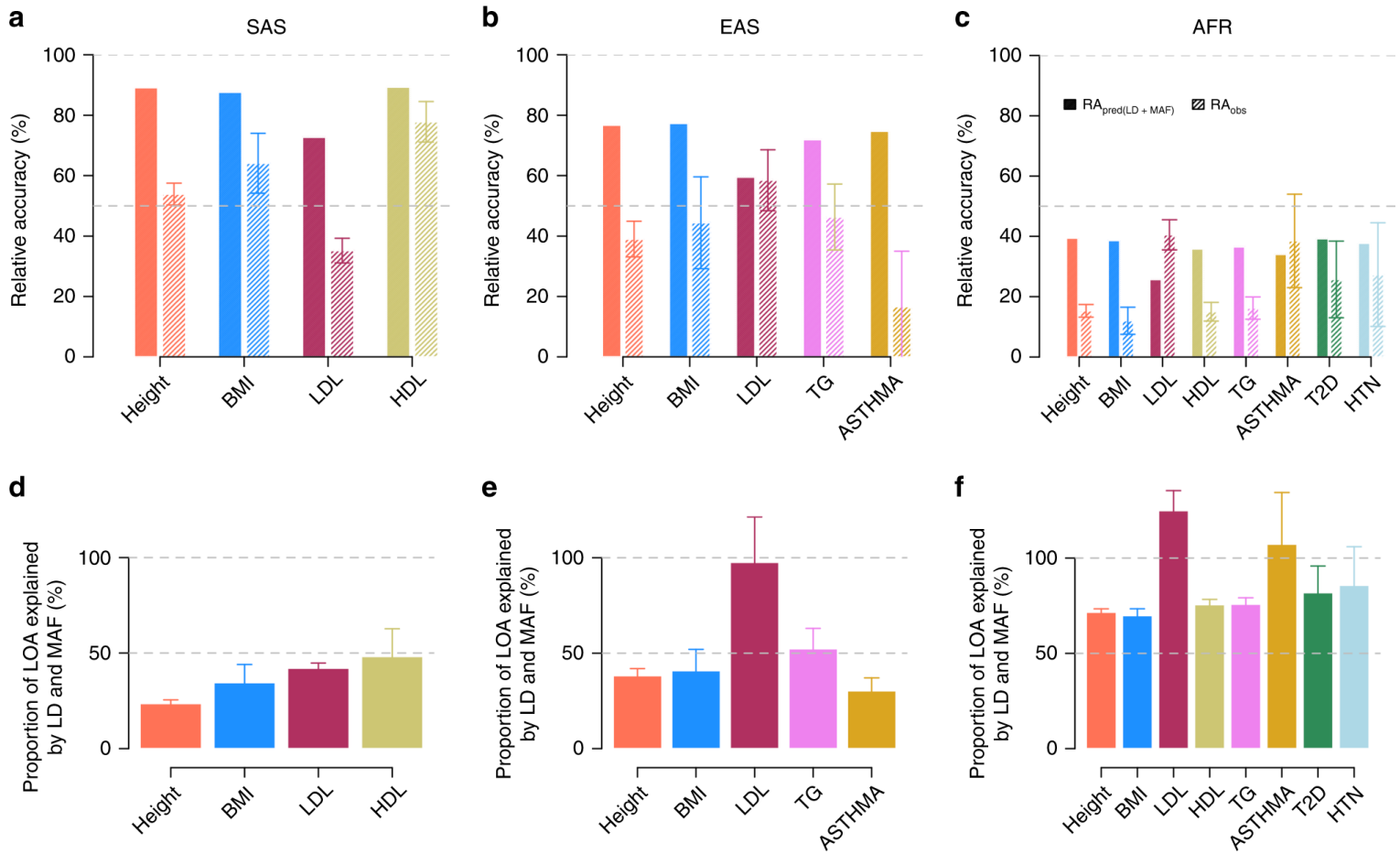
- Within-family prediction
- PGIs based on within-family GWAS



Source: Okbay et al. (2022)

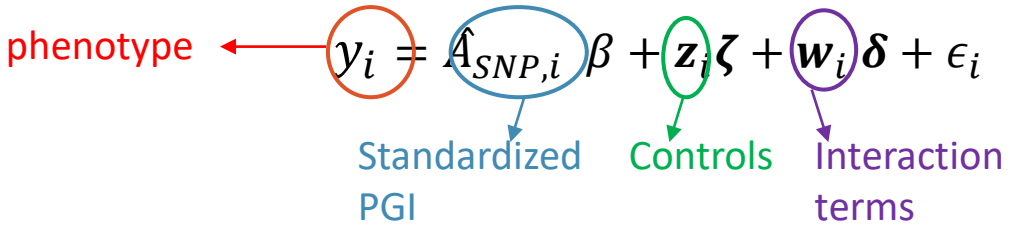
LIMITATIONS & PITFALLS

- Current polygenic indices far less predictive in non-European-descent samples.
- For example, for the EA4 PGI:
  - $R^2 \approx 17\%$ for European-ancestry individuals in Add Health, 13% in HRS.
  - $R^2 \approx 2.3\%$ for African-ancestry individuals in Add Health, 1.3% in HRS.
- Methodologies are being developed to improve cross-ancestry predictive power (e.g. PRS-CSx)
- PGIs incorporating functional annotation (Mega-PRS, SBayesRC) do better



Source: Wang *et al.* (2020)

MEASUREMENT ERROR



$\hat{\beta}$ is the expected increase in y corresponding to 1 SD increase in the PGI. Two issues with this:

1. $\hat{\beta}$ will be attenuated due to classical measurement error
2. The SD of the PGI depends on the amount of measurement error in it – how to compare results from different studies using differently constructed PGIs?

Solution: Scale everything to resemble a regression using A_{SNP} instead of $\hat{A}_{SNP}$

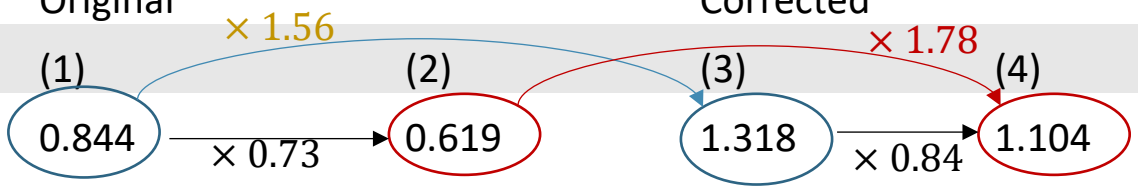
- Very straightforward in a simple regression with only the PGI (scale by $\sqrt{\frac{h^2}{R^2}}$ where R^2 is the predictive power of the PGI) but not so interesting.
- More interesting but a little more complicated when there are controls or interactions correlated with PGI

PGI-correct Python tool

https://github.com/JonJala/pgi_correct

Papageorge, N. W. & Thom, K. Genes, Education, and Labor Market Outcomes: Evidence from the Health and Retirement Study. *J. Eur. Econ. Assoc.* **18**, 1351–1399 (2020).

Panel A. Association Between EA and the PGI, Without and With Controls for Parental EA				
	Original		Corrected	
	(1)	(2)	(3)	(4)
EA PGI	0.844	0.619	1.318	1.104
	(0.026)	(0.024)	(0.041)	(0.042)
Father's EA	-	0.154	-	0.112
		(0.010)		(0.010)
Mother's EA	-	0.176	-	0.141
		(0.011)		(0.012)
# Obs.	8,537	8,537	8,537	8,537



- For EA in the HRS, $\hat{h}^2_{SNP} \approx 0.25$ and $\hat{R}^2 \approx 0.10$, according to the rule of thumb, coefficients should be expected to have increased by a factor of 1.58 ($\approx \sqrt{0.25/0.10}$).
- (2)→(4) increase is larger due to the positive correlations between the PGI, the controls, and the dependent variable.

➔ The correction deflates estimates of how much covariates mediate the effect of the PGI

PGI Repository

- First release includes 47 phenotypes in 11 datasets.
- PGIs based on GWAS including 23andMe for many phenotypes.
- Second release about to come out
  - 20 datasets
  - 60 phenotypes: anthropometric, cognitive, reproductive, biomarkers, health, psychiatric, substance use
  - Parental PGIs based on imputed parental genotypes in datasets with sib or parent-offspring pairs

Resource | Published: 17 June 2021

Resource profile and user guide of the Polygenic Index Repository

HRS Health and Retirement Study
Dunedin Multidisciplinary Health & Development Research Unit DMHDRU
MIDUS
biobank uk Improving the health of future generations
elsa English Longitudinal Study of Ageing
ncds National Child Development Study
BCS70 1970 British Cohort Study
Karolinska Institutet Swedish Twin Registry
FINNGEN
MCTFR Minnesota Center for Twin and Family Research
E-RiSK Environmental Risk Longitudinal Twin Study
TwinLife
generation scotland
the twin project @ the university of texas
Add Health The National Longitudinal Study of Adolescent to Adult Health
SOEP The Socio-Economic Panel
UNIVERSITY OF TARTU 1632
estonian genome center
Wisconsin Longitudinal Study
Children of the 90s Avon Longitudinal Study of Parents and Children
Millennium Cohort Study

QUESTIONS?

Practical

- Five steps:
  1. GWAS sumstats
  2. Genotype data
  3. SBayesR
  4. Making PGIs
  5. Prediction
- It is ok if you don’t finish it all, feel free to work on the remaining parts in your own time.

https://qimr.az1.qualtrics.com/jfe/form/SV_7OEO4h4KZgDZMc6

Reading materials

1. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e1003348 (2013). (**Theoretical framework, power**)
2. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–15 (2013). (**Complexities of interpretation**)
3. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019). (**SBayesR methodology**)
4. Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol. Psychiatry* **90**, 611–620 (2021). (**Overview of methods**)
5. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019). (**Poor cross-ancestry portability of PGIs and consequences**)
6. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 1–9 (2020). (**Factors contributing to poor cross-ancestry portability of PGIs**)
7. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* 1–8 (2022). (**Methodology to improve cross-ancestry prediction: PRS-CSx**)
8. Becker, J. *et al.* Resource profile and user guide of the Polygenic Index Repository. *Nat. Hum. Behav.* (2021) doi:10.1038/s41562-021-01119-3. (**Measurement correction**)
9. Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet. 2024 565* **56**, 767–777 (2024). (**Methodology to incorporate functional annotations into PGIs**)
10. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022), **Supplementary Note Section 7**. (**Contribution of confounders to PGI predictive power**)